Studies in History and Philosophy of Science xxx (xxxx) xxx

Contents lists available at ScienceDirect

Studies in History and Philosophy of Science

journal homepage: http://www.elsevier.com/locate/shpsa

How uncertainty can save measurement from circularity and holism

Sophie Ritson [a], Kent Staley [b,*]

[a] School of History and Philosophy of Science, University of Sydney, Australia
[b] Department of Philosophy, Saint Louis University, St. Louis, MO, USA

ARTICLE INFO

Keywords:
Measurement
Uncertainty
Underdetermination
Duhem
Circularity
Holism
Sensitivity

ABSTRACT

Measurement results depend upon assumptions, and some of those assumptions are theoretical in character. This paper examines particle physics measurements in which a measurement result depends upon a type of assumption for which that very same result may be evidentially relevant, thus raising a worry about potential circularity in argumentation. We demonstrate how the practice of evaluating measurement uncertainty serves to render any such evidential circularity epistemically benign. Our analysis shows how the evaluation and deployment of uncertainty evaluation constitutes an in practice solution to a particular form of Duhemian underdetermination that improves upon Duhem's vague notion of "good sense," avoids holism, and reconciles theory dependence of measurement with piecemeal hypothesis testing.

1. Introduction

Philosophers of science have long associated the claim that observations or experimental results in science are in some way *theory-laden* with a logical/epistemological problem regarding the possibility of scientific knowledge: reasoning from theory-laden observations may involve *circularity*. The circularity worry is that the very conclusion being aimed at (a theoretical claim) has been assumed or somehow relied upon in the justification or determination of one's premises.

The term theory-ladenness appears in discussions of various ways in which theoretical concerns enter into the production of observational or experimental evidence. This paper will not address all of these. The target here is a kind of potential circularity that arises in the context of producing measurement results that are then invoked in evidential arguments to support theoretical hypotheses. For convenience, we will label this *evidential circularity*. In the relevant cases, the measurement results represent conclusions drawn from data with the help of additional assumptions, including some of a theoretical character, that are relied upon in producing the specific result produced. In that sense, such results are theory dependent, but not all cases of theory dependence involve evidential circularity. The problem of evidential circularity arises when a measurement result is used to make an evidential argument in support of a theoretical claim that has been assumed in the production of the measurement result itself.

Cases of evidential circularity involve something more specific than what 'theory-ladenness' conveys. It involves a certain kind of dependence of a measurement result upon theoretical assumptions. Such terminology raises the question of what is meant by the term 'theoretical'. It turns out not to matter much, because the target of this investigation is ultimately the problem of evidential circularity itself. This problem enters into philosophy of science as a kind of offspring of the issue of theory dependence, but can be characterized independently of its parent, at least to the extent that our argument will not depend on any particular way of delimiting what counts as theoretical.

The inspiration for this analysis comes from a recent paper on measurement in experimental High Energy Physics (HEP), by Pierre-Hugues Beauchemin, a practitioner of measurement and the evaluation of measurement uncertainty in his role as a member of the ATLAS collaboration at the CERN Laboratory's Large Hadron Collider (LHC). Measurement in HEP is, according to him, a process of generating observational facts that is essentially theory dependent. Practices of evaluating the uncertainty and sensitivity of measurement results serve to avoid "potential circularity problem [s] traditionally implied by theory-ladenness" (Beauchemin, 2017, p. 275).

To state Beauchemin's argument roughly, evaluations of measurement uncertainty weaken the conclusions drawn from measurements (by adding larger "error bars") to the extent that the statement of that measurement result depends on uncertain theoretical assumptions. Measurement results that are weaker in this sense are correspondingly less sensitive; their larger error bars make them compatible with a broader range of theoretical alternatives and consequently less useful for deciding amongst those alternatives. By thus rendering results that

* Corresponding author.
E-mail address: kent.staley@slu.edu (K. Staley).

https://doi.org/10.1016/j.shpsa.2020.10.004
Received 16 October 2020; Accepted 23 October 2020

suffer from significant circularity problems less capable of deciding amongst competing hypotheses, uncertainty evaluation helps prevent such results from being deployed in vicious arguments in support of theoretical claims.

We find ourselves in broad agreement with the outlines of Beauchemin's account. Moreover, we welcome such a contribution to the philosophical understanding of the epistemology of HEP from a physicist engaged in analytic work crucial to the production of knowledge in that field. We propose to clarify the epistemological import of the insights that can be gleaned from Beauchemin's work. We will show how, from a careful analysis of his account of the evaluation of uncertainty and sensitivity in measurement, one can reconcile two claims that have generally been regarded as opposed to one another: (1) determining the outcome of an experiment that might serve as relevant evidence with respect to theoretical claims at least sometimes requires assuming theoretical claims that are closely related to and possibly identical to those that are to be empirically adjudicated; (2) experimentation enables targeted and piecemeal assessment of theoretical claims, such that a particular experimental result can be said to support or cast doubt upon a specific hypothesis independently of other claims that might reside within the same theory as that hypothesis. The first type of claim has been associated with the work of Pierre Duhem and work on confirmational holism, while the second type of claim has been embraced by New Experimentalists such as Hacking (1983), Ackermann (1985), Franklin (1986), and Mayo (1996).

Beauchemin's paper provides us with a thoroughly worked out example because it is a case in which the derivation of a measurement result from experimental data does in fact require an assumption of the same kind as that which the result may be used to test. This creates the potential for evidential circularity. We also present a second example that allows us to make this point in an especially vivid way. It may turn out that, among the competing hypotheses that the result is used to test, a favored hypothesis is the very one that was assumed in producing the result. We will show that the surprising upshot of taking into account the uncertainty and sensitivity considerations introduced by Beauchemin is that such seemingly blatant circularity need not be vicious. This runs contrary to a commonly encountered response to claims about the theory-dependence of observations in science, which is that theory dependence is not problematic so long as the theory on which an observation depends is in some sense independent from that which the observation is used to test or confirm, but that it becomes problematic when this condition is not met (Hacking, 1983; Kosso, 1989).

But the implications of this analysis go beyond the issue of evidential circularity itself. We relate the methodology of uncertainty and sensitivity assessment, by which problems of theory dependence are handled in HEP measurement experiments, to the problem of underdetermination posed by Duhem (1954 (1914)). Like Beauchemin, Duhem characterizes theory-testing in physics in a manner that emphasizes the dependence of experimental results on physical theory broadly speaking. Such dependence threatens the ability of the investigator to draw sharp conclusions about particular hypotheses of interest, unless one can somehow differentiate amongst the multitude of assumptions on which experimental results depend. For Duhem, solving the resulting problem calls for the physicist to exercise "good sense" and only in this way could one escape from a disabling holism that would prevent one from deciding theoretical questions on the basis of experimental results. The methodology described by Beauchemin, we argue, goes beyond the vagaries of good sense to provide a systematic and practical framework for both *identifying* those assumptions, theoretical or otherwise, on which a given result depends and, by quantifying the extent of dependence of the result on such assumptions, *discriminating* amongst those hypotheses that can and cannot be evaluated on the basis of that result at an acceptable level of epistemic risk. Contrary, then, to Beauchemin's own conclusion that his account supports "a more holistic view of the structure of science," we conclude that practices of uncertainty evaluation respond to underdetermination without leading to holism. We thus reconcile *at the level of practice* the theory dependence of measurement results with piecemeal, non-holistic empirical assessment, where previous reconciliations have offered theoretical responses in the form of accounts of confirmation (Glymour, 1980).

Our paper proceeds as follows: In section two we review relevant aspects of two examples. The first, a measurement of quantities crucial for understanding the production of the W boson, is also discussed in Beauchemin's paper. The second measurement, of the W and Z production cross sections, is included to clearly outline aspects of our own analysis. Drawing on both examples, we discern different types of theory dependence, not all of which threaten to introduce evidential circularity. Section three examines in greater detail the issues of evidential circularity and holism as problematic offspring of theory dependence. First, we show how the issue of potential circularity does indeed arise in the context of the kind of measurements discussed in section two. We then clarify the notion of holism that seems most germane to the kind of theory dependence under discussion, and distinguish the issue of holism from that of non-separability, where the latter is the claim that isolated hypotheses alone are not sufficient for conducting an experimental test. Section four presents solutions to the two problems posed in section three. We first explain how the evaluation of systematic uncertainty, understood as the outcome of a robustness analysis applied to a measurement model and put to service in the consideration of sensitivity, serves to render the threat of evidential circularity due to theory dependence epistemically benign. Second, we utilize Duhem's articulation of the underdetermination problem to clarify how this same methodology provides a principled and practically implementable escape from holism that takes full account of theory dependence, reconciling the latter with the pursuit of targeted, piecemeal experimental testing of some of the very same theoretical hypotheses on which the experimental result in hand depends. Duhem's own solution to underdetermination, in terms of *bon sens*, has been criticized as being too vague to provide illumination. We see the methodology discussed here as providing detail and structure to Duhem's response, transforming Duhem's insistence that there is a solution in principle, to a demonstration of how a solution works in practice. Section five summarizes our findings.

1.1. Clarification: how we approach Beauchemin's paper

Before proceeding, we will interject a clarifying note regarding our approach to Beauchemin's paper. Our paper appears to be a response to another paper that is not at all well-known. Our effort cannot, therefore, be justified by the influence that Beauchemin's work has exerted in philosophy of science.

Beauchemin's paper is unique in being the first serious philosophical investigation to directly target the particular constellation of issues that interest us: theory dependence, circularity, uncertainty, and sensitivity in measurement experiments. He advances philosophical claims regarding these issues, some of which we find broadly appealing, but they are not the object of our arguments here. Our approach is to utilize two important features of his paper as resources in developing our own argument regarding these same issues: (1) Beauchemin articulates a general account of how practices of evaluating measurement uncertainty enable the deployment of sensitivity considerations to avoid vicious circularity when using theory-dependent measurement results to test theoretical hypotheses. (2) Beauchemin provides a richly detailed description of a particular example of a measurement result from the ATLAS group to illustrate that general account.

Our discussion of feature (1) goes beyond merely replicating Beauchemin's account, insofar as Beauchemin's use of the term 'circular' appears to be a broad notion that includes both the specific relation evidential circularity, as well as other patterns of theory dependence. We develop a narrower construal to be able to clarify, and articulate with greater specificity, how the methodology invoked by Beauchemin functions epistemologically. This articulation provides the basis on

which we then develop our own distinct treatment of the issues of Duhemian underdetermination and holism that occupies the latter part of our paper.

We will not fully replicate the discussion in feature (2) of Beauchemin's paper, but instead summarize just enough of the illustration to explicate the methodology involved and to enable a clearer understanding of our arguments. We also present a brief discussion of a second ATLAS measurement, in which evidential circularity and the way in which it is addressed through uncertainty evaluation is especially striking. Our argument then proceeds to show how the methodology described allows, in an important class of cases, theory-dependent measurements to be used to target, in a piecemeal way, particular theoretical hypotheses, even the very same hypotheses upon which the results themselves depend.

2. Two measurement results

Beauchemin's analysis is based on the example of a measurement at the LHC by ATLAS of a process involving production and decay of the *W* boson. This section elucidates relevant aspects of Beauchemin's discussion of this example as well as an additional example in which the emphases of our own analysis stand out especially clearly.

2.1. Measuring a differential cross section for the W boson

The study of processes involving the production and decay of the *W* boson in association with hadronic jets is important both as a test of quantum chromodynamics (QCD) and for understanding backgrounds relevant to the study of other Standard Model (SM) processes and in searches for Beyond Standard Model (BSM) physics. Beauchemin discusses a measurement presented by ATLAS as one part of a comprehensive study of the production of *W*'s in conjunction with hadronic jets (ATLAS, 2012a). Such events can involve different numbers of jets (*jet multiplicity*). For several multiplicities, ATLAS measures the *W cross section* as a function of a number of kinematic features. The fact that the measurement is of the dependence of the cross section on some quantitative feature of the decay event, rather than measuring the total cross section of the process, is expressed by the term differential cross section. Our discussion focuses specifically on a measurement of the differential cross section σ_w of the production of W bosons decaying to an electron (or positron) e and antineutrino (or neutrino) ν in conjunction with at least one hadronic jet, with respect to the transverse momentum (component of the momentum transverse to the collider beam) P_T of the leading (i.e., most energetic) jet. This measured quantity, or measurand, we may denote $d\sigma(W \rightarrow e\nu + \geq 1jet)/dP_T$.

To obtain this measurement, ATLAS must identify events that are instances of the process in question, based on outputs from the ATLAS detector, and quantify the rate of occurrence of such events while also measuring the transverse momentum of the leading jet in each such occurrence. Events to be included in the sample are selected on the basis of selection criteria (*cuts*) applied to the measured characteristics of each event. Because no set of cuts can guarantee that only instances of the target process will be chosen, events that satisfy the cuts are considered *candidate events*, some of which will be instances of non-target processes (i.e., *background*). ATLAS reports the main sources of background to the measurement of $W \rightarrow e\nu + jet$ processes to consist of "multijet QCD events, other leptonic decays of gauge bosons [such as $Z \rightarrow ee$] and … $t\bar{t}$ [i.e., top quark—anti-top quark] production" (ATLAS, 2012a).

Each step enumerated by Beauchemin, in the transformations that take place from electronic signals in the detector to the assignment of numerical values to the measurand (Beauchemin, 2017, pp. 289–290), constitutes a complex scientific problem of its own, demanding the skilled application of some combination of experimental, computational, technological, and theoretical expertise. Each step also depends upon assumptions, theoretical or otherwise, of which some are uncertain in the sense that they could be replaced with plausible alternatives. These introduce uncertainty into the measurement result. The evaluation of this uncertainty is a quantitative matter. We will say more about the methodology for such evaluation in Section 4. For now, suffice it to say that any specific assumptions will result in attributing a specific (range of) value(s) to the measurand; the quantification of uncertainty requires determining *how much* that attribution changes when varying among the assumptions that one might possibly (and plausibly) make, and incorporating that 'how much' into the interval of values ultimately attributed to the measurand (i.e., the size of the "error bars"). Consequently, the statement of the evaluated uncertainty constitutes an essential part of the statement of the measurement result.

This evaluation of uncertainty is also intrinsically connected to the sensitivity of a measurement result to features of the underlying physics potentially targeted for testing by that measurement. The determination of sensitivity requires comparing, for a given quantity, (1) the difference in predicted values of that quantity based on competing hypotheses of interest and (2) the uncertainty on the measurement result for that same quantity (303). A measurement result is considered more sensitive to the theoretical possibilities implicated in the competing hypotheses the more that (1) exceeds (2).

Sensitivity, thus, constitutes a relationship between a measurement result and a theoretical question (framed in terms of a choice between competing hypotheses) that the measurement result might be used to answer. The *learning goals* of the experiment provide context for sensitivity evaluation. Generically, learning goals characterize the kinds of inferences investigators aim to be able to draw on the basis of data being generated. An analysis of sensitivity will understand these goals to include answering a theoretical question by discriminating amongst competing hypotheses by comparing their predictions to a measurement result. Theoretically predicted values will typically have their own uncertainty that will bear on the determination of (1). Greater uncertainties therefore weaken the potential for measurement results to be used to settle theoretical questions. On the theoretical side (1), uncertainties attached to predicted quantity values will tend to blur the differences between what competing theoretical hypotheses have to say about the quantity of interest. On the measurement result side (2) greater uncertainty blunts the ability of a result to be used to answer the theoretical question at hand. Simply put, all else being equal, greater measurement uncertainty means less sensitivity to the physics questions the measurement result may be used to answer.

The learning goal of measuring the differential cross section in question, according to Beauchemin, is to "obtain a better understanding of some aspects of the strong interaction … that are uncertain, such as parton distribution functions (PDFs)" (p. 291). These PDFs are important to our analysis, so a few words to explain them are in order.

What happens in a collision between protons at the LHC depends on details of the internal structure and dynamics of the protons involved. The term 'parton' refers generically to both quarks and gluons that together constitute protons. At any given time, the momentum of a proton will reflect the contributions of its partons, but the details of how that momentum is distributed among the partons cannot be uniquely determined from QCD principles alone. PDFs describe the probability of finding a parton of a given flavor within the proton carrying a given fraction of the total momentum of the proton, for a specified hard interaction energy scale. Theoretical predictions of cross sections like σ_W depend upon PDFs. Consequently, careful measurements of such cross sections can be used to test hypotheses about PDFs.

Such a measurement requires enumerating candidate events for the process in question. Counting is conducted on the basis of cuts that serve to operationalize the concept of a candidate event by imposing requirements in terms of key physical properties of the event (such as having at least one jet with $P_T \geq 30GeV$).

Cuts are chosen so as to enhance the signal to background ratio, which is crucial to the determination of the sensitivity of the measurement result. Because a theoretical understanding of the signal underlies the strategy that guides the choice of cuts, and because of the role of

theory-based simulation in translating that strategy into numerically defined cuts, Beauchemin concludes that the application of cuts to the data is "highly theory-laden" (292). To illustrate theory dependence due to simulation, however, we have chosen to focus on *background estimation*.

Cuts serve to differentiate between signal events that involve the target physical process and background events that do not, but they do so imperfectly. Some signal events will fail to pass the cuts imposed and some background events will survive the cuts and be counted as candidates. The latter constitute the residual background. Measuring a cross section requires the reliable estimation, for a given data set, of the expected size of this residual background, and we focus on two approaches to this task. The *Monte Carlo approach* relies more heavily on simulation than the *data-driven approach*. Both, however, rely on theoretical assumptions about the very process that ATLAS seeks to measure.

Background estimation begins with theoretical expectations regarding the kinds of processes that will contribute to the background. For the $W(\rightarrow e\nu) + jets$ process, these are $Z(\rightarrow ee) + jets, W(\rightarrow \tau\nu) + jets$, top, diboson, and dijet events (294). For most of these background processes, the Monte Carlo approach is applied, meaning that simulated events of these types are generated and then passed through a simulation of the ATLAS detector. The simulated data is then subjected to the same reconstruction software applied to non-simulated data. The cuts applied to the experimental data are then applied to the simulated data "as if they were real data events" (294). Finally, the data selected in the simulation are normalized to compensate for any difference in the sizes of the experimental and simulation data sets.

Beauchemin emphasizes the ways in which the theoretical understanding of these backgrounds processes necessary for this procedure applies to the $W(\rightarrow e\nu) + jets$ process itself. Although the decay of the W boson differs in the target process and the background processes, "[t]he sensitivity to PDF, to parton shower, to non-perturbative QCD corrections, to higher order QCD corrections, etc. is the same in both $W(\rightarrow \tau\nu) + jets$ and $W(\rightarrow e\nu) + jets$ processes. It is also highly similar in $Z(\rightarrow ee) + jets$ events" (294). Estimating these backgrounds therefore requires assumptions about physics processes that figure in the detection of both signal and background. The uncertainties introduced by such assumptions are relevant to the measurement result through their effects on both background and signal, and both must be quantified for the purpose of reporting a credible result.

To estimate other backgrounds, ATLAS uses a "data driven" approach using events reconstructed from actual data rather than simulated data. The QCD dijet background consists primarily of events in which either a hadronic jet from a light quark passes the cuts used to identify electrons and a mismeasurement of energy results in a large amount of missing transverse energy (the signature of a neutrino), or a hadron containing a bottom or charm quark decays to an electron. Because "the mechanisms by which a jet fakes an electron are difficult to simulate reliably" (ATLAS, 2012a, p. 5), this background is estimated by means of a background-enriched data sample obtained by reversing the application of some of the cuts used to define candidate events – in this case the cuts on electron candidate objects (Beauchemin, 2017, p. 294).

Estimating the dijet background requires the generation of two "templates" that characterize the distribution of the missing transverse energy E_T^{miss}. The first template is drawn from the background-enriched data. The second template, however, is based on "Monte Carlo simulation samples of all the other processes contributing to the data signal sample, including the $W(\rightarrow e\nu)$ + jets process itself" (294). The two templates are then fitted to the data to obtain a normalized E_T^{miss} distribution used to derive the dijet background estimate. Because the normalization requires simulation of the signal (and the leptonic background) as well as the data-derived template, Beauchemin concludes that "rather than eliminating the use of W + jets theoretical assumptions in the full W + jets background estimate, data-driven techniques introduce a further dependence on this process in the inference of the amount of dijet events contributing to the data sample" (p. 294).

Beauchemin emphasizes two points regarding these background estimation procedures. First, "theories, models, and assumptions are mandatory inputs to obtain measurement results." Second, "there is a real danger of circularity here: the same process to be measured is also used as input to obtain a measurement of this process" (295). Presumably we are to understand from the second point that measuring the $W(\rightarrow e\nu)$ + jets process depends upon assumptions (as manifested in simulation models) that are descriptive of the very same $W(\rightarrow e\nu)$ + jets process. Although this might be sufficient to indicate that Beauchemin's broad notion of circular theory dependence is in play, what remains to be clarified in the discussion to come is whether and how these features present the risk of evidential circularity, i.e., *argumentation in support of a theoretical claim that depends on that same claim as premise*.

2.2. Measuring the total cross section for the W and Z bosons

We have been discussing the measurement experiment targeted by Beauchemin's analysis in order to put on display a case in which the broad notion of circular theory dependence has been laid out in detail by someone with an insider's understanding of the analysis involved, but in which the stricter notion of evidential circularity that is our concern may also be identified. We now present a brief discussion of another ATLAS measurement in which evidential circularity and the way in which it is addressed through uncertainty evaluation are even more easily seen, which also has the benefit of yielding a more easily understood measurement result. This measurement, of the W and Z production cross sections, involves the very same features of dependence on simulation (and hence PDF) as the result just discussed, while being more explicit about the uncertainty assessment of this dependence and the use of the results in pursuit of the learning goal of testing PDF hypotheses.

LHC's Run 2 began in April 2015, delivering collisions at a center of mass energy of 13 TeV, a significant increase from the 7 (then 8) TeV collisions of Run 1. The results of the differential cross section measurement just discussed were based on a relatively small data set taken at 7 TeV. Our next example is based on results published in 2016, partway into Run 2 (ATLAS, 2016). Noting that theoretical predictions of W and Z production cross sections "depend on the parton distribution functions and are thus sensitive to the underlying dynamics of strongly interacting particles," ATLAS goes on to describe measurement of these quantities as "a unique opportunity to test models of parton dynamics" at the higher Run 2 energy (601).

The paper presenting these results (henceforth "ATLAS, 2016") reports two kinds of cross section measurements. The total cross section is defined as $\sigma^{tot} = N/(L \cdot A \cdot C)$. N quantifies the number of candidate events. The *luminosity* L is proportional to the accumulated number of proton-proton collisions on which the measurement is based. A, the *acceptance*, represents the fraction of Monte Carlo events at the "generator" level (simulating physics processes of proton-proton collisions at the relevant energy) that satisfy the geometric and kinematic requirements imposed on the data. These latter constitute the *fiducial acceptance* of the experiment. The *correction factor* C is defined as the "ratio of the total number of generated events which pass the selection requirements after reconstruction to the total number of generated events within the fiducial acceptance" (601–602).

The *fiducial cross section* is defined as $\sigma^{fid} = \sigma^{tot} \cdot A$. A measurement of the fiducial cross section thus seeks to measure the production rate of the relevant process within the fiducial region of the experiment, whereas the total cross section measurement extrapolates to the production rate of the process in total. Because the fiducial cross section is more directly measured, the systematic uncertainty on a measurement of fiducial cross section is smaller than for the corresponding total cross section measurement.

Assumptions of a theoretical character play an important role in these measurements, particularly through their dependence on simulation in numerous respects. In addition to the dependence of factors C and

A on Monte Carlo, the contributions of some backgrounds (top quark events and electroweak events involving two bosons or single bosons decaying to tau leptons) are estimated directly from simulated data samples. Estimating the background in the $W^{\pm}$ measurement from "multi-jet" events (which includes events with heavy quarks and hadrons misidentified as leptons) is handled in a "data-driven" way that is similar to the previously described data-driven method of estimating the QCD dijet background in the differential cross section measurement. Although the estimate is based on actual data rather than simulated data, the estimation procedure requires fitting the data using templates drawn in part from simulation (603).

The contribution of PDF assumptions is listed separately in the discussion of systematic uncertainties in ATLAS, 2016, where it is treated as a contribution to the uncertainty on the correction factor C. In Table 1 it can be seen that PDF contributes a relatively small amount to the overall systematic uncertainty.

ATLAS, 2016 then uses these cross section measurements to test PDF hypotheses. This can be seen in the two plots in Fig. 1. Here the ratios of fiducial cross sections are used as the quantities of interest: $R_{W^+/W^-} = \sigma_{W^+}/\sigma_{W^-}$ and $R_{W/Z} = \sigma_{W^\pm}/\sigma_Z$. These plots compare values of these ratios to the theoretical predictions from different PDF sets, which are represented by filled and unfilled triangles, circles, and squares, along with error bars to represent uncertainties. The central (red) line represents the calculated value of the ratio represented, while the inner (yellow) band around that line represents the statistical uncertainty on that value. The outer (green) band represents the total uncertainty obtained by adding statistical and systematic uncertainties in quadrature. Stated numerically, the results are $R_{W^+/W^-} = 1.295 \pm 0.003 \pm 0.010$ and $R_{W/Z} = 10.31 \pm 0.04 \pm 0.20$ (ATLAS, 2016, p. 607).

The importance of systematic uncertainty for sensitivity can be seen clearly. For the comparison of the measurement of R_{W^+/W^-} to theoretical predictions, the systematic uncertainty renders one prediction compatible ("within uncertainty") with the measurement result that would otherwise be incompatible. Compare the measurement of $R_{W/Z}$, for which the systematic uncertainty is several times larger than the statistical uncertainty, rendering the measurement result insensitive to the differences between theoretical predictions. Although from Table 1 we can see that in this case the contribution of PDF assumptions to the overall systematic uncertainty is relatively small, the example illustrates how the evaluation of such uncertainty matters to the inferences to be drawn from the result produced.[1]

Let us take stock of theory dependence as it has figured in our discussion thus far. We can note first that, as summarized by Beauchemin, the transformations involved in arriving at a measurement result involve inputs from theory that are varied in their content, but also in the way in which they influence the analysis and its results. The variety of theoretical inputs is in fact greater than our limited discussion conveys. For example, detector simulation, which plays a role in the determination of the unfolding matrix, involves models of nuclear physics (p. 297). To more clearly define the problem of evidential circularity, however, we will first articulate a generic category of model dependence.

The generic relationship of model dependence is obtained when the result of a measurement depends on a model assumption, particularly in cases where such assumptions are of a theoretical character. The notion of dependence here is to be understood in terms of inputs to the procedure P whereby the measurement result is actually produced. To say that a measurement result R arrived at by procedure P depends on a particular assumption A is simply to say that the removal without replacement of A among the inputs to P would prevent P from producing R. It is not to say that R could not be produced without A, because there may be another procedure P′ that would generate R without A.[2] The term 'theory dependence' will be applicable when the model in question is in some sense theoretical, but we do not intend to legislate the application of this label in any way. As has been well-documented, experimental results in HEP depend on the deployment of a rich variety of different kinds of models (Cartwright, Shomar, & Suárez, 1995; Karaca, 2013, p. 1999; M. Morrison, 1999). The centrality of modeling to measurement in particular is prominent in Tal's contributions (2012, 2019). We are happy to allow model dependence to cover a broad range of cases of model dependence. Although noting the prevalence of model dependence in the context of HEP measurement is neither novel nor profound, our specification of this relationship provides a schema that we can use to define evidential circularity.

3. Circularity and holism

3.1. Circularity

When the support that an argument lends to its conclusion depends upon a premise that is identical in content to the conclusion itself, the argument must be, it would seem, epistemically defective. After all, for a person accepting the conclusion on the basis of such an argument to be supported logically in their acceptance, they must have already accepted the very proposition the acceptance of which is supposedly warranted by the argument in question. This is precisely the condition denoted by the term "vicious circularity."

We will next identify the specific context in which evidential circularity arises in HEP measurements, and show how it exemplifies the pattern of circularity just mentioned. Applied to that context, the apparatus of evaluating uncertainty and sensitivity allows empirical argumentation to exhibit simultaneously the attributes of circularity and cogency. That is to say, circularity of argumentation arising from theory dependence need not be vicious and need not prevent the conferral of genuine empirical support for the theoretical conclusion responsible for such circularity.

As mentioned in section 2.1, an important theoretical motivation for the cross section measurement described above is to obtain a "better understanding" of strong interaction processes described by QCD, and to do so by testing the predictions of different hypotheses essential for the task of providing QCD with predictive content, such as parton distribution functions (PDFs). PDF hypotheses, however, are also essential for the purposes of simulating QCD processes, and QCD simulations play multiple roles in generating measurement results in HEP.

We can apply the schema for model dependence to show that the potential for circularity is evident: Producing the measurement result R depends on making an assumption A about PDF. Once R has been produced, it can be used to test hypotheses about PDF, including A. It could very well happen that in using R to test between A and a competing hypothesis A', R turns out to favor A over A'. Any argumentation in support of the theoretical conclusion A that depends upon this test outcome thus relies on R as a premise. If challenged to support R by making explicit all of the assumptions that investigators relied on to arrive at R, investigators will include A among their assumed premises. In a single argumentation context, then, A plays both roles: conclusion and premise.

Consider the results presented in the left-hand plot of Fig. 1, for example. The result R: $R_{W^+/W^-} = 1.295 \pm 0.003 \pm 0.010$ would appear to lend support to the conclusion that one ought to favor the hypothesis H_1: "The PDF set CT14nnlo is adequate for the analysis of LHC data" over

[1] Lest it be thought that we have chosen an unusual example, ATLAS has published many other papers that explicitly invoke measurement results as tests of PDF and that evaluate uncertainty due to PDF assumptions relied upon in arriving at those results. In some cases the PDF uncertainty is listed separately (ATLAS, 2011; ATLAS, 2017), in others it is folded into the uncertainty attributed to the choice of simulation (ATLAS, 2012a; ATLAS, 2018).

[2] Accordingly, statements in this paper to the effect that particular results are dependent on, for example, simulation should not be understood as claims that such results would be impossible without simulation.

Table 1
Systematic uncertainties, in percentages, on the correction factor C. Individual contributions are added in quadrature to calculate the total (ATLAS, 2016).

$\delta C/C$	$Z\to e^+\nu$	$W^+\to e^+\nu$	$W^-\to e^-\bar{\nu}$	$Z\to\mu^+\mu^-$	$W\to\mu^+\nu$	$W^-\to\mu^-\bar{\nu}$
Lepton trigger	0.1	0.3	0.3	0.2	0.6	0.6
Lepton reconstruction, identification	0.9	0.5	0.6	0.9	0.4	0.4
Lepton isolation	0.3	0.1	0.1	0.5	0.3	0.3
Lepton scale and resolution	0.2	0.4	0.4	0.1	0.1	0.1
Charge identification	0.1	0.1	0.1	–	–	–
JES and JER	–	1.7	1.7	–	1.6	1.7
E_T^{miss}	–	0.1	0.1	–	0.1	0.1
Pile-up modeling	<0.1	0.4	0.3	<0.1	0.2	0.2
PDF	0.1	0.1	0.1	<0.1	0.1	0.1
Total	1.0	1.9	1.9	1.1	1.8	1.8

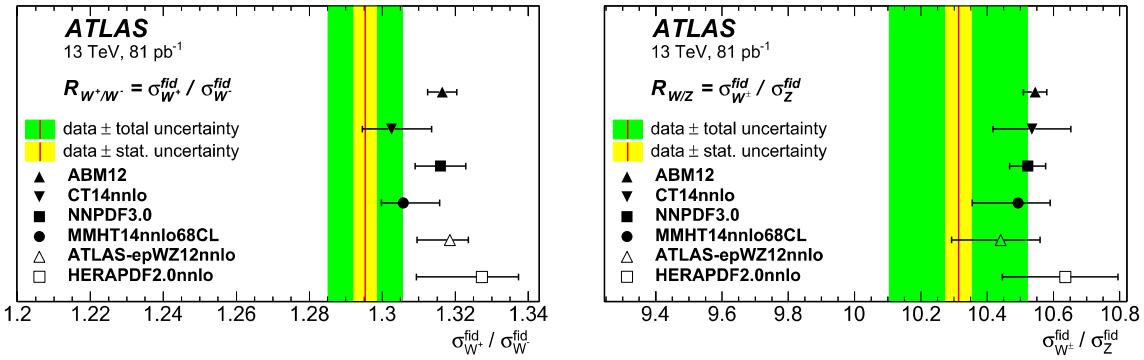

Fig. 1. Measurements of the fiducial cross section ratios for W^+ to W^- (left) and $W^\pm$ to Z (right). Both are compared to theoretical predictions based on different PDF sets. The inner (yellow) band is the statistical uncertainty and the outer (green) band includes the systematic uncertainty (ATLAS, 2016). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

the hypothesis H_2: "The PDF set ABM12 is adequate for the analysis of LHC data." Such an argument in support of H_1 would invoke the result R as a premise. If we ask what kind of argument can be given in support of R, we will note that it follows not from the data alone, but by additional assumed premises. What kind of premises? Well, the data are LHC data, and deriving a measurement result from them requires the use of a PDF set. If the PDF set chosen for such an analysis is CT14nnlo, then an assumed premise is H_1: "The PDF set CT14nnlo is adequate for the analysis of LHC data."[3]

This pattern of dependence on PDF hypotheses that are potential targets of empirical argumentation is not restricted to background estimation and is not restricted to the particular examples under discussion. PDF assumptions are crucial to any QCD simulation of deep inelastic scattering (the relevant process for LHC collisions). Moreover, PDFs play more than one role in simulating event generation. They are needed for computing the cross section of the basic scattering process and for determining the initial state radiation, i.e., the radiation from the incoming partons of gluons, which may in turn radiate new quark-antiquark pairs. Thus, any of the transformations of data leading to a measurement result that depend on simulation also depend on PDF and constitute a potential source of circularity in argumentation using the measurement result to draw conclusions about PDF.

It is important to distinguish this type of evidential circularity from a slightly different pattern of theory dependence to which the term circularity may be sometimes applied, in which a broader theoretical framework is assumed in the testing of a hypothesis that resides within that broader framework. For example, evidence of the existence of a Higgs-like boson emerged from an analysis of LHC data that depended strongly on the assumptions of the Standard Model. Yet the Higgs mechanism from which the prediction of such a boson derives is part of the Standard Model (ATLAS, 2012b; CMS, 2012; Franklin, 2017). Some of Beauchemin's own uses of the term circularity seem to involve something like this pattern (as in the quotation from p. 295 in section 2.1 above).[4]

Regarding this type of theory dependence as exemplifying circularity (even if of a benign sort) involves a contested assumption about relations of support, however, which is that evidence supporting a hypothesis that resides within a broader theory may constitute evidence for that broader theory. While such an assumption can find clear motivation in philosophical approaches such as Bayesian confirmation theory (Howson & Franklin, 1985), a severe-test approach, for example, will reject it on the grounds that a hypothesis may pass a severe test with given data although the theory within which that hypothesis resides does not (Mayo, 1996, 2018). From the latter point of view, no circularity would be involved in the Higgs example so long as the parts of the Standard Model that are assumed in the analysis of the Higgs data do not include the assumptions implicated in the Higgs hypothesis itself, for it is only the latter that is evidentially supported by (severely tested with) those data. These considerations do not apply in the kind of case we are discussing. Here it is the very same hypothesis for which a measurement result may serve as evidence that is assumed in the derivation of that result.

[3] Whether the specific PDF set CT14nnlo was relied upon in producing the measurement results in (ATLAS, 2016) is not specified in the paper, but simulations used to evaluate efficiency and some background contributions did rely on the CT10 PDF set, a kind of ancestor to CT14 produced by the same group and adhering to the same general approach. PDF construction is a data-driven undertaking, but different sets result from reliance on different kinds of data as well as different methodologies.

[4] Thanks to an anonymous referee for raising this issue and suggesting the example.

3.2. Holism

The term 'holism' appears only in the last sentence of Beauchemin's paper, which declares that his discussion "points toward a more holistic view of the structure of science, at least in modern HEP" (310). This word can signify a broad range of views, however. In this section we investigate possible connections between theory dependence of the sort that potentially engenders evidential circularity and holistic theses about science. We aim to show how the practices of uncertainty and sensitivity evaluation demonstrate how an acknowledgement of theory dependence, even of a sort that threatens evidential circularity, is compatible with a denial of strong holistic theses about theory evaluation in physics.

One way of distinguishing holist theses about empirical assessment of theories is in terms of what Quine labelled the "unit of empirical significance" (Quine, 1951). A local, piecemeal approach holds out for the possibility of isolating particular hypotheses as the targets of testing, such that falsification rules out the acceptability of the targeted hypothesis while leaving unaffected the antecedent status of any auxiliary assumptions required for conducting the test. At the opposite end of the spectrum would be a global holism that regards "all of science" as being subject simultaneously to evaluation in the light of any given observations. (Whether such a view has been seriously advocated within philosophy of science may be debatable, but at least some of Quine's more provocative comments point to such a view (Quine, 1951).)

The versions of holism most germane to the issues under discussion here tie support for particular hypotheses to support for the broader theories within which those hypotheses reside. Specifically, consider the holist thesis that any evidence E that provides empirical support for any hypothesis H that resides in theory T to some degree supports T in its entirety, which is to say that E also supports (to some degree) any other hypothesis H_i that resides in T.[5]

Of particular relevance is the so-called Duhem problem. Duhem's account of theory testing in physics can be separated into a 'non-separability' thesis and a 'holism' thesis. The non-separability thesis states that you can only ever subject to test a hypothesis that is accompanied by a group of auxiliary hypotheses, theories, and assumptions.

"To seek to separate each of the hypotheses of theoretical physics from the other assumptions upon which this science rests, in order to subject it in isolation to the control of observation, is to pursue a chimera" (Duhem, 1954 (1914), pp. 199–200).

For Duhem, the non-separability thesis introduces holism as a problem.

"[T]he physicist can never subject an isolated hypothesis to experimental test, but only a whole group of hypotheses; when the experiment is in disagreement with his preconditions, what he learns is that at least one of the hypotheses constituting this group is unacceptable and ought to be modified but the experiment does not designate which one should be changed" (Duhem, 1954 (1914), p. 187).

Duhem argues that when only the resources of logic are brought to bear on a test that is necessarily directed at a group of hypotheses, theories, and assumptions, any result of that test bears immediately upon the group of hypotheses, theories, and assumptions as a whole. Therefore, the experimenter is not able, on the basis of the result alone, to attribute the evidential weight of the result wholly or partially to any particular hypothesis, theory, or assumption (at least initially). This holds for the case of a discrepant result, which Duhem outlines above, where the experimenter cannot determine which hypothesis ought to be changed or discarded. In the positive case the experimenter faces the same problem, whether to accept as trustworthy the auxiliary assumptions needed to draw a conclusion regarding the hypothesis of interest. In each case, the choice is underdetermined by applying the resources of logic to the experimental result. Duhem details an additional consequence: that it is sometimes the case that the group of auxiliary hypotheses, theories, and assumptions, required for an experiment, includes the very same theory or hypothesis that the experiment is designed to test (1954 (1914), pp. 188–190). This introduces the possibility of vicious circularity.

Both Duhem and Beauchemin commit to the non-separability thesis. However, Beauchemin's commitment is implicit and can be found, for example, where he claims that meaningful physics conclusions can only be drawn following the addition of a number of statistical inferences that transform the large number of events required. Beauchemin argues that each such transformation requires the assumption of theoretical content in order "to confer epistemic value to measurement results" (Beauchemin, 2017, p. 309). Beauchemin therefore commits that for HEP experiments a group of hypotheses, theories, and assumptions is required for a measurement result.

Non-separability leads to the threat of holism and the potential for circularity. As was outlined in section 2.1, the potential for evidential circularity arises from the fact that simulating QCD processes requires assumptions concerning PDFs. Simulation-dependent measurement results therefore require PDF assumptions, and it could be the case that the measurement result *R* is later cited as evidence for a PDF assumption used to arrive at *R*. In 2.2 the cross section measurement result is explicitly used to test PDF hypotheses. Consequently, the experimenter needs to distinguish among the assumptions required for the measurement result to determine how strongly the value assigned to the measurement result depends on any particular assumption. Absent a method to do so, non-separability would appear to lead to holism: the measurement result would bear upon the group of hypotheses, theories, and assumptions as a whole. In the following section we will identify how the evaluation of uncertainties allows for the needed differentiation amongst assumptions, thus rendering circularity non-vicious and forestalling holism.

4. Uncertainty: undoing circularity and avoiding holism

4.1. Circularity in empirical support and its undoing

Measurements based on experimental data in HEP, then, are theory dependent in a way that generates a potential for evidential circularity in the arguments that support theoretical conclusions. Yet HEP is a field that appears to succeed regularly in producing measurement results that researchers regard not only as trustworthy but as relevant evidence for drawing conclusions at the level of physical theory. We can now reconcile these two facts about HEP. In short, although every HEP measurement result that utilizes QCD simulation requires investigators to make some assumption about PDF, it is possible to evaluate, for a given PDF assumption, how much the value assigned to the measurand depends on that assumption. Doing so involves a determination of how much the value assigned to the measurand changes when changes are made to the PDF assumption itself. If the change is large enough, then the result will have too large an uncertainty to enable discrimination amongst the hypotheses being tested. If the changes are sufficiently small, then the result will have a correspondingly small uncertainty, such that it may be consistent with one hypothesis and inconsistent with others. Even if the hypothesis that passes the test is one that was also assumed in arriving at the result itself, this would not undermine the use of the result as supporting that hypothesis. This is because the small uncertainty reflects the fact that the result does not depend very much on the particular PDF assumed; other plausible PDF assumptions would lead to results within the same small uncertainty interval.

[5] A hypothesis *H* resides in a theory *T* when *H* contributes to the specification of the content of *T*. Exactly how hypotheses contribute to theory content is disputed between syntactic and semantic views of theory structure, regarding which we are here neutral.

Incorporating uncertainties thus changes the argumentation involved in arriving at a theoretical conclusion. We can reconstruct (in an abstract manner) the impact of uncertainties on argumentation[6] through three stages of measurement evaluation and theory assessment as follows (Staley, 2020):

- Stage 1 (measurement without uncertainty): Data X is used as input to a model M. That model is defined by assumptions that attribute definite values to model parameters and does not characterize the variance of the data. This allows the computation of a determinate value for the measurand $\mu = \widehat{\mu}$.
- Stage 1 (theory assessment): The result $\mu = \widehat{\mu}$ is not useful for theory assessment. The model used to produce the result does not account for the degree to which future measurements of the same quantity would vary in their outputs, even when performed in the same way. Possible flaws in model assumptions are not considered.
- Stage 2 (measurement with statistical uncertainty): Data X is used as input to a model M_0, which includes all assumptions of M, but adds distributional assumption(s) to account for variance. This allows for determination of a measurement result $\mu = \widehat{\mu} \pm u$, where u is the statistical uncertainty.
- Stage 2 (theory assessment): The result $\mu = \widehat{\mu} \pm u$ is a candidate for theory assessment, but remains epistemically flawed. Suppose H and H' make predictions about the value of μ. Any conclusions about H or H' based on $\mu = \widehat{\mu} \pm u$ will depend on the correctness of the assumptions of M_0. An argument in support of H or H' that depends on a false model assumption will be unsound. If H is a model assumption in M_0, then an argument in support of H based on agreement between $\mu = \widehat{\mu} \pm u$ and the value of μ predicted by H will be viciously circular.
- Stage 3 (measurement with full uncertainty): Data X is evaluated by model M_0, along with evaluations by a family of variants of M_0: M_0^i, $i = 1, 2, \ldots$. Each M_0^i involves alternative(s) to the assumptions of M_0 and yields results determining a range $\mu = \widehat{\mu} \pm u_1 \pm u_2$, where $u_1$is the statistical uncertainty and u_2 is the systematic uncertainty reflecting the variation generated by the variant models M_0^i. The relevance of the result $\mu = \widehat{\mu} \pm u_1 \pm u_2$ depends on the aptness of the choice of the M_0^i, which should reflect the extent to which alternative model assumptions are plausible or could have been reasonably chosen.
- Stage 3 (theory assessment): The result $\mu = \widehat{\mu} \pm u_1 \pm u_2$ may provide the basis for a cogent empirical argument supporting theoretical conclusions. If H and H' both make predictions about the value of μ, and $\mu = \widehat{\mu} \pm u_1 \pm u_2$ agrees with the prediction from H and disagrees with the prediction from H', then, ceteris paribus, the result provides a reason to prefer H to H'. An objection based on a possible flaw in a model assumption can be countered to the extent that such a flaw has been anticipated in one of the alternatives M_0^i: If the flaw is actual, the measurement result (as computed with a model that corrects that flaw) still falls within the range $\mu = \widehat{\mu} \pm u_1 \pm u_2$. An objection based on circularity may also be countered. Suppose the objector points out that the hypothesis H that is being claimed to have support from the measurement result was relied upon in producing the measurement result. Provided that the dependence of the result on H has been explored in the alternative models M_0^i, the following reply is available: Although H was used in arriving at the result, the interval $\mu = \widehat{\mu} \pm u_1 \pm u_2$includes estimates of μ produced by relying on alternatives to H (possibly including H').

As argued at length in (Staley, 2020), the core of any methodology for evaluating systematic uncertainty involves robustness analysis. Wimsatt's influential discussion describes robustness analysis as involving (1) the analysis of a "variety of independent derivation, identification, or measurement processes," (2) determination and analysis of that which is invariant or identical "in the conclusions or results of these processes," (3) determination of the scope of such invariance as well as the conditions of invariance, and (4) analysis and explanation of relevant failures of invariance (Wimsatt, 1981, p. 126).

In this case, robustness analysis is applied to a model of the measurement process (M_0), alternatives to which are generated by varying model assumptions that are subject to uncertainty. Although these alternative models differ from M_0 only slightly, they are independent in the relevant sense for the purposes of the analysis undertaken: each alternative model involves at least one assumption that might be correct in case its counterpart in M_0 is incorrect. The result of the process is constituted by the calculated value $\widehat{\mu}_i$ of the measurand result derived from a particular model variant M_0^i, and the invariant feature of these results is that they all lie within the interval $\widehat{\mu} \pm u_1 \pm u_2$.

That interval, the product of stages 2 and 3 in the above reconstruction, constitutes a kind of weakening of the measurement result, relative to the empirically unusable computation of a determinate value for the measurand $\mu = \widehat{\mu}$ that is achieved in stage 1. The robustness analysis and weakening work together to manage (but not eliminate) the epistemic risk assumed by the investigator.

Multiple risks threaten the reporting and use of measurement results, including the following: (1) One might draw an erroneous conclusion about the value of the measurand on the basis of some feature of the data that will not be present in future samples, because of the variability of the data. The statistical component of the uncertainty (stage 2 in the reconstruction above) addresses this risk. (2) The attribution of values to the measurand might be biased by some flaw in the model of the measurement process, such as a theoretical assumption, and hence lead to an erroneous conclusion about the value of the measurand. The systematic component of measurement uncertainty addresses this risk. (3) In using a result to test hypotheses that predict values for the quantity measured, the agreement between the predictions of a hypothesis and the measurement result (and disagreement of the result with predictions from competing hypotheses), might be an artifact of assumptions relied upon in arriving at the result that would vanish under plausible alternative assumptions. A special case of this is the one in which the hypothesis apparently supported by the measurement result and a hypothesis assumed in arriving at that result are one and the same.

Such risks cannot in general be eliminated, but they can be managed. We speak of managing here because it is a matter not simply of reducing risk of error, but of making judicious choices regarding the extent to which inferences are exposed to risk. Implementing the robustness methodology just described involves tradeoffs between weakening one's inferences about the value of the measurand (which also reduces sensitivity) and exposure to risk.

Robustness analysis aids in this management task because it is a useful strategy to distinguish, in Wimsatt's words, the "ontologically and epistemologically trustworthy" from the "unreliable, ungeneralizable, worthless, and fleeting" (Wimsatt, 1981, p. 128). Critics have called into question the ability of robustness analysis to achieve this ambitious aim (Odenbaugh & Alexandrova, 2011; Stegenga, 2012). Stegenga, for example, raises a concern about the possibility of a systematic approach to robustness analysis because of its dependence on judgments about kinds and degrees of independence. These in turn depend upon "background assumptions which we are uncertain about," and Stegenga despairs of the prospects for "a satisfactory and general criterion of identifying problematic background assumptions" (Stegenga, 2012, p. 220).

With or without a criterion, however, identifying problematic background assumptions is precisely the task undertaken in uncertainty evaluation, and no responsible reporting of measurements in HEP would

[6] For reasons of economy, this reconstruction focuses on uncertainties in measurement results. Theoretical predictions also come with uncertainties that are crucial for reaching conclusions about theoretical hypotheses based on such predictions, as reflected in Fig. 1.

be possible without it. Although the sense in which the variation of measurement model assumptions introduces independent assumptions is minimal, it is well suited to the problem the physicist seeks to solve. This problem, as Beauchemin depicts it, is to "account for errors in measurement procedures that are unknown but which affect the results without being a priori correctable" (299). Treating the absence of an effective decision procedure for identifying and accounting for all possible errors in a measurement procedure as an obstacle to reporting results in a manner that accounts for errors that can be treated through a robustness analysis would block the road of inquiry.

Importantly, the logic of this methodology also plays a constraining role that complements the argumentative strategy just outlined. If a measurement result does strongly depend on a particular theoretical assumption (including one that could introduce evidential circularity), the resulting large uncertainty bounds will render it compatible with competing hypotheses and incapable of being invoked as supporting one hypothesis over another (as illustrated in the right hand plot in Fig. 1). In this way, uncertainty and sensitivity considerations protect against the potentially vicious features of evidential circularity – that argumentation might not provide a good reason to accept a conclusion independently of having already accepted it.

This point allows for an expanded understanding of the way in which robustness analysis applied to a model of the measurement procedure supports the management of epistemic risk: Not only does it enable an epistemically warranted weakening of the claimed measurement result to reduce the risk of reporting erroneously the value of the measurand, it reduces the risk of relying on a theory-dependent measurement result in an evidential argument for a theoretical claim that is subject to a vicious form of circularity. Such an epistemically flawed means of defending a theoretical claim could be present even were the measurement result itself correct, but sensitivity considerations prevent this from happening, provided that the uncertainty estimate attached to the result accurately represents the extent to which the result is sensitive to the particular assumptions upon which it depends.

4.2. *Undoing circularity and 'good sense'*

In this section, we show how uncertainty practices can transform the epistemic risk of assigning evidential weight to measurement results. We examine these practices in the light of Duhem's claim that it is 'good sense' that breaks the underdetermination and can account for historic progress in science. In this discussion we here highlight how the epistemic risk introduced by potential underdetermination and holism, classic problems in philosophy of science, can be managed in practice.

The holism that follows from Duhem's non-separability thesis can be avoided, as discussed, if the experimenter can discriminate amongst the assumptions required for the measurement result to determine how strongly the value assigned to the measurement result depends on any particular assumption. As Duhem points out, any discrimination is initially underdetermined, as the logic of the experimental test is that the result bears upon the group of assumptions and theories as a whole. Consequently, there is no indication, absent the investigation of uncertainty, as to which assumptions strongly influence the measurement result from those that do not, or which hypothesis or assumption should be revised following a negative result or supported following a positive result.

Duhem does argue for the existence of a solution to the problem: 'good sense'.

"Pure logic is not the only rule for our judgments; certain opinions which do not fall under the hammer of the principle of contradiction are in any case perfectly unreasonable. These motives which do not proceed from logic and yet direct our choices, these 'reasons which reason does not know' and which speak to the ample 'mind of finesse' but not to the 'geometric mind,' constitute what is appropriately called good sense." (Duhem, 1954 [1908], p. 217).

As many authors have pointed out, Duhem's solution is too vague to be of use as it does not specify criteria or a methodology to determine discrimination or revision. As Gillies has argued, "Duhem's theory of good sense [is] more in the nature of the problem, or a starting point for further analysis, than of a final solution to the difficulty with which it deals" (Gillies, 1993, p. 108). Rather than offering an epistemically justified method, Duhem points to evidence for the existence of a process of discrimination by gesturing to the history of science where many choices, that were underdetermined due to non-separability in experimental testing, were ultimately supported by empirical evidence (Duhem, 1954 [1908], p. 217–218). This, however, is a post-hoc justification.

Duhem's claim to the existence of a solution departs from what has been attributed to Quine as his (in)famous denial of a solution to the problem. Quine claimed that:

"any statement can be held true come what may, if we make drastic enough adjustments elsewhere in the system. Even a statement very close to the periphery can be held true in the face of recalcitrant experience by pleading hallucination or by amending certain statements of the kind called logical laws. Conversely, by the same token, no statement is immune to revision." (Quine, 1951, p. 40, p. 40).

The denial of the solution is a consequence of the holist 'web-of-beliefs' picture often associated with Quine, where possibility to always adjust within a system prevents discrimination amongst the group of statements (or hypotheses and assumptions) which make up the system. It is important to note here that holism follows from non-separability only if one also denies the possibility of discrimination. Duhem does not deny the possibility of discrimination (in principle escaping holism), however offers no method. The uncertainty practices identified by Beauchemin show how discrimination can be guided by method, so that the escape is not merely in principle, but in practice.

In order to work towards an 'in practice' solution, we adopt an approach advocated by Mayo (1997) who argued that Duhem's problem is a problem regularly faced by scientists who also very often develop methodologies, using what she calls "error statistics" (p.229), which seek to solve Duhemian problems. Mayo's discussion highlights the potential of examining the methods of working scientists in which local epistemic solutions can be located, rather than pursuing a global solution. By similarly starting from an analysis of methods adapted to the particular epistemic problems posed by high energy physicists, we identify a robustness analysis method whereby the epistemic risk introduced by Duhem's problem is managed.

We identify two important feats of the methodology outlined by Beauchemin which, together, provide an in practice, or epistemic, solution to Duhem's problem. The first of these feats is identification of the assumptions on which the result is dependent. As was outlined in section 4.1, the intention of robustness analysis is to locate the "epistemically trustworthy" (Wimsatt, 1981). The systematic uncertainty evaluation in stage 3 identifies the space of possible assumptions on which the result is dependent. By identifying a definite group of relevant assumptions, the scope of the potential holism is determined not to be global. This alleviates (without eliminating) the concern of the possibility of a problematic assumption remaining unknown (as there is no effective procedure to generate the alternative model assumptions). This also aligns with the significant body of work coming from the 'New Experimentalists', who showed the piecemeal nature of experimental testing (for example, see Mayo (1996)).

Note that in delimiting the group of relevant assumptions, we can also see an epistemic justification to a claim made by Duhem. Duhem claimed that the experimenter "makes use also of a whole group of theories accepted by him as beyond dispute" (Duhem, 1954 (1914), p. 185). Duhem here alludes to a local rather than a global scope for holism by placing limits on the size of the group. The robustness analysis method ensures what Morrison has called 'moderate scope' for the holistic thesis (J. Morrison, 2017, p. 3) by demonstrating the finiteness of the group, or the piecemeal nature of the experimental test, in practice rather than in principle.

At this point, identification has only limited the holism to its local version. Also required is discrimination amongst the local group of assumptions to escape holism altogether. This feat is achieved by determining the size of the contribution to the uncertainty of specific assumptions, thereby discriminating amongst the group of assumptions on the basis of how much the result cited depends on specific assumptions. The discussion in 4.1 focuses on the possibility to evaluate how much the measurand depends on any particular PDF assumption. Stages 2 and 3 show the general structure of how the uncertainty and sensitivity practices limit epistemic risk in theory testing and support.

The ability to discriminate amongst the group of assumptions manages the epistemic risk introduced by the possibility of evidential circularity due to the non-separability of the local group of assumptions required for experimental test in one of two ways. In the case where the hypothesis that is being tested is assumed in arriving at the result and, importantly, where the result strongly depends on the very same assumption, the uncertainty interval will be too large to allow for the experiment to confirm or disconfirm the hypothesis that is being tested. This allows for the eschewal of viciously circular argumentation. The alternative case, where the hypothesis that is being tested is one that is assumed in arriving at the result and where the result is determined to not be strongly dependent on the very same assumption, the uncertainty interval will be small enough to allow for the experiment to confirm or disconfirm the hypothesis that is being tested. This method allows for epistemically risky models of the measurement process to be discarded. It also enables theory-dependent measurements to target, in a piecemeal way, particular theoretical hypotheses, even the very same hypotheses upon which the results themselves depend.

Identification also ensures the viability of discrimination because, if global holism is assumed, every evidential use of measurement results would include inescapably viciously circular reasoning, as the scope of the group of assumptions and theoretical content in any experimental test is the 'whole of science'. Whilst there have been very few attempts to claim and argue for a thesis of global holism, the concern remains that the local group of assumptions may be larger than supposed as a result of a lack of awareness of assumptions (such as implicit assumptions which may not be initially apparent to the experimenter). This highlights the importance of the identification of the local group of hypotheses that are required for an experimental test or measurement. Identification is crucial for the effectiveness of discrimination as a strategy to manage epistemic risk.

The robustness analysis solution corresponds nicely to the negative claims made by Duhem, that logic cannot provide a solution to his problem. Non-separability introduces epistemic risks that cannot be eliminated. However, we have here provided an epistemically justified account of how such risks can be managed, and holism escaped, in practice. In this account the 'good sense' of the experimenter is in making judicious choices regarding the extent to which their inferences are exposed to risk.

5. Conclusion

We have sought to identify an epistemological problem that arises in the context of analyzing data for the purposes of deriving a measurement result, and that exemplifies a widely discussed feature of scientific inquiry – that experimental results or observations are dependent upon theoretical assumptions – the philosophical implications of which remain a matter of debate. Our arguments here cannot resolve that wider debate, but they do show that theory dependence that engenders evidential circularity in argumentation need not constitute an epistemic defect.

It is easy to miss this point. That circularity is an epistemologically threatening consequence of theory dependence is such a widely held assumption that it is rarely articulated. To be sure, there is good reason to worry about circularity in argumentation, but we have argued that the practices of evaluating uncertainty and sensitivity employed in measurements in HEP enable investigators to differentiate between benign and vicious instances of circularity and to reduce significantly the risk posed by the latter. We have further argued that these same practices constitute a means of responding to Duhemian problems of underdetermination as they arise in measurement experiments in HEP.

The response to problems of evidential circularity and underdetermination that we have investigated here does not take the form of a philosophical theory of confirmation or empirical support. It is instead a practical response. A proper philosophical appreciation of this requires understanding the epistemological basis of the diverse and widespread practices of uncertainty evaluation employed by scientists using data to generate measurement results.

Perhaps because these practices seem so mundane and remote from the theoretical heights of theories of confirmation, philosophers have long neglected them. We hope to have contributed to the gradual elucidation of the epistemological efficacy of scientific inquiry that comes from attending to the details of scientific practice.

CRediT authorship contribution statement

Sophie Ritson: Conceptualization, Investigation, Writing - original draft, Writing - review & editing. **Kent Staley:** Conceptualization, Investigation, Writing - original draft, Writing - review & editing.

Acknowledgements

This research grew out of discussions at the 2018 Summer School on the Philosophy, History, and Sociology of Particle Physics, organized by Florian Boge and Adrian Wüthrich for the Epistemology of the LHC research unit. Three anonymous referees commented generously and insightfully on previous drafts of this paper and we are grateful for their help in improving the paper. We are especially grateful to Hugo Beauchemin for his helpful feedback on an earlier draft and for helping us to better understand the relevant physics. Sophie Ritson's work was supported by the Austrian Science Fund (FWF): I 2692-G16.

References

Ackermann, R. (1985). *Data, instruments, and theory: A dialectical approach to understanding science*. Princeton, NJ: Princeton University Press.

ATLAS. (2011). Measurement of the W charge asymmetry in the $W\to\mu\nu$ decay mode in *pp* collisions at $\sqrt{s} = 7$TeV with the ATLAS detector. *Physics Letters B, 701*(1), 31–49. https://doi.org/10.1016/j.physletb.2011.05.024

ATLAS. (2012a). Study of jets produced in association with a *W* boson in *pp* collisions at $\sqrt{s} = 7$TeV with the ATLAS detector. *Physical Review D (Particles, Fields, Gravitation and Cosmology), 85*(9). https://doi.org/10.1103/PhysRevD.85.092002

ATLAS. (2012b). Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Physics Letters B, 716*(1), 1–29. https://doi.org/10.1016/j.physletb.2012.08.020

ATLAS. (2016). Measurement of $W^{\pm}$ and Z-boson production cross sections in *pp* collisions at $\sqrt{s} = 13$ *TeV* with the ATLAS detector. *Physics Letters B, 759*, 601–621. https://doi.org/10.1016/j.physletb.2016.06.023

ATLAS. (2017). Precision measurement and interpretation of inclusive W+, W- and Z/Y* production cross sections with the ATLAS detector. *The European Physical Journal C, 77*(6), 367. https://doi.org/10.1140/epjc/s10052-017-4911-9

ATLAS. (2018). Measurement of differential cross sections and W^+/W^- cross-section ratios for boson production in association with jets at $\sqrt{s} = 8$ TeV with the ATLAS detector. *Journal of High Energy Physics, 2018*(5), 77. https://doi.org/10.1007/JHEP05(2018)077

Beauchemin, P. H. (2017). Autopsy of measurements with the ATLAS detector at the LHC. *Synthese, 194*(2), 275–312. https://doi.org/10.1007/s11229-015-0944-5

Cartwright, N., Shomar, T., & Suárez, M. (1995). The tool-box of science. In W. Herfel, W. Krajewski, I. Niiniluoto, & R. Wojcicki (Eds.), *Theories and models in scientific process* (pp. 137–150). Amsterdam: Rodopi.

CMS. (2012). Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Physics Letters B, 716*(1), 30–61. https://doi.org/10.1016/j.physletb.2012.08.021

Duhem, P. M. M. (1954). *The aim and structure of physical theory (P. Wiener, trans.)* (1914). Princeton: Princeton University Press.

Franklin, A. (1986). *The neglect of experiment*. New York: Cambridge University Press.

Franklin, A. (2017). The missing piece of the puzzle: The discovery of the Higgs boson. *Synthese, 194*(2), 259–274. https://doi.org/10.1007/s11229-014-0550-y

Gillies, D. (1993). The Duhem thesis and the Quine thesis. In *Philosophy of science in the twentieth century* (pp. 98–116). Oxford: Blackwell Publishers.

Glymour, C. (1980). *Theory and evidence*. Princeton, NJ: Princeton University Press.

Hacking, I. (1983). *Representing and Intervening: Introductory topics in the philosophy of natural science*. Cambridge: Cambridge University Press.

Howson, C., & Franklin, A. (1985). A Bayesian analysis of excess content and the localisation of support. *The British Journal for the Philosophy of Science, 36*(4), 425–431. https://doi.org/10.1093/bjps/36.4.425

Karaca, K. (2013). The strong and weak senses of theory-ladenness of experimentation: Theory-driven versus exploratory experiments in the history of high-energy particle physics. *Science in Context, 26*(1), 93–136. https://doi.org/10.1017/S0269889712000300

Kosso, P. (1989). Science and objectivity. *The Journal of Philosophy, 86*(5), 245–257. https://doi.org/10.2307/2027109

Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. Chicago: University of Chicago Press.

Mayo, D. G. (1997). Duhem's problem, the Bayesian way, and error statistics, or "What's belief got to do with it?". *Philosophy of Science, 64*(2), 222. https://doi.org/10.1086/392549

Mayo, D. G. (2018). *Statistcal inference as severe testing: How to get beyond the statistics wars*. Cambridge: Cambridge University Press.

Morrison, M. (1999). Models as autonomous agents. In M. S. Morgan, & M. Morrison (Eds.), *Models as Mediators: Perspectives on natural and social sciences* (pp. 38–65). Cambridge: Cambridge University Press.

Morrison, J. (2017). Evidential holism. *Philosophy Compass, 12*(6), 1–16. https://doi.org/10.1111/phc3.12417

Odenbaugh, J., & Alexandrova, A. (2011). Buyer Beware: Robustness analyses in economics and biology. *Biology and Philosophy, 26*(5), 757–771.

Quine, W. V. O. (1951). Two dogmas of empiricism. In *From a logical point of view* (2nd ed., pp. 20–46). Cambridge, MA: Harvard University Press.

Staley, K. W. (2020). Securing the empirical value of measurement results. *The British Journal for the Philosophy of Science, 71*(1), 87–113.

Stegenga, J. (2012). Rerum concordia discors: Robustness and discordant multimodal evidence. In E. Trizio, T. Nickles, & W. Wimsatt (Eds.), *Boston Studies in the philosophy of science, 292* pp. 207–226). Dordrecht: Springer Netherlands, 2012.

Tal, E. (2012). *The epistemology of measurement: A model-based account*. University of Toronto (PhD Thesis).

Wimsatt, W. C. (1981). Robustness, reliability, and overdetermination. In M. B. Brewer, & B. E. Collins (Eds.), *Scientific inquiry and the social sciences* (pp. 124–163). San Francisco: Jossey-Bass.