When "Replicability" is More than Just "Reliability": The Hubble Constant Controversy

Vera Matarese[*] C.D. McCoy[†]

8 April 2022

Abstract

We argue that the epistemic functions of replication in science are best understood by their role in assessing kinds of experimental error. Direct replications serve to assess the reliability of an experiment through its precision: the presence and degree of random error. Conceptual replications serve to assess the validity of an experiment through its accuracy: the presence and degree of systematic errors. To illustrate the aptness of this view, we examine the Hubble constant controversy in astronomy, showing how astronomers have responded to the concordances and discordances in their results by carrying out the different kinds of replication that we identify, with the aim of establishing a precise, accurate value for the Hubble constant. We contrast our view with Machery's "re-sampling" account of replicability, which maintains that replications only assess reliability.

1 Introduction

The replicability crisis, much discussed in connection with certain scientific disciplines, like psychology and medicine, which are said to be engulfed in it, has provoked an expanding philosophical debate on the concept of replication and its place in the epistemology of science. This debate has centered on three inter-related issues. The first issue concerns the status of replicability in the epistemology of science. While the essential replicability of experiment has traditionally been thought of as a pillar supporting the objectivity of science (Dunlap, 1926; Popper, 2002), some have questioned whether pervasive non-replicability necessarily impugns the credibility of those sciences in which it occurs. Norton (2015) and Leonelli (2018) are among those who claim to oppose tradition by proposing that non-replicability need not represent an epistemic failure in certain scientific contexts. The second issue concerns the meaning and interpretation of replication. Although the terminology used in connection with replicability varies considerably by discipline, and even by individual author, one common version distinguishes between direct replications, often described as (near) exact duplications of the original experiment, and conceptual replications, often described as experiments which test the same hypothesis of a previous experiment but change the methods involved. Dissatisfaction with distinctions of this kind has led philosophers, such as Machery (2020), and scientists, such as Nosek et al. (2022), to offer improved accounts of what replication is and what kinds there are. The third issue concerns the explication of the epistemic functions or roles of replicability in science. While it is usually the case that replications are regarded as some kind of "validation" or "confirmation" of the results of an earlier experiment, different authors have thematized

[*] Institute of Philosophy and Center for Space and Habitability, University of Bern, Bern, Switzerland. email: vera.matarese@csh.unibe.ch

[†] Underwood International College, Yonsei University, Seoul and Incheon, Republic of Korea. email: casey.mccoy@yonsei.ac.kr

this function in different ways. Fletcher (2021), for example, unifies replicability's role by claiming that replications serve to undercut the underdetermination of hypotheses by empirical evidence.

In this paper, we regard the third issue as primary, and take our proposed resolution of that issue to shed light on the first two issues. In our view, the epistemic functions of replication are to be understood by attending to the kinds and roles of error (or uncertainty) in experimental science. Our claim is that different kinds of experimental replications serve to assess different kinds of error (or uncertainty). This position answers directly to the third issue. As there are essentially two kinds of error that experiments can be used to assess — random (or statistical) error and systematic error — we therefore concur with the general view that two broad (idealized) categories of replications should be distinguished, and hence our response to the third issue also provides a response to the second. Although the popular terminology for these two categories may not be particularly apt ("direct replications" need not be especially "direct," and in general there need be nothing necessarily "conceptual" about "conceptual replications"), we nevertheless choose to use it due to its increasing familiarity in the scientific and philosophical literature. For us, however, "direct replications" have the ideal function of assessing the precision (and thereby the reliability) of previous experiments, where precision describes the presence and degree of random errors; "conceptual replications" have the ideal function of assessing the accuracy (and thereby the validity) of previous experiments, where accuracy describes the presence and degree of systematic errors.

Real experiments, of course, may blend together aspects of both kinds of idealized replications (and other, non-replicability-related function as well), yet often enough in practice a replication is clearly performed for the sake of one of these functions rather than the other. In such cases, it is natural to describe the experiment as a replication of the appropriate kind. We emphasize that a general norm of replicability covering these two kinds need not mandate the actual carrying out of any particular experimental replication, at any particular time. Nevertheless, such a general, two-fold norm is demanded epistemologically, since experimental results, and the hypotheses to which they are evidentially relevant, cannot be regarded as both reliable (precise) and valid (accurate) without some reason to think that the experiment is replicable in these two ways. Thus, non-replicability is only an issue for a science to the extent that there is insufficient reason to trust that replications would appropriately confirm previous results. Thus, our proposal answers to the first issue as well.

Although our proposal may strike some readers familiar with experimental methodology as nothing more than a commonplace, it seems to us that the familiarity of its basic ideas nevertheless belies full comprehension of these ideas' epistemological significance, especially in connection to replicability. We are, of course, far from the only authors suggesting that considerations of experimental error are relevant to debates on replicability. Bird (2021) and Machery (2021), for example, have recently drawn attention to the possibility that the prevalence of certain kinds of experimental errors may be responsible (at least to some extent) for the replicability crises that some sciences are undergoing. Our interest, while provoked by the replicability crisis, is not so much to connect considerations of experimental error to these crises but rather to articulate an account of replicability in science based on such considerations (which may then be applicable to the assessment of putative replication crises).

Moreover, there is a need to make these basic connections explicit to contest some digressive moves in the recent philosophical literature on the topic of replication. For example, several philosophers have lately taken the disciplinary peculiarities of specific sciences to favor "localized" standards of replication in science (Leonelli, 2018; Guttinger, 2020; Fletcher, 2021). According to this "new localism," as Guttinger (2020) dubs it, "replicability itself should not be treated as a universal standard," essentially because there are scientific contexts where he (and others) claim it does not apply (or at least applies differently). According to our proposal, however, replicability is methodologically significant for all experimental science. A generalized "localism", as such, should therefore be steadfastly resisted as a matter of epistemological principle. Indeed, it is precisely the erosion of epistemic standards that should be (and often is) a matter of great concern for sciences suffering from a replicability crisis. Granted, the local realization of basic epistemic values is necessarily shaped by the relevant epistemic, methodological, and practical contexts. Indeed, there are perfectly good practical reasons not to insist on certain replications being carried out (obviously so in the extreme case of singular events), in a variety of circumstances. Such contextuality does not undermine the basic imperatives of the epistemic values of reliability and validity. A scientist who wants to claim that her theory is true or her experimental results are correct must be able to justify their validity; a scientist who wants to claim that her theory is coherent or her experimental results are reliable should be able to likewise defend their reliability. Naturally, if she does not wish to claim these things, then the relevant epistemic standards need not apply. Thus, for sciences where theory and experiment are in development or exploratory, and no knowledge claims are being advanced, there is clearly no need to enforce a standard of replicability (although this is certainly not to say that replicability plays an insignificant role in developmental and exploratory science).

Another of our motivating concerns arises from the fact that contemporary discussions on replication, both in the philosophical and scientific literature, tend to be informed primarily by scientific cases from the disciplines most affected by the crisis, in particular psychology. We see two worrying consequences that can easily arise from this circumstance. One is that an intimate acquaintance with the peculiarities of a specific science can easily distort the general epistemological issues that are at stake (potentially leading, e.g., to the "localism" just mentioned). Two, there is substantial risk in grounding a discussion on the functions of replications on disciplines that are going through a replicability crisis — the more so when those disciplines also rely on the use of qualitative and inexact methods. It is precisely in the confusing circumstances of such a crisis that the functions of replication are most obscured.

For these reasons, we think it profitable to turn for guidance to disciplines which have long incorporated replication into their experimental methodology, especially those that do so in a constructive way. While there is a good number of such examples across a wide range of sciences, we are compelled to narrow our focus here to a single, instructive case (which we believe is also of considerable independent interest). It concerns recent efforts to measure the value of the Hubble constant, a cosmological parameter that quantifies the rate of expansion of the universe. At present, there is a surprising discordance in results from three major experiments measuring the value of the Hubble constant. This circumstance has in effect given astronomy a mini-"replicability crisis" of its own. The story of how the discordance emerged and how it is being solved will not only help us illustrate key planks in our replicability platform but also help us reveal shortcomings in competing accounts of replicability.

2 Replication: Its Nature and Kinds

Although one finds a variety of classifications of replications in the scientific literature, it is relatively common for philosophers (and many scientists from certain disciplines) to distinguish between two categories: "conceptual" replications and "direct" replications (cf. (Romero, 2019)). As said in the introduction, we choose to use this conceptual/direct terminology to distinguish kinds of replications in terms of experimental function. It is normally used, however, to indicate which aspects of the experiment change from the original experiment to the replication. Thus, a (successful) conceptual replication is usually described as re-obtaining the results of a previous experiment by different methods or procedures; a (successful) direct replication is usually described as re-obtaining them by (substantially) the same method or procedure.

Recently, Machery (2020) has offered a forceful critique of the latter way of distinguishing kinds of replication, arguing that the category of conceptual replications, as just described, is incoherent and should be abandoned. While replications that "change methods" are normally referred to as conceptual replications, Machery argues that we should not give them a special designation in virtue of that alone. According to his "re-sampling" account of replication, all components of an experiment (the experimental units, the treatments, the methods, etc.) can be treated either as "fixed" factors or "random" factors. Components that are regarded as random factors in one experiment can, in a subsequent replication, be re-sampled from the same population or from a different one. Insofar as conceptual replications are thought to involve a change of method, this change could be a change from one fixed method to another, a re-sampling from the same population of methods, or a sampling of methods from a different population of methods. If these are replications that take methods as random factors and sample from the sample population, then Machery argues that there is no need to oppose them against direct replications, for they are functionally identical. That is, direct replications are also normally understood as acts of sampling experimental units (most of the time) from the sample population. In the other two cases, replications that either treat methods as fixed factors or else take them as random factors but sample them from a different population, Machery claims that what are normally called "conceptual replications" are "extensions" or other experiments entirely, and not replications at all.

Whatever the classification one adopts, whether the usual direct/conceptual distinction or the one introduced by Machery (between replications and extensions), we would emphasize that the description of an experiment as being of some particular kind of replication depends on some specific interpretive choices. This is especially evident in the case of Machery's account. For Machery, if some aspect of the original experiment is regarded as a fixed factor, modifying that fixed factor in a novel experiment makes that experiment an "extension" (or another experiment entirely). If a component of the original experiment is regarded as a random factor, then re-sampling that component from the same population in a new experiment makes that experiment a "replication." Since it matters how the components are regarded, these are not facts about the experiments but interpretive choices made by the experimenters.

Implicit in the interpretation of an experiment as one kind or another in the context of Machery's account is the idea that the predecessor experiment targeted a particular hypothesis (or set of hypotheses) with a certain degree of generality, and likewise for the successor experiment. Thus, for Machery, an extension is so-called because it targets an extension of the original hypothesis (or set of hypotheses) to a more general one. This broadening of scope requires treating the relevant factors which change as fixed factors (or as random factors but with a change of population) rather than random ones (with sampling from the same population). A replication, by contrast, maintains the original hypothesis, and this requires treating the relevant factors which change as random factors (with samplings from the same populations). How one conceptualizes the hypothesis under test thus fully determines whether the components of an experiment are fixed or random (with the same population or a different one). Similarly, the "same results" clauses in the conventional distinction implicitly indicates the common targeting of the same hypothesis.

A given concrete experiment does not dictate its interpretation in these respects. Different scientists may have different hypotheses in mind when assessing the relevant experiments' impact on those hypotheses. Indeed, in principle, any hypothesis which is evidentially dependent on the experimental results can be fairly regarded as the "target hypothesis" of the experiment. Because of the hypothesis's role in dictating the interpretation of an experiment as one kind of experiment or another, we see that the underlying idea of such distinctions between kinds is that experimentation's fundamental function is hypothesis testing (or hypothesis confirmation, in the sense used by philosophers of science). Naturally, this is not to say that scientists always perform experiments minded to test some hypothesis or set of hypotheses (although they frequently do); it is just to say that the salient epistemic function of experimentation can be naturally and easily regarded as such in any case whatsoever.

Given this understanding, we take as our basic standard for some experiment being a replication that it may be interpreted as targeting the same hypothesis for evidential appraisal as another experiment. This is evidently the case on Machery's account as well. Nevertheless, for him, the only appraisal that a replication can make is an appraisal of reliability (Machery, 2020, 547, 556, 559, 561, 565). It is on just this point that we part company with Machery's account, for we maintain that there are scientific experiments (like some of those carried out in the experimental program to measure the Hubble constant) which are clearly replications in a relevant sense and also have the function of appraising validity.[1]

[1] We do not yet enter a discussion on the meaning of reliability and validity. The reader may consult Machery's discussion of reliability and validity (Machery, 2020, 554–555), which is applicable to our claims in this section.

Such experiments cannot be captured coherently on Machery's account of replicability, for they are neither re-samplings for appraising reliability, nor extensions testing a more general hypothesis, nor other experiments entirely.

While we fully agree with Machery's critique of a distinction in replications based on a mere difference between "different methods" and "same methods," it would be a serious mistake to discard a distinctive category of replications related to so-called "conceptual replications" entirely, for something of the kind is essential to capture experimental practice, understood as hypothesis confirmation, in science. Both components, reliability and validity, are essential aspects of confirming a scientific hypothesis (just as logical validity and truth are essential to assessing deductive argumentation). Without reason to think that the experimental evidence supporting a hypothesis is reliable and valid, that hypothesis cannot be regarded as confirmed. The means of assessing these qualities is carrying out replications of the corresponding type.

We postpone, for the moment, filling out how specific kinds of error relate to reliability and validity in favor of first laying out the relevant details of our case of interest, the experimental efforts over the past decade in astronomy to measure the Hubble constant. When we subsequently fill out how experimental error connects to epistemology, we will then be able to develop an interpretation of these experimental efforts which indicates how different kinds of replications are performed in order to make progress in experimental knowledge. By carrying out different experimental procedures at different stages of a dynamic experimental context, the astronomers involved have sought to use replications to assess the reliability and validity of their results for confirming a specific value of the Hubble constant. These assessments not only provide a justification for their conclusions but also give guidance to the experimenters on which experiments they should perform next as part of their respective research programs.

Particularly important at the present stage of investigation in astronomy, given the discordance in the major program's results for the value of the Hubble constant mentioned previously, is the employment of distinct methods (i.e., conceptual replications) to measure the Hubble constant, the aim of which is to expose the underlying cause of the discordance. These efforts reveal an important dimension in the class of conceptual replications, involving the degree of independence from the original experiment. Along this dimension, conceptual replications range from those that are maximally independent from the original experiment to those that are only partially independent. The degree of independence of a replication has tremendous epistemic and methodological significance, since fully independent replications can only cross-check results, whereas partially independent replications can expose overlooked sources of discordance. We return to this consideration in more detail below after presenting the details of our case.

3 The Hubble Constant Controversy

In the context of expanding universe models of cosmology, the Hubble constant (H_0) is a quantity that represents the present rate of spatial expansion of the universe. Determining its value has been one of the most important goals of experimental research in astronomy for nearly a century, since Hubble first (inaccurately) measured its value as 500kms^{-1}Mpc^{-1} (Hubble, 1929).[2] Finding an accurate value for H_0 has always had a great deal of theoretical significance in cosmology; its value has implications for what the material (and non-material) components of the universe are, what the age of the universe is, and what its eventual fate is. It also gives a convenient way to determine distances to astronomical objects like stars and galaxies. Especially in the past three decades, considerable progress has been made in narrowing its range. Nevertheless, during the last decade, a discrepancy between different measurement results has kindled a major controversy in astronomy and cosmology, which has led to a proliferation of many different experimental programs and theoretical alternatives to the standard cosmological model in

[2] This unit is kilometer-inverse seconds-inverse megaparsec. A parsec is a common unit of distance in astronomy equivalent to 3.26 light years or 206,000 astronomical units (au) (one au is essentially the mean distance between the Earth and the Sun).

the hopes of finding some resolution to the discordance.[3]

Broadly speaking, the standard experimental approaches to measuring the Hubble constant can be divided into two different groups based on their method of obtaining a value for H_0.

First, there are those programs which measure H_0 by inferring its value from measurements of other related cosmological parameters within a given cosmological model. In the context of the current standard model of cosmology, the ΛCDM model, probing certain "global" features of the early universe (near the time of the Big Bang), especially the cosmic microwave background (CMB) radiation (the "after-glow" of the Big Bang), allows one to infer a value for the Hubble constant. While a number of space missions have studied the CMB in recent decades, the best results have come from the European Space Agency's Planck satellite, in operation during the last decade. The latest results from the Planck team give a value for H_0 of $67.4\text{kms}^{-1}\text{Mpc}^{-1}$, with an uncertainty of less than 1% (Planck Collaboration, 2020).

Second, there are those programs which measure H_0 by inferring its value from measurements of "local" features (astronomical objects like galaxies, stars, etc.) of the late universe (relatively recent times). The relevant measurements are used to build up a "cosmic distance ladder" of intergalactic distances. A distance ladder will involve a variety of different astronomical objects and techniques, among them geometric direct distance measurements (e.g., parallax), standard candles (e.g., Cepheid variable stars, Type Ia supernovae), eclipsing binaries, etc. With a cosmic distance ladder in hand, one can use the velocity-distance equation ($D = vH_0$), which relates the distances D of galaxies to their recession velocity v relative to Earth (determined by measuring their redshifts), to calculate a best fit value for the Hubble constant H_0 (which must have, according to the velocity-distance equation, units of inverse time, although it is usually quoted, as in this article, in the preferred unit $\text{kms}^{-1}\text{Mpc}^{-1}$). The most consistently rigorous results obtained over the past decade have been by the SH0ES team led by Riess. Their best measurement gives a value of $73.2\text{kms}^{-1}\text{Mpc}^{-1}$, with an uncertainty of 1.8% (Riess et al., 2021), revealing a significant discrepancy (4.2σ) with the Planck result.

Given this discordance of results between measurements obtained, on the one hand, by looking at the early universe and, on the other, by looking at the late universe, the obvious question on physicists' mind is: what accounts for it? One possibility, enticing for many theoretical cosmologists, is a failure of the ΛCDM model itself, which requires that the Hubble constant measured locally and the Hubble constant measured via the CMB give the same value. It does so because the ΛCDM assumes that the universe is spatially homogeneous and isotropic, which means that the current expansion rate of space should be the same everywhere and not differ based on distance from our local position in the Milky Way.

The sober-minded judgment of many astronomers, based on long experience with erroneous experimental results, is that the well-confirmed ΛCDM model is not at fault. This conclusion is also supported by the extensive exploration of model changes that could account for the two approaches' discordant results, so far yielding only physically improbable models[4] The most plausible explanation for the discordance, then, is that one of the two experimental results is somehow wrong. Since building a reliable, accurate cosmic distance ladder is much more complicated, involving tracking a variety of different sources of error, than the measurements of the CMB in the Planck experiment, one prevailing suspicion is that the result obtained by the SH0ES program has not properly taken into account all the relevant error. Nevertheless, as the SH0ES program itself has emphasized, careful checking of systematic errors and repeated measurements (carried out by the SH0ES team and others) have consistently corroborated its results time and time again.

In the last few years, the Carnegie-Chicago Hubble (CCH) program, led by Freedman, has complexified the controversy by obtaining a different result from both SH0ES and Planck for the Hubble constant:

[3] There are numerous reviews covering these developments. The reader may usefully refer to (Freedman and Madore, 2010; DiValentino et al., 2021; Shah et al., 2021).

[4] Cf. (Shah et al., 2021, sec. 4); "despite many papers, no compelling theoretical solution to the Hubble tension has yet emerged" (Efstathiou, 2020). See (DiValentino et al., 2021) for an exhaustive review of proposals of new physics to explain away the Hubble tension.

69.06kms^{-1}Mpc^{-1}, also with a small uncertainty (Freedman et al., 2019). They adopt more or less the same local, late-universe approach as the SH0ES team, based on building up a cosmic distance ladder, but by relying on a different kind of standard candle for a key part of the ladder. For our purposes, it is this discrepancy, between the SH0ES team and the CCH team, which has the most interesting consequences for the topic of replicability, in particular because of the relative independence of their cosmic distance ladders.

Because of their importance, let us say a little more about cosmic distance ladders and their relation to the Hubble constant.[5] Given the complications of building a reliable and accurate cosmic distance ladder, it would be convenient if we could simply infer distances from surer data. If one somehow knew the value of the Hubble constant accurately and had accurate measurements of galaxy redshifts (from which one infers their recession velocity), then the velocity-distance equation would conveniently give all their distances with a single measurement technique (i.e., just by measuring redshifts and inferring recession velocities). However, it is the Hubble constant's value that we want to determine, so we need to use velocities and distances to determine it. Unfortunately, there is no single technique available that can accurately give the distances to all galaxies. Hence, it is necessary to build a cosmic distance ladder with a variety of techniques, "rung by rung," as it were.

As mentioned above, astronomers have identified a variety of techniques over the years for measuring distances to galaxies and other astronomical objects. In general, the precision and accuracy of all kinds of distance measurement decrease with distance. In general, different techniques are also applicable at different distances. These different techniques must therefore be calibrated to one another over distances where the techniques are both applicable. In this way, in a complete cosmic distance ladder, each step or "rung" of the ladder relies upon the previous step for calibration. Simplifying the complexities somewhat, three different measurement techniques, covering different but overlapping distance ranges, make up the typical cosmic distance ladder.

For small distances (less than roughly 5 kpc for the most advanced experiments), astronomers rely on parallax, which involves measuring the angular shift of a nearby star (i.e., within our galaxy) against the background of (essentially) fixed stars from opposite points in the Earth's orbit around the Sun.

For farther distances, astronomers make use of what are called "standard candles." Different standard candles apply at different distance scales. A standard candle is a kind of star (or astronomical object) whose intrinsic brightness is known in advance. The star's apparent brightness is then measured and compared to its intrinsic brightness in order to derive a distance (as brightness decreases with distance squared).

For intermediate distances (roughly between 100 pc and 50 Mpc, i.e., from within the Milky Way to nearby superclusters of galaxies), Cepheid variable stars have long provided astronomers with a fairly reliable standard candle. Cepheids are present in galaxies in a range of distances, from the neighboring Magellanic Clouds (dwarf galaxy companions to the Milky Way) to nearby galaxies in the Local Group of galaxies. Cepheids are hot and massive stars that brighten and dim periodically according to Leavitt's law, which proportionally relates the pulsation period of the star with its intrinsic brightness: the longer the period, the brighter the star. As their (mean) intrinsic brightness can be deduced from their pulsation period, comparison with their observed brightness yields a distance.

For even larger distances (up to 1 Gpc), astronomers predominantly rely on the standard candles known as type Ia supernovae. Type Ia supernovae occur when extremely dense stars (white dwarfs) explode after stealing sufficient mass from their binary system companions (an aging red giant in one standard model) to trigger a runaway fusion reaction. Despite being rare, one-time events, they are thought to be particularly good standard candles, since at peak brightness all supernovae of this type are supposed to have the same intrinsic brightness (because they always form when the white dwarf reaches the same amount of mass), which allows one to infer the distance to their host galaxy.

[5] More details are available in, among many other places, (Freedman and Madore, 2010).

The two experimental programs we have mentioned so far, SH0ES and CCH, are both focused on developing a precise and accurate cosmic distance ladder, but in measuring the value of the Hubble constant they rely on different standard candles for the intermediate distances. The SH0ES program relies principally on Cepheids (we are simplifying somewhat, since real distance ladders incorporate as many distance indicators as possible). The CCH program has instead favored a relatively new technique for measuring distance, based on a standard candle known as the Tip of the Red Giant Branch (TRGB).

Stars at the tip of the red giant branch are (low to intermediate mass) stars which have branched off from the main sequence of stellar evolution to evolve as red giants, and have reached a limit in growth in size and luminosity: the tip of the red giant branch. As they grow along the red giant branch, these stars produce more and more helium at their core, increasing in size and luminosity, until eventually their helium cores are able to undergo nuclear fusion. At this point, their previously increasing brightness reverses direction abruptly as their temperature drops from this "helium flash." The corresponding rapid drop in brightness creates an apparent discontinuity that can be easily detected and used to infer distance.

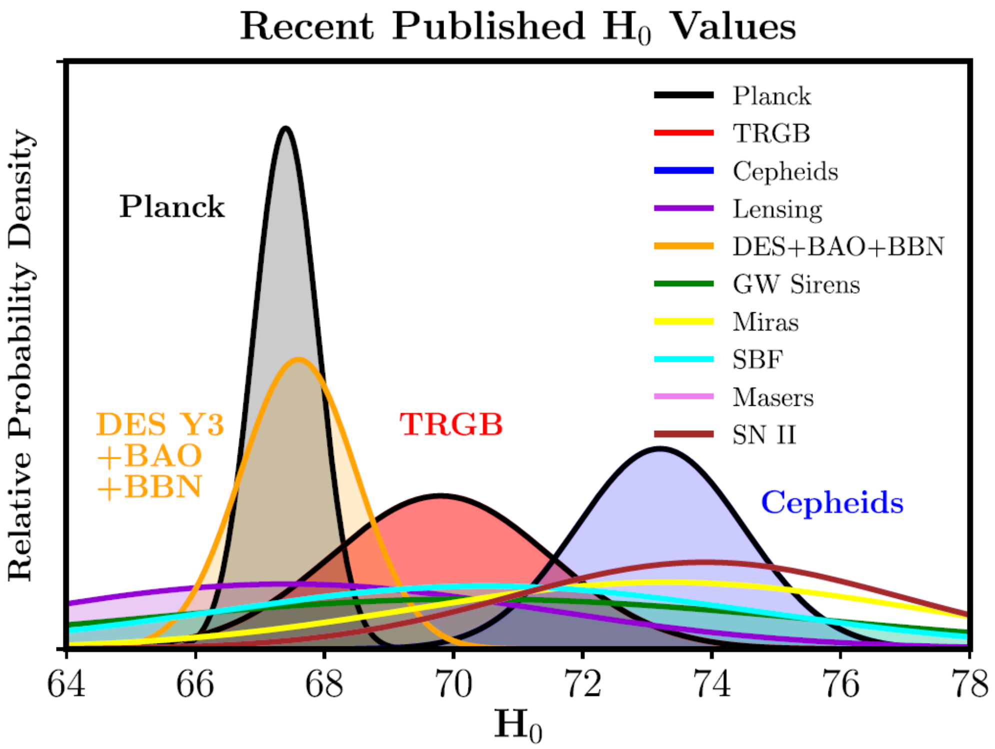

Figure 1: Probability density functions for several current methods for measuring H_0. Reproduced from fig. 10 in (Freedman, 2021) under the terms of the Creative Commons Attribution 4.0 license.

To sum up the main points of the case, we have highlighted three experimental programs, Planck, SH0ES, and CCH, which have produced discordant results for the value of the Hubble constant, 67.4, 73.2, and 69.06kms^{-1}Mpc^{-1} respectively, each with a small range of error, thereby putting each in some tension with the others (see fig. 1). While the "early universe" Planck method is largely independent of the "late universe" cosmic distance ladder methods of SH0ES and CCH, these latter programs also partially differ in their use of intermediate distance standard candles. Understanding how these degrees of independence and dependence function and relate to error in the general experimental context will be key to how we understand replication.

4 Error: Its Nature and Kinds

In the previous section, we showed how the three major experimental programs use different methods and procedures to determine the value of the Hubble constant but end up with significantly incompatible results. Taking as hypothesis that the Hubble constant has a unique value, these results therefore represent an experimental falsification of that hypothesis. Falsification of a hypothesis, of course, need not be grounds for its rejection. Rather, the falsification (or "tension" in results, if one prefers) exhorts scientists to begin a novel phase of research to identify what is responsible for the falsification. That source could be in a number of places: the theoretical framework, the apparatus, the observations, the data processing, etc. (Hon, 1989). Accordingly, scientists have searched widely for possible explanations of the discrepancy, from the exploration of alternative cosmological models, to the identification of a variety of insufficiently acknowledged errors, to efforts to re-analyze the data produced in the experiments.

It is important to emphasize, from an error analysis point of view, that the incompatibility of results is a consequence of the lack of agreement of results inclusive of all errors which have been acknowledged. The results from the TRGB-based method do show some degree of overlap with both the Cepheids-based method and the Planck method, which implies some degree of tension but also some degree of compatibility. The latter two approaches, however, are incompatible to a very high degree. Supposing that all three experimental programs have identified all relevant sources of errors and correctly incorporated them into their results, then the only reasonable conclusion to draw is that there is no unique value of the Hubble constant (in which case there is a "problem of definition" of the Hubble constant, i.e., a problem with the ΛCDM model or its background theory). But the programs may not have identified all sources of error, and they may not have correctly incorporated them into their results. If there really is a unique value of the Hubble constant, then at least one of the results is not correct.

In traditional error analysis (Taylor, 1997; Bevington and Robinson, 2003; Rabinovich, 2005), kinds of error are classified into two kinds: random (or statistical) error and systematic error. Random errors arise from an indeterminative source of deviations from the mean value of the measured quantity. Such sources cause different experimental outcomes under repetition. To the extent that there is variability in experimental results caused by a source of random error, there is a corresponding lack of precision. Systematic errors arise from a determinative source of error that causes a departure from the true value of the quantity being measured, but where the caused departure is realized consistently under repetition. To the extent that there is a departure from the true value being measured that is caused by a source of systematic error, there is a corresponding lack of accuracy. Thus, if there are discordant results between an experiment and a replication thereof, then in general the discordance could be due either to an incorrect assessment of random error (by one or more of the experiments) or else to an incorrect assessment of systematic error.

Besides this fundamental distinction between kinds of errors, it is also useful to acknowledge a second distinction between kinds of error, namely between known sources of error and unknown sources of error. While in principle the sources of both random and systematic errors could be described as known or unknown, the distinction is only practically relevant for systematic errors. This is because random error is estimated altogether and at once based on the variability in the outcomes of the experiment; there is little advantage to be found in separating "components" of random error into individual sources . By contrast, systematic errors (by definition) do not show up in the experimental outcomes, because they affect the results exactly in the same way under repetition. An experimenter must therefore strive to identify all possible sources of systematic error (preferably in advance of the experiment), and either eliminate their influence on the experiment, remove them (by correcting for them in the results), or put bounds on them and include them as residual systematic errors in the results. Such acknowledged errors can be described as "known systematic errors." However, the possibility invariably remains that some relevant sources of systematic error have not been identified and incorporated into the error analysis. These errors can be described as "unknown systematic errors."

Traditional error analysis tends to presuppose that all systematic errors have been acknowledged and either reduced, corrected, or bounded, focusing instead on techniques for analyzing and estimating statistical errors (Rabinovich, 2005, 118). The problem of how to address systematic errors, especially unknown systematic errors, is left to experimenters as a practical (and presumably discipline specific) problem. Nevertheless, a general approach to uncovering unknown systematic errors is widely acknowledged and well-known in experimental practice: carry out methodologically independent experiments that measure the same thing — that is, carry out conceptual replications.

As a case in point, the CCH team's emphasis on the TRGB method is motivated precisely by concerns over the accuracy and precision of Cepheids as standard candles (Freedman et al., 2019). One issue is that Cepheids often cannot be found in galaxies inhabited by type Ia supernovae, which limits calibration between the two distance measures. Stars at the tip of the red giant branch, by contrast, are relatively common and can be found widely in any type of galaxy. Another issue is that Cepheid distance measurements involve several sources of systematic error (reddening, metallicity, crowding, etc.) that are challenging to model accurately. The TRGB method, by contrast, is held to be one of the most precise and accurate ways to measure distances at intermediate distance scales. Like Type Ia supernovae, there are relatively few systematic errors to worry about, as the intrinsic brightness of stars at the tip of the red giant branch is determined by the helium-flash phenomenon they undergo.

All three mentioned experimental programs, Planck, SH0ES, and CCH, have invested significant effort into identifying systematic errors, mitigating them, correcting them, and including residual systematic errors in their results. Nevertheless, it remains possible that errors have been overlooked or incorrectly handled. Thus, we can identify three possible, independent resolutions of the Hubble discordance which are furnished by error considerations: (1) One (or more) of the experiments under-estimates its random error; in this case, decreasing the precision of the results to correctly account for it would allow for overlapping results and hence compatibility. (2) One (or more) of the experiments under-estimates its known (residual) systematic error; decreasing the accuracy of the results to correctly account for it would allow for overlapping results and hence compatibility. (3) One (or more) of the experiments has not accounted for unknown systematic error; correcting for this error would recover compatibility. In the first two cases, increasing the "width" of the error in the results restores compatibility; in the third case, the erroneous results are "shifted" so that they overlap with the correct results.

Each of teams involved in measuring the Hubble constant so far insists on the correctness of its analysis. Each has also blamed the discordance on errors committed by the other teams, accusing them of either not calculating their acknowledged systematic errors well or neglecting some possible source of systematic error. For example, the SH0ES team remarks that "systematic uncertainties in CMB radiation measurements may play a role in the tension" (Riess et al., 2016, 1), thereby suggesting that there are problems with the Planck experiment. Freedman maintains that that "crowding/blending effects are not an issue for the TRGB ... and metallicity effects are better understood from theory and more easily addressed empirically for TRGB stars than for Cepheids" (Freedman, 2021, 20), hinting that the SH0ES team has not correctly handled some of its systematic errors. Finally, the Planck team concludes that "the tension between base ΛCDM and the SH0ES H_0 measurement is intriguing and emphasizes the need for independent measurements of the distance scale" (Planck Collaboration, 2020, 26), subtly indicating that outstanding problems with the SH0ES team's construction of their cosmic distance ladder are the likely culprit for the discordance.

To be sure, each team has done high quality experimental work, overcoming many technical challenges along the way to their results, which are at the limit of what is currently experimentally possible in astronomy. Nevertheless, none is presently in the position to argue that their result for the Hubble constant is correct and the others are mistaken. First of all, there is (at present) no clear evidence that one team or another is to blame for the discordance (although suspicion may perhaps fairly fall somewhat more on one than another). Second of all, until the discordance is resolved, it remains reasonable to suppose that inadequately handled systematic errors affect any of the results, given that the history of experimentation in general shows it to be quite likely that difficult experiments to perform will not have sufficiently had their systematic errors adequately handled. In sum, so long as it is not clear where the unaccounted for error lies, the discordance represents a problem and a challenge for all experimental programs aiming to measure the Hubble constant accurately.

5 The Methodology of Replicability

We now turn to how these kinds of error relate to the different functions of replicability. To provide a suggestive illustration, we only need to look at the recent history, stretching back over the last two decades, of efforts to measure the value of the Hubble constant. The experimental results of the three main programs over the last two decades are depicted in fig. 2. Stepping back in time to the 2000s, one can see that the CMB-based and Cepheid-based measurements of the value of the Hubble constant were consistent, as there is substantial overlap in the results (although it is also clear that there is a substantial amount of error in the results for both experiments).

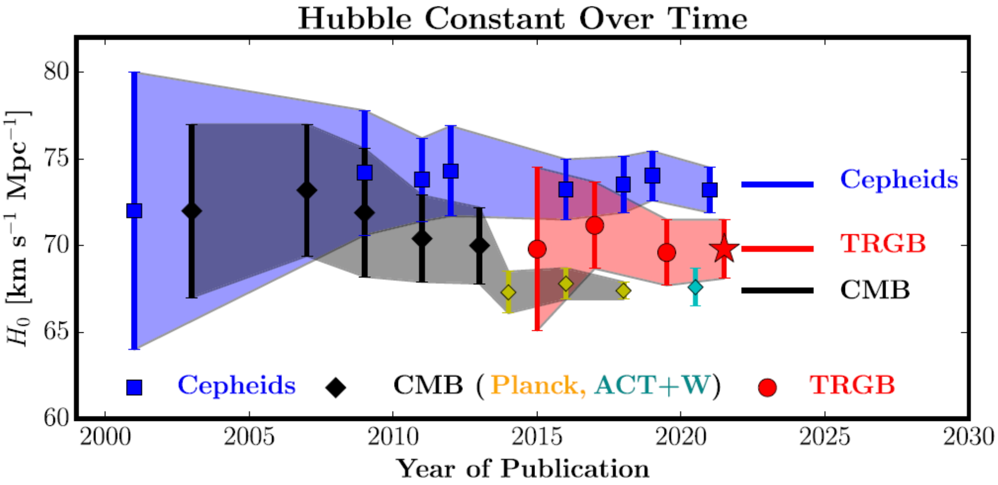

Figure 2: Summary of Hubble constant values in the past two decades based on Cepheids, the TRGB method, and the CMB. Reproduced from fig. 11 in (Freedman, 2021) under the terms of the Creative Commons Attribution 4.0 license.

Consistency in independent results (even with large error) induces some degree of confidence in their accuracy via a "triangulation" argument. Because the results were produced by independent means and those independent means have different possible sources of systematic error (Kuorikoski and Marchionni, 2016), it is unlikely that independent experiments testing the same hypothesis come to the same result unless that result is accurate (Dawid, 2021). It is in this way that a conceptual replication has the basic function of assessing accuracy.

We can therefore understand the CMB experiments in the early 2000s as conceptual replications of the Cepheid-based experiment: they are checks on the accuracy of results. Given the substantial amount of error quoted in those results, however, it is clearly desirable to improve the precision and accuracy of the experiments to see if this compatibility can be sustained under more severe testing. As one can see from fig. 2, both the CMB-based and Cepheid-based measurements (and the TRGB-based method, once it began to be employed) have substantially reduced their known error over the years in repetitions. For example, (Riess et al., 2016, fig.1) indicates the reductions in different sources of identified systematic error in successive measurements carried out by the SH0ES team. Even though these experiments are not "perfect" direct replications of their predecessors, due to improvements in the amount of error, we are still inclined to call them direct replications because their replicatory function is by and large that of a direct replication, namely, checking precision.

Moving up to the present time, we see the present discordance also depicted in fig. 2. Consider, though, the counterfactual possibility that substantial overlap in the CMB-based and Cepheid-based measurements had actually continued to the present, along with the steady improvements in error. Would this concordance be a strong argument for a unique, accurate value of the Hubble constant? Plausibly, yes, it would. The two experimental methods are substantially independent; independence of method entails different systematics; concordant experimental results despite different systematics is an argument for accuracy (a triangulation argument). Such arguments are defeasible, of course. Their success therefore depends additionally on a compelling argument that all relevant sources of error have been identified. In this case, one can "argue from error" (Mayo, 1996) not only that the results are accurate but also rebut the salient defeaters to the triangulation argument (i.e., incorrectly accounted for or unaccounted errors).

This counterfactual scenario has not happened, of course. The discordance that emerges in the 2010s between CMB-based measurements and Cepheid-based measurements of the Hubble constant suggests different methodological priorities compared to the scenario just sketched (where the priority would only be on continuing to improve the accuracy and precision of the different methods). The challenge in the actual scenario becomes one of identifying the cause of the discordance. Based on the discussion above, if the error analysis has been correctly carried out by each team (something that has been checked and re-checked, both by the teams and by independent researchers), then the only possibilities are that there are unknown systematic errors in one or more of the experiments, or else that the measured quantity, the Hubble constant, does not exist as described in the ΛCDM model (and background theory). Setting aside the latter possibility (which in most physicists' estimation is less likely), the overriding question for the teams, then, is, "how to ferret out unknown systematic errors?"

Theory, for sure, may give guidance, and the "error repertoire" of the experimental practice may also yield clues. Yet the most decisive approach is performing further, complementary experiments with the aim of revealing the source of the problem. These complementary experiments can come in two forms: one may further test assumptions that feed into the different experiments ("sub-experiments"), and one may perform a novel, independent experiment targeting the same hypothesis (or set of hypotheses), that is, what we call a conceptual replication. The degree of dependence and independence between an original experiment and a conceptual replication thereof plays a crucial role here. If the goal is not merely to cross-check previous results but also to isolate and identify unknown systematic errors, then experiments which differ in some respects but are otherwise the same can give experimenters positive guidance on where unknown sources of systematic error might be hidden. If the results of partially independent experiments are discordant, one has reason to suspect that there are overlooked systematic errors in one or both of the experiments where they are independent.

It is precisely in this way that the CCH program is of particular importance in the current experimental context. As a late universe program focused on constructing a cosmological distance ladder, it substantially shares systematics with the SH0ES program, agreeing in near and far distance measurements with SH0ES, but differing by the use the TRGB method rather than Cepheids to connect the near distance rungs of the distance ladder to SN1a supernovae. Indeed, the CCH program was developed specifically to shed light on the discordance between the Planck and the SH0ES results in this way (Freedman et al., 2019, 2–3).

The CCH program's choice of where to allow for independence from SH0ES is motivated by the conjecture that there are improperly handled systematic errors in the Cepheid photometry which are responsible for the discordance. Indeed, there are well-known difficulties in assessing the magnitude of Cepheid brightness, ranging from accurately accounting for their metallicity to accurately accounting for their reddening due to intervening dust. These problems led Freedman to replace the relatively error-prone Cepheids with TRGB stars as the standard candles used for intermediate distance. Naturally, the TRGB method involves systematic errors too. According to Freedman (2021), though, since physicists have a good theoretical and experimental handle on TRGB stars, they can calculate their brightness easily, without problems from dust and metallicity, and can have a higher degree of confidence that their systematic uncertainties have been correctly and fully accounted for. That does not mean, of course, that there might be overlooked systematic errors in this method too.

Some other relevant counterfactual scenarios are worth considering. First, if the SH0ES and CCH results had been strongly convergent, then experimenters could have concluded that the source of the discordance is probably not to be found in the intermediate distance standard candles' systematic errors. Instead, it would have to be something tied to the early universe method or something common to the late universe methods. Second, if the CCH results had been strongly convergent with the Planck results instead, then attention would surely have shifted to the Cepheids as likely culprit. The actual CCH results, however, are in some degree of tension with both the SH0ES and Planck results (fig. 1). This scenario, unfortunately, gives somewhat less guidance to experimenters than they might have hoped. Nevertheless, the variance between the CCH results and the SH0ES results does suggest that special scrutiny of the intermediate distance standard candles is and was warranted.

Yet, what about the fact that two highly independent experiments, Planck and SH0ES, have discordant results? Does that not also and already provide programmatic guidance to experimenters? After all, is it not the case that the tension between the Planck and SH0ES results already plausibly leads one to suspect that there may be unknown systematic errors in one or both of the experiments? To some extent, yes, but here is where degrees of independence and dependence make a difference. The high degree of independence of the Planck and SH0ES experiments allows one to infer only that there may be unaccounted for systematic errors affecting the experiment(s), but without any suggestion of exactly where. One can only go back to each individual experiment and check for the likely culprits. By instead carrying out an experiment involving only partial independence from the SH0ES experiment, the CCH's experiment is potentially able to offer a much more informative clue as to the source of the discordance than what is suggested by the discordance between Planck and SH0ES.

This might suggest that more informative, partially independent experiments are always better, but that is not so. If the Planck and SH0ES experiments had given consistent results even under the more severe testing in recent years, then, because of their high degree of independence, there would be a stronger confirmation (by triangulation) of the common result than if, say, SH0ES and CCH experiments had consistent results (which could only give a weaker such argument). These examples demonstrate that there is in fact a spectrum of possible conceptual replications that experimenters can perform, which have differing ramifications based on whether results are concordant or discordant.

6 Revisiting the Re-Sampling Account of Replicability

The Hubble constant example illustrates the importance of maintaining both notions of replicability, direct and conceptual, due to their distinct functions. Direct replications assess the reliability of an experiment by checking its precision; conceptual replications assess the validity of an experiment by checking its accuracy. We have shown how the teams involved in measuring the value of the Hubble constant have carried out both direct replications (albeit with steadily decreasing known errors) and conceptual replications (with, in the case of the CCH program, the special aim of uncovering the source of the discordance in results).

As discussed above, Machery (2020) is motivated to discard the category of conceptual replications based on his criticism of the common distinction between direct and conceptual replication, that is, the one based only on whether the experimental targets of an experiment are changed (leaving everything else fixed) or some different method is implemented. While we do agree with Machery that the distinction between direct and conceptual replication which he criticizes is not apt, we claim that it would be a mistake to divide up experiments into the three categories determined by his account of replication: replications (re-samplings of random factors of any kind), extensions (changes in fixed factors or changes in the populations of random factors to test more general hypotheses), or other experiments entirely (different hypotheses).

The basic problem of his account of replication is that re-sampling type experiments can only be used to assess reliability. As our account and case study show, experiments can also have the function of assessing the accuracy of experimental results too. Machery's categories, however, leave no place for experiments that can be used for this purpose. If such experiments involve a change in fixed factors (or a change of population for a random factor), then according to Machery's categories these must be either an extension or a different experiment entirely. Yet neither of these kinds of experiment target the same hypothesis as the original experiment, which would be required if we aim to check accuracy. Instead, they target a more general hypothesis or an independent one. Therefore, Machery must consider such experiments, which maintain the same test hypothesis, as re-samplings. Yet re-sampling experiments can only check reliability, not accuracy. We conclude, contra Machery, that it is essential to carve out a distinct category of replications for those experiments that assess the accuracy of an original experiment.

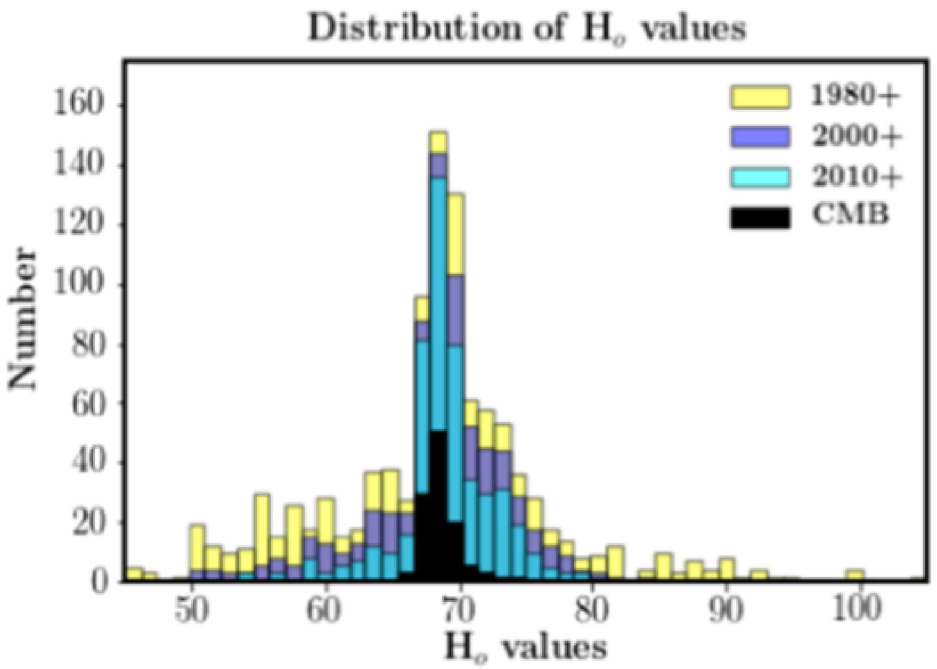

Figure 3: Summary of experimental results for Hubble constant values in the past four decades. Reproduced from fig. A2 in (Freedman, 2021) under the terms of the Creative Commons Attribution 4.0 license.

To make the point more concrete, consider what would result from treating the various experiments measuring the Hubble as re-sampling experiments, as Machery would evidently have us do. In that case, we should aggregate their results as one aggregates samples in normal sampling experiments. However, the problem with doing that for the Hubble constant experiments is that it would “hide” the discordance between the different kinds of experiment. Consider the fig. 3), which is an aggregation of all experiments that have provided a value for the Hubble constant over the past few decades. It appears that there is not only a strong agreement in its value, indeed in a fairly normal-looking distribution, but the result is also very precise. Clearly, if we regard different experiments measuring the Hubble constant simply as re-samplings, then there should be no controversy about the Hubble constant at all.

Regarding all experiments that target a common hypothesis as re-samplings thus obscures the very discordances that experimenters productively use to assess accuracy and find systematic errors. If we “stratify” our “samples” according to type of experiment (based on shared degrees of dependence and independence), then we instead see the strong, mostly non-overlapping “bumps” for the best results from CMB, TRGB-, and Cepheid-based experiments (as in fig. 1 above). When we recognize these experiments as (partially-) independent conceptual replications, we are able to acknowledge the discordances which must be resolved by further experimentation and analysis of sources of error.

7 Conclusion

Several recent contributions to the philosophical literature on replication have attempted to topple replication from its long-standing place in scientific epistemology in a number of ways: by dissolving a methodologically well-founded distinction found in experimental practice between direct and conceptual replications, by indexing the meaning of experimental replication to particular disciplines, and by skeptical arguments based on the limitations of different kinds of replications. We have defended the place of replication in scientific epistemology by identifying the epistemic functions of two different kinds of replication, functions which we claim hold across any experimental science whatever. Replication is a crucial experimental practice, because it is by replicating experiments that scientists are able to secure the needed reliability and validity of experimental knowledge.

In proposing this way of understanding replication, we are in part influenced by those philosophers of science who have emphasized the epistemic relevance of error analysis in experimentation (especially (Mayo, 1996)). Experience with carrying out experimental programs shows error to be both the experimenter’s friend as well as her enemy. Regarded as enemy, the experimenter devises ways to eliminate, limit, or circumvent it; she must seek out it and its sources. If after handling all known errors the experimenter’s diligent search turns up no more further sources of error, then she has grounds to conclude that her results validly represent what she sought to measure (“arguing from error,” as Mayo (1996, 7) calls it). However, in the mind of the experimenter, there is no experiment without error. Much like the air resistance that keeps the dove aloft, as in Kant’s famous metaphor, it is precisely the confrontation with error that allows experimenters to secure empirical knowledge. It is her friend, for it is by identifying and targeting errors in a program of “severe testing” that any hypothesis may emerge as confirmed or corroborated.

The significance of error to experiment thus leads us to make it the root concept of our account of replicability. The twin notions of reliability and validity are values determined by the presence and absence of errors of two basic kinds: systematic errors, which give rise to inaccuracy, and random (or statistical) errors, which give rise to imprecision. Although to some extent these differing kinds of error can be superficially represented in the same way (as “quoted” error or uncertainty), they are fundamentally different kinds of error that not only require different techniques and methods to handle properly but have different methodological ramifications and epistemic significance. It would be a mistake to conflate them, and thereby conflate accuracy and precision, and thereby conflate conceptual and direct replications, just as it would be a mistake to dispense with one in favor of the other.

Our Hubble constant case has also highlighted a significant distinction among kinds of conceptual replicability worth the attention of philosophers of science. At one end of this spectrum of possible conceptual replications are those that are minimally independent of their predecessor experiments. In our case study, this kind of experiment is exemplified by the CCH program. If the results of such an experiment are at variance with its predecessor, one gains valuable information about possible sources of unknown systematic error. At the other end of this spectrum are those conceptual replications that are maximally independent of their predecessor experiments. While this kind of experiment cannot illuminate sources of unknown systematic errors in case of discordant results, such experiments do provide a strong argument (via triangulation) for the accuracy of results in case of concordant results.

References

Bevington P. R., and K. D. Robinson. 2003. *Data Reduction and Error Analysis for the Physical Sciences.* New York: McGraw Hill.

Bird, A. 2021. "Understanding the Replication Crisis as a Base Rate Fallacy." *The British Journal for the Philosophy of Science* 72: 965–993.

Dawid, R. 2021. "The Role of Meta-Empirical Theory Assessment in the Acceptance of Atomism." *Studies in History and Philosophy of Science* 90: 50–60.

Di Valentino, E., O. Mena, S. Pan, L. Visinelli, W.-Q. Yang, A. Melchiorri, D. F. Mota, et al. 2021. "In the realm of the Hubble tension—a review of solutions." *Classical and Quantum Gravity* 38: 153001.

Dunlap, K. 1926. "The experimental methods of psychology." In *Psychologies of 1925*, edited by Carl Murchison, 331–351. Worcester: Clark University Press.

Efstathiou, G. 2020. "A Lockdown Perspective on the Hubble Tension (with comments from the SH0ES team)." *ArXiV Preprint: 2007.10716* .

Feest, U. 2019. "Why Replication Is Overrated." *Philosophy of Science* 86, 5: (2019) 895–905.

Fletcher, S. C. 2021. "The role of replication in psychological science." *European Journal for Philosophy of Science* 11: 23

Freedman, W. L. 2021. "Measurements of the Hubble Constant: Tensions in Perspective." *The Astrophysical Journal* 919: 16.

Freedman, W. L., and B. F. Madore. 2010. "The Hubble Constant." *Annual Review of Astronomy and Astrophysics* 48: 673–710.

Freedman, W. L., B. F. Madore, D. Hatt, T. J. Hoyt, I.-S. Jang, R. L. Beaton, C. R. Burns, et al. 2019. "The Carnegie-Chicago Hubble Program. VIII. An Independent Determination of the Hubble Constant Based on the Tip of the Red Giant Branch." *The Astrophysical Journal* 882: 34.

Guttinger, S. 2020. "The limits of replicability." *European Journal for Philosophy of Science* 10, 10.

Hon, G. 1989. "Towards a typology of experimental errors: An epistemological view." *Studies in History and Philosophy of Science* 20: 469–504.

Hubble, E. 1929. "A relation between distance and radial velocity among the extra-galactic nebulae." *Proceedings of the National Academy of Sciences* 15: 168–173.

Kuorikoski, J., and C. Marchionni. 2016. "Evidential Diversity and the Triangulation of Phenomena." *Philosophy of Science* 83: 227–247.

Leonelli, S. 2018. “Re-Thinking Reproducibility as a Criterion for Research Quality.” In *Research in the History of Economic Thought and Methodology*, edited by L. Fiorito, S. Scheall, and C. E. Suprinyak, Bingley: Emerald Publishing Ltd., 129–146.

Machery, E. 2020. “What Is a Replication?” *Philosophy of Science* 87: 545–567.

———. 2021. “A mistaken confidence in data.” *European Journal for Philosophy of Science* 11: 34.

Mayo, D. 1996. *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.

Norton, J. D. 2015. “Replicability of Experiment.” *Theoria* 30: 229–248.

Nosek, B. A., T. E. Hardwicke, H. Moshontz, A. Allard, K. S. Corker, A. Dreber, F. Fidler, et al. 2022. “Replicability, Robustness, and Reproducibility in Psychological Science.” *Annual Review of Psychology* 73: 719–748.

Planck Collaboration. 2020. “Planck 2018 results VI. Cosmological parameters.” *Astronomy & Astrophysics* 641: A6.

Popper, K. 2002. *The Logic of Scientific Discovery*. New York: Routledge.

Rabinovich, S. G. 2005. *Measurement Errors and Uncertainties*, 3rd Ed. New York: Springer.

Riess, A. G., S. Casertano, W. Yuan, J. B. Bowers, L. Macri, J. C. Zinn, and D. Scolnic. 2021. “Cosmic Distances Calibrated to 1% Precision with Gaia EDR3 Parallaxes and Hubble Space Telescope Photometry of 75 Milky Way Cepheids Confirm Tension with ΛCDM.” *The Astrophysical Journal Letters* 908: L6.

Riess, A. G., L. M. Macri, S. L. Hoffmann, S. Casertano, D. Scolnic, A. V. Filippenko, B. E. Tucker, et al. 2016. “A 2.4% determination of the local value of the Hubble constant.” *The Astrophysical Journal* 826: 56.

Romero, F. 2019. “Philosophy of science and the replicability crisis.” *Philosophy Compass* 14: e12633.

Shah, P., P. Lemos, and O. Lahav. 2021. “A buyer’s guide to the Hubble constant.” *The Astronomy and Astrophysics Review* 29: 9.

Taylor, J. R. 1997. *An Introduction to Error Analysis.* Sausalito: University Science Books.