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

Vera Matarese[†] and C. D. McCoy[‡]

Abstract

We propose that the epistemic functions of replication in science are best understood by relating them to kinds of experimental error/uncertainty. One kind of replication, which we call "direct replications," principally serve to assess the reliability of an experiment through its precision: the presence and degree of random error/statistical uncertainty. The other kind of replication, which we call "conceptual replications," principally serve to assess the validity of an experiment through its accuracy: the presence and degree of systematic errors/uncertainties. To illustrate the aptness of this general 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 replication, which maintains that replications only assess reliability.

Keywords

Hubble constant – Direct Replication – Conceptual Replication – Reliability – Validity – Resampling Account of Replication.

1 Introduction

The replicability crisis, much discussed in recent years 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.

1. The epistemological status of replication. While the replicability of experiment has traditionally been thought of as a pillar supporting the objectivity of science (Dunlap, 1926; Popper, 2002), some have questioned whether non-replicability necessarily impugns the credibility of those scientific contexts in which it occurs. Norton (2015), Leonelli (2018) and Guttinger (2020) are among those who maintain that non-replicability need not always represent an epistemic failure.
2. The meaning and interpretation of replication. Although the terminology used in connection with replicability varies considerably by discipline, and even by author, one common distinction is 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 like this one 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.

[†] Department of Philosophy, University of Perugia, Perugia, Italy. Email: vera.matarese@unipg.it
[‡] Underwood International College, Yonsei University, Seoul/Incheon, South Korea.
Email: casey.mccoy@yonsei.ac.kr

3. The epistemic functions or roles of replicability in science. While replications are usually regarded as some kind of "confirmation" of an earlier experiment, different authors have thematized its epistemological function in different ways. Fletcher (2021), for example, unifies replicability's role in science by claiming that replications serve to undercut the underdetermination of hypotheses by empirical evidence, while Matarese (forthcoming) offers a functional concept of replication that is also context-sensitive, and so respectful of the varieties of scientific standards across different sciences.

In this paper, we regard the third issue as primary, and take our proposed resolution of that issue, which follows the lead of (Matarese forthcoming), to shed light on the first two issues. In our view (and others, e.g., (Feest 2019)), the epistemic functions of replication can be understood by attending to the kinds and roles of uncertainty (or error) in experimental science.[1] Our claim is that different kinds of experimental replications serve to assess different kinds of uncertainty, thereby unifying the concept of experimental replication through the concept of experimental uncertainty and error analysis. This position answers directly to the third issue. As there are essentially two kinds of uncertainty that experiments can be used to assess — statistical uncertainty (or random error) and systematic uncertainty — we 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 familiarity in the scientific and philosophical literature. For us, however, "direct replications" have the (ideal) function of assessing the precision (and thereby testing the reliability) of previous experiments, where precision describes the presence and degree of statistical uncertainties; "conceptual replications" have the (ideal) function of assessing the accuracy (and thereby testing the validity) of previous experiments, where accuracy describes the presence and degree of systematic uncertainties. In this way we provide a more explicit epistemological rationale for the categories of replication referred to by scientists.

Real experiments, of course, may blend together aspects of both kinds of idealized replications, 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. 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, failing to carry out replications is only an issue for a science when there is insufficient reason to trust that replications would appropriately confirm previous results if performed. Thus, our proposal answers to the first issue as well.

[1] For reasons that we do not have the space to go into, it is preferable to use the more general term "uncertainty" rather than "error." However, we will use the terms interchangeably for ease of exposition and consistency with our references. By doing so, we do not intend to dismiss the importance of the distinction nor take a position on controversial philosophical issues related to it.

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 uncertainty 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 uncertainties 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 uncertainty to these crises but rather to articulate a general epistemological account of replicability in science based on such considerations (which may then be applicable to the assessment of putative replication crises).

Another of our motivating concerns arises from the fact that contemporary philosophical discussions on replication tend to be informed primarily by scientific cases from the disciplines most affected by the crisis, in particular psychology. We see a potentially worrying consequence that can easily arise from this circumstance, namely that there is substantial risk in grounding a discussion on the functions of replications on the practices of disciplines that are going through a replicability crisis, for those disciplinary practices do not afford us the means to identify what correct functioning looks like. Indeed, 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 an epistemically progressive 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. 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 noteworthy discordance in results from three major experiments measuring the value of the Hubble constant. While this case is of considerable independent interest,[2] it is particularly relevant to our interests in this paper, as it constitutes a localized replicability failure in a field with strong replicability standards, that is, a field which has the resources to manage and potentially resolve localized replicability failures through further experimentation.[3] This localized replicability failure

[2] See especially the recent papers by both Gueguen (2023), who aims to evaluate whether the Hubble constant discordance constitutes a crisis for the ΛCDM model and to provide a methodological guide to managing uncertainty in astrophysical measurements, and Smeenk (2022), who argues that taking the ΛCDM model seriously has allowed astronomers to use tensions in experimental results like that over the Hubble constant to make concrete scientific progress.

[3] While Gueguen (2023) too connects her discussion of the Hubble constant controversy to replicability and draws out some similar conclusions to ours concerning the case study, there are several major differences in our views and aims which should be pointed out for those familiar with her article. Firstly, our two-fold typology of replications (direct and conceptual) is based in a fundamental epistemological distinction between reliability and validity, whereas Gueguen identifies four different kinds of replication (48–9) – direct, methodological, systematic, conceptual – without grounding the distinction in epistemic functions. Gueguen's kinds, by contrast to ours, are not intended to provide "clean-cut separations between different types of replication" (48). Secondly, and relatedly, where our aim is explicitly epistemological (providing a general account of replication in science), Gueguen's aim is narrower: to provide (philosophically motivated) methodological guidance to astrophysicists on whether they should regard the Hubble constant controversy as a crisis or not (33). Thirdly, Gueguen (34–5) maintains that there is a

provides a valuable case study for understanding the methodological practices that are crucial for resolving discordances in experimental results. Most importantly for our purpose, the story of how the discordance emerged and how it is being solved will help us illustrate key planks in our replicability platform and also 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 to distinguish between two categories: "conceptual" replications and "direct" replications (Romero, 2019; Matarese, 2022). In this paper, 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. That is, 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 critique of the latter way of distinguishing kinds of replication, arguing that the category of conceptual replications is confused and should be abandoned. While those replications that "change methods" are conventionally referred to as "conceptual replications," Machery argues that we should not give them a special designation in virtue of that feature 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 (Machery borrows terminology from statistical experiment design). Any component that is regarded as a random factor in one experiment can, in a subsequent experiment, be "re-sampled" from the same population (of experimental units, treatments, methods, etc.) from which the original experiment "sampled," where the purpose of treating a component as "random" is to make a generalization to a population. Any component that is regarded as a fixed factor, however, is not regarded as drawn from a population, and hence no generalization to a population is intended. Thus, insofar as conceptual replications are thought to involve a change of method, Machery explains that this change could be regarded as (1) a change from one fixed method to another, (2) as a re-sampling from the same population of methods, or (3) as a sampling of methods from a different population of methods. If an experiment is performed where the only difference between it and a previous experiment is that a different method is used, and we regard the choice of method as a sample from a population of possible methods (i.e., as a random factor, as in case (2)), then Machery argues that there is no reason to distinguish this "conceptual replication" from a "direct replication", for they are functionally identical. Both are *re-samplings* which are, according to Machery, checks on the reliability of the experiment (with respect to its different components). Indeed, direct replications are normally understood as re-samplings of experimental units from the sample population of experimental

methodological priority to tracking "unknown systematic errors" (by using "systematic replications") over reducing "known uncertainties." As we will show below (Sec. 5), on our view there is no such general priority, since the epistemic context determines methodological priority. Despite these differences, we do recommend Gueguen's paper as a useful complement to our own, particularly for the more detailed technical development of the case study.

units. In the other two cases, namely, replications that either (1) treat methods as fixed factors or else (3) take them as random factors but understand them to sample from a different population, Machery argues that what are normally called "conceptual replications" should be regarded as "extensions" of the original experiment (or else as other experiments entirely), that is, not kinds of replication at all, since a change in a fixed factor is not replicating the original experiment in Machery's preferred sense, namely, by re-sampling it as a check on its reliability.[4]

Whatever the classification one adopts, whether the usual direct/conceptual distinction or the one introduced by Machery (between replications and extensions), we emphasize that the description of an experiment as being of some particular kind depends on specific interpretive choices on the parts of the experimenters. This relativity to interpretation is evident in the case of Machery's description of random and fixed factors, and is indeed something that he emphasizes throughout. 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, whether a factor is regarded as fixed or random is not a fact about the experiment but an interpretive choice made by experimenters.

Although Machery emphasizes the role of interpretation mainly for the purpose of individuating experiments, we claim that implicit in the interpretation of an experiment as some kind or another is the idea that an experiment targets a particular hypothesis (or set of hypotheses). A given concrete experiment obviously does not dictate its interpretation, whether in terms of the hypotheses targeted for test or for how it is individuated from other experiments. 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 hold that the underlying idea of such distinctions between kinds is that experimentation's fundamental function is hypothesis testing (or confirmation). Naturally, this is not to say that scientists always perform experiments intending to test some specific 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 of the role of hypotheses in interpreting experiments, we take as our basic standard for some experiment being a *replication* that it may be *interpreted as targeting the same hypotheses for evidential appraisal as another experiment*. While this standard holds for replications on Machery's account as well, he maintains that the only epistemic appraisal that a replication can make is an appraisal of reliability, since for him replications are re-samplings, and re-samplings can only check for reliability (as Machery insists throughout his paper). While Machery's identification of replications with an epistemic function rather than some epistemically irrelevant operational category is commendable, we have some reservations about the identification of a single epistemic function of replicability at work in experimental practice.

[4] Although Machery distinguishes cases (1) and (3), it seems to us that case (3) is just a special case of (1), since we may regard populations as a kind of fixed factor.

We maintain, by contrast, that there are scientific experiments (like some of those carried out in projects to measure the Hubble constant) which are replications according to this basic standard and also have the function of appraising validity.[5]

Although Machery does allow that there are experiments which check for validity, he maintains that these must be what he calls "extensions," which involve a change in fixed factor or sampling from a distinct population. He argues that this rigid conceptual separation of replications from extensions is apt, for he claims that there is no "common measure" to compare extensions checking validity and replications checking reliability: indeed, he remarks that "it is strange to think that there can be a meaningful comparison between these two goals [of reliability and validity]" (Machery, 2020, 563). In fact, despite the fact that they are different epistemic functions (which Machery rightly stresses) there *is* a meaningful comparison to make between them and a common measure as well: experimental uncertainty. Because of this common measure and the crucial cooperation between experiments checking viability and those checking reliability (as the Hubble constant case will illustrate), we believe it makes much sense to regard both as kinds of replications. Moreover, as we will show, within particular experimental contexts it is possible to establish an epistemic superiority of one kind of replication over the other. For sure, the specific terminology one uses is not at issue for us; rather, what is crucial is understanding the methodological and epistemological role of these kinds of experiments in a progressive experimental program.

We postpone, for the moment, filling out how specific kinds of uncertainty 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 uncertainty connects to the epistemology of experiment, 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 experimental programs.

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 background 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 500 kms^{-}

[5] We do not yet enter a discussion on the meaning of reliability and validity, since for us these are to be related to experimental uncertainties. The reader may wish to consult Machery's discussion of reliability and validity (Machery, 2020, 554–555), which is entirely applicable to our claims in this section. We neglect discussion of internal vs. external validity, however, since this distinction is not relevant to our concerns in this paper.

$^{1}Mpc^{-1}$ (Hubble, 1929).[6] 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 the hopes of finding some resolution to the discordance.[7]

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.

1. 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 (i.e., 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. Recent results from the Planck team give a value for H_0 of 67.4 $kms^{-1}Mpc^{-1}$, with an uncertainty of less than 1% (Planck Collaboration, 2020).

2. Programs which measure H_0 by inferring its value from measurements of local features (astronomical objects like galaxies, stars, etc.) of the late universe (i.e., 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 $kms^{-1}Mpc^{-1}$). 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 $kms^{-1}Mpc^{-1}$, 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 is what

[6] This unit is kilometer per second per megaparsec. A parsec is a common unit of distance in astronomy equivalent to 3.26 light years.

[7] There are numerous reviews covering these developments. Among them, the reader may usefully refer to (Freedman and Madore, 2010; Di Valentino et al., 2021; Shah et al., 2021).

accounts for it. One possibility, enticing for many theoretical cosmologists, is a failure of the ΛCDM model itself, which requires that the Hubble constant inferred from features of the late universe and the Hubble constant inferred from features of the early universe give the same value.

The sober-minded judgment of many astronomers, based on long experience with apparently discrepant 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. The most plausible explanation for the discordance, then, is that one of the two experimental results is somehow erroneous.[8] Since building a reliable, accurate cosmic distance ladder is much more complicated than the measurements of the CMB in the Planck experiment, as it involves tracking a variety of different sources of uncertainty, a prevailing suspicion among some astronomers is that the result obtained by the SH0ES program has not properly taken into account all the relevant uncertainties. Nevertheless, as the SH0ES program itself has emphasized, careful checking of systematic uncertainties 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 adopted 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. This has further complicated the controversy by obtaining a different result from both SH0ES and Planck for the Hubble constant: 69.06 $kms^{-1}Mpc^{-1}$, also with a small uncertainty (Freedman et al., 2019). 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 to the issue, let us say a little more about cosmic distance ladders and their relation to the Hubble constant. 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 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 measurements decrease with distance. In general, different techniques are also applicable at different distances. These different techniques must therefore be

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

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.

1. For the smallest 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.

2. 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 other 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).

    A. 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 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.

    B. 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.

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 for ease of exposition, by the way, 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.

That concludes our introduction to the case study. To sum up the main points, 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 69.06 $kms^{-1}Mpc^{-1}$ respectively, each with a small range of uncertainty, thereby putting each in some tension with the others (see Fig. 1, which depicts results for Planck, SH0ES [Cepheids], and CCH [TRGB]). 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 uncertainty in the general experimental context will be key to how we understand replication.

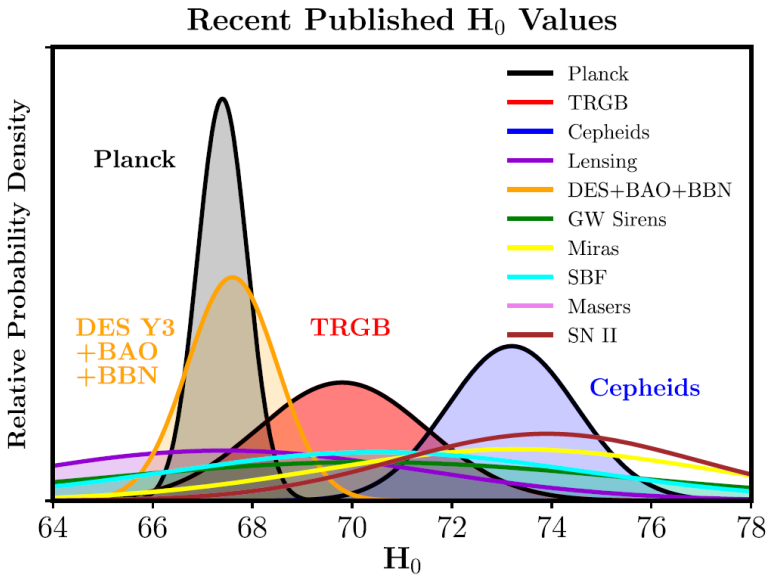

Figure 1: Probability density functions for several current methods for measuring H_0. Reproduced from fig. 10 in (Freedman 2021) under the terms of the Creative Commons Attribution 4.0 license.

4 Uncertainty: 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 incompatible results. Taking as a background assumption that the Hubble constant has a unique value, these results therefore represent an apparent experimental falsification of that hypothesis.

Apparent 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 apparent 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 uncertainties, 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 uncertainties 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 uncertainties 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 uncertainty, and they may not have correctly incorporated them into their results.*

In traditional error analysis (Taylor, 1997; Bevington and Robinson, 2003; Rabinovich, 2005), kinds of error are classified into two kinds: random error (statistical uncertainty) and systematic error (or uncertainty). Random errors (statistical uncertainties) arise from a source of indeterminate deviations from the mean value of the measured quantity. That is, such sources cause unpredictable 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 (or uncertainties) are said to arise from a source of error that causes a determinate departure from the "true" value of the quantity being measured, where the caused departure is realized predictably under repetition (not necessarily constantly but in some way determinately). 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.[9]

Besides this fundamental distinction between kinds of uncertainties (errors), it is also important to acknowledge a second distinction between kinds of uncertainty, namely between *acknowledged* sources of uncertainty and *unacknowledged* sources of uncertainty. While in principle the sources of both statistical and systematic uncertainties could be described as acknowledged or unacknowledged, the distinction is only practically relevant for systematic uncertainties. This is because statistical uncertainty is estimated altogether and at once based on

[9] Again, various considerations arising from simulations and in metrology motivate a preference for a more general concept of uncertainty (and even a rejection of the idea of a measurement error with respect to a "true value"), yet bringing in these considerations in adequate detail here is not possible given the space available. A useful entry point to the relevant literature is (Boumens, Hon, and Peterson, 2014).

the variability in the outcomes of the experiment; there is little advantage to be found in separating "components" of statistical uncertainty into individual sources. By contrast, systematic uncertainties (by definition) do not show up in the experimental outcomes, because their sources affect the results exactly in the same way under repetition. An experimenter must therefore strive to identify all possible sources of systematic uncertainty (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 uncertainties in the results. Such acknowledged uncertainties can be fairly described as "known systematic uncertainties" (even if all one can do is put estimates on them), as Gueguen (2023) does. However, the possibility almost inevitably remains that some relevant sources of systematic uncertainty have not been identified and incorporated into the analysis. These uncertainties can then be fairly described as "unknown systematic uncertainties."

Traditional error analysis tends to presuppose that all sources of systematic uncertainty have been acknowledged and either reduced, corrected, or bounded, focusing instead on techniques for analyzing and estimating statistical uncertainties (Rabinovich, 2005, 118). The problem of how to address systematic uncertainties, especially unacknowledged sources thereof, is left to experimenters as a practical (and discipline specific) problem. Nevertheless, a general approach to uncovering the existence of unacknowledged sources of systematic uncertainties is 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 uncertainty (reddening, metallicity, crowding, etc.) that are challenging to model accurately. The TRGB method, by contrast, is thought 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 sources of systematic uncertainty to worry about, as the intrinsic brightness of stars at the tip of the red giant branch is determined precisely by the helium-flash phenomenon they undergo.

All three mentioned experimental programs, Planck, SH0ES, and CCH, have invested significant effort into identifying sources of systematic uncertainty, mitigating them, correcting for them, and including residual systematic uncertainties in their results. Nevertheless, it remains quite possible that there are sources which have been overlooked or incorrectly handled. Thus, we can identify three possible, independent resolutions of the Hubble discordance which are furnished by uncertainty considerations:

1. One (or more) of the experiments under-estimates its statistical uncertainty; 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 uncertainty; *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 sources of unacknowledged systematic uncertainty; *identifying and incorporating sources of unacknowledged systematic uncertainty* into the analysis would recover compatibility.

Refer again to Figure 1: In the first two resolutions, we can see that increasing the "width" of the uncertainty in the results restores compatibility, while in the third, the erroneous results can be seen as "shifted" so that they overlap with the correct results.

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. Second of all, until the discordance is resolved, it remains reasonable to suppose that inadequately handled sources of systematic uncertainty 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 had all their sources of systematic error adequately handled. In sum, so long as it is not clear where the unaccounted uncertainty 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

According to our account of a replication, namely, that it is an experiment that can be interpreted as assessing the validity or reliability of the same set of hypotheses, we regard all the experiments attempting to measure the Hubble constant's value as replications.[10] However, these experiments differ in their epistemic functions, as we will demonstrate in this section, in particular due to their differing degrees of dependence and independence with respect to one another, as well as their different sources and degrees of uncertainty. We will now show how the kinds of uncertainty sketched in the previous section relate to the different functions of replicability – validity and reliability – to ground the functional distinction between direct and conceptual replication described in the introduction.

An illuminating illustration of how the different kinds of replication become salient in different epistemic contexts is found in the recent history, stretching back over the last two decades, of

[10] Plainly, our account of replicability is quite inclusive, encompassing experiments that may be very different from the original experiment of which they are regarded (by our account) as replications. While it may seem counterintuitive to regard so many experiments as replications, our goal is to defend an account of replication that is centered on the epistemic function of experiments, and in this respect, it is appropriate to regard an experiment as a replication when it has the function of a replication. Some experiments may have a stronger replicatory function than others, and for that reason deserve better the name "replication," but conventional choices like this are no concern in an epistemological analysis like our own. In any case, the inclusiveness of our terminology is actually reflected in the practice of those scientific disciplines where the term and category of "conceptual replications" is used, since in those disciplines "conceptual replication" is used with a wide scope for experiments, experiments which according to our account would indeed be appropriately regarded as replications.

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 Figure 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 from the "error bars" that there is a substantial amount of uncertainty in the results for both experiments).

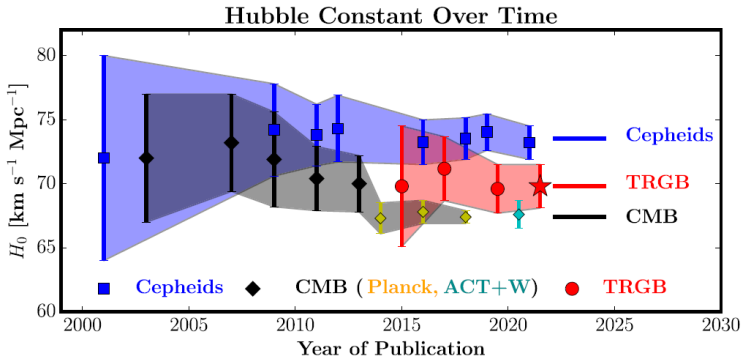

Figure 2: Summary of Hubble constant values in the past two decades based on Cepheids, the TRGB method, and the CMB. Reproduced from fig. 11 in (Freedman, 2021) under the terms of the Creative Commons Attribution 4.0 license.

Consistency in independent results (even with large amounts of error) induces some degree of confidence in their accuracy via a "triangulation" or "robustness" argument (Kuorikosi and Marionni, 2016; Beauchemin, 2017; Staley, 2020). Because the results were produced by independent means and those independent means have different sources of systematic uncertainty, it is unlikely that such independent experiments testing the same hypothesis come to the same result unless that result is accurate. It is in this way that a conceptual replication has the basic function of assessing accuracy.

As on our account conceptual replications have the (ideal) function of assessing validity by checking for the existence of unacknowledged sources of systematic uncertainty, we can characterize the CMB experiments in the early 2000s principally as successful conceptual replications of the earlier Cepheid-based experiment. Given the substantial amount of uncertainty quoted in those results, however, it is clearly a priority to improve both the precision and accuracy of the experiments to see if this compatibility can be sustained under more severe testing. As one can see from Figure 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 uncertainty over the years in experimental repetitions.[11] Even though these experiments are not "perfect" direct replications of their predecessors, due to improvements being made in the amount of uncertainty involved in the experiments, we are still inclined to call them direct replications because their principle *replicatory* function is by and large that of a direct

[11] See, e.g., fig. 1 of (Riess et al., 2016), which indicates the reductions in different sources of identified systematic uncertainty in successive measurements carried out by the SH0ES team.

replication, namely, a check on the reliability of the experiment by comparing it to a past version with (substantially) the same sources of systematic uncertainty.

Moving up to the present time, we see the present discordance depicted in Figure 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 limiting the uncertainty of the result. Would this concordance be a strong argument for a unique, accurate value of the Hubble constant? Yes, it plausibly would: the two experimental methods are substantially independent, independence of method entails different sources of systematic error, and concordant experimental results (despite different sources of systematic error) is a triangulating argument for accuracy. Such arguments are defeasible, of course. Their success therefore depends additionally on making a sufficiently compelling argument that all relevant sources of uncertainty have been identified (Mayo (1996) calls an argument of this kind an "argument from error" or "learning from error"). If such an argument can be mustered, then one has particularly strong reasons to conclude that the results are accurate (i.e., there are no remaining incorrectly or un-accounted for uncertainties).

In the event, this hypothetical scenario has not happened. 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 and performing further direct replications). 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, then the only possibilities are that there are unacknowledged or mis-analyzed sources of systematic uncertainty 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 remains less likely), the overriding question for the teams, then, is, "how to ferret out issues with systematic uncertainties?"

Theory, for sure, may give guidance, and the "error repertoire" of the experimental practice (the stock of suspected sources of error with which an experimentalist is familiar) may also yield clues. Yet the most decisive approach is usually performing further, complementary experiments with the aim of revealing the underlying source of the problem. They 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 are calling a conceptual replication. The degrees of dependence and independence between an original experiment and a conceptual replication thereof play a crucial role here. If the goal is not merely to cross-check previous results but also to isolate and identify unacknowledged sources of systematic uncertainty, then experiments which differ in some respects but are otherwise the same can give experimenters positive guidance on where unacknowledged sources might be hidden. If the results of partially independent experiments are discordant, one has reason to suspect that there are overlooked sources of systematic uncertainty 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 pointed out also by Gueguen (2023)). As a late universe program focuses on constructing a cosmological distance ladder, it substantially shares the same sources of systematic error 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. The CCH program's choice of where to allow for independence from SH0ES is motivated by the conjecture that there are improperly handled sources of systematic uncertainty in the Cepheid photometry which are responsible for the discordance, which is what 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 uncertainties too. Nevertheless, according to Freedman (2021), since physicists have a good theoretical and experimental handle on TRGB stars, they can calculate their brightness easily and can have a higher degree of confidence that their systematic uncertainties have been correctly and fully handled.

Some other relevant counterfactual scenarios are worth considering at this juncture as well. 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 uncertainties. 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. As it happens, though, the actual CCH results 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 unacknowledged sources of systematic uncertainty in one or both of the experiments? To some extent, yes, but it is here 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 sources of systematic uncertainty 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.

These considerations might seem to suggest that more informative, partially independent experiments are always better, but that is not so. Partially independent experiments are only half of the story. Consider that if the Planck and SH0ES experiments had given consistent results even under the more severe testing of recent years, then because of their high degree of independence there would be a stronger confirmation 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 epistemic and methodological ramifications based on whether results are concordant or discordant. There is no general priority of one kind of replication or another – priority depends on the developing epistemic context of the experimental program.

Before concluding this section, it is worth making a comparison between the application of our account to the Hubble constant controversy and Gueguen's (2023) discussion of the Hubble constant controversy, which shares a few topical commonalities. Whereas our aim in this paper has been to develop a novel account of experimental replication, Gueguen's aim is quite different: she is narrowly focused on providing philosophically informed guidance to physicists on whether the Hubble constant controversy signifies a potential breakdown in currently accepted physical theory or instead points to the persistent presence of an unknown systematic error responsible for the discrepancy.

While carrying out this project, Gueguen does, however, introduce a certain categorization of replication, the purpose of which it to better appreciate the different kinds of experiments carried out within the Hubble program (sec. 3.4.2.2.). As she herself points out, "it is important to note that these categories are better conceived of as covering a spectrum and revealing different aspects of replication than as clean-cut separations between different types of replication" (48). In particular, she introduces a typology of four types of replication: direct, methodological, systematic and conceptual. According to her description,

1. Direct replications are those experiments that reproduce exactly the original study but on a different statistical set.
2. Methodological replications are those replications that constitute a re-analysis of an experiment performed by another team.
3. Systematic replications are those consisting in systematically varying one variable at a time, leaving all the other variables fixed.
4. Conceptual replications are those that involve a change in the methodology and for this reason, showing a relatively high degree of difference, are able to check for robustness.

These four categories, in line with the traditional literature on replicability, are practically oriented, focusing on operational aspects such as the number of variables changed, the actors involved in the replication, and the manner of execution (whether it is systematic or otherwise). This approach, while useful for her specific purposes, clearly does not employ any epistemologically principled distinctions. Why carry out a direct replication? What could a methodological replication uncover? How is a systematic replication a replication? These epistemologically motivated questions are not answered by Gueguen (nor are we suggesting that she needed to do so, given her aims). Were one to interrogate categories like these, we anticipate that one would find a confusion of practical aims and epistemic ones inherent in the categorization, and a confusion of various epistemic aims even within individual categories. It is not our purpose to offer such a critique here, however, since we take it as given that epistemological analyses are philosophically valuable, and that an epistemological analysis of

replication in particular is valuable (in agreement with Machery and many other philosophers working on the topic).

6 Revisiting the Re-Sampling Account of Replicability

We have argued that the Hubble constant case illustrates the importance of two notions of replicability, which we have been calling "direct" and "conceptual," due to their distinct epistemic functions in experimental practice. 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 also shown how the teams involved in measuring the value of the Hubble constant chose to carry out direct replications and conceptual replications depending on the evolving status of the collective experimental program of measuring the value of the Hubble constant.

As discussed above, Machery (2020) is motivated to discard the category of conceptual replications based on his criticism of a certain common distinction made 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 *this* distinction between direct and conceptual replication which he criticizes is not apt, we disagree that the only function of experimental replication is to check reliability.

It is instructive to consider what would result from treating the various experiments measuring the Hubble as merely re-sampling experiments checking reliability. 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 Figure 3, which is an aggregation of all experiments that have provided a value for the Hubble constant over the past few decades. It appears from this figure 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.

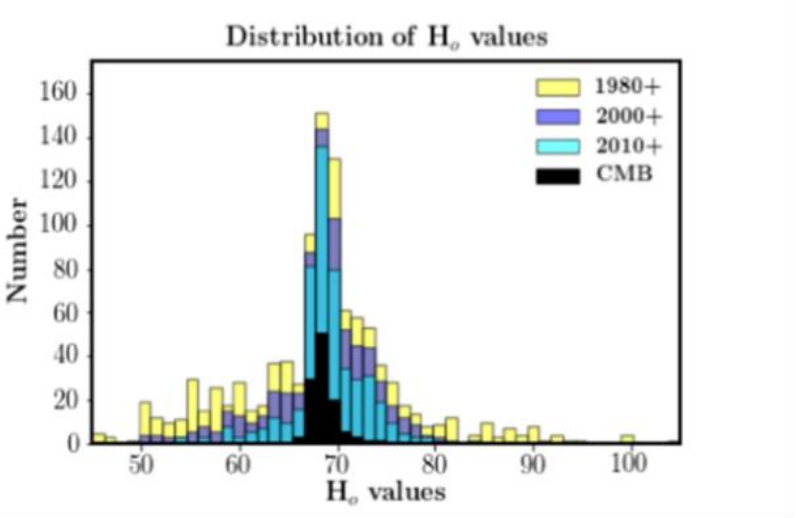

Figure 3: Summary of experimental results for Hubble constant values in the past four decades. Reproduced from fig. A2 in (Freedman, 2021) under the terms of the Creative Commons Attribution 4.0 license.

Regarding all experiments that target a common hypothesis as re-samplings obscures the very discordances that experimenters productively use to assess accuracy and find sources of systematic uncertainty. 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 uncertainty.

Certainly, Machery will agree with us on the value of treating the different experiments separately and not as mere re-samplings; he would just have us regard them as "extensions," which he says serve to check for validity. There is, however, a dilemma that arises. If an extension changes a fixed factor, and since a fixed factor is precisely a factor beyond which the experimenter does not intend to generalize, it follows that an extension must have a different hypothesis than the original experiment: a hypothesis that acknowledges the fixed factor as an assumption.[12] How, then, can one check the validity of the original experiment with an extension, having as it does a different hypothesis from the hypothesis of the original experiment? On the other hand, if we regard the successor experiment as targeting the same hypothesis, then it is because we are treating its factors as random. According to Machery, though, experiments that change random factors are re-samplings, which can only check for reliability, not validity. Thus, it is not so clear to us how on Machery's account one can coherently interpret an experiment as a check on validity: extensions are incommensurable with the original experiment and replications only check for reliability.

[12] Cf. Choi's (2023) discussion of Machery's account in his reply to a criticism of the re-sampling account by Matarese (2023). The same is of course true in case the extension involves a change in population.

Perhaps an explicit account of extensions (which Machery does not give in his paper) may resolve this apparent issue. Our purpose here, in any case, has not been to criticize Machery's account but rather to emphasize the important role of checks on validity as part of experimental practice. Because we see checks on validity and checks on reliability as the two fundamental means of justifying experimental results, we have chosen to offer an account of experimental replication that integrates these methods together through the common currency of experimental uncertainty. While Machery is perfectly within his rights to prefer to restrict the term "replication" to just those experiments that check reliability, we prefer to see the practice of replication as more than just reliability: as having the goal of justifying our empirical knowledge in both of its basic aspects, reliability and validity.

7 Conclusion

Several recent contributions to the philosophical literature on replication have attempted to topple replication from its long-standing place in scientific epistemology, whether 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 or particular statistical approaches, or 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. Replication is a crucial experimental practice because it is by replicating experiments that scientists are able to secure the needed reliability and validity of empirical 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 (or uncertainty) 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 further sources of error, then she has grounds to conclude that her results validly represent what she sought to measure. 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 critical, severe testing that any hypothesis may emerge as confirmed or corroborated.

The significance of uncertainty to experiment thus leads us to make it the basic concept of our account of replicability. The twin notions of reliability and validity are values determined by the presence and absence of uncertainties of two basic kinds: systematic uncertainty, which gives rise to inaccuracy, and statistical uncertainty, which gives rise to imprecision. Although to some extent these differing kinds of uncertainty can be superficially represented in the same way (as "quoted" uncertainty), they are fundamentally different kinds of uncertainty 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, for they are cooperative concepts in experimental practice.

Our Hubble constant case has also highlighted a significant distinction among kinds of conceptual replicability worth the further 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 unacknowledged systematic uncertainty. 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 unacknowledged sources of systematic uncertainty in case of discordant results, such experiments do provide a strong argument for the accuracy of results in case of concordant results.

References

Beauchemin, P.-H. 2017. “Autopsy of measurements with the ATLAS detector at the LHC.” *Synthese* 194: 275–312.

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.

Boumens, M., G. Hon, and A. C. Petersen. 2014. *Error and Uncertainty in Scientific Practice*. London: Pickering and Chatto.

Choi, H.-H. 2023. “In Defense of the Resampling Account of Replication.” *Journal of Theoretical and Philosophical Psychology* 43: 249–251.

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: 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.

Gueguen, M. (2023). “A Crack in the Track of the Hubble Constant.” In *Philosophy of Astrophysics: Stars, Simulations, and the Struggle to Determine What is Out There* (pp. 33-55). Cham: Springer International Publishing.

Guttinger, S. (2020). The limits of replicability. *European Journal for Philosophy of Science*, 10(2), 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.

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

Matarese, V. 2022. “Kinds of Replicability: Different Terms and Different Functions.” *Axiomathes* 32: 647–670.

Matarese, V. 2023. “Against the Resampling Account of Replication.” *Journal of Theoretical and Philosophical Psychology* 43: 108–115.

Matarese, V. Forthcoming. “A new concept of replication.” *Inquiry*: 1-26. DOI: 10.1080/0020174X.2023.2278032.

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.

Smeenk, C. (2022). Trouble with Hubble: Status of the Big Bang Models. *Philosophy of Science* 89: 1265–1274.

Staley, K. 2020. “Securing the Empirical Value of Measurement Results.” *The British Journal for the Philosophy of Science* 71: 87–113.

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