IOP Concise Physics

Chapter 11

Conclusion

In the preceding ten chapters, we have presented ample evidence, covering a 400-year history, that null experiments and null results play significant roles in physics. We began with Galileo's experiments on falling bodies and concluded with tests of the weak equivalence principle in General Relativity, the search for physics beyond the Standard Model, and the search for neutrinoless double beta decay, all in the 21st century.

During that 400 year period, null results have refuted theories, confirmed theories, provided evidence for potential new theories to explain, introduced new experimental techniques, corrected previous incorrect or misinterpreted results, and were used to explore previously unstudied phenomena. In short, null experiments play many of the roles that ordinary experiments play in physics. Although there are many similarities between null experiments and other experiments, there are some subtle differences which are discussed below.

11.1 How do we know it is null result?

A reasonable start is to say that a null result has a zero value. The raw data, however, rarely, if ever, yield a zero result but rather the raw data once corrected for disturbing effects yield a value that is consistent with zero taking into account what are considered to be reasonable appraisals of experimental uncertainties. We have also seen examples of experiments that yield non-zero results, but nevertheless confirm null hypotheses. This was illustrated in chapters 9 and 10 in the search for physics beyond the Standard Model and in the search for neutrinoless double beta decay. In both episodes, events were found that mimicked those expected from supersymmetry or from neutrinoless double beta decay. They did not, however, exceed, in any significant way either the predictions of Standard Model processes or, in the case of neutrinoless beta decay, the calculated background.

These two episodes are similar, but not identical. In the searches for physics beyond the Standard Model, there was a well-confirmed theory that was used, in

doi:10.1088/2053-2571/ab3918ch11

combination with estimates of both systematic and statistical uncertainties, to calculate the number of events expected even if there were no supersymmetric particles. For neutrinoless double beta decay, no such theory existed. The experimenters used their knowledge energy peaks expected, combined with estimates of background based on their raw data to conclude that there was no evidence of neutrinoless double beta decay. It is clear that the estimated systematic and statistical uncertainty played crucial roles in these episodes. They are also essential in the other episodes we have discussed.

11.1.1 The appraisal of systematic and statistical uncertainty

How does one estimate possible systematic corrections? One technique is to *amplify* a suspected systematic uncertainty and then determine its corresponding effect on the raw data. The earliest instance of the use of this approach was employed by Newton when he amplified the effect of the postulated Cartesian aether and demonstrated that the systematic error caused by such an aether would be of no consequence when considering his pendulum experiments. We have seen this approach used many times in the experiments described here, including the sequence of experimental tests of the Fifth Force hypothesis and later the weak equivalence principle. In those instances, effects such as temperature gradients, tilt of the apparatus and other possible masking or mimicking effects were amplified and it was shown that, at the level expected in the actual experiment, they did not have any significant effect on the final result. In a slight variant of this technique, Dayton Miller claimed that amplification of temperature gradients and mechanical deformation had no significant influence on his positive result. It was later shown by Shankland *et al* and by Roberts that Miller had underestimated these effects and that his results were, in fact, consistent with zero, or, at the very least, with a small upper limit.

Another technique is to rotate the experimental apparatus to test for asymmetrical effects as was done by the Eöt-Wash group, as well as by Thieberger. Hall used similar orientation variations in his falling body experiments when he rotated the positional 'beaks' and the receiving pans.

A more direct method is simply to eliminate or reduce the presence of the suspected systematic disturbances. A straightforward example was the attention paid to reducing the lack of rigidity and the friction caused by the supporting hinges for the pendulum experiments of Newton, Bessel and Potter, as well as by attention to maintaining a fixed center of inertia for the pendulum bobs. More elaborate were the efforts of Kennedy and Joos to carefully insulate their apparatus from both temperature and mechanical variations. Though here, as Miller insisted, one had to be wary of insulating the apparatus from the very interaction—a possibly entrained ether—that the experiment was supposed to interact with.

Sometimes, however, the amplification of what otherwise might be considered an interfering effect is required in order to magnify the size of the sought after effect. What we have in mind here was the need, discovered upon analysis, for a large nearby mass in order to create a horizontal gravitational attraction sufficient to reveal

a differential effect in the acceleration of different substances. This was the case in experiments on the Fifth Force.

Finally, when a *comparison* of different aspects or elements of an experiment is involved the Eöt-Wash group designed 'the test bodies to *appear identical in all respects except for baryon content*'. Echoing Newton's preparation of his pendulums, so that they were 'exactly like each other with respect to their weight, shape, and air resistance' (Newton 1999, p 807). The Eöt-Wash group following suit stated that:

> We minimize false signals by designing the test bodies to *appear identical in all respects except for baryon content*. Each body was a cylinder 1. 908 cm high and 1.905 cm in diameter and had a mass of 10.04 g. The external dimensions of the bodies were identical to within ± 0.0025 cm and their masses were equal to ± 4.6 mg. The difference in density between Be and Cu was accommodated by fabrication of the Cu bodies as cylindrical shells fitted with endcaps. (Stubbs *et al* 1987, p 1071)

11.1.2 Sensitivity, calibration and surrogate signals

An important question for all forms of experimentation is how do we know the experimental apparatus would have detected the predicted effect had it occurred? In previous work, Franklin (2004) discussed strategies used to establish this sort of counterfactual. These included the use of *surrogate signals*. We observed an example of this in the search for neutrinoless double beta decay where the experimenters showed that they could observe peaks due to other physical process that mimicked the events predicted by double beta decay, albeit at different energies. Because these peaks were at known energies they could be used to *calibrate* the energy scale of the experiments.

The replications by Kennedy and Illingworth of the Michelson–Morley experiment made an analogous use of a surrogate signal that was used to both calibrate the experimental apparatus and confirm that the apparatus had sufficient sensitivity to detect the signal in question. Here, the surrogate signal was created by placing small weights on one corner of the supporting marble slab. As discussed in section 7.4, using such a signal, Illingworth very cleverly devised a method of calibration whereby fringe displacements were effectively measured in terms of the weights needed to restore the fringe location to its initial zero value.

A different form of surrogate signal, namely, a known and externally supplied twisting force, was used to calibrate the torsional balances in the Roll *et al* solar experiments, as well as in the many Eöt-Wash replications and tests of WEP (see, for example, Adelberger *et al* 1990: ('The most important calibrations were of the torsional constant κ of our fiber, and of the angular deflection θ', pp 3275–6) and Roll *et al* 1964: 'This was accomplished by using a micrometer head to rotate the telescope of the optical lever through a known angle relative to the balance [where] the torsion constant was obtained from the torsional oscillation period, knowing the moment of inertia of the torsion balance', p 474.)

Another way of ensuring that the experimental apparatus is able to effectively detect the signal in question is to maximize its sensitivity. So, for example, and, as noted above, in the many attempts to determine the presence of the Fifth Force, it was realized that '[t]o obtain good sensitivity for short-range interactions … one should place the balance near a topographic feature such as a hill or cliff' (Adelberger 1990, p 3268).

Similarly, as emphasized by Potter, the test masses employed in a pendulum (torsional or otherwise) experiment should maximize what were believed to be the relevant differences in physical properties. Such maximization was consistently employed in the various replications by the Eöt-Wash group. So, for example, '[f]or maximum sensitivity, the test bodies should have relevant properties (binding energy per unit mass, atomic charge Z, neutron-to-proton ratio N/Z, etc.) that differ by the greatest practical amount' (Gundlach *et al* 1997, p 2523). Such maximization of the relevant physical differences in the test masses also played a central role in the solar experiments by Roll *et al* as discussed above in section 5.2.

11.1.3 Idealization and approximation

The variation between predicted and experimentally determined values is a function not only of the systematic and statistical uncertainty but also a function of the *idealizations and approximations* employed[1]. Thus it should not be surprising that variations between predicted and observed values persist even after systematic and statistical uncertainty is taken into account. How then should any such *residual* variation be dealt with? The most straightforward response is to develop *less idealized and approximate* analyses of the experimental situation.

In chapter 3, we briefly mentioned in this regard *increasing more accurate* analyses of pendulum motion. Similarly, as discussed in section 5.3, there was the use made of *more realistic* models of the Earth. As summarized by Eckhardt:

> To the zeroth order, the Earth is a sphere held together by gravitation, and the plumb line is directed toward the center of the sphere. To the first order, the Earth is an ellipsoid of revolution held together by gravitation but deformed by the centrifugal forces of its rotation, and the plumb line is not, in general, directed toward the center of the ellipsoid. The ellipsoid is an equipotential surface: *Horizontal gravitational forces … on passive gravitational masses are exactly balanced by opposing centrifugal forces … on inertial masses* …. in the absence of local mass inhomogeneities the Eötvös experiment is quite insensitive to any intermediate-range (small compared with the Earth's radius) coupling of any nature … *in effect there would be no horizontal 'fifth force' component, so the torsion balance would sense nothing*. (Eckhardt 1986, p 2868, emphasis added)

[1] There is no hard and fast distinction between idealizations and approximations other than, to our mind at least, that idealizations tend to be global and somewhat gross misdescriptions of the phenomena whereas approximations carry the connotation of being more closely associated with better behaved mathematical approximations, which converge in the limit to more realistic values.

In this case, the use of a more realistic model of the Earth served to *delegitimize* any test for the Fifth Force that was not conducted near to a large mass. On the other hand, as emphasized by Fischbach *et al* (1986b, p 2869), the use of an idealized spherical earth approximation doesn't matter once the experiment is made close to a nearby large mass because the anticipated effect is sufficiently amplified so as to overwhelm any small uncertainty due to use of the spherical earth model.

A variation on this theme is provided by the simplifying assumption, made by Gauss and Laplace, of a uniform gravitational field and by Roever's replacement of that assumption with a more realistic analysis of the gravitational field that took into account meridional curvature. But while Roever's more realistic analysis led to a southerly deflection nearly five times greater, that increase was not sufficient to explain away Hall's still considerably larger deflection values.

Sometimes the uncertainty associated with the idealizations employed may be so unruly and unmanageable as *to discredit the entire process* as was the case with the intermediate range force 'gap' noted in (Su *et al* 1994, p 3279) where this 'gap' was not rectified until more accurate descriptions and analyses of the relevant geological environment were developed. A dramatic example of the discrediting of an experiment and its idealized analysis is that of the collapse of Galileo's argument regarding what he claimed was shown by his supposed isochronal pendulum. While highly ingenious, the discordance between the idealized claim of being isochronal and the actual pendulum performance was sufficient to render Galileo's argument irredeemably unsound.

Finally, we note that the uncertainty associated with the use of certain idealizations and approximations may *simply disappear* with the emergence of a new theory as was in the case regarding the uncertainty and disagreement regarding how to deal with light reflecting off the moving (with respect to the ether) mirrors of the Michelson–Morley interferometer. Here, the Special Theory of Relativity eliminated the uncertainty because there was no ether and hence no complicating relative motion with respect to the interferometer mirrors and the ether (see section 8.1, Michelson *et al* (1928), and Miller (1933, pp 238–9)).

11.1.4 Sensitivity with respect to data analysis

One feature of science is the increasingly sophisticated use of statistics in the analysis of the experimentally determined raw data, and the occasional upheaval of what had hitherto been accepted interpretations of that data. Our initial example of such upheaval was Fischbach's reanalysis of the Eötvös data, where the status of the revised non-null result is still unresolved (see section 5.1).

We also saw this sort of sensitivity in data analysis in the reexamination of Miller's data by Shankland *et al*, and later in a more sophisticated form by Roberts. There the previously unexplained inconsistency of phase among the epochs was resolved because, as shown by Roberts, the reported displacements were not statistically significant precisely because Miller's data upon analysis had nothing to say about phase and could only place a restriction on the maximum velocity of the Earth's motion through the ether.

Finally, as discussed in chapter 10, there is the case of the Klapdor-Kleingrothaus *et al* Bayesian reanalysis of their 2001 data, which claimed it demonstrated the existence of neutrinoless double beta decay. After a very extensive set of experimental replications and increasingly more sophisticated statistical analyses, the Klapdor-Kleingrothaus *et al* Bayesian reanalysis has been set aside.

11.2 The roles of theory

11.2.1 Theories of the phenomena

In our studies, we have seen that the existence of well-articulated and well-confirmed theories of the phenomena under investigation have allowed null experiments to play its varied roles. Changes in such theories have also allowed these experiments to play different roles. Thus, the experiments on free fall initially refuted Aristotle's theory of falling bodies and later confirmed Newton's Law of Universal Gravitation and his laws of mechanics. With the advent of General Relativity, these experiments, while not refuting Newton's mechanics, became important confirmations and ever more stringent tests of that theory.

In the case of neutrinoless double beta decay, the null experiments, which at the present time have not yet conclusively decided the important issue of whether the neutrino is a Dirac or Majorana particle, should, in the near future, decide the issue.

Even in the absence of a detailed theory, the searches for physics beyond the Standard Model have provided both constraints on theories of supersymmetry and also provided evidence for other possible theories to explain.

11.2.2 Theories of the apparatus

The existence of well-developed, highly precise and relatively stable descriptions of the experimental apparatus employed and what it is that's being determined is obviously of the utmost importance when it comes to replication, and here we have in mind not just what might be described as an *exact replication* but also for an *improved replication*. A clear sense of what constitutes an improvement clearly depends on how well the original experiment is described and understood. Examples in this regard are: Replications of the Michelson–Morley experiment; beginning with Michelson's improvements, Miller's various developments, and the later Kennedy, Illingworth and Joos replications; the post 1991 Eöt-Wash *improved replications* of their tests of WEP and determination of the upper limits on non-Newtonian gravitation. We note here that Michelson received the 1907 Nobel Prize in Physics 'for his optical precision instruments and the spectroscopic and metrological investigations carried out with their aid'. Recently, in fact, the LIGO-VIRGO collaboration used a very large Michelson interferometer (each arm was 4 km long) in the first direct detection of gravitational waves.

Despite the relative stability in physics of descriptions of apparatus and data, the *refinement* of an experiment may often be a function of changes to the overriding theory.

11.3 Replication in physics and the social sciences

The general features of experimental physics described above stand in obvious contrast with the situation in the social sciences, and to a lesser degree with medical research. So it should be no surprise that there is very little in experimental physics that corresponds to the 'crisis' of replication in the social sciences. But how could it be otherwise given the relatively narrow (and well controlled) subject focus of physics versus the highly complicated and not well understood behavior of human agents? You don't need to get informed consent from an electron or proton, or even a neutrino in order to experiment on them, but not so for human subjects.

Still, it is nevertheless worthwhile to take at least a brief look at what similarities there may be and how the experimental methods employed in physics may usefully apply in psychology research. In our Introduction, we drew attention to what has been described as a 'crisis' of failed replications in the social sciences, especially psychology. There have been many proposed explanations of this failure of replication including shortcomings in statistical methodology as well as the more elusive difficulties involved in the selection of test subjects and the specification of the test environment[2]. It is this latter proposed explanation, especially as it applies in psychology, that we wish to focus attention on. This because, as will be seen, physics also has had to deal with its own version of this problem.

In psychology, the argument is that 'the failure to reproduce results might reflect contextual differences—often termed 'hidden moderators'—between the original research and the replication attempt' (Van Bavel *et al* 2016, p 6454). In other words, the problem is that while two experiments may be nominally similar, the existence of unaccounted for 'hidden moderators'[3] creates the risk that the despite their similarities, the experiments are different in causally relevant ways—hence there will be a failure of replication. Examples of such hidden moderators include culture, location, and population. Obviously, these are broad terms without clearly defined boundaries. But that's the point; namely, that they serve to mark off—albeit imprecisely—a potential range of factors some of which may be obvious in their influence while others may be hidden and unrecognized. More specific candidates for such hidden moderators 'range from obvious but sometimes overlooked factors, such as the race or gender of an experimenter, temperature, and time of day, to the more amorphous (e.g. how the demeanor of an experimenter conducting a first-time test of a hypothesis she believes is credible may differ from that of an experimenter assessing whether a study will replicate)' (Van Bavel *et al* 2016, p 6458).

The problem therefore for research in psychology is how to identify with sufficient specificity such 'hidden moderators' and, having done so, how to deal with them in a way that raises the likelihood of successful replication and thus subsequent confirmation. That too, as we have seen, is also a problem in physics. Our question

[2] For an overview of the various proposed explanations of the replication 'crisis' in psychology in particular, see National Academies of Sciences (2019, pp 122–4 National Academies of Sciences, Engineering and Medicine), Van Bavel (2016 #1224, pp 6454–65), and Zwaan *et al* (2018, pp 2–3).

[3] In physics and philosophy of science, these are often referred to as 'confounding factors'.

is whether the methods used in physics to deal with systematic uncertainty, i.e. the physics variant of uncertainty, due to 'hidden moderators', are of any relevance for research in psychology and more specifically the high rate of replication failure.

As a starting point on the physics side, consider, for example, the extensive *series of replications* by the Eöt-Wash group of their tests of WEP. What was involved here was not exact replication but rather replication *with improvement* where there were standards of what constituted such improvement. Also, here we note that the driver of this series of replications was the specific goal of *narrowing the error interval* around the zero value—and where as discussed in section 5.3 there was a well-grounded theoretical basis for extending the raw data to well defined limit boundaries for a non-Newtonian gravitational component. To successfully achieve this goal of narrowing the error interval involves, in large part, as we have seen, the management of *systematic uncertainty*[4].

In physics the management of systematic uncertainty begins with the identification of the likely systematic effects and the resultant uncertainty. So too in psychology where the hidden moderators must somehow be unearthed and thus made available for investigation and experimental management. In this regard, it has been suggested that in psychology 'failed replication attempts represent an opportunity to consider new moderators, even ones that may have been obscure to the original researchers, and to test these hypotheses formally' (Van Bavel *et al* 2016, 6457). Rolf Zwaan *et al* have made a similar recommendation:

> If a … replication fails to obtain the same result as the original study, researchers may question whether the initial result was a false positive (and this will be especially true after multiple failed direct replications) or whether there is a misunderstanding about the understanding of the essential features required to produce an effect. This will likely prompt a more critical evaluation of the similarities between the original study and the replication. (Zwaan *et al* 2018, p 4)

There is, however, a significant difference in the role played by replication in physics as opposed to the role played in psychology. In physics, as we have seen, the aim is not to produce what may be described as 'exact' replications, but rather to created *improved* replications. Moreover, the concept of replication in physics can be naturally expanded so as to regard, for example, as *mutual replications* the floating sphere experiment of Thieberger and the Eöt-Wash torsional pendulum experiment. Such expansion is justified *by the common underlying theoretical basis and aim*, namely, testing the Fifth Force hypothesis.

In psychology, however, the emphasis has been on producing 'exact' replications. This emphasis is due to the realization that a great many of the experiments in psychology cannot be replicated even using the same experimental procedures and methods of data analysis. But the existence of 'exact' replications in psychology is

[4] It may also involve taking more data to reduce the statistical uncertainty.

problematic because it has been claimed with some justification that 'there is no such thing as exact replication' in the field of psychology (see Anderson *et al* 2016, p 1037, and the supporting citations therein). Similarly, Van Bavel *et al* argue that while the methods section of an experiment's description 'should include enough detail to permit a direct replication, this seemingly reasonable demand is rarely satisfied in psychology, because human behavior is easily affected by seemingly irrelevant factors' (Van Bavel 2016, p 6455).

Assume for the moment that it is indeed the case that 'there is no such thing as exact replication.' If so then one can readily imagine a situation where there exists a set of failed replications all ostensibly about the same subject matter. In such cases, their applicability must be restricted to the *specific* situation in which they were performed, and that is at least part of the sense in which the failure of replication—now understood as *never being able to get beyond the specifics of the individual research*—poses a 'crisis' for psychology. Such a dreary conclusion, however, depends on whether or not further scrutiny *prompted by the discordance* leads to a better appraisal of the 'hidden moderators' at work such that some subset of the original set of failed replications are judged to be superior in the sense of involving more of the otherwise hidden moderators[5].

In order to add some specificity to how such discordance might work in psychology, we'll briefly review (Dijksterhuis and Van Knippenberg 1998), which was 'one of the most well-cited' experiments dealing with what's known as the 'priming' effect. The motivating background derived from various earlier studies showing that 'brief exposure to a category or construct can mentally activate related categories or constructs', and even 'directly affect overt behavior'. Thus, experimental subjects 'are faster to recognize the word doctor after initially seeing the word nurse,' and will walk more slowly to a nearby elevator after being exposed to words that 'related to stereotypes of older adults' than subjects who have been exposed to words that were 'neutral'' (O'Donnell *et al* 2018, p 269).

In Dijksterhuis and Van Knippenberg (1998), one set of experimental subjects were 'primed' by being asked to imagine what their daily life would be like as a 'professor', and the other set what their life would be like as a 'soccer hooligan'. After writing a paragraph about their imaginings, the subjects were given an unrelated trivia test. The 'professors' scored significantly better than the 'soccer hooligans'.

Despite the accolades, including having been cited more than 800 times, the experiment could not be replicated by any of the 23 labs involved in the replication attempt despite the efforts made to reproduce the original experimental procedures as closely as possible but where the trivia questions were modified according to the country where the replication attempts were conducted. The following were offered as explanations for the failed replication: (1) experimental subjects were likely to realize the purpose of the experiment because of the ubiquity of the professor-priming effect in modern psychology courses; (2) the original study was conducted in

[5] For a discussion of discordant results in physics, see Franklin (2002a, chapters 7–10) and for a general discussion of repetition in physics, see Franklin (2018).

The Netherlands, where the then existing social cultures of professors, hooligans, and experimental participants have changed, and in any case are not likely to be matched in the other countries where the replications were attempted. For details, see O'Donnell *et al* (2018, pp 276–8).

In sum, while the discordance between the original research and its failed replications did serve to reveal what were otherwise 'hidden moderators,' the bottom line is that this experiment was unable to reach beyond the specifics of the original research environment. And one might well expect the same fate for any updated replacement. Moreover, this episode of failed replication serves to highlight a *fundamental difference* between research in physics and psychology which *significantly constrains* the efficacy of psychology experimentation. In physics, unless the experimentation is exploratory, there is a clearly understood purpose and associated target value both of which have significant theoretical underpinning. Thus, in the case of the Eöt-Wash series of replications the purpose was to test the weak equivalence principle and at the same time to lower the limit boundaries for any non-Newtonian gravitational component. Highly developed theoretical underpinning was also present, and in fact required, in the high-energy physics cases discussed earlier in order to specify the background and the corresponding test signal. Given such well-defined aims and target values, the experimental problem in physics is to make a determination of such target values *after having taken into account systematic uncertainty*. But in psychology, there aren't such clearly defined aims and associated test values. Thus, while there may be analogues for amplification and the use of calibration signals in psychology, they are not likely to be as well developed and robust as their exemplars from physics.

An aggravating variant of above sort of problem for psychology research is the selection of suitable experimental *surrogates* for the more general psychological concepts in question. So, for example, in Finkel *et al* (2002), the basic question was 'what motivates partners to forgive?' The problem here was that while there were many studies that reported 'an association between relationship commitment and willingness to forgive transgressions' the direction of causation was undetermined. In order to determine the direction, the experimenters 'used a priming task to *experimentally manipulate* commitment (low or high) and then assessed forgiveness responses' (Cheung *et al* 2016, p 751, emphasis added). The experimental subjects 'were primed by writing responses to open-ended prompts that guided them to think about either their dependence and commitment to their partner (high commitment) or their independence and lack of commitment to their partner (low commitment).' They were then asked to react to 'descriptions of 12 hypothetical betrayals committed by their partner and indicated how they would react' (Cheung *et al* 2016, p 751).

Simplifying the complicated taxonomy used to categorize the responses, the reported experimental result was that 'forgiveness' measured higher for those primed for high commitment than for those primed for low commitment. But once again the experimental results could not be replicated. 'The findings from [Finkel *et al* 2002] provide no evidence for (or against) the causal role of commitment in the forgiveness process' (Cheung *et al* 2016, p 761). This time, however, there was no ready answer

for the failure of replication and the best that could be done was to suggest 'the possibility that a different manipulation might reveal a causal effect of subjective commitment on forgiveness' (Cheung *et al* 2016, p 761).

Underlying all this, however, is the assumption of the adequacy of the test surrogates used to measure commitment and forgiveness, and the possibility of manipulation of such commitment. In short, the problem in psychology exemplified in this example is to somehow attach the psychological traits at issue with a tangible experimental procedure all in the absence of anything like the firm theoretical basis for such attachment that exists for experimentation in physics. In other words, Finkel *et al* (2002) requires the assumption that 'priming' experimental subjects by having them write responses to certain prompts serves as a reliable pathway to understanding human forgiveness. The absence of a firm theoretical basis for such an assumption thus creates a fundamental *disanalogy* between the systematic uncertainty of experimental physics and the problem of the existence of 'hidden moderators' in psychology research. And because of this disanalogy, there is a greater risk of replication failure in psychology. There is too much distance and slack between human forgiveness and experimentally feasible priming procedures. There is, of course, the existence of an experimental culture where concepts such as 'priming' are frequently employed. Add as well, some common sense intuition about human nature and one does have something of an underlying theoretical basis. But even so, it falls far short of what exists in physics.

Thus, while there may be some genuine similarities between experimental procedure in physics and psychology in their treatment of systematic uncertainty and hidden moderators, such similarities will, in the case of psychology research, be constrained and limited. While the mental lives of electrons, protons, and neutrinos, and the many other entities of theoretical physics are nothing like those exhibited by human subjects, such simplicity does have the benefit of making their theoretical analysis more direct and certain.

References

Adelberger E G, Stubbs C W, Heckel B R, Su Y, Swanson H E, Smith G, Gundlach J H and Rogers W F 1990 Testing the equivalence principle in the field of the Earth: particle physics at masses below 1 μeV? *Phys. Rev.* D **42** 3267

Anderson C J *et al* 2016 Response to Comment on 'Estimating the reproducibility of psychological science' *Science* **351** 1037

Cheung I *et al* 2016 Registered replication report: study 1 from Finkel, Rusbult, Kumashiro and Hannon (2002) *Perspect. Psychol. Sci.* **11** 750–64

Dijksterhuis A and Van Knippenberg A 1998 The relation between perception and behavior *J. Pers. Soc. Psychol.* **74** 865–77

Eckhardt D H 1986 Comment on 'Reanalysis of the Eötvös experiment' *Phys. Rev. Lett.* **57** 2868

Finkel E J, Rusbult C E, Kumashiro M and Hannon P A 2002 Dealing with betrayal in close relationships does commitment promote forgiveness? *J. Pers. Soc. Psychol.* **82** 956–74

Fischbach E, Talmadge C and Aronson S H 1986b Response to Eckhardt *Phys. Rev. Lett.* **57** 2869

Franklin A 2002a *Selectivity and Discord* (Pittsburgh, PA: University of Pittsburgh Press)

Franklin A 2004 Doing much about nothing *Arch. Hist. Exact Sci.* **58** 323–79

Franklin A 2018 *Is It the Same Result? Replication in Physics* (Bristol: IOP Publishing)

Gundlach J H, Smith G L, Adelberger E G, Heckel B R and Swanson H E 1997 Short-range test of the equivalence principle *Phys. Rev. Lett.* **78** 2523

Michelson A A, Lorentz H A, Miller D C, Kennedy R J, Hendrick E R and Epstein P S 1928 Conference on the Michelson–Morley experiment held at Mount Wilson February 1927 *Astrophys. J.* **68** 341

Miller D C 1933 The ether-drift experiment and the determination of the absolute motion of the earth *Rev. Mod. Phys.* **5** 203

National Academies of Sciences, Engineering and Medicine 2019 *Reproducibility and Replicability in Science* (Washington, DC: The National Academies Press)

Newton I 1999 *The Principia, Mathematical Principles of Natural Philosophy* (Berkeley, CA: University of California Press)

O'Donnell M *et al* 2018 Registered replication report Dijksterhuis and Van Knippenberg (1998) *Persp. Psychol. Sci.* **13** 268–94

Roll P G, Krotkov R and Dicke R H 1964 The equivalence of inertial and passive gravitational mass *Ann. Phys.* **26** 442–517

Stubbs C W, Adelberger E G, Raab F J, Gundlach J H, Heckel B R, McMurry K D, Swanson H E and Watanabe R 1987 Search for an intermediate-range interaction *Phys. Rev. Lett.* **58** 1070

Su Y, Heckel B R, Adelberger E G, Gundlach J H, Harris M, Smith G L and Swanson H E 1994 New tests of the universality of free fall *Phys. Rev.* D **50** 3614

Van Bavel J J, Menda-Siedlicki P, Brady W J and Reinero D A 2016 Contextual sensitivity in scientific reproducibility *Proc. Natl Acad. Sci.* **113** 6454–9

Zwaan R, Etz A, Lucas R E and Donnellan M B 2018 Making replication mainstream *Behav. Brain Sci.* **41** e120