Synthese (2017) 194:313–332
DOI 10.1007/s11229-015-0749-6

Experimenter's regress argument, empiricism, and the calibration of the large hadron collider

Slobodan Perovic[1,2]

Received: 8 April 2014 / Accepted: 10 April 2015 / Published online: 13 May 2015

Abstract H. Collins has challenged the empiricist understanding of experimentation by identifying what he thinks constitutes the experimenter's regress: an instrument is deemed good because it produces good results, and *vice versa*. The calibration of an instrument cannot alone validate the results: the regressive circling is broken by an agreement essentially external to experimental procedures. In response, A. Franklin has argued that calibration is a key reasonable strategy physicists use to validate production of results independently of their interpretation. The physicists' arguments about the merits of calibration are not coextensive with the interpretation of results, and thus an objective validation of results is possible. I argue, however, that the in-situ calibrating and measurement procedures and parameters at the Large Hadron Collider are closely and systematically interrelated. This requires empiricists to question their insistence on the independence of calibration from the outcomes of the experiment and rethink their position. Yet this does not leave the case of in-situ calibration open to the experimenter's regress argument; it is predicated on too crude a view of the relationship between calibration and measurement that fails to capture crucial subtleties of the case.

Keywords High energy physics · Higgs boson · Experiments · Experimenter's regress · Calibration

✉ Slobodan Perovic
perovicslobodan@gmail.com

[1] Department of Philosophy, University of Belgrade, Cika Ljubina 18-20, 11000 Belgrade, Serbia

[2] Department of History and Philosophy of Science, University of Pittsburgh, Cathedral of Learning, 4200 5th Avenue, Pittsburgh, PA 15260, USA

1 Empiricist philosophy versus philosophical sociology of experimentation

Various empiricist accounts (in a broad sense) have implicitly or explicitly identified different roles that experiments play in science. More recently, philosophers studying scientific experimentation have given up the view assumed in the logical empiricist understanding of science, namely that theory-testing plays an exclusive or at least the main role in experiments. A number of other roles, ranging from exploration, theory formation, and the discovery of physical or biological entities, have been pointed out and studied. These roles are often interrelated with the theory-testing. And a number of experimental strategies have been identified and explored in the last three decades, especially in the philosophical and historical study of physics.

What the majority of empiricist-minded accounts of experimentation have in common, despite their differences in understanding the structure and the role of experiments, is the view that the social forces and the social context in which experiments are designed and performed are not crucial to the validation process of experimental results by the application of various experimental procedures. Collins (1985, 1994, 2002) has challenged such a view over the last three decades. His account of the so-called experimenter's regress emerges from such criticism.

Collins explicitly rejected what he took to be the dominant empiricist view of experimentation, offering instead a philosophically motivated sociological analysis:

> One must distance oneself from the standard view of experimentation in science and escape from the railroad of common sense to see the conventional nature of this reconstruction of what really went on in an experiment. […] Scientists and others tend to believe in the responsiveness of nature to manipulations directed by sets of algorithm-like instructions. This gives the impression that carrying out experiments is, literally, a formality. (Collins 1985, pp. 75–76)

He purported to show in his elaborate case studies that there is no deep distinction between epistemological criteria and the social forces that guide experimental research, where philosophical analysis could single out the former as an independent ground for validation of the experimental results (Collins 2002, p. 157). Following this assessment, he nevertheless optimistically sets a desired neutrality of scientific enterprise as its regulative goal, even though this can never be fully achieved. (Collins 2002, p. 158).[1]

The argument is partially based on an assessment of replication as a crucial piece of any argument for the validity of the results of an experiment. The replication of an

[1] The broader philosophical motivation for advocating such a view, especially the claim that epistemological criteria cannot be disentangled from the social context, is motivated in part by Wittgenstein's mature philosophical views and the understanding of epistemological criteria it condones (Collins 2002). Yet focusing on Wittgesteinian traits of Collins's argument does not answer the question of whether Franklin's reading of the argument as hopelessly social constructionist is correct. For gesturing to late Wittgenstein's philosophy implies little more than a vague view that social practices and sound epistemological criteria are entangled. Wittgenstein himself spent decades trying to understand either how they could be entangled, or why they cannot be disentangled even for the purposes of the analysis—depending on which interpretation of his philosophy one subscribes to.

experiment is never a guarantee of its validity, the argument goes, since experimental techniques do not consist of clear-cut procedures and rules. Whether a repeated experiment is done properly can be checked only with another test, and this further test can only be checked with yet another one, and so on. Experimenters have different strategies for circumventing such a regress, perhaps the most prominent being calibration and the use of surrogate signals, but in any case "[i]t is crucial to separate the simple idea of replicability from the complexities of its practical accomplishments" (Collins 1985, p. 19). Thus, although replication is understood as an axiom of scientific practice, it is non trivial. And although this becomes very apparent in the case of complex, and especially controversial experiments—upon some of which Collins has focused his analysis—, it defines the experimental practice as such. In disputed cases of experimental work it becomes apparent that replicability is not a matter of pre-established experimental procedures; it "really isn't a matter of experiment" (Collins 1985, p. 20). The test or replicability *as such* does not bring about agreement on the validity of results.

The point about replicability is complemented by the claim that experimental practice is a matter of skills, the transfer of which is an inherently social phenomenon. "Experimental ability […] cannot be fully explicated or absolutely established" (Collins 1985, p. 73). The proper workings of the apparatus are defined solely by the ability of the experimenter to produce proper experimental outcome (Collins 1985, p. 74). Given this, the "standard view", according to which a scientist can become trained in experimental procedures and then go on to apply them successfully to repeat various experiments, is simplistic and misleading. Thus, there is neither an external objective procedure nor a set standard that can properly fix the apparatus and then deliver valid results: fixing the apparatus properly is intrinsically related to the ability to deliver the proper result. What counts as a correct outcome of an experiment depends on whether the phenomenon that one searches for exists. And in order to find out whether the phenomenon exists one needs to build a good apparatus to check. "But we won't know if we have built a good detector until we have tried it and obtained the correct outcome! But we don't know what the correct outcome is until […] and so on *ad infinitum*" (Collins 1985, p. 84). This is what Collins labels *experimenter's regress*.

Typically, in an experiment universally accepted criteria of the quality of an experiment emerge. But the fact that disagreements do not occur very often does not mean that the real nature of the agreement on the proper workings of the apparatus and proper outcomes is not revealed by analysis of disputes—such as the dispute over the gravity waves experiments on which Collins has elaborated. A crucial test cannot on its own break the regress circulating between attempts to establish good apparatus and deliver good results on its own. It is recognized as a crucial test, and its procedures as valid, if some way is found to break the circle beforehand. A wide agreement on a criterion independent of the outcome of the experiment must be established, but the experimental procedures do not provide the required objectivity and universality on their own.

Franklin has criticized Collins's argument as detrimental to our confidence in scientific knowledge (Franklin 1990, 1994, 1997, 2013), as it casts principled doubt on the validity of widely-accepted experimental results, and not only on the extent of certainty that characterizes them and that scientists regularly debate. Thus Collins's

analysis suggests that the experimental process "leads to the experimenters' regress in which a good detector can only be defined by its obtaining the correct outcome, whereas a correct outcome is one obtained using a good detector" (ibid. 471). And if there are no criteria for obtaining a valid result or a good apparatus independent of the outcome of the experiment, then the experimenters' regress is a vicious circle that only social forces can break.

I will argue later on that understanding experimental practice along the lines of Collins's argument does not necessarily lead to such an unwelcome conclusion. The blurred line between the measurements and calibration and other techniques that provide a well-functioning apparatus is in fact sometimes a matter of actual practice that does not imply a vicious experimenter's regress.

Franklin, however, offers a staunch empiricist alternative to what he perceives to be Collins's social constructivist view of experimental practice. He argues that experimental physicists have developed a number of strategies in order to assess the results their instruments deliver, which provide reasonable grounds for validation of the results. Over more than two decades Franklin has compiled a list of such procedures through an analysis that makes the structure and the use of the experiments apparent. A list of these strategies can be found in Franklin (2013, p. 242). They range from intervention and confirmation in different experiments, use of theoretical explanations and statistical arguments, to calibration, reproduction of artifacts that are known to be present, and the use of the results themselves to argue for the validity of the results. We will deal with the latter three when they are applied simultaneously; they all use predictions of known phenomena to argue for the proper operation of the apparatus and thus validity of the results (ibid. 190).

Franklin's list of strategies (Franklin 2013, pp. 242–243) is not exhaustive, and nor does it provide necessary and sufficient conditions for the validity of experiments. The point of the analysis is that the strategies are used in practice and that they are central for accepting or rejecting the experimental results. This is why the case studies of experimental practice play a major role in Franklin's empiricist argument. And the methodological disagreement with Collins on what exactly should be analyzed in the case studies is not surprising: Franklin insists that the best form of the arguments is presented in the published work, while Collins (1985, pp. 169–174) insists that interviews with the scientists reveal the key aspects of the practice that the polished, and thus impoverished published accounts leave out. Ideally, one should weigh up the evidence based on both kinds of sources.

2 The epistemological significance of calibration

Among the strategies emphasized by Franklin and other empiricist-minded authors that are supposed to validate the results, calibrating procedures constitute an indispensable and central component of any experimental process. They are supposed to ensure proper functioning of the instrument and enable experimenters to rely on them in delivering valid results. While the nature of such procedures varies from one experiment to the next, as well as across disciplines, and despite the fact that they can become very complex in large complex instruments such as the LHC, some authors, including

Franklin and Collins, believe that certain epistemic properties are common to even such diverse procedures that range across experimental contexts.

Broadly conceived, *any combination of experimental techniques that ensures the proper functioning of the apparatus based on already-known phenomena* may be characterized as calibration. This seems to be a thread that connects a number of different techniques for ensuring that the apparatus works properly. Many such techniques and the ways in which they are combined have been described by Franklin (ibid.), as well as Collins. Franklin (1994) sometimes identifies calibration in a somewhat narrower sense, where simply a properly functioning apparatus is supposed to detect artifacts that are known to be present in advance, but he goes on to relate calibration with a broad notion of checking the apparatus by using surrogate signals in more complex ways (1997). As we will see shortly, he labels (Franklin 1997, p. 36) 'extended calibration' different combinations of various strategies, e.g. calibration in the above-described narrow sense and the use of the results to validate these very results. He correctly points out that "[c]alibration is more than just the use of a surrogate signal to test whether an apparatus is operating properly. In an extended sense, calibration also includes aspects of the general issue of how one argues for the validity of an experimental result" (Franklin 1997, p. 33). Through the analysis of calibrating techniques at the LHC we will refine the notion of calibration and its different aspects following this view.

We will see that the term calibration is used in this broader sense in practice, including the experiments at the LHC.[2] The techniques used in other disciplines, such as biology or cognitive science, seem to rely on such a characterization as well. We cannot provide necessary and sufficient conditions of what constitutes calibration in this broad sense, and reasonable disagreements on whether a particular experimental technique counts as calibration are possible. And calibration may only be a label used by scientists to aggregate diverse techniques that have essentially different goals and properties after all. But the commonly-used characterization described above, and its specific instances, as much as it may be unsatisfying once we look at the intricacies of particular calibrating techniques, is the best starting point from which we can begin to understand the role these techniques play in the experimental process. In order to avoid reducing our analysis to a squabble over words, actual practice should be the guide for understanding the common thread among various calibrating techniques and for the way we define it. And a general characterization of calibration should be assessed primarily by the kind of insight it gives us into the nature and the structure of the experimental practice, and especially by the extent to which it clarifies whether and how exactly such experimental techniques validate the results.

Similarly to Franklin, Collins initially gives a straightforward preliminary definition of calibration as a familiar procedure of standardizing an apparatus (Collins 1985, p. 101), and also states that "[c]alibration is the use of a surrogate signal to standardize an instrument" (Ibid. 105). He invokes the calibration of a new voltmeter that still lacks scale: typically another voltmeter is used in order to determine gradients by producing known voltages and observing the behavior of the needle. Collins points out that this

[2] We will see that a number of techniques applied at the LHC can be defined as calibration precisely in light of sharing this particular key aspect.

seemingly simple example conceals the fact that even in such a simple procedure something more significant occurs: "the standardized voltages are used as a surrogate for as yet unmeasured signals" (ibid. 102). The assumptions in the calibration, such that the unknown voltages act the same way on the new instrument as standard voltages, need not be reflected upon in simple cases, but they can become much more significant in more complex experiments. For example, in particle physics, we never directly see a particle that we expect the apparatus will produce when appropriate change in energy level occurs. Rather we rely on a measurement itself (or on a simulation) and its comparison to previous measurements to recover the expected energy indicative of the known particle. (Similarly, in the voltmeter case, we compare the result of applying the current to the new voltmeter to the results of previous measurements on other apparatus.) In this rather basic sense, measurement and calibration are already inextricably intertwined. In addition, calibration procedures that initially try to produce well-known phenomena often turn into novel measurements, if nothing else because of improved detection apparatus and techniques.[3] This was also the case at the LHC, as we will see shortly.

Collins goes on to argue, however, that there is a more substantial and problematic connection between calibration and measurement, along the lines of his general argument. Thus, calibration is not a standard procedure that can as such break the experimenter's regress. In Weber's gravity waves search upon which Collins focuses, electromagnetic waves were used to mimic gravity waves to check whether the apparatus was working properly. A number of assumptions are taken on board when calibration is performed, and the limitations of such a technique depend on the extent to which the signals can be treated as proper surrogates. In other words, they act as restrictions on interpretation of the outcomes (Collins 1985, p. 103). Thus, the assumptions of what constitutes good calibration are coextensive with the assumptions of what constitutes good outcome. Thus, for instance, whether one will be satisfied with the electromagnetic waves as surrogates of gravity waves will depend on what one thinks gravity waves are, and thus what could be expected to be detected with the apparatus (Collins 1985, pp. 104–105). Typically, the assumption of the near-identity between the surrogate signal and the signal potentially produced by an investigated phenomenon is uncontroversial, so it and its role are hard to notice.

> Calibration can only be performed provided this assumption is not questioned too deeply. In fact, the questioning is constrained only by what seems plausible within the state of the art of the science in question. But the very act of using a calibration surrogate may help to establish the limits of plausibility. (Ibid.)

Thus, whether an instrument is well calibrated depends on what one thinks of the results it delivers, and *vice versa*. "Calibration is not simply a technical procedure for closing debate" on the nature of the phenomenon at stake, "by providing an external criterion of competence. In so far as the calibration works this way, it does so by controlling interpretive freedom. It is the control on interpretation which breaks the circle of the experimenters' regress, not the 'test of a test' itself" (Collins 1994).

[3] Franklin and Howson (1984) explain why experimenters prefer to vary rather than simply reproduce well known phenomena.

Collins's conclusion is that an internal agreement's being hammered out using effective rhetorical devices, and its being coextensive with the agreement on the results, is the reason why a calibrating procedure gets accepted in the first place. Any decision made about the adequacy of calibration, and thus the validity of the results, is not based on entirely objective epistemic criteria external to the interpretation of the results. For instance, appealing to theoretical reasons to judge the validity of an experimental procedure such as calibration, and thus breaking the regress, is on a par with appealing to other reasons external to the experimental process.

Franklin has argued extensively against such a view of the role of calibration, while developing an account of calibrating strategies along the lines of his above-described empiricist-minded view. Contrary to Collins, he states that "[i]t is usually the case that calibration of the apparatus is independent of the phenomenon one wants to observe" (Franklin 1994, p. 485). The role of calibration is to ensure the validity of the result: "Calibration does not guarantee a correct result; but its successful performance does argue for the validity of the result" (Franklin 1997, p. 76). Franklin exemplifies this by discussing, among other cases, various examples where an apparatus in particle physics produces a known phenomenon recorded with a different detector in similar circumstances. In the cases he deems typical, "the calibration of the apparatus did not depend on the outcome of the experiment in question" (Franklin 1994, p. 487). He thinks this is typical of calibrating procedures in physics, while he regards as atypical those procedures that one could justifiably question (e.g. in the case of the controversial search for gravity waves).[4]

This point of the disagreement on the role of calibration, and consequently the nature of the validation of experimental results in the most complex physics experiments to date, is epistemically most relevant. We will analyze innovative calibrating techniques at the LHC and look into the intricacies of those techniques that require conceptual refinements. It turns out that there is an interesting relationship between calibration and measurement, requiring us to rethink both the experimenter's regress argument and empiricist-minded accounts of experimentation. In fact, as I will argue in Sect. 6, the experimenter's regress argument is predicated on too crude an account of the relationship between calibration and measurement to capture the case of an in-situ calibrating procedure and the theoretically—and technically-motivated agreement driving it. The case is best accounted for by a broad empiricist approach, identifying reasonable experimental strategies for validating results with a relaxed version of a principle insisting on the independence of calibration from measurement outcomes.

3 The experiment: the mass of the top quark, the Higgs boson, and calibration of the LHC

The Standard Model (SM) of particle physics predicts existence of the top quark but it does not predict the exact value of its mass (Mt). Its mass is a free parameter the value of which has to be determined empirically. This fundamental parameter of the Standard Model is interrelated with many other fundamental parameters of the SM.

[4] He also suggests strategies other than calibration that brake the experimenter's regress.

For example, its value can significantly affect predictions for determining the mass of the W and Z bosons, via radiative corrections (higher order loop effects) to the value of these masses. Since Mt is one of the key inputs to electroweak fits, consistency criteria with other measured and/or predicted Standard Model parameters can be used to restrict the range of values that Mt can take. However, nothing in the SM can determine exactly what Mt can be.

The Mt can be measured via the production of a single top-quark in weak interactions, or via the production of top quark and anti-quark pairs ($t\bar{t}$) in hadronic strong interactions (i.e. collisions of protons) (CMS Collaboration 2012, 2013; ATLAS Collaboration 2012; Borjanovic et al. 2005; Jovicevic 2013). The produced $t\bar{t}$ pairs rapidly decay to final state formed by so-called jets, some of which come from b-quarks, in addition of leptons, depending on the decay channel considered. The pair production is a preferred final state for Mt measurement because it offers more constraints, the parameters are well known or can be reliably estimated and as such provide a variety of ways to simultaneously measure the value of Mt.[5] The analysis of $t\bar{t}$ pairs produced in pp collisions begins with a comparison between kinematic properties of the particles that result from the pairs' decay, i.e. "top events", on the one hand, and theoretically predicted (via simulations) kinematic properties for different values of Mt on the other. Finally, the top quark mass is reconstructed through an analysis of all suitable lepton+jets top-quark events of interest, after background subtraction.

Now, "[a]part from the detailed measurements of the top quark properties, the huge amount of events … allow[s] to use top quark for calibration and commissioning purposes" (Ibid.). As we will see very shortly, the key to calibration in order to obtain a consistent value for Mt are jet energy scales and so-called b-tagging, the procedures that interest us most and that we will explain in some detail.

Thus, four different tasks have been performed with respect to top quark occurrences at the LHC. First, the initial calibration in a broad sense, of the LHC based on past Mt data, mainly those obtained at Tevatron. The second was the precision measurement of its mass. Third, top quark mass reconstructions were used for continuous fine-tuning of the energy scale of quark jets. And finally, the value of Mt was used as the key constraint in estimating the mass of the Higgs boson from its coupling with the top quark and other particles.

The past measurements, among which those performed at Tevatron were especially valuable, indicated that the top quark exists within a fairly narrow mass-energy range. In order to calibrate the LHC and thus validate the process of determining whether a novel phenomenon at a particular energy is an artifact, noise, or a genuine phenomenon (e.g. whether the expected signatures of the Higgs boson are genuine), past data concerning a well-known phenomenon such as the top quark are used as calibration values. Thus, this particular kind of calibration aims at an energy range where in the past, on other colliders—mainly Tevatron—the top quark has turned up. A similar procedure was undertaken with the data on Z bosons (Jovicevic 2013). Now, after the top quark has turned up within the expected mass-energy range, the measurements of

[5] In addition, there is more statistical data coming from this channel, allowing for a more precise disentanglement of signal and backgrounds as harsher cuts can be applied.

that value continued. This was one of the key results of the LHC experiments (CMS Collaboration 2013, p. 4; ATLAS Collaboration 2012).

One of the goals of these measurements, other than determining Mt as one of the fundamental values of the SM, was the application of the results in attempts to determine the mass of the Higgs by estimating it based on, or rather constraining it by, the top quark mass values and the measurements of the electroweak parameters (Van Mulders 2010, p. 13; Jovicevic 2013): "[t]he Higgs boson mass can be constrained by the top quark mass because of the large Yukawa coupling between the two elementary particles" (Van Mulders 2010, p. 14).[6] And the fit to the electroweak data, assuming the SM, is a function of the top quark mass and the Higgs boson mass.

4 Extended calibration

The use of the previously determined values of Mt was far from being the only kind of calibration performed at the LHC. In fact, calibration, broadly understood in the way we characterized it earlier, permeated the entire experimental process, from the calibration of the electronics of detectors, to the offline reconstruction of the events, to the analysis of data. The constraints on the kinematics that we have mentioned previously, which are used to derive desired quantities like masses of particles, are used for calibration of the detectors. In addition, the simulations of the events are corrected and tuned to the data obtained in the experiment. Finally, as Franklin (2013) points out, the LHC reproduced almost the entire zoo of elementary particles discovered in the twentieth century, including the top quark. Analogously to the calibration and measurements based on Mt, these calibrating procedures, which initially reproduced various well-known particle properties (e.g. the mass of Z boson, its transverse momentum, W/Z+ jets cross sections, etc.), have turned into precision measurements of such properties due to improved detecting apparatus and techniques.

Franklin characterized this extensive production of known particles as an instance of *extended calibration*. But calibration can be extended in various atypical ways. Thus, he states that "calibration in an extended sense", which involves checking the backgrounds with various procedures that may mask or mimic the desired effects, "although difficult, is not impossible" (Franklin 1997, p. 75). For instance, he discusses the measurement of the Ke2+ mesons branching ratio as one such case. Thus "the decay positron resulting from the decay [of mesons] was to be identified by its momentum, its range in matter, and by its counting in a Cherenkov counter set to detect positrons. The proper operation of the apparatus was shown, in part, by the the results themselves"—if the sought-after phenomena existed they would have to be detected by the apparatus, since there was no plausible source of background mimicking or masking the sought-after events (Franklin 1994, p. 486). The result confirmed not only that the phenomena did not exist but also that the apparatus was working properly, as masking effects could have occurred only if the apparatus was *not* working properly. And the occurrence of Kμ2+ could potentially mask the occurrence of Ke2+

[6] Because the Higgs boson couples with the top quark, in combination with the W boson mass, we can determine the Higgs' mass.

mesons that the physicists aimed to detect. Typically, masking signals are not used to calibrate the machine because of uncertainty about whether they or the targeted event has been detected. But in this very unusual case they were used to check whether the machine was operating properly and to calibrate the momentum resolution, because the probability of the occurrence of the desired event (Ke2+) was extremely high in comparison to the occurrences of the masking event (Kμ2+). Despite the unusual use of the results in the calibration procedure, which was enabled by unusual experimental circumstances, Franklin nevertheless lists this case with those that he thinks demonstrate independence of calibration from the studied phenomenon. What physicists thought of the phenomenon they were searching for did not impact the validity of the calibrating procedure—the procedure was valid irrespective of the sort of data that would be produced by the experiment, and was also calibrated on positron data produced by another apparatus.

The calibration of the LHC with the Tevatron data on Mt includes an extended calibration where the construction of the results and calibration procedures are not independent, and nor do they deal with entirely independent processes; they are interrelated in an iterative procedure to a much higher degree than the empiricist-minded account of Franklin's type assumes. Yet, as we will argue in Sect. 6, the close relation between calibration and measurement need not necessarily worry the empiricists, as the iteration between the two does not exhibit the vicious regression.

5 The commissioning calibration of the LHC, in-situ calibration, and the top quark mass (Mt) reconstruction

The calibrating procedure starts at the commissioning phase (Fig. 1), but continues along with the measurements and the reconstruction of Mt. "The top quark mass measured at the Tevatron will be used as an input to estimate the most probable jet energy corrections that fit the data at the Large Hadron Collider" (Van Mulders 2010, p. 15). However, only initially, during the *commissioning phase*, are Tevatron data alone used for the calibration of the jet energy scale (CMS collaboration 2008; Acharya et al. 2008) and as such could validate the data model obtained by the subsequent measurements of Mt. This matches the calibration procedures Franklin focuses on.

During the second, much longer stage (Fig. 1), *there is a dynamic dependence between the measurement outcomes and the calibration*. In a nutshell, the calibration of the jets is a set of procedures continuously supplied by updated parameters within the measurement model, which includes the latest results obtained at ATLAS and CMS (Fig. 1). The in-situ calibration is done through jet energy scale corrections for the top quark (quark/anti-quark interactions) and through b-tagging.[7] These are two instances of techniques for calibrating the jets that are subsidiary to the measurements. In one sentence, before we unpack the details: "Assuming the Standard Model prediction B(t $\rightarrow$ bW) $\sim$ 1, the heavy flavour content of $t^{-}t$ events is well predicted, which allows to

[7] "Both the commissioning as well as the measurement of b-tagging efficiencies and jet energy corrections from data are crucial for the searches for new physics phenomena." (Van Mulders 2010, p. 15)

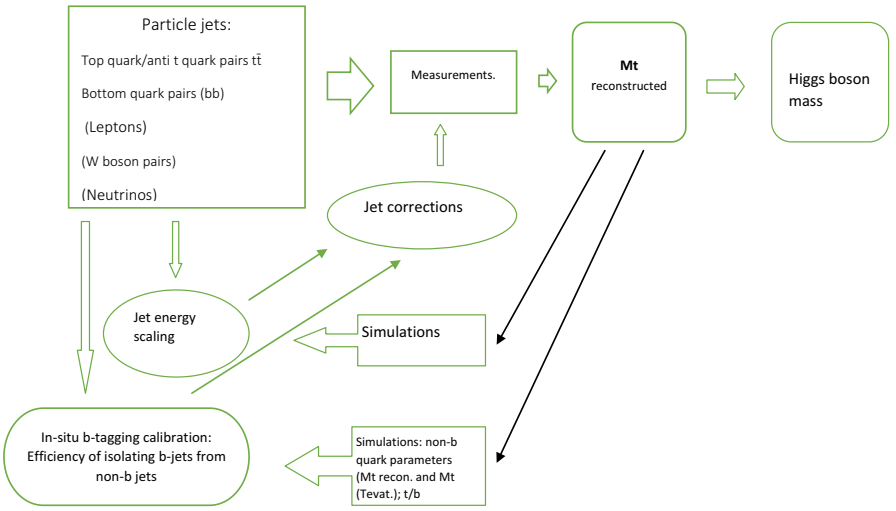

Diagram 1 A schematic representation of calibrating procedures at the Large Hadron Collider

calibrate and measure the efficiency of b-tagging algorithms and jet energy corrections from the produced top quark pair events" (Van Mulders 2010, p. 15).

The calibrating techniques that the physicists label 'in-situ calibration' (Schouten 2007, Chap. 5; Van Mulders 2010; Jovicevic 2013) are certainly more complex than calibration as standardization of simple experimental instrumentation. But it would be hard not to qualify these techniques as a form of calibration based on the broad definition of calibration we introduced earlier. They are actually a prime example of extended calibration and its complexities.

It should be pointed that sometimes, in the literature, the notion of in-situ jet corrections refers to data production, while calibration designates constants plugged into Monte Carlo simulations. But this is done within the context of a particular section of the analysis, and it is not surprising that the terms are mostly used interchangeably, and that the notion of in-situ calibration is used to capture the entire process—since it involves a complex interrelation of Monte Carlo simulations and measurements, i.e. data-derived and Monte Carlo-derived corrections and calibration constants. Thus, for example, the in-situ measurements cross-check the Monte Carlo values (Schouten 2007, p. 63), and conversely, the quality of jet calibration is cross-checked by Monte Carlo simulations (ibid. 69) This sort of hybrid of measurement and multifaceted calibration is well illustrated by jet energy scaling and b-tagging—to which we will turn shortly. One can insist on delimiting measurement and calibration for the purposes of analysis and finding out the relationships between different aspects of the process. But, as we shall see, measurements of that sort are performed for the purposes of the calibration, while plugging in the calibration constants is part of a subsidiary to the measurements. In addition, when one is performing a fitting routine with the Mt as a constraint, then one is using a statistical or simulation tool for calibration. To try and distinguish the use of the tool from the calibration in the context of understanding calibration is uninformative at best. The main point is that there is a much closer

connection between such processes than we see in the commissioning calibration, and the use of data and plugging in of the calibrating constants is much more pervasive. Simplifying the picture by forcing a strict division between two categories (in order to bypass the possible problem of the experimenter's regress) would not satisfyingly capture the nature of the process and its role. This situation is analogous to the blurred line between simulation and experiment at ATLAS and CMS that M. Morrison points out (Morrison 2014). Trying to pinpoint such a line is as problematic as attempting to define the in-situ procedures simply in terms of measurements.

As far as the jet energy scale goes, it "calibrates the measured calorimeter level jet energy to the particle level." (Schouten 2007, p. 44) On the one hand, the jet energy levels, crucial in Mt measurements, are corrected with the help of continuous in-situ calibration performed on the jets, and thus uncertainties concerning the resulting Mt are decreased. (Borjanovic et al. 2005; Blyweert 2012) On the other hand, the calibration procedure itself relies on constraints that include previous measurements of Mt performed both at Tevatron and LHC. (CMS Collaboration 2012, 2013; Cacciari et al. 2008)

"*The (mis)calibration of the Jet Energy Scale (JET) appears as an important source of systematic uncertainty on Mw and Mt.* The a priori knowledge of jet-energy calibration is about 10 %" (Acharya et al. 2008, p. 3).[8] In order to address this major source of systematic uncertainty, following the commissioning, the CMS collaboration has been producing, over the course of several years, data based on the application of the measurement model of Mt constrained by the mass of the Z and W bosons.[9] The outcome is *a reconstruction of Mt*. (Borjanovic et al. 2005; CMS Collaboration 2012) The top quark decays into a b quark and a W boson. Thus, the $t\bar{t}$ pair's final state results in a) both W bosons decaying hadronically, into jets, or b) into lepton plus jets when only one W boson decays hadronically, or c) into dileptons when both W bosons decay leptonically. The kinematic properties of these sets of final states are analyzed via twenty-four parameters. Fourteen parameters are inferred from measurements, namely three momenta of two charged leptons and two (light (u, d, s) and b) quark jets, and a missing energy of two neutrinos (since neutrinos do not interact strongly nor electromagnetically and cannot be detected). These are later reconstructed in simulations based on data obtained. Nine parameters are constrained: the world-value of the mass of the W boson, equalizing of the mass of top quark and the top anti-quark, and the masses of the six final state particles of those listed above (which are treated as the world values). Thus a number of hypotheses, each concerning one particle at a time (i.e. velocity of a particle, its mass being constrained), can be tested. As such, the jet energy scales are determined from reconstructions of this cluster of particle couplings i.e. information from which the mass of the Mt (as well as W boson and Mb) is derived.

Now, while the measurements of Mt and its reconstruction based on the constraining parameters is being performed, the top quark is used as the basic "instrument" for the

[8] Early commissioning calibration is done *in-situ*, based on the parameters of each detector, as well as with cross-checking detectors, where a parameter for a detector is calibrated against the parameters provided by other detectors.

[9] In order to extract the pole mass of the top quark the mass of W boson is constrained by world average value, a value obtained from all relevant experiments (with its corresponding uncertainty).

in-situ calibration at LHC, since the instrument has a large top quark/top anti-quark cross-section. "Semileptonic $\mathbf{t\bar{t}}$ events link all these items together [W and Z bosons, t and b quarks and anti-quarks, and leptons] and can therefore be used to make what is commonly referred to as an *in-situ* calibration" (Acharya et al. 2008). The measured energy of jets is influenced by various experimental and theoretical elements (e.g. jet's electromagnetic fraction, jet flavors, underlying events that contribute to its energy, etc.) with respect to which the corrections for each jet are estimated and finally, based upon these, the jets are corrected. The jet corrections are continuously being improved in order to increase the precision of the measurements of Mt and its reconstruction. And the more precise values of Mt, along with other above-mentioned constraints, are used in subsequent measurements.[10]

The second element central to calibration leading to Mt measurements is b-tagging. It relies on the fact that "[a] good way of identifying top quark events is tagging b-quark jets" (Jovicevic 2013, p. 56). This is necessary for identifying the top-quark with a small background and for reducing the combinatoric of top-quark kinematic reconstruction in the semi-leptonic or all-hadronic channels. Since almost every top-quark produces a b-quark in its decay, the jets that these quarks produce must be identified as b-jets via identification criteria called b-taggers. Such b-jet identification algorithms rely on the reconstruction of secondary vertices, i.e. by the presence of B-hadrons, which have a longer lifetime than other hadrons that do not contain a b-quark. The presence of a secondary vertex inside a jet thus means that this jet is b-tagged (Gutierrez 2007; CMS Collaboration 2013; Borjanovic et al. 2005).

The technique of isolating b-jets from other quark jets is used to calibrate b-tagging (identifying) algorithms it is the same technique that is used regardless of the $t\bar{t}$ decay channel. There are many discriminators for b-taggers that determine whether there is a secondary vertex in the examined hadron jet (Van Mulders 2010, Sects. 74–84). Secondary vertices are obtained from the convergence of tracks away from other objects reconstruction in vertex where the collision leading to the event is taken to have occurred. The tracks are reconstructed using the same procedure as that used in the primary Mt measurements, based on clustering and pixel signals that charged particles produce and on the estimates of their position and uncertainty (due to the uncertainty of the momentum in the inner detector). The b-tagging is obtained by testing the hypothesis that there is a secondary vortex in the collected signals similarly as what the simulation of b-quark features. By applying various selection cuts on the value of each discriminator used to identify secondary vertices, different figures of merit in terms of b-jet purity and efficiency can be obtained. This is where the constraint parameters enter the calibration process, including Mt from the previous reconstruction of Mt. The efficiency for b-tagging is estimated using $t\bar{t}$ events and the purity-efficiency working point is selected depending on which point yields the best precision for the measurement of Mt to be produced. Previous measurements can be used to further improve the knowledge of the parameters to be measured in the current Mt measurement, and precision steadily increases as more data becomes available.

[10] Similarly, the Monte Carlo simulations are tuned to data in order to provide appropriate corrections of parton collisions. (Van Mulders 2010, p. 83)

The latest reconstructed Mt is one of the constraint parameters in the current round of calibration and jet energy corrections (energy scales) rely on the reconstructed value of Mt that includes both Tevatron and the latest LHC value as a constraint. Thus, the jet energy corrections are constrained by the values obtained in the latest reconstruction of Mt, also including MW and estimates of b-jets tagging efficiency. It would be circular, of course, to use the Mt reconstruction from the current measurement and average it with the Mt value that is used as a constraint in the calibration. Rather, the procedure is analogous to a very long string of steadily improving measuring apparata where each new apparatus uses the best results of the previous one for the calibration relying on various constraints and parameters.[11]

Now reliance on the Tevatron data for the measured parameters, including the value of Mt, is steadily decreasing as the LHC measurement data grows, and the precision of the instrument improves, thus warranting the increased significance of the LHC component in the measurement model.[12] This uncertainty is initially minimized by the commissioning calibration to some extent. The long-term application of various calibration procedures and the vast amount of data recorded by ATLAS and CMS decrease uncertainty to a level that enables determining Mt—a level of precision that matches and exceeds that of Tevatron.

Statistically speaking, systematic uncertainties cannot be reduced with more samples. So why is this possible in the case of determining Mt? In this case, as in other similar cases, seemingly systematic uncertainties with respect to the so-called 'nuisance parameters' that are used in the discovery analysis (e.g. b-jets in the case of reconstructing Mt), "typically are reduced as we take more data, since the subsidiary analysis that calibrate them also benefit from data" (Cousins 2013, p. 16), as is the case in the identification of pure b-jets and jet energy scales and the resulting Mt reconstruction. The subsidiary analysis enables extended long-term calibration, without which the physicists would not be able to increasingly reduce systematic uncertainties with the increasing amount of samples.

The in-situ procedure was much improved at the LHC because of the structure of the apparatus. The Tevatron signal to background ratio is twenty times lower than that of the LHC so cross-check of different channels of decay (dileptonic, leptons plus jets, all jets) when measuring the mass is much more accurate. Also, physicists working on Tevtaron were not at all certain that the calibration technique based on b-taggers would be successful. "It is the author's hope", one said, "that by measuring the mass as a function of jet rapidity, or b-jets momentum, or the b-jets angle with the beam, etc. the LHC experiments will be able to reduce the systematic errors substantially" (Gutierrez 2007, p. 174) which is precisely what they did.

[11] For instance, a mercury-based thermometer can be calibrated with a constant volume gas thermometer, a completely different apparatus that works on a different physical principle. But it can also be calibrated against other mercury based thermometers by using containers made of materials with different thermal coefficients. The commissioning calibration is more closely analogous to the former case, while the in-situ calibration to the latter.

[12] "Currently the most precise measurements of Mt are performed by the Tevatron experiments, but the ATLAS and CMS results are getting more and more precise." (Blyweert 2012, p. 5)

6 Rethinking the empiricist account of experimentation

So how exactly should we characterize such a calibrating procedure in order to determine where it stands in relation to Collins's argument and Franklin's empiricism? The in-situ calibration is situated somewhere in-between the commissioning phase, which looks pretty much like standard calibrating procedures described by Franklin on the one end of a range, and the final phase of calibration via Mt that left the Tevatron data out of the measurement model on the other. It is a process that involves the results of the actual and earlier experiments in a complex way; one that does not quite fit the categories of calibration offered by Franklin's or Collins's approaches. They simply do not deal with the possibility of an on-going relationship between calibration and measurement such as we find in the LHC case. If we analyzed the case with their categories of calibration in mind, it would be easy to overlook both the nature of the challenge that the calibration in the LHC case addresses and the calibrating procedures that meet the challenge.

So what does this tell us about calibration and its relationship with the measurements more generally? First, does the close interrelation between calibration and measurement we find in the in-situ procedures indicate that the empiricists are wrong? Or to put it more precisely, is there anything in the procedures that distinguishes them from the kind of other reasonable strategies of validating results that Franklin and others have described? And second, as far as Collin's arguments go, is a systematic interdependence between calibration and measurement we described an argument for the inevitability of the experimenter's regress and its breaking by a wide external agreement rather than by independent calibrating procedures?

Franklin's general empiricist-minded point of view, which is supported by pointing to various experimental strategies for validating results, is generally on the right track for understanding experimentation. Yet Franklin's view that calibration of the apparatus does not depend on the outcome of the experiment (Franklin 1994, p. 487) and that "[i]t is usually the case that calibration of the apparatus is independent of the phenomenon one wants to observe" (Franklin 1994, p. 485), requires thorough rethinking in light of our case analysis.[13] The accounts of the practice in the cases that Franklin has analyzed do not offer a satisfying framework.

In the case of the LHC, the measurement outcome and the calibration are complex *reconstructions*, not a one-shot production of data and unrelated parameters. Both the outcome, such as measurements of particle masses, and the calibration techniques, such as those based on b-tagging and fine-tuning of jet energy scales, reconstruct and utilize shared parameters such as Mt, b-quark, $t\bar{t}$ branching, etc. And these reconstructions are long-term iterative procedures. In fact, the calibration (utilizing jet energy scales, b-tagging, Monte Carlo tuning, etc.) and measurements of desired phenomena (Mt) are *systematically co-extensive*. Thus, as we have seen in the previous section, the precision measurements of Mt count on the improvement of b-tagging efficiency, while the efficiency of b-tagging will rely on improved reconstructions of Mt. And the entire

[13] Although Franklin may not object to any of the points I will make individually, his account does not emphasize them and does not assimilate them into a general account or into conclusions concerning the relation between the calibration and the outcome.

in-situ calibration is a *subsidiary of the measurement*, not a fully independent process. Since these procedures are so closely interrelated, neither they nor their validity can be understood independently. Each new stage of calibration in this case depends on the previous outcome, as much as the outcomes depend on the calibration. The case suggests to us at very least that the idea that the calibration of the apparatus does not depend on the outcome of the experiment should be accepted only very cautiously and conditionally. The dynamics of the calibration in the LHC case is such that the point is valid without crucial caveats only in the commissioning phase.

So, in light of our discussion of this case, how far do we need to depart from the empiricist tenet concerning independence of calibrating procedures, if at all? Could, theoretically speaking, a very careful analysis of the presented case disentangle calibrating procedures from measurement outcomes? Looking at the case itself this claim seems very hard to substantiate. In fact, *the more detailed our knowledge of the procedure we have, the less adequate the tenet on which such an expectation is predicated seems. This has become particularly apparent, as we have pointed out, in the way data-derived and Monte-Carlo derived corrections and calibration constants are used in in-situ procedures* (See Sect. 5).[14]

Does the lack of squarely-independent calibrating procedures means that the described experimental procedure is open to the viciousness of the experimenter's regress? And, given the interdependence between measurements and calibration characterizing it, is it only external social forces, i.e. the implicit agreement between the experimenters, that breaks it? The categories Franklin's empiricist-minded analysis offers to us are not satisfying enough for the analysis of in-situ calibration and measurements it validates. But the fact that the reasons for accepting the calibrating procedures as a validation of the outcome cannot be neatly separated from the outcomes themselves does not support the experimenter's regress argument in our case.

First, as we have seen, Collins claims that there are no formal rules that guide calibration that could make a validating procedure epistemically independent of the outcome of the measurement. It may very well be true that there could be no formal criteria of that sort. Yet there are theoretical reasons, as well as technical reasons—those concerning particular processes occurring in the apparatus—that justify calibration. In fact, a requirement for the existence of formal rules as an independent epistemic basis for calibration is fundamentally misleading in the LHC case. The theoretical and technical reasons justifying the calibration in the LHC case are developed in an iterative process—not in a single shot, so to speak—where one theoretical advance and an aspect of the apparatus are constantly used as a leverage for another. Thus, the reconstructions of the outcome and calibration share parameters (e.g. reconstructed Mt and MW, Mt (Tevatron)), and these parameters leverage the iterations between the calibration and the outcomes. And the significance of various parameters changes over time; for instance, the Mt measured at Tevatron becomes increasingly less relevant as the experiment progresses, while the b-tagging efficiency becomes increasingly more

[14] This insight illustrates why the case-studies that identify various strategies of validation, rather than vacuous conceptual discussions and distinctions, are the most reliable and effective way of refining the empiricist viewpoint.

valuable for the Mt reconstructions. It is hard to see how theoretical and technical reasons behind such procedures could be formalized, and in particular how formalization would capture the role, the nature of the parameters shared between measurements and calibrating procedures, and their interrelations.

Perhaps one can stick to such a requirement of formal rules as the only acceptable independent epistemic basis for calibration. And perhaps there is an ingenious formalization, including possibly dynamic logics and similar formal tools that would prove Collins wrong on the terms he sets out, even with respect to the in-situ calibration. Yet we do not necessarily need to go that far to address the argument: the requirement of formalization itself seems to be predicated on a picture of calibration as a one-shot and rather simple procedure that we find only in the commissioning phase, and with respect to which we could at least generally comprehend what the requirement consists of. Thus, one can certainly proclaim any sort of calibration, including the in-situ type, that does not meet the formalization of calibrating procedures requirement, supposedly prone to experimenter's regress. But the very nature of the requirement indicates that the whole experimenter's regress argument was constructed to address kinds of procedures different from the case above, the complexities of which we have discussed. As it is, the requirement is predicated on a far too simple view of the interrelation between calibration and measurements.

Second, let us assume, for the sake of argument, that this sort of entanglement of calibration and outcomes is prone to experimenter's regress and that this is not obvious only because a wide agreement between experimentalists external to the calibration procedures has been reached and that this agreement brings with it a number of implicit assumptions. But what sort of agreement on calibrating precedures do we actually have, in our case? It is a comprehensive, multilevel theoretical and technical agreement developed over the course of the experiment. The broadest agreement concerns the acceptance of the background theories, Quantum Field theory and Quantum Chromodynamics. A pre-set agreement Collins describes when developing his argument seems to be analogous, in our case, simply to a starting point that applies, throughout the experiment, mainly to these background theories. A more specific and dynamic level of agreement concerns phenomenological theories that provide properties of the particles and predictions for their detection, e.g. branching rations for $t\bar{t}$ interactions or energy scales of b-jets.[15] In the commissioning phase, such specific theoretical and technical agreements are not likely to be questioned—at least initially. But in the in-situ calibration, the agreement on the exact properties of say, b-taggers, used in calibration is developed over the course of the experiment, and is increasingly better justified by the incoming results on Mt. Thus, we do not have an open-ended inter-subjective agreement that pre-establishes calibrating procedures, but one based on ever more substantiated theoretical and technical reasons. It is not clear why we should characterize such a circle as regressive, which allegedly can be broken only by an agreement external to the experimental process. Despite broad assumptions of the background theories taken on board, it enables progressively more insight into the phenomena and an empirically better-substantiated implicit agreement on their nature.

[15] See Karaca (2013) for an insightful discussion of the distinctive layers of theories used in high-energy physics experimentation.

To sum up, it seems justified then that we stick to the—broadly speaking—empiricist account of experimentation with its appeal to reasonable validating strategies. Yet the case of in-situ calibration cannot be accommodated by the empircist view unless we relax its tenet of independence of calibrating procedures and allow for the kind of entanglement between the two we find in the in-situ procedures we analyzed above. Relaxing this tenet—crafted by Franklin precisely to answer Collin's argument—does not leave the experimenter's regress argument unanswered. In fact, one need not appeal at all to the tenet of independent calibration to defend a broad empiricist approach to experimentation against the experimenter's regress argument; the argument is predicated on too simplistic categories to capture the actual iterative and dynamic relation between in-situ calibration and measurements. Thus, the empiricist can take a relaxed stance towards the form of calibration that counters Collin's argument, at least with respect to the case of in-situ calibration and similar cases.

Even so, the independence of calibration may still be a standard in many important experiments that Franklin discusses, which raises a possible objection from the empiricist point of view. This leads us to a related concluding methodological remark I wish to make.

Our analysis is limited to a specific sort of calibration, albeit ubiquitous in recent high-energy particle physics research. So, is independence of calibrating procedures from the results of subsequent measurements a standard, with occasional exceptions, one of which is the case we analyzed? The case, however, concerns a technique that is ubiquitous in, and very important to current experimentation in high energy physics. And it is not surprising that calibration in the case of new instruments, instruments working in completely new regimes, and experiments with new phenomena will be extended one way or another. Yet Franklin argues that "although experiments may find new phenomena it is not usual to have an experiment in which a new type of apparatus is used to search for a hitherto unobserved phenomenon" (Franklin 1994, p. 485). The fully independent calibration procedures that an empiricist deems to be at the core of result validation become more widespread as more experiments are performed on instruments that have already been tried out. But it is not clear why one would consider these procedures core procedures given the importance of the breakthrough experimentation that works with novel instruments and phenomena. The calibrating procedures in very well-known and novel circumstance may be identical in name, but their functions with respect to data generation and interpretation may differ substantially. And it is up to the case analysis to clarify these subtle distinctions and forms of calibration and appropriately extend the empiricist account. Indeed, this is what the in-situ calibration demonstrates—by identifying a validating strategy that requires us to rethink and relax the tenet insisting on the independence of calibration.

Thus, the kind of experimental procedures we have characterized may be ubiquitous, and a principle of separation between calibration and the measurement of the outcomes may hold in specific experimental circumstances. The motivation for insisting on a strict separation of calibration from measurement outcomes vanishes in the case at stake and similar cases due to the inaptness of the categories on which the experimenter's regress argument is predicated to capture the calibration-measurements relations. Whether there are reasons for insisting on a strict independence of calibrating procedures as a rule in experimentation other than countering the experimenter's

regress is open to discussion. A consideration of the merit of general empiricist tenets, however, as well as a possible revisiting of the case we looked at in light of such consideration is a topic for another paper. But let us make a final, cautionary methodological remark relevant to such possible considerations.

An advocate of the strong empiricist view of experimentation may require one to define an epistemic core of calibration strategies even if necessary and sufficient conditions defining such a core cannot be produced. And one may insist that such a core is defined precisely by the tenet on the independence of calibration procedures from the experimental outcomes. Yet the danger of the assumption that such a core of calibrating procedures exists and as such ensures validity of the results may lead to our overlooking the details of experimental practices that do not square with such a narrow empiricist framework. Irrespective of whether we initially lean towards empiricism or social constructivism in our analysis of experimentation, it may be epistemically prudent and empirically more acceptable (in terms of the analysis of relevant cases) to predicate one's analysis first and foremost on a view of calibrating techniques, as well as experimental procedures in general, as procedures constructed, often anew, in multifaceted symbiosis with the measurements, generation, and interpretation of results.

Acknowledgements This work was supported by the project "Dynamic Systems in nature and society: philosophical and empirical aspects" (evidence # 179041) financed by the Ministry of Education, science and technological development of the Republic of Serbia. I am grateful to Helge Kragh, Brian Hepburn, Samuel Schindler, John Norton, and graduate students in the course I taught at the Department of the History and Philosophy of Science at the University of Pittsburgh for commenting on the very first draft of the manuscript; Allan Franklin and Harry Collins for their encouragement and gentle criticism; and especially two anonymous referees for an outstanding effort to improve my argument.

References

Acharya, B.S., Cavallari, F., Corcella, G., Di Sipio, R., & Petrucciani, G. (2008). Commissioning ATLAS and CMS with top quarks. arXiv preprint arXiv:0805.3816.

ATLAS Collaboration. (2012). Measurement of the t-channel single top-quark production cross section in pp collisions at with the ATLAS detector. *Physics Letters B717*(4), 330–350.

Borjanovic, E., et al. (2005). Investigation of top mass measurements with the ATLAS detector at LHC. *The European Physical Journal C*, *39*(s2), s63–s90.

Blyweert, S. (2012). Top quark mass measurements at the LHC. arXiv preprint arXiv:1205.2175.

Cacciari, M., et al. (2008). Updated predictions for the total production cross sections of top and of heavier quark pairs at the Tevatron and at the LHC. In *JHEP09*, p. 127.

CMS Collaboration. (2008). Measurement of jet energy scale corrections using top quark events. In *CMS PAS TOP-07-004*.

CMS Collaboration. (2012). Measurement of the top-quark mass in $t\bar{t}$ events with dilepton final states in pp collisions at $\sqrt{s} = 7$ TeV. *The European Physical Journal C*, *72*, 2202.

CMS Collaboration. (2013). Determination of the top-quark pole mass and strong coupling constant from the tt production cross section in pp collisions at $\sqrt{s} = 7$ TeV. arXiv:1307.1907v1. http://arxiv.org/pdf/1307.1907v1.pdf.

Collins, H. M. (1985). *Changing order: Replication and induction in scientific practice*. Beverly Hills, CA: Sage Publications.

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

Collins, H. M. (2002). The experimenter's regress as philosophical sociology. *Studies in History and Philosophy of Science Part A*, *33*(1), 149–156.

Cousins, R. D. (2013). The Jeffreys–Lindley paradox and discovery criteria in high energy physics. arXiv preprint arXiv:1310.3791.
Franklin, Allan. (1990). *Experiment, right or wrong*. Cambridge: Cambridge University Press.
Franklin, Allan. (1994). How to avoid the experimenters' regress. *Studies in the History and Philosophy of Science*, *25*, 97–121.
Franklin, A. (1997). Calibration. *Perspectives on Science*, *5*(1), 31–50.
Franklin, A. (2013). *Shifting standards: Experiments in particle physics in the twentieth century*. Pittsburgh: University of Pittsburgh Press.
Franklin, A., & Howson, C. (1984). Why do scientists prefer to vary their experiments? *Studies in History and Philosophy of Science Part A*, *15*(1), 51–62.
Gutierrez, G. (2007). Top quark mass: Past, present and future. *AIP Conference Proceedings*, *928*, 169.
Jovicevic, J. (2013). *Probing the standard model higgs boson in the WW decay mode with the ATLAS detector at the LHC*. Licentiate Thesis in physics. University of Stockholm.
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*(01), 93–136.
Morrison, M. (2014). Bridging the great divide: Simulation, experiments, and validation experiments. In *Presented at the philosophy of scientific experimentation 4, Center for Philosophy of Science, Pittsburgh, 2014*.
Schouten, D. W. (2007). *Jet energy calibration in atlas* (Doctoral dissertation, Simon Fraser University). http://hep.phys.sfu.ca/theses/DougSchouten_msc.pdf.
Van Mulders, P. (2010). *Calibration of the jet energy scale using top quark events at the LHC, doctoral thesis*. Brussels: Vrije University.