Two Senses of Experimental Robustness: Result Robustness and Procedure Robustness

Koray Karaca

ABSTRACT

In the philosophical literature concerning scientific experimentation, the notion of robustness has been solely discussed in relation to experimental results. In this article, I propose a novel sense of experimental robustness that applies to experimental procedures. I call the foregoing sense of robustness 'procedure robustness' (PR) and characterize it as the capacity of an experimental procedure to maintain its intended function invariant during the experimental process despite possible variations in its inputs. I argue that PR is a precondition for what I call 'result robustness' (RR), which refers to the traditional sense of experimental robustness, namely, the existence of convergent experimental results obtained through different and independent means of detection. Furthermore, I argue, PR and RR constitute useful experimental strategies in the context of high-energy physics experiments, but these strategies are not without limitations.

1 Introduction

The concept of robustness was introduced into the philosophical literature by Wimsatt ([2007]). In Wimsatt's account, various things studied by science, such as properties, entities, theoretical, and experimental results, are robust if they remain invariant under what he calls a 'robustness analysis' whose

doi:10.1093/bjps/axy031

general structure is essentially based on the following two procedures: '1. To analyze a *variety* of *independent* derivation, identification, or measurement processes. 2. To look for and analyze things that are *invariant* over or *identical* in the conclusions or results of these processes' (Wimsatt [2007], p. 44). In this analysis, while the first procedure is concerned with processes through which robustness is obtained, the second procedure is concerned with corresponding types of invariances associated with these processes. Wimsatt also provides a comprehensive list of activities that illustrate different types of robustness processes and associated invariances that are common in scientific practice. Wimsatt's ([2007], p. 45) list includes the following activities that are particularly relevant to the practice of scientific experimentation: 'Using different sensory modalities to detect the same property or entity [and using] different experimental procedures to verify the same empirical relationships or generate the same phenomenon'. The sense of robustness underlying these activities denotes the existence of convergent experimental results[1] obtained through different and independent means of detection. In this article, I shall call this sense of robustness 'result robustness' (RR).[2]

Wimsatt ([2007], p. 63) points out that RR enables what Campbell and Fiske ([1959], p. 81) call 'convergent validation', namely, 'confirmation by independent measurement procedures'. According to Campbell and Fiske, due to the possibility of spurious convergence, convergent validation alone is not sufficient to validate test results obtained through different measurement procedures. In their account, spurious convergence can be avoided if convergent test results also possess what they call discriminant validity. The latter is ensured if measurement procedures are not correlated with each other to the extent that they yield convergent results when they should not yield such results, as in the case of the absence of the claimed phenomenon or in cases where the data are not relevant to the investigation of the claimed phenomena (Campbell and Fiske [1959], p. 81; see also Staley [2004], p. 473). Since independence of multiple means of detection is essential to RR, Wimsatt regards discriminant validity as a safeguard against pseudo-RR that arises from the spurious convergence of test results in cases where there is a lack of sufficient independence between different means of detection. In his account, 'method bias, a common cause of failures of discriminant validity, is a kind of failure of the requirement for robustness that the different means of detection used are

[1] Experimental results are said to be convergent if they agree with each other to the extent allowed by the standards of researchers.

[2] The term 'RR' can also be used to refer to the sense of robustness that denotes the existence of convergent theoretical results obtained by 'using different assumptions, models, or axiomatizations' (Wimsatt [2007], p. 45). In the context of models, this sense of robustness, which Wimsatt attributes to Levins ([1966]), has received considerable attention in recent years; see, for example, (Weisberg [2006]; Lloyd [2010].)

actually independent, in this case because the method they share is the origin of the correlations among traits' ([2007], p. 63).

Due to its implications for theory testing, RR has been a subject of considerable interest in the literature of the philosophy of science. The various case studies (see, for example, Hacking [1983]; Franklin and Howson [1984]; Trout [1993]; Culp [1994]; Franklin [1998], [2002]; Staley [2004]) have shown that researchers in diverse fields, such as biology, physics and psychology, appeal to RR to validate experimental results and thereby to test (confirm or disconfirm) theoretical results, thus suggesting that in experimental practice RR is taken to have a particular epistemic value in theory testing.[3] This also explains why RR has been an attractive notion for philosophers of science and widely discussed in the philosophical literature.[4] In contrast, philosophers of science have not so far paid due attention to other senses of experimental robustness. In this article, I will propose a novel sense of experimental robustness that applies to experimental procedures. To this end, I will first illustrate the different uses of RR in the context of two experiments from the history of high-energy physics (HEP). I will then introduce the aforementioned novel sense of robustness and illustrate its use in the case of the ATLAS experiment, which is one of the two general-purpose HEP experiments currently running at CERN's Large Hadron Collider (LHC), where the Higgs boson was discovered in 2012 (ATLAS Collaboration [2012c]).[5]

2 Robustness of Experimental Results

In this section, I will draw on the case studies by Franklin ([1998]) and Staley ([2004]) to illustrate the use of RR as an experimental strategy to validate experimental results in the practice of HEP.

2.1 The case of the K^+_{e2} branching ratio experiment

Franklin ([1998]) has illustrated the use of RR in the case of the K^+_{e2} branching ratio experiment that was performed by Bowen *et al.* ([1967]) to test the prediction of the V–A (vector minus axial-vector) theory of weak interactions concerning the K^+_{e2} branching ratio, that is, the fraction of all K^+ mesons decaying into a positron (e^+) and an electron-neutrino (ν_e) in the way represented by the decay process: $K^+ \rightarrow e^+ + \nu_e$. If the weak interaction is purely

[3] This claim has been contested in the philosophical literature; see, for example, (Hudson [1999], [2009]; Stegenga and Menon [2017]; Schupbach [2018]).

[4] For a comprehensive philosophical discussion concerning the use of robustness in scientific practice in general, see, for example, (Soler *et al.* [2012]).

[5] The other LHC experiment is the CMS experiment (CMS Collaboration [2012]).

axial-vector in character as predicted by the V–A theory,[6] the ratio of K_{e2}^{+} to $K_{\mu2}^{+}$ decays is predicted to be 2.6×10^{-5},[7] corresponding to a branching ratio of 1.6×10^{-5}. However, if the interaction is purely pseudo-scalar,[8] the aforementioned ratio is predicted to be 1.02. The predicted ratio would be much greater, if the interaction contains a small admixture of pseudo-scalar interaction added to the dominant pure axial-vector interaction. Therefore, as Bowen *et al.* ([1967], p. 1314) pointed out, 'even a rough measurement of the K_{e2}^{+} rate [could] provide an additional test of the applicability of V–A theory to strangeness nonconserving weak interactions and give a stringent test of the presence of any [pure pseudo-scalar] amplitude in the weak interaction of the kaon'. Since the predicted ratio for the axial-vector is so small and the expected background of the events that might mask or mimic K_{e2}^{+} events is so large, it was necessary to use different types of selection criteria for the acquisition and analysis of data (for details, see Franklin [1998]). The branching ratio of K decay was calculated by normalizing the K_{e2}^{+} events to the known K^{+} decay rates by using two different methods. Even though these methods were based on very different selection criteria, they yielded convergent results for the branching ratio (R), namely, $R = \left(2.0^{+1.8}_{-1.2}\right) \times 10^{-5}$ and $R = \left(2.2^{+1.9}_{-1.4}\right) \times 10^{-5}$, and the final result about the K_{e2}^{+} branching ratio was taken to be their average, namely, $R = \left(2.1^{+1.8}_{-1.3}\right) \times 10^{-5}$.

The sense of robustness obtained in the K_{e2}^{+} branching ratio experiment is RR, in that the final experimental result is taken to be the average of the aforementioned convergent results obtained by using two different data analysis methods. Franklin ([1998]) points out that Bowen *et al.* also showed that the above experimental results were stable over reasonable variations of the selection criteria. Since the final result was in agreement with the V–A theory's prediction of 1.6×10^{-5}, it was taken by Bowen *et al.* to be a confirmation of this theoretical prediction. Therefore, Franklin's analysis suggests that in the case of the K_{e2}^{+} branching ratio experiment, RR was used as an experimental strategy to test a theoretical prediction.

The aforementioned stability analysis in the K_{e2}^{+} branching ratio experiment was to ensure that the convergence of the test results was not spurious. However, this was not sufficient to rule out the possibility of pseudo-RR, because, as Franklin's discussion suggests, Bowen *et al.* did not perform an analysis as to what extent the normalization methods used to obtain the convergent test results were correlated to each other. This indicates that Bowen *et al.* did not demonstrate the discriminant validity of experimental results.

[6] An axial-vector is a vector-like object that is invariant under space inversion.

[7] The decay process relevant to $K_{\mu2}^{+}$ is: $K^{+} \rightarrow \mu^{+} + \nu_{\mu}$, where μ^{+} and ν_{μ} stand for the anti-muon and muon-neutrino, respectively.

[8] A pseudo-scalar is a scalar-like quantity that changes its sign under parity inversion.

Therefore, the K_{e2}^{+} branching ratio experiment illustrates a case where RR was claimed without actually proving it.[9]

2.2 The case of the CDF experiment

Staley has introduced an important dimension into the use of RR as an experimental strategy. What is key to his account of RR is the following distinction between what he calls first-order evidence and second-order evidence:

> If some fact E constitutes first-order evidence with respect to a hypothesis H, then it provides some reason to believe (or indicates) that H is the case. If a fact E is second-order evidence with respect to a hypothesis H, then it provides some reason to believe (or indicates) that some distinct fact E' is first-order evidence with respect to H. (Staley [2004], p. 469)

To illustrate his account, Staley has offered a case study concerning the discovery of the top quark in the CDF experiment at the Tevatron Collider at Fermilab.[10] He notes that since, according to the SM, the top quark would nearly always decay into a W boson and a b quark, three different counting experiments, namely, dilepton, soft lepton tagging, and secondary vertex (SVX) tagging, were designed in the CDF experiment in order to detect the decay products of the W boson and the b quark ([2004], pp. 477, 478).[11] The dilepton counting experiment searched for events yielding a pair of energetic leptons, at least two energetic jets, and a neutrino, while the other two counting experiments searched for lepton plus jet events.[12] Staley also notes that the CDF Collaboration combined the results of the three different counting experiments to obtain a statistical significance of 2.6×10^{-3} as regards the strength of the first-order evidence for the top quark. The CDF Collaboration also developed three different SVX algorithms for tagging b quarks, namely, $d - \phi$, jet probability and jet vertexing algorithms. However, only the jet vertexing SVX algorithm was used as a source of data in obtaining the above statistical significance. Staley points out that

> [the] agreement between the outcomes of the three [SVX] algorithms is presented [by the CDF Collaboration], not as direct evidence for the top quark claim, but instead as evidence that [the jet vertexing SVX

[9] This illustrates a common problem of robustness arguments in science as previously pointed out in the philosophical literature; see, for example, (Stegenga and Menon [2017]; Schupbach [2018]).

[10] The DZero experiment was the other experiment at the Tevatron Collider that discovered the top quark.

[11] In a counting experiment, the events passing the selection criteria are counted and their number is compared to the expected number of events in cases where the particle being searched for exists or does not exist.

[12] In the SM of elementary particles, the following elementary particles are called leptons: electron, muon and tau, and their respective neutrinos. Leptons are spin 1/2 particles that interact through electromagnetic and weak interactions, but not through strong interaction.

> algorithm] tags the kind of events it is intended to identify [thus providing] evidence in support of [. . .] the reliability of jet vertexing as a procedure for *b*–tagging. ([2004], pp. 480–1)

Since the jet vertexing algorithm is the only SVX type of algorithm whose results were used in obtaining the first-order evidence for the top-quark discovery, the convergent results from the three different SVX algorithms constitutes a second-order evidence for the top quark hypothesis, in that they support the validity of the results of the jet vertexing algorithm on which the result of the SVX counting experiment was based. Therefore, Staley's analysis suggests that in the case of the CDF experiment, RR is appealed to in order to provide a second-order evidence for the top quark hypothesis.

Staley also points out that in addition to convergent validation, the

> CDF [Collaboration] employed discriminant validation in demonstrating that the algorithms failed to correlate in their 'mistags', i.e., instances of tagging a secondary vertex that is not a result of a *b*-quark decay [. . .] In other words, the results of the tagging algorithms tended *not* to agree when applied to events that did not contain *b* quarks. ([2004], pp. 480–1)

Discriminant validation was to ensure that the results of the tagging algorithms would not agree if the data were not relevant to the testing of the top quark hypothesis. If the data were not relevant in this sense, this would give rise to the spurious convergence of the results of the different aforementioned algorithms. Thus, discriminant validation was used by the CDF Collaboration to rule out the possibility of pseudo-RR.

Furthermore, Staley argues that the fact that the statistical significance of the aforementioned result obtained by the CDF Collaboration is based upon the result of three independent counting experiments cannot be evidence for the claimed evidential strength of the combined result, which is indicated by the above mentioned statistical significance. His consideration is that the 'validity of the significance estimate for the combined result requires that each of the assumptions for each of the individual counting experiments be valid. The fact that some of those assumptions are independent of one another does not help to show that this requirement has been met' ([2004], pp. 481, 482). However, the 'use of convergent results from counting experiments resting on distinct assumptions, when combined into a single result [. . .] secures the first-order evidence claim', because a flaw in one of the counting experiments that undermines the combined result does not necessarily affect the others ([2004], p. 487). Therefore, Staley's analysis suggests that in the case of the CDF experiment, RR is used as an experimental strategy to provide different levels of experimental support, in that it is appealed to in order to secure both the first and second-order evidence in support of the top quark hypothesis.

3 Robustness of Experimental Procedures

The discussion so far indicates that in HEP experiments, RR is regarded as a criterion of validity for experimental results. Since experimental procedures are necessary to obtain experimental results, I suggest that validity should also be required for experimental procedures. However, in the literature of the philosophy of science, the relationship between validity and robustness has not yet been discussed in the context of experimental procedures. In this section, I shall deal with the question of how to characterize the sense of robustness relevant to the validity of experimental procedures. To this end, I shall first note that what is essential to RR is the invariance of an experimental result across different means of detection. In the practice of scientific experimentation, for example in the context of HEP experiments, this kind of invariance is of epistemic significance in the sense that an experimental result is taken to be valid if it possesses this kind of invariance. Therefore, given that RR is a type of experimental robustness, the foregoing considerations suggest that different senses of experimental robustness should be characterized in terms of invariances that lend validity to their corresponding aspects, which I shall call the ‘robust aspects’ of an experiment. For example, the particular robust aspect relevant to RR is an experimental result that remains invariant across different means of detection.

The criterion of validity for an experimental procedure is different from that for an experimental result. While the former criterion concerns the fulfilment of the intended function (or purpose) of an experimental procedure, which is the specific function that a procedure is designed to serve in an experiment, the latter criterion concerns the reproducibility of an experimental result through different means of detection. Therefore, in accordance with the characterization of experimental robustness proposed above, I suggest that the sense of robustness relevant to an experimental procedure should be characterized in terms of the invariance of the procedure’s maintenance of its intended function under possible variations in its inputs, which I take to be the particular robust aspect relevant to the robustness of an experimental procedure. This kind of invariance is of epistemic significance in that an experimental procedure is considered to be valid to the extent that it maintains its intended function invariant in this sense. Thus, I shall define the sense of robustness that applies to experimental procedures as the capacity of an experimental procedure to maintain its intended function invariant despite possible variations in its inputs. I shall call this sense of robustness ‘procedure robustness’ (PR). It should be noted that PR comes in degrees in the sense that the higher the degree of robustness of an experimental procedure, the greater its capacity to correctly perform its intended function despite variations in its inputs. Note that RR also comes in degrees in the sense that its extent indicates the extent of convergence of experimental results obtained through different means of detection.

The account of PR proposed above applies to scientific experimentation in general, as it does not presuppose any set of experimental procedures that are solely used within a particular science. Therefore, it can be used to characterize the sense of robustness that applies to procedures used in experiments across the physical and biological sciences, including data acquisition and analysis procedures as well as experimental procedures concerning instrumentation, such as calibration procedures. According to the proposed account, if an experimental procedure is not sufficiently robust, it is not sufficiently sensitive to the variations in its inputs. The lack of sufficient PR thus indicates that the procedure is biased in performing its intended function and does not have the necessary capacity to correctly perform it. Therefore, I suggest that in experiments across the physical and biological sciences, PR should be required for all procedures used in an experiment in order to ensure that the results of the experiment are not biased by the lack of sufficient sensitivity of its procedures to the variations in their inputs. For instance, an experiment might yield a null result due to some its procedures lacking sufficient PR. If an experiment has yielded a null result even after all of its procedures have been shown to be sufficiently robust, this would indicate that there might be some other possible reasons for the null result, such as the possibility of the absence of the phenomenon or effect being searched for.

The above discussion suggests that if the different experimental procedures used to obtain an experimental result are not sufficiently robust, they have the potential to yield spurious convergent results, such as null results about a real effect or phenomenon. This means that PR is necessary to avoid cases of spurious convergence of experimental results and thus acts as a safeguard against the risk of pseudo-RR. It is important to note that PR alone cannot ensure discriminant validity of experimental results, because the lack of sufficient procedural sensitivity is one of the possible ways in which different experimental procedures could be correlated to each other and yield spurious convergent results.[13] Therefore, PR should be seen as a precondition for RR, meaning that it is necessary but not sufficient for discriminant validation.

In the technical literature, researchers have proposed various definitions of robustness to characterize the behaviour of complex systems. For instance, Steven Gribble ([2001], p. 21) regards robustness as a design requirement for complex systems in general. In his account, robustness denotes the ‘ability of a [complex] system to continue to operate correctly across a wide range

[13] Wimsatt ([2007], p. 63) alludes to this point in his discussion of Campbell and Fiske’s account of validation, where he remarks that ‘discriminant validity can be regarded as an attempt to guarantee that the invariance across test methods and traits is not due to their insensitivity to the variables under study’.

of operational conditions, and to fail gracefully outside of that range'. Similarly, Jean Carlson and John Doyle ([2002], p. 2539) conceive of robustness as 'the maintenance of some desired system characteristics despite fluctuations in the behaviour of its component parts or its environment'. Some researchers have offered similar definitions of robustness in relation to complex biological systems. For example, Hiroaki Kitano ([2004], p. 826) has argued that robustness is a fundamental feature of complex biological systems in that it 'allows a system to maintain its functions despite external and internal perturbations'. Stelling *et al.* ([2004], p. 675) regard robustness as an inherent property of complex biological systems and define it as 'the ability to maintain performance in the face of perturbations and uncertainty'.

The above short overview shows that in the technical literature robustness is taken to be a system property and characterized in terms of the capacity of a complex system to maintain its proper functioning despite the internal and external variations in the operational conditions of the system. I shall call this sense of robustness 'system robustness'. The definition of PR proposed above is inspired by this definition of system robustness in characterizing robustness in terms of the capacity to fulfil an intended function despite variations in relevant conditions. Since an experimental procedure needs to be implemented by an experimental system, system robustness applies to experimental systems, rather than experimental procedures. Therefore, if an experimental system lacks sufficient robustness in the foregoing sense, this does not necessarily mean that the procedure to be implemented by this experimental system would lack sufficient PR. Rather, it only implies that the procedure would not perform its intended function due to its inadequate implementation by an experimental system that lacks sufficient system robustness.

In the next section, I will discuss the problem of data selection in the ATLAS experiment. This discussion will set the stage for Section 5 where I will characterize the robustness of the ATLAS data selection procedure.

4 The Data Selection Problem in the ATLAS Experiment

The ATLAS experiment is designed as a multi-purpose HEP experiment with the following set of objectives (see ATLAS Collaboration [2003], Section 4): (i) to test the prediction of the Higgs boson by the standard model (SM) of elementary particle physics as well as the conclusions of a wide range of theoretical models that have been proposed as possible extensions of the SM, such as supersymmetric and extra-dimensional models, which are often called models beyond the SM (BSM models) in the literature of HEP; and (ii) to search for unforeseen physics processes, which are the processes that have not been predicted by the present HEP models, including possible deviations from the SM at low energies.

In order to determine what kinds of collision events are to be considered interesting for the process of data selection in the ATLAS experiment,[14] the decay signatures—namely, stable decay products—predicted by the SM and the BSM models need to be taken into account, as these signatures are relevant to the intended objectives of the ATLAS experiment. Most BSM models of interest in the ATLAS experiment predict the existence of some novel heavy particles with mass around or above the energy threshold of $O(100)\ GeV$, including heavy gauge bosons $W'$and Z', heavy supersymmetric particles, and heavy gravitons. The decay signatures of the heavy particles predicted by the BSM models include high transverse-momentum (p_T) particles—namely, photons and leptons—and jets of the SM as well as high missing and total transverse energy (E_T).[15] The Higgs boson predicted by the SM is also a heavy particle whose signatures include high p_T signatures consisting of photons and leptons. The above considerations suggest that if the predictions of the SM and BSM models are true, then the aforementioned high p_T and E_T signatures could be produced at the LHC. It is also possible that these high p_T and E_T signatures could be produced as a result of unforeseen physics processes. Therefore, the collisions events containing the aforementioned high p_T and E_T signatures are relevant to the intended objectives of the ATLAS experiment and thus considered interesting for the process of data selection (ATLAS Collaboration [2003], Section 4; Ellis [2010]).

Figure 1 shows theoretical expectations about the occurrence rates for different types of processes predicted by the SM and BSM models. As shown in this figure, the events considered interesting (located in the lower part of the figure) in the ATLAS experiment are (theoretically) expected to have much lower production rates than those of the events produced through the SM related processes. As a result, the event production at the LHC is dominated by the types of events associated with the SM, giving rise to a large background of well-known collision events. Since these background events are not relevant to the intended objectives of the experiment, they are not considered interesting for the process of data selection. The above discussion indicates that at the LHC, the interesting collision events are distinguished from the rest of the collision events by the aforementioned high p_T and E_T decay signatures. Therefore, in order for the ATLAS experiment to achieve its objectives, the selection of interesting events should be performed by using selection criteria that consist mainly of the above-mentioned signatures. Otherwise, the process

[14] In present-day HEP, the collision events considered relevant to the intended objectives of an experiment are called 'interesting events'.

[15] Transverse-momentum is the component of the momentum of a particle that is transverse to the proton-proton collision axis, and transverse-energy is obtained from energy measurements in the calorimeter detector. High p_T and E_T refer to the p_T and E_T values that are around or above the threshold of $O(10)\ GeV$ for particles, and $O(100)\ GeV$ for jets.

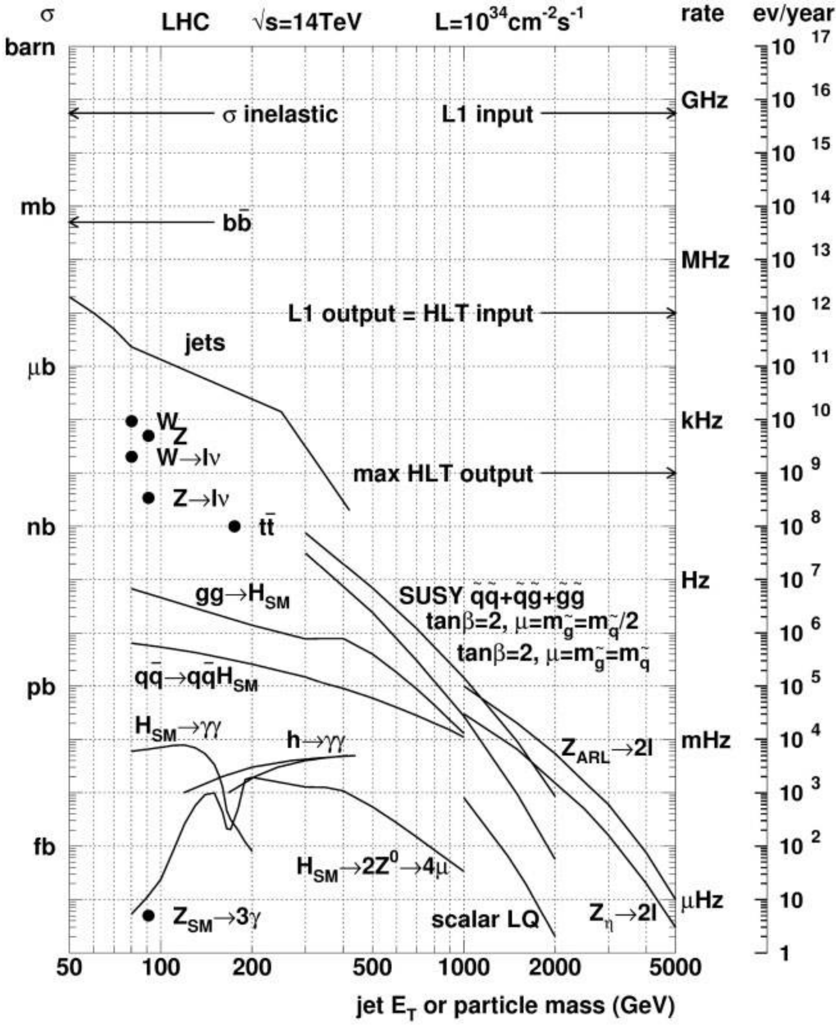

Figure 1. Expected cross-section and production rates (for a luminosity of $10^{34} cm^{-2} s^{-1}$) at the LHC for various processes in proton–proton collisions, as a function of the centre-of-mass energy (Ellis [unpublished], p. 4).

of data selection would be biased against the selection of interesting events, because it would be dominated by the well-known and abundant events of the SM model. As a result, the ATLAS experiment would fail to achieve its intended objectives concerning the testing of the SM and the BSM models.

The existence of a large background of events containing the well-known processes of the SM is not the only difficulty with the acquisition of interesting events at the LHC. Another important difficulty is that due to the technical limitations both in terms of data-process rate and data-storage capacity, only a minute fraction (at a ratio of approximately 5×10^{-6}) of the types of events occurring at the LHC could be selected for further evaluation during the stage of data analysis. Given that the types of collision events relevant to the testing of the SM and BSM models span a wide range of high p_T and E_T signatures in terms of both types of signatures, namely, particles and jets of the SM and threshold energies, the foregoing difficulty has the potential to give rise to a selection bias against certain types of interesting events, if the set of data selection criteria is not appropriate to select the entire range of different types of interesting events that are relevant to the intended objectives of the ATLAS experiment. This poses a significant threat to the discovery potential of the ATLAS experiment, in that it has the potential to cause the data selection procedure to fail to select certain types of interesting events, as a result of which the ATLAS experiment would fail to achieve some of its intended objectives.

5 The Robustness of the Data Selection Procedure in the ATLAS Experiment

The discussion in the previous section shows that the ATLAS data selection procedure is aimed at selecting the sets of events that are relevant to the entire set of objectives of the experiment. In order for the ATLAS experiment to achieve its aforementioned objectives, it is necessary that the selection procedure fulfil this function over the entire range of collision events resulting from the proton-proton collisions occurring at the LHC. Drawing upon the account of PR proposed in Section 3, I shall characterize the robustness of the ATLAS data acquisition procedure as the capacity of this procedure to maintain its intended function—namely, to select only interesting events—invariant despite variations in types of signatures and associated energy thresholds in collision events, which are the inputs to this procedure. Robustness in this sense is a safeguard against the possibility of selection bias resulting from the limited capacity of the set of data selection criteria, often called the trigger menu, with respect to the diversity of interesting events in terms of types of signatures and threshold energies. This means that the particular feature of the ATLAS data selection procedure that makes it robust is the sensitivity of the trigger menu, which amounts to its capacity to select the range of types of interesting events that will serve the entire range of objectives of the ATLAS experiment. Therefore, the extent of robustness of the ATLAS data selection procedure depends on the extent of sensitivity of the trigger menu, which in turn depends on the extent to which the trigger menu is diversified in terms of types of selection signatures and energy thresholds that are appropriate for the various objectives of the experiment.

Since the ATLAS experiment is mainly aimed to test the SM's prediction of the Higgs boson and the predictions of the BSM models, the robustness of the data selection procedure in the aforementioned sense requires the trigger menu to be diversified in terms of selection signatures composed of only high p_T and E_T types of signatures relevant to the aforementioned predictions. These signatures are referred to as inclusive triggers (ATLAS Collaboration [2003], Section 4), because they constitute the main set of selection criteria in the trigger menu used in the ATLAS experiment. Since the aforementioned data selection bias poses a threat to the capacity of the ATLAS experiment to achieve its intended objectives, PR is necessary to ensure that the process of data selection is performed without jeopardizing any of the intended objectives of the experiment.

The previous discussion indicates that PR imposes a criterion on the trigger menu of the ATLAS experiment in the sense that the selection criteria need to be established in such a way that the trigger menu is sensitive to the detection of interesting events. As I shall illustrate in what follows, this can be achieved by diversifying the trigger menu in terms of high p_T and E_T types of selection signatures and associated energy thresholds (see also Karaca [2017]). To this end, I shall consider Table 1 that shows a sample trigger menu that contains major inclusive triggers used in the ATLAS experiment. In this table, the left column is for the selection signatures, while the right column is for the physics processes predicted by the SM and the BSM models to which the selection signatures on the left column are relevant. Note also that each selection signature given in the left column of Table 1 is represented by the label '*NoXXi*.' Here, '*N*' denotes the minimum number of objects—namely, particles, jets and transverse energy—required for a particular selection, and '*o*' denotes the type of signature, for example, '*e*' for electron; 'γ' for photon; 'μ' for muon; 'τ' for tau; '*xE*' for missing E_T; '*E*' for total E_T; and '*jE*' for total E_T associated only with jet(s). The label '*XX*' above denotes the threshold of E_T (in units of GeV), that is, the lowest E_T at or above which a given selection criterion operates, and '*i*' denotes whether the given signature is isolated or not.

In order to diversify the trigger menu with respect to the testing of the SM's prediction of the Higgs boson (H), the decay processes relevant to this prediction need to be considered in order to identify the associated high p_T and E_T types of signatures. In the SM, the Higgs boson could decay into the W and Z bosons, respectively, as follows: $H \rightarrow WW^*$ and $H \rightarrow ZZ^*$, where a '$*$' denotes an off-shell boson, namely, one that does not satisfy classical equations of motions. The W and Z bosons produced in these processes could subsequently decay into leptons, including electrons and electron-neutrinos (ν), as well as into top quarks, as indicated in the first line of the right column in Table 1 (ATLAS Collaboration [2012c]). The top quark could decay into a bottom quark, and a W boson that could subsequently decay into an electron

Table 1. A sample of main data selection criteria used in the ATLAS experiment (ATLAS Collaboration [2003], p. 38)

Selection signature	Examples of physics coverage
$e25i$	$W \rightarrow$ ev, $Z \rightarrow$ ee, top production, $H \rightarrow$ WW$^{(*)}$/ZZ$^{(*)}$,W',Z'
$2e15i$	$Z \rightarrow$ ee, $H \rightarrow$ WW$^{(*)}$/ZZ$^{(*)}$
$\mu 20i$	$W \rightarrow \mu$v, $Z \rightarrow \mu\mu$, top production, $H \rightarrow$ WW$^{(*)}$/ZZ$^{(*)}$,W',Z'
2μ 10	$Z \rightarrow \mu\mu$, $H \rightarrow$ WW$^{(*)}$/ZZ$^{(*)}$
$\gamma 60i$	Direct photon production, $H \rightarrow \gamma\gamma$
$2\gamma 20i$	$H \rightarrow \gamma\gamma$
$j400$	QCD, SUSY, new resonances
$2j350$	QCD, SUSY, new resonances
$3j165$	QCD, SUSY
$4j110$	QCD, SUSY
$\tau 60i$	charged Higgs
$\mu 10 + e15i$	$H \rightarrow$ WW$^{(*)}$/ZZ$^{(*)}$, SUSY
$\tau 35i + xE45$	qqH($\tau\tau$), $W \rightarrow \tau$v, $Z \rightarrow \tau\tau$, SUSY at large tan β
$j70 + xE70$	SUSY
$xE200$	New phenomena
$E1000$	New phenomena
$jE1000$	New phenomena
$2\mu 6 + \mu^+\mu^- +$ mass cuts	Rare B-hadron decays ($B \rightarrow \mu\mu$X) and $B \rightarrow J/\psi$ (ψ')X

and an electron-neutrino. These considerations suggest that the events containing at least one electron with high E_T have the potential to contain the aforementioned decay processes of the Higgs boson. This means that selection signatures consisting of at least one electron with high E_T are appropriate for the testing of the SM's prediction of the Higgs boson. Table 1 contains such a selection signature, namely, $e25i$, which requires at least one isolated electron with an E_T threshold of 25 GeV. Since the signatures predicted by some BSM models for the new heavy gauge bosons $W^{'}$ and $Z^{'}$ include leptons, this type of selection signatures are appropriate for the testing of these predictions (see, for example, ATLAS Collaboration [2015]) as well as for the study of the top quark related processes in the SM (ATLAS Collaboration [2016b]). Furthermore, the following decay in the SM: $H \rightarrow \gamma\gamma$, where the Higgs boson decays into two photons, indicates that selection signatures consisting of at least two photons with high E_T are also appropriate for the testing of the SM's prediction of the Higgs boson. Table 1 contains such a selection signature, namely, $2\gamma 20i$, which requires at least two isolated photons each of which has an E_T threshold of $20\,GeV$.

In order to illustrate how the trigger menu is diversified with respect to the testing of the BSM models considered, I shall consider the minimal supersymmetric extension of the SM (MSSM),[16] which is currently the most studied BSM model in the HEP literature. The signatures predicted by the MSSM for the squarks and gluinos are high E_T jets and missing high E_T,[17] indicating that selection signatures consisting of various combinations of these signatures are appropriate for the testing of the MSSM (ATLAS Collaboration [2012a]). For example, Table 1 contains selection signatures that are appropriate for the testing of the MSSM, namely, $j400$, $2j350$, $3j165$, and $4j110$, which consist of different numbers of high E_T jets. As shown in Table 1, these selection signatures are also appropriate for the study of the QCD processes. Moreover, selection signatures consisting of both jets and missing E_T are also relevant to the testing of the MSSM. Such a selection signature, given in Table 1, is $j70 + xE70$, which denotes the requirement of at least one jet with an E_T threshold of $70\,GeV$ and a missing E_T at or above $70\,GeV$.

The signatures predicted by the MSSM for charginos or neutralinos are high E_T leptons and high missing E_T, indicating that selection signatures consisting of various combinations of these signatures are appropriate for the testing of the MSSM (ATLAS Collaboration [2012a]). Such a selection signature is shown in Table 1, namely, $\mu 10 + e15i$, which denotes at least one muon with an E_T threshold of $10\,GeV$ and one isolated electron with an E_T threshold of $15\,GeV$. Since the signatures predicted by the SM for the Higgs boson also include leptons, this type of selection signatures are also appropriate for the testing of the SM's prediction of the Higgs boson. The trigger menu is further diversified for the testing of the predictions of the MSSM by means of types of selection signatures that consist of various combinations of high E_T types of selection signatures predicted by the MSSM (see, for example, ATLAS Collaboration [2016a]).

Extra-dimensional models are another class of BSM models that the ATLAS experiment is aimed to test. These models predict signatures consisting of a certain number of high E_T jets, leptons and photons (such as di-lepton, di-jet, and di-photon) as well as high missing E_T. Since selection signatures consisting of various combinations of these signatures are appropriate for the testing of extra-dimensional models, they are included in the trigger menu of the ATLAS experiment.[18] Such a selection signature, which is

[16] The MSSM predicts the existence of supersymmetric particles: for each bosonic (fermionic) particle in the SM, a fermionic (bosonic) superpartner with the same internal quantum numbers and mass is predicted.

[17] Squarks and gluinos are the supersymmetric particles predicted by the MSSM to be the superpartners of quarks and gluons.

not included in Table 1, is $2\gamma20$, which denotes at least two photons each of which has an E_T threshold of $20\,GeV$.[19]

The types of selection signatures discussed so far are motivated by the specific theoretical predictions of the current models of HEP. In order to increase the sensitivity of the trigger menu to the detection of unforeseen high E_T processes, general-purpose selection signatures that are not necessarily motivated by the predictions of the current models of HEP are included in the trigger menu. Table 1 illustrates such selection signatures, namely, $xE200$, $E1000$, and $jE1000$, where $xE200$ denotes a missing E_T at or above $200\,GeV$, and $E1000$ and $jE1000$, respectively, denote a total E_T at or above $1000\,GeV$ and a total E_T, only due to jets, at or above $1000\,GeV$. The types of selection signatures appropriate for the selection of interesting events relevant to the current HEP models—namely, the SM and the BSM models—are also appropriate for the search for unforeseen physics processes, because the same types of interesting events might be produced at the LHC as a result of unforeseen processes at high energies.

Since inclusive triggers consist of only high p_T and E_T types of signatures, they are not appropriate for the search for novel p_T and E_T processes at the low-energy scale, that is, below $10\,GeV$. The previous HEP experiments, namely, the DZero and CDF experiments at the Tevatron Collider at Fermilab, did not detect any deviations from the SM or any novel processes at the low-energy scale, even though they probed energies up to $2\,TeV$. However, this does not imply that at the LHC, where energies up to $13\,TeV$ are currently probed, novel p_T and E_T processes would not be produced at the low-energy scale, because the current collision rate ($\sim 40\,MHz$) at the LHC is considerably higher than the one ($\sim 2.5\,MHz$) at the Tevatron Collider. In HEP experiments, the higher the collision rate, the greater the chance to detect novel processes. This is due to the fact that since novel processes are expected to occur in rare collision events, collision rate, rather than collision energy, is a decisive factor in the production of novel processes at particle colliders.

The above considerations indicate that the trigger menu needs to be sufficiently diversified in terms of low p_T and E_T types of selection signatures in order to increase its sensitivity to the detection of novel processes at low energies. These selection signatures are referred to as pre-scaled triggers, because they are obtained by pre-scaling inclusive triggers with lower p_T and E_T thresholds ($< 10\,GeV$) (for details, see ATLAS Collaboration [2003], Section 4.4.2). Here, pre-scaling means suppressing the number of events that a trigger

[18] For a detailed discussion of selection signatures that takes into account the differences between the different models with extra dimensions, see (ATLAS Collaboration [2012b]).

[19] For an analysis based on events selected according to this criterion, see (ATLAS Collaboration [2012b]).

could accept by what is called a pre-scale factor, so that the selection process is not swamped by the events containing vastly abundant low p_T and E_T types of signatures. Therefore, the above considerations indicate that pre-scaled triggers further increase the robustness of the data selection procedure, as they permit the selection of events that have the potential to serve the detection of novel processes at the low-energy scale, such as possible deviations from the SM (ATLAS Collaboration [2016b]).

The discussion in this section indicates that the trigger menu in the ATLAS experiment is diversified in terms of types of selection signatures, so that the data selection procedure becomes robust against possible variations in types of signatures.[20] The range of possible variations spans a wide variety of signatures, namely, elementary particles—leptons and photons—jets, missing or total energy, as well as associated energy thresholds, ranging from a few GeVs up to the TeV energy scale. The extent to which the trigger menu is diversified in terms of possible types of signatures and associated energy thresholds determines the extent to which it is sensitive to detecting different types of interesting events, which in turn determines the extent of robustness of the ATLAS data selection procedure against possible variations in types of interesting events.

As the previous discussion shows, since the signatures associated with the novel processes predicted by the SM and the BSM models are theoretically well known, it is possible to diversify the trigger menu so as to render the ATLAS data selection procedure optimally robust for the detection of the events relevant to the testing of the aforementioned models. The general-purpose high E_T selection signatures and pre-scaled triggers are included in the trigger menu in order to make it sensitive, respectively, to the detection of events relevant to possible unforeseen physics processes at the GeV energy scale and at energies lower than the GeV scale. However, theoretically speaking, since it is not known what types of signatures are associated with unforeseen physics processes, it is not possible to determine the space of the selection criteria relevant to the detection of these signatures. This means that it is completely unknown what portion of this space of selection criteria is covered by the general-purpose high E_T selection signatures and prescaled triggers. As a result, even though the extent of robustness of the ATLAS data selection procedure ensures the acquisition of a wide range of types of interesting events, the data selection procedure is not optimally robust against possible variations in types of interesting events that are solely due to the existence of unforeseen physics processes. However, despite this limitation, the ATLAS data selection procedure is exploratory in that by virtue of its extent of

[20] The trigger menu is constantly updated in terms of types of signatures and energy thresholds in order to further increase the sensitivity of the data selection procedure.

robustness, it serves to extend the range of possible experimental results, thereby increasing the discovery potential of the ATLAS experiment. Experimental exploration in this sense (Karaca [2017]) is an important desideratum for the process of data selection in the ATLAS experiment, as it is necessary for the experiment to achieve its various objectives. Therefore, PR enables the ATLAS data selection procedure to fulfil its intended function and thus lends validity to this procedure.

In this section, I have shown that in the ATLAS experiment, PR is sought for the data selection procedure. In the same experiment, PR should also be sought for the other experimental procedures used in the production of experimental results, because it is a precondition for the robustness of experimental results.[21] At the LHC, RR is required for the convergent validity of experimental results in the sense that the two different multi-purpose LHC experiments, namely, the ATLAS and CMS experiments, should yield convergent results about the particular phenomenon or effect under study. The requirement of RR was satisfied in the Higgs boson search at the LHC, as the ATLAS and CMS experiments yielded convergent results about the mass of the Higgs boson, namely, 126.0 GeV and 125.3 GeV, respectively (ATLAS Collaboration [2012c]; CMS Collaboration [2012]).[22] This was essential to the CERN's discovery claim concerning the existence of the Higgs boson. At this point, the important question is whether and how the requirement of discriminant validity was satisfied in the Higgs boson search in the LHC experiments in order to safeguard against the risk of pseudo-RR. I shall not deal with this question in this article, as it requires a detailed examination of the data analysis procedures used in the ATLAS and CMS experiments, which is not necessary for the argument of this article.

In this section, I have also argued that in the ATLAS experiment, PR is essential to overcome the problem of data selection bias that arises from the application of selection criteria to the collision events in real time during data acquisition.[23] This problem of data selection bias was also encountered in early HEP experiments. One well-known case noted by Franklin is the experiment (Benvenuti *et al.* [1974]) where weak neutral currents were discovered (for details, see Galison [1989], Chapter 4):

[21] Interesting events could also be missed out due to deficiencies in the technical implementation of the data selection procedure. Therefore, in order for the ATLAS experiment to achieve its various objectives, the data acquisition system (ATLAS Collaboration [2003]) needs to be robust in the sense of system robustness in order to correctly implement the data acquisition procedure.

[22] RR can be obtained in the case of a single experiment, as in the K_{e2}^{+} branching ratio experiment and the CDF experiment, as well as in the case of multiple experiments. The latter case is illustrated by the Higgs boson search at the LHC.

[23] A similar but different problem of data selection bias can arise from the application of selection criteria to the acquired data during data analysis. In the context of HEP experiments, the latter problem of data selection bias has been extensively discussed by Franklin ([2002], Chapters 1–6).

> When the experiment was initially conceived, it was a rule of thumb in particle physics that weak neutral currents did not exist. The initial design included a muon trigger, which would be present only in charged current interactions. In a charged-current event a neutrino is incident and a charged muon is emitted, in a neutral-current event there is a neutrino in both the initial and final states, and no muon is emitted. Thus, requiring a muon in the event trigger would preclude the observation of neutral currents. After discussion with theorists, who pointed out that the then recently proposed Weinberg-Salam unified theory of electroweak interactions predicted neutral currents, the trigger was changed so that neutral currents could be observed. In its original form, the experiment could not have detected those currents. Fortunately, the design was changed before the experiment was performed. (Franklin [2015], pp. 159, 160)

Another relevant HEP experiment is the one where 'a SLAC team [missed] the discovery of pions even though their apparatus was producing it in a required way' (Perovic [2011], p. 39). In this experiment, 'the experimenters were based on an incorrect estimate of the energy of the alpha-particles they were producing, as they thought the apparatus could not reach the 95 MeV needed for the production of pions' (Perovic [2011], p. 39).

These historical cases illustrate that the lack of sufficient PR due to the use of inappropriate selection criteria could give rise to bias in the selection of data and thus undermine the discovery potential of the experiment.

6 Conclusions

In this article, I have introduced PR as a novel sense of experimental robustness and distinguished it from RR. I have argued that both PR and RR serve validation purposes in scientific experimentation. While PR serves as a criterion of validity for both experimental procedures and results, RR serves as a criterion of validity for experimental results.

In the context of HEP experiments RR and PR have different epistemic values and thus constitute different experimental strategies. In the case of the ATLAS experiment, PR acts an exploratory data selection strategy to deal with the problem of data selection and thereby serves to increase the discovery potential of the experiment for novel physics processes predicted by the SM and BSM models as well as for unforeseen physics processes. Therefore, in the context of the ATLAS experiment, PR has an epistemic value in the sense that it determines the scope of the experimental inquiry, which in turn determines the extent of experimental knowledge to be gained by performing this experiment. This is unlike the cases of the CDF experiment and the K_{e2}^{+} branching ratio experiment where RR has a validatory value, in that it is taken by experimenters to be a measure of the correctness of experimental results.

Therefore, in these two cases, obtaining RR constitutes an experimental strategy to validate experimental results.

In the context of HEP experiments, PR and RR are useful strategies, but they are not without limitations. The use of RR as a validation strategy serves the convergent validation of experimental results, while it gives rise to the risk of pseudo-RR that has the potential to undermine the credibility of results. The case of the CDF experiment shows that discriminant validation is used as a backup strategy to defeat the risk of pseudo-RR in cases where RR is appealed to in order to validate experimental results. However, in the case of the K^+_{e2} branching ratio experiment, discriminant validation was not used to back up RR, thus illustrating an example of a HEP experiment where the risk of pseudo-RR was not properly addressed. The fact that discriminant validation was not used in the K^+_{e2} branching ratio experiment performed in the sixties, while it was used in the CDF experiment performed in the nineties, can be interpreted as an indication that over the years experimenters in HEP have become more cautious in using RR as a validation strategy due to the risk of pseudo-RR. Unlike the aforementioned cases, the case of the ATLAS experiment shows the limitation of PR as an experimental strategy in the sense that it does not ensure that the data selection procedure is optimal for the discovery of unforeseen physics processes. But on the other hand, there seems to exist no better strategy than PR to deal with the problem of data selection encountered in the ATLAS experiment.

Acknowledgements

I would like to thank Allan Franklin, Kent Staley, and Miles MacLeod for helpful discussions on robustness. I am grateful to two anonymous BJPS referees, as well as one of BJPS editors, for helpful comments and suggestions. Earlier versions of this article have been presented at conferences at the University of Hannover, the University of Konstanz, the University of Milan-Bicocca, the University of Nancy, Virginia Tech, and Ryerson University. I would like to thank the audiences at these conferences for their feedback.

Department of Philosophy
University of Twente
Enschede, The Netherlands
karacak@gmail.com

References

ATLAS Collaboration [2003]: 'High-Level Trigger, Data Acquisition, and Controls', Technical Report CERN-LHCC-2003-022, ATLAS-TRD-016, available at <inspirehep.net/record/629896>.

ATLAS Collaboration [2012a]: 'Further Search for Supersymmetry at $\sqrt{s} = 7$ TeV in Final States with Jets, Missing Transverse Momentum, and Isolated Leptons with the ATLAS Detector', *Physical Review D*, **86**, p. 092002.

ATLAS Collaboration [2012b]: 'Search for Extra Dimensions Using Diphoton Events in 7 TeV Proton–Proton Collisions with the ATLAS Detector', *Physics Letters B*, **710**, pp. 538–56.

ATLAS Collaboration [2012c]: '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**, pp. 1–29.

ATLAS Collaboration [2015]: 'Search for High-Mass Diboson Resonances with Boson-Tagged Jets in Proton–Proton Collisions at $\sqrt{\mathrm{s}} = 8$ TeV with the ATLAS Detector', *Journal of High Energy Physics*, **2015**, pp. 1–39.

ATLAS Collaboration [2016a]: 'Search for Supersymmetry at $\sqrt{\mathrm{s}} = 13$ TeV in Final States with Jets and Two Same-Sign Leptons or Three Leptons with the ATLAS Detector', *European Physical Journal C*, **76**, p. 259.

ATLAS Collaboration [2016b]: 'Measurement of $\mathrm{D}^{*\pm}$, $\mathrm{D}^{\pm}$ and $\mathrm{D_s}^{\pm}$ Meson Production Cross Sections in pp Collisions at $\sqrt{s} = 7$ TeV with the ATLAS Detector', *Nuclear Physics B*, **907**, pp. 717–63.

Benvenuti, A., Cheng, D. C., Cline, D., Ford, W. T., Imlay, R., Ling, T. Y., Mann, A. K., Messing, F., Piccioni, R. L., Pilcher, J., Reeder, D. D., Rubbia, C., Stefanski, R. and Sulak, L. [1974]: 'Observation of Muonless Neutrino-Induced Inelastic Interactions', *Physical Review Letters*, **32**, pp. 800–3.

Bowen, D. R., Mann, A. K., McFarlane, W. K., Franklin, A. D., Hughes, E. B., Imlay, R. L., O'Neill, G. K. and Reading, D. H. [1967]: 'Measurement of the K^{+}_{e2} Branching Ratio', *Physical Review*, **154**, pp. 1314–22.

Campbell, D. T. and Fiske, D. W. [1959]: 'Convergent and Discriminant Validation by the Multitrait–Multimethod Matrix', *Psychological Bulletin*, **56**, pp. 81–105.

Carlson, J. M. and Doyle, J. [2002]: 'Complexity and Robustness', *Proceedings of the National Academy of Science*, **99**, pp. 2538–45.

CMS Collaboration [2012]: 'Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC', *Physics Letters B*, **716**, pp. 30–61.

Culp, S. [1994]: 'Defending Robustness: The Bacterial Mesosome as a Test Case', *Philosophy of Science*, **1994**, pp. 46–57.

Ellis, N. [unpublished]: 'The ATLAS Years and the Future', available at <www.ep.ph.bham.ac.uk/general/outreach/dowellfest>.

Ellis, N. [2010]: 'Trigger and Data Acquisition', CERN Yellow Report CERN-2010-001, pp. 417–49, available at <lanl.arxiv.org/abs/1010.2942>.

Franklin, A. [1998]: 'Selectivity and the Production of Experimental Results', *Archive for History of Exact Sciences*, **53**, pp. 399–485.

Franklin, A. [2002]: *Selectivity and Discord*, Pittsburgh: University of Pittsburgh Press.

Franklin, A. [2015]: ‘The Theory-Ladenness of Experiment’, *Journal for General Philosophy of Science*, **46**, pp. 155–66.

Franklin, A. and Howson, C. [1984]: ‘Why Do Scientists Prefer to Vary Their Experiments?’, *Studies in History and Philosophy of Science Part A*, **15**, pp. 51–62.

Galison, P. [1989]: *How Experiments End*, Chicago, IL: University of Chicago Press.

Gribble, S. D. [2001]: ‘Robustness in Complex Systems’, in *Proceedings of the Eighth Workshop: Hot Topics in Operating Systems*, Elmau: IEEE, pp. 21–6.

Hacking, I. [1983]: *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*, Cambridge: Cambridge University Press.

Hudson, R. [1999]: ‘Mesosomes: A Study in the Nature of Experimental Reasoning’, *Philosophy of Science*, **66**, pp. 289–309.

Hudson, R. [2009]: ‘The Methodological Strategy of Robustness in the Context of Experimental WIMP Research’, *Foundations of Physics*, **39**, pp. 174–93.

Karaca, K. [2017]: ‘A Case Study in Experimental Exploration: Exploratory Data Selection at the Large Hadron Collider’, *Synthese*, **194**, pp. 333–54.

Kitano, H. [2004]: ‘Biological Robustness’, *Nature Reviews Genetics*, **5**, pp. 826–37.

Levins, R. [1966]: ‘The Strategy of Model Building in Population Biology’, in E. Sober (*ed.*), *Conceptual Issues in Evolutionary Biology*, Cambridge, MA: MIT Press, pp. 18–27.

Lloyd, E. A. [2010]: ‘Confirmation and Robustness of Climate Models’, *Philosophy of Science*, **77**, pp. 971–84.

Perovic, S. [2011]: ‘Missing Experimental Challenges to the Standard Model of Particle Physics’, *Studies in History and Philosophy of Modern Physics*, **42**, pp. 32–42.

Schupbach, J. N. [2018]: ‘Robustness Analysis as Explanatory Reasoning’, *British Journal for the Philosophy of Science*, **69**, pp. 275–300.

Soler, L., Nickles, T., Trizio, E. and Wimsatt, W. C. [2012]: *Characterizing the Robustness of Science: After the Practice Turn in Philosophy of Science*, Dordrecht: Springer.

Staley, K. W. [2004]: ‘Robust Evidence and Secure Evidence Claims’, *Philosophy of Science*, **71**, pp. 467–88.

Stegenga, J. and Menon, T. [2017]: ‘Robustness and Independent Evidence’, *Philosophy of Science*, **84**, pp. 414–35.

Stelling, J., Sauer, U., Szallasi, Z., Doyle, F. J., III and Doyle, J. [2004]: ‘Robustness of Cellular Functions’, *Cell*, **118**, pp. 675–85.

Trout, J. D. [1993]: ‘Robustness and Integrative Survival in Significance Testing: The World’s Contribution to Rationality’, *British Journal for the Philosophy of Science*, **44**, pp. 1–15.

Weisberg, M. [2006]: ‘Robustness Analysis’, *Philosophy of Science*, **73**, pp. 730–42.

Wimsatt, W. C. [2007]: ‘Robustness, Reliability, and Overdetermination’, in his *Re-Engineering Philosophy for Limited Beings, Piecewise Approximations to Reality*, Cambridge, MA: Harvard University Press, pp. 43–71.