A Crack in the Track of the Hubble Constant

Marie Gueguen

Chapter 3
A Crack in the Track of the Hubble Constant

Marie Gueguen

Abstract Measuring the rate at which the universe expands at a given time–the 'Hubble constant'– has been a topic of controversy since the first measure of its expansion by Edwin Hubble in the 1920s. As early as the 1970s, Sandage and de Vaucouleurs have been arguing about the adequate methodology for such a measurement. Should astronomers focus only on their best indicators, e.g., the Cepheids, and improve the precision of this measurement based on a unique object to the best possible? Or should they "spread the risks", i.e., multiply the indicators and methodologies before averaging over their results? Is a robust agreement across several uncertain measures, as is currently argued to defend the existence of a 'Hubble crisis' more telling than a single 1% precision measurement? This controversy, I argue, stems from a misconception of what managing the uncertainties associated with such experimental measurements require. Astrophysical measurements, such as the measure of the Hubble constant, require a methodology that permits both to reduce the known uncertainties and to track the unknown unknowns. Based on the lessons drawn from the so-called Hubble crisis, I sketch a methodological guide for identifying, quantifying and reducing uncertainties in astrophysical measurements, hoping that such a guide can not only help to re-frame the current Hubble tension, but serve as a starting point for future fruitful discussions between astrophysicists, astronomers and philosophers.

M. Gueguen (✉)
Institut de Physique de Rennes 1, Université de Rennes 1, Rennes, France
e-mail: mgueguen@uwo.ca

N. Mills Boyd et al. (eds.), *Philosophy of Astrophysics*, Synthese Library 472,
https://doi.org/10.1007/978-3-031-26618-8_3

3.1 Introduction

From the realization in the end of the 1920s by Edwin Hubble that a relation of proportionality exists between the recessional velocities of galaxies and their distance; to the crisis around the Hubble constant that currently undermines the standard model of cosmology, the history of this constant has been that of the chase of a fleeing number that kept escaping the scientists' net. Among the most remarkable episodes of this track: the *Hubble war* in the 1970s, opposing Sandage and de Vaucouleurs, arguing both about the correct methodology to adopt for measuring the Hubble constant and about its actual value;[1] the dispute between Sandage and his colleague Wendy Freedman at the Carnegie Observatories of Pasadena in the 1980s, the latter defending a much higher value than the former, probably lying in the middle of the range spanned by Sandage and de Vaucouleurs; the disagreement since 2014 between two opponents that nobody had seen coming: the distant and the local universe, the former with a Hubble value of $67.4\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$, the latter one approaching 75, and finally the so-called and on-going Hubble 'crisis' that this persisting disagreement and its apparent confirmation by the publication of new local measures in 2019 have seeded.

The agitated history of the Hubble constant mirrors how fundamental this parameter has been for the development of our modern, precision cosmology. As one may remember from the famous words of A. Sandage, modern cosmology can be considered as the "search of two numbers": the values of the Hubble constant and of the q_0 parameter, which characterizes the deceleration in the expansion (Sandage 1970). But this troubled history certainly also reflects how tedious and delicate the task of measuring the Hubble constant is, and, as a result, how difficult it has been to assess the accuracy of its past measures. The determination of the Hubble constant requires one to find stellar objects (a) whose luminosity is known on theoretical grounds, (b) sufficiently far away from us to be freely moving (i.e., located in the so-called 'Hubble flow'), and (c) bright enough to be detected even that far away. But no single technique allows for the measurement of the distance satisfying all these properties. Hence, doing so necessitates deploying a 'cosmic ladder' from the nearby universe to the Hubble flow, where each of the rungs leads to a proliferation of systematic errors that must be constantly tracked and eliminated.

Yet, recent developments in astrophysics have led many scientists to consider that we know enough to consider the Hubble tension as a Hubble crisis. Such takes are grounded in the idea that (1) measurements of the Hubble constant have reached a sufficient precision for the discrepancy between early and late universe measures to become meaningful, and that (2) the robustness of the high value inferred from independent late universe techniques guarantees that no systematic errors will explain away this discrepancy. Here, I argue that in the context of highly uncertain measurements, methodologies that favor tracking the unknown unknowns

[1] See Guralp (2020).

always have epistemic priority over robustness arguments in the sense that they constrain the appropriate domain and timing for applying such arguments. On this basis, I contend that the current Hubble constant crisis is yet another avatar of a methodological confusion between the possible roles that different kinds of replication can play. The form of replication that robustness constitutes cannot be considered as evidence of a crisis when the necessary condition of systematic replication, which promotes tracking down unknown unknowns, is not successful.

Section 3.2 introduces the reader to the different ways of measuring the Hubble constant. In Sect. 3.3, I reconstruct the reasons that have been provided to justify the idea of a Hubble crisis and how they relate to the notion of robustness. Finally, Sect. 3.4 clarifies the roles that robustness and replication can play and contends that the use of robustness to establish a conclusion as dramatic as a Hubble crisis at this stage of the investigation is misguided: a day may come when a Hubble constant crisis arises and our cosmological model crashes down, but it is not this day.

3.2 How to Track the Hubble Constant

Different methods have been developed since Edwin Hubble's first attempt at measuring the Hubble constant *via* the cosmic ladder. One consists in inferring the Hubble constant from the early universe, for instance from the cosmic microwave background. The Santa Barbara conference of 2019 has seen new techniques based on the local universe blossom and reach a precision that makes their comparison with the most mature techniques genuinely informative. In this section, I briefly introduce some background for each of these techniques, such as to facilitate the philosophical interpretation of their concordance and of its signification for a cosmological crisis.

3.2.1 Jack and the Magic Bean: Building a Cosmic Distance Ladder in the Local Universe

At first sight, measuring the Hubble constant seems quite straightforward. The expansion law that must be solved in order to measure its value takes the following form:

$$c\frac{\delta\lambda}{\lambda_0} = H_0 D_0 \tag{3.1}$$

where c is the speed of light, $\frac{\delta\lambda}{\lambda_0}$ the redshift of the observed spectral lines of the galaxies compared to what would be expected only taking into account their distance, and D_0 their present distance. In other words, H_0 is known when the distance of an object and its redshift are known. Determining the redshift of a given

stellar object is done by comparing the observed spectral lines of galaxies to the 'laboratory' ones. The distance, on the other hand, is determined on the basis of two pillars: the choice of a standard candle on one hand, its apparent magnitude and its absolute magnitude on the other hand. A standard candle is an object that has a known intrinsic luminosity, referred to as its 'absolute magnitude' M-the brightness we would measure if we were standing 10 parsec away from it.[2] The apparent magnitude m of an object corresponds to its brightness as it appears to us, taking into account its distance and the effects that interstellar dust or bright stars nearby could have on it. The distance modulus μ is equal to $m - M$ and is related to the distance d in parsecs as follows:

$$\mu = 5log_{10}d - 5 \tag{3.2}$$

The problem, as we mentioned above, is that the relevant objects for measuring the Hubble constant must be located in the 'Hubble flow', and that determining the apparent magnitude of an object so far away requires a bundle of different methods that each comes with its own difficulties. There are, indeed, many phenomena that can alter the apparent magnitude of an object along our line of sight beyond its distance. It may, for instance, appear much fainter than it should, due to the absorption of part of its spectrum by the dust surrounding it –a phenomenon referred to as 'extinction'; or brighter than predicted, due to crowding effect by nearby stars. Such phenomena, among many others, must be accounted for and the apparent magnitude calibrated on this basis. Thus, objects in the Hubble flow cannot be *directly* probed. They require developing a ladder that will allow for the calibration of distances and magnitudes one rung at a time. Each rung is built on a different object, on a different technique, and on the information provided by the former rungs. Needless to say that in such a case, any error done in the first steps of the process has important repercussions on the final value found for the Hubble constant.

Cosmic distance ladders built to measure the Hubble constant are usually based on three rungs. The first rung 'anchors' the ladder in the sense that it serves as a zero-point calibration for extinction and crowding effects. Anchors must be sufficiently close to measure their distances with geometric methods, either through trigonometric parallaxes or Detached Binary Eclipses (DEBs). Anchors usually include the Large Magellanic Cloud, the Milky Way and NGC 4258. The second rung consists of determining the distances of galaxies known as 'calibrators', hosting both the selected standard candle and Type I Supernovae. Based on the calibration done in the first rung, the difference between apparent and absolute magnitudes allows for the determination of the distance of galaxies hosting both our

[2] Astronomers use the notion of "magnitude" to measure the brightness of an object. Magnitude is defined on a logarithmic scale, and the brighter an object is, the lower its magnitude. For instance, the absolute magnitude of the Sun is 4.8, but the faintest objects visible by the Hubble telescope have an apparent magnitude of 30.

standard candle and Type I Supernovae. Ideal standard candles consist of objects whose luminosity does not depend on their mass or composition. They usually fall into one of these two categories: either stars whose luminosity varies according to a known period-luminosity law, or extremely luminous objects whose brightness is due to a well-known and well-described phenomena. They typically include, among many other examples, Cepheid stars, whose average intrinsic luminosity varies depending on the period at which they pulse, a relationship well-documented and empirically verified by Henrietta Levitt.[3] Another favored one are Type Ia Supernovae, which correspond to a rare but extremely bright explosion –around five million times the brightness of the Sun! –, that of a dying white dwarf star exceeding its critical mass. Their brightness is perfectly suited for exploring the Hubble flow. The combination of the two provides both the anchors needed for building the first rungs of the cosmic ladder, and an access to regions where galaxies are freely moving. One starts by calibrating distances to nearby Cepheids in the LMC, Milky Way or NGC 4258, before gauging distances to much farther away Cepheids. Finally, the distance of Type I Supernovae in the Hubble flow is determined, on the basis of the second rung calibration.

This picture of a "Jack and the Magic Bean" astronomer climbing the cosmic ladder to catch supernovae, as beautiful as it is, is however anything but simple. As we mentioned above, each rung comes with many traps and errors propagating from one rung to the others. Maybe surprisingly, the zero-point calibration of the ladder is one of the trickiest part of the process and the largest source of systematic uncertainties. Uncertainties associated with the anchors carry over as systematic errors and have a huge possible impact on the determination of H_0. Yet, these sources of uncertainties are not only important, but impossible to reduce otherwise than by improving the accuracy of observational tools. Among them, one can include the distance to these anchors, extinction, but also difficulties related to converting the I-band photometric system of space telescopes to ground-based telescopes,[4] both needed. The second major source of uncertainties comes from the fact that it is usually difficult to find a statistically significant sample of galaxies that host both the relevant standard candles and Type I Supernovae. In the case of the Cepheids, four decades of research have allowed astronomers to build a sample of only 19 host galaxies[5]. But more generally speaking, standard candles are rarely really 'standard', as their luminosity may actually depend on their age or metallicity.[6] Random velocities of specific galaxies can be perturbed by local gravitational perturbations, thereby complicating the task of determining their redshift if the statistical sample of standard candles is not big enough to average

[3] Leavitt (1908).

[4] See Freedman et al. (2019, p. 11).

[5] Riess et al. (2022) have succeeded in more than doubling this sample in 2022, with a Hubble value now at $72.53 \pm 0.99\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$.

[6] In astronomy, the metallicity of a star corresponds to the heavy elements it contains, a 'heavy' element being any element other than hydrogen or helium.

away these perturbations. In sum, each rung comes with its ensemble of systematic and statistical proliferating errors, each of which could significantly distort the final value inferred for H_0. Significant progress has allowed for the improvement of the measurement of the Hubble constant based on Cepheids up to 1%,[7] after many years of rigorous investigation to explore and reduce the systematics associated with this technique.

3.2.2 Hubble Constant in the Early Universe

What characterizes measures of the Hubble constant based on the primordial universe is their model-dependence. One cannot infer the value of the Hubble constant from the early universe without already assuming a cosmological model. This feature holds both for measures inferred from the cosmic microwave background (henceforth CMB) or from the Baryon Acoustic Oscillations (BAO). For space reasons, I chose to limit the introduction to the early universe measures of H_0 to the CMB measure, but a detailed and accessible introduction to the BAO determination of H_0 can be found for instance in Fong (2011).

A couple words on the CMB first. When the primordial universe got cold enough for the first neutral hydrogen atoms to form –the epoch of 'recombination', photons decoupled from matter and started to free-stream across the universe. These photons have been propagating ever since, and the relic of this radiation is what is referred to as the 'cosmic microwave background' (henceforth CMB). This fossil electromagnetic radiation offers an extraordinary window into the early universe, as it provides a map of how matter was distributed across the universe at the time of decoupling and thus informs us about fundamental parameters of the ΛCDM model, included its matter density Ω_m. Assuming the standard cosmological model, and the dark energy density Ω_λ and spatial curvature k that characterise this model, secondary parameters such as the Hubble constant can be derived through the following equation:

$$H_z = H_0\sqrt{\Omega_m(1+z)^3} + \Omega_\Lambda + \Omega_k(1+z)^2 \tag{3.3}$$

A nearly exact geometrical degeneracy exists however, both for Ω_Λ and k, that make different cosmological models based on different Ω_Λ and spatial curvature k compatible with the anisotropies mapped by the CMB.[8] Hence the need and importance of a model-independent measure of H_0 that could waive this degeneracy, such as measures in the local universe.

[7] Riess et al. (2019).

[8] See e.g. Fong (2011) or Efstathiou (2020). The Planck 2018 results released in Collaboration (2020) seem to possibly break this degeneracy however.

3.3 A Tale of Two Values: The Hubble Crisis

Now, here lies the problem: the results delivered by the two techniques, the one based on the early universe and the Cepheids-based one, do not agree. In other words, the expansion of the universe as measured from our local universe is much faster than that predicted on the basis of the CMB, by almost 8%. The difference between the two is significant, close to 5σ: the value announced by the SH0ES team in 2019, led by A. Reiss and working with Cepheids, was $74.03 \pm 1.42\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$,[9] when the value obtained from the CMB after the last release of Planck results[10] is $67.4 \pm 0.5\mathrm{km\,s^{-1}\,Mpc^{-1}}$. Until now, this difference was not considered as too alarming. As we saw, the measure of the Hubble constant using the cosmic ladder with Cepheids as standard candles is a delicate task. Cepheids are young stars, thereby living in the dusty and crowded center of galaxies–an environment that maximizes extinction and crowding effects, whose period-luminosity depends on their age and metallicity, and whose nature itself is a problem, inasmuch as variable stars necessitate many exposures during several observational campaigns, adding new sources of systematic errors to account for. No wonder then that the first measure of H_0, based on Cepheids, was off by an order of magnitude: Edwin Hubble estimated the constant value around $500\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$. The complexity of the first technique, the multiple systematic and statistical errors that could affect each rung of the ladder, and the degeneracy associated with H_0 in the early universe context could legitimately lead us to think that, as the accuracy of the measures improve in the future, the results would have converged, especially as these values did progress a lot over the last two decades, and globally toward a possible convergence.[11]

The conference that was held in Santa Barbara in 2019 has however shaken this confidence and consolidated the gap that one was hoping to bridge. Over the last decade, many new techniques have indeed been developed to measure the Hubble constant independently of the Cepheids and the CMB, in order to break the tension between the two–especially as it becomes less and less clear how the precision of these measurements could further be improved. Four new measures were released during the conference that did not resolve the tension, but on the contrary corroborated the high value found by the SH0ES team, turning an apparent disagreement into a genuine problem, and generating a strong feeling of crisis among cosmologists.[12] Part of the goal of this paper is to elucidate the reasons that underlie such a feeling and whether these reasons are justified.

[9] See Riess et al. (2019).

[10] See Collaboration (2020).

[11] See Di Valentino et al. (2021) for an exhaustive plot of the current Hubble constant situation, and figure 17 of Freedman et al. (2019) to see the evolution of these values over the last two decades.

[12] A nice overview of the different reactions heard during this conference can be found in the paper "Cosmologists debate how fast the universe is expanding"(https://www.quantamagazine.org/print), written for *Quanta Magazine* by N. Wolchover.

3.3.1 The Blossoming of New Measurement Techniques

The main reason driving this sense of crisis is the fact that these new techniques are considered as independent measurements, that is, measurements that differ enough from the Cepheids measure to exclude possible common sources of systematic errors. Let us briefly review three of these techniques[13] to assess the extent to which this claim is justified.

- H0LiCOW: $H_0 = 73.3 \pm 1.7 \pm 1.8 \mathrm{km\,s^{-1}\,Mpc^{-1}}$
  The H0LiCOW[14] project uses strong gravitational lensing, i.e., the distortion of spacetime produced by supermassive objects, to measure the Hubble constant. This lensing phenomena permits the researchers to obtain multiple images of a same object, based on the different paths that the electromagnetic radiation follows given the curvature of spacetime. The idea of H0LiCOW is to study the light emitted by 5 quasars, i.e, extremely luminous active galactic nuclei whose magnitude varies. Massive galaxies between us and these objects act as magnifying and distorting lenses that multiply the images of the lensed target. Since light takes a different path for each of the images, the oscillation of the luminosity is delayed for each of these pictures. Thus, given that the distance travelled by light depends on the expansion of the universe, the time delay between each image allows to calculate the Hubble constant.
- MIRA variables:[15] $H_0 = 73.3 \pm 3.9 \mathrm{\,km\,s^{-1}\,Mpc^{-1}}$
  This method is a variant of the cosmic distance ladder, but using red pulsating stars called MIRAs as standard candles. The main sequence of the life of a star consists of converting the hydrogen contained in its heart into helium through nuclear fusion. At some point of the life of the star, after hydrogen and helium in the core are fully exhausted, the fusion starts in the outer shell which will expand, sometimes up to 1 AU; then cool down and shrink, before expanding again. This cycle is what makes MIRA-type stars good standard candles, because this period-luminosity relationship is fully determined by the mass and radius of the star. Note that this cycle, that lasts at minimum 100 days, is even longer than the Cepheids', whose fluctuations usually range from 1 to around 50 days.
- The Megamasers Cosmology Project:[16] $H_0 = 73.9 \pm 3.0 \mathrm{\,km\,s^{-1}\,Mpc^{-1}}$
  Megamasers and gravitational lensing are especially interesting techniques, as both offer the possibility of a direct measure of the Hubble constant, skipping the rungs of the ladder altogether. The characteristic of interest of masers is the equivalent of the laser effect in the microwave domain. The rough idea is

[13] I leave aside the result obtained by the Surface Brightness Fluctuations method, whose error bar is so large that its agreement/disagreement with other results is not meaningful. See more on this in Potter et al. (2018).

[14] See Wong et al. (2020).

[15] See Huang et al. (2020).

[16] See Pesce et al. (2020).

the following: the spontaneous emission of a photon generated by an atom's transition to its fundamental state triggers a cascade of similar emissions. The incident photon provokes the deexcitation of another atom; and thus the emission of another photon. Hence, the initial photon is so to speak photocopied up to a very powerful electromagnetic beam. Megamasers are typically located around 1pc of the center of a galaxy, close to the active galactic nuclei that can stimulate the surrounding gas clumps or water clumps (in the cases of water masers). Hydrogen and oxygen atoms composing the water maser absorb the galactic nuclei energy and radiate it in the form of a microwave 22 Hz beam that can be detected by Very Long Baseline Interferometry (VLBI). From this stable radiation can be inferred the velocities of gas clumps and water clouds orbiting the nuclei, their radius from the nuclei and distance to the galaxy, and their host galaxy's redshift.

These new techniques are all very promising, but also very recent. Beyond the many known sources of errors and unknown unknowns to uncover associated with them, their youth comes with a high price, that of relying on limited (and so possibly biased) statistical sample. Take for instance the H0LiCOW or the Megamaser Cosmology Project results: the former was based on only 7 lensed quasars in their 2019 paper, and the latter had only 4 megamasers– including NGC4258 which is used as an anchor for cosmic ladder techniques. One has to grant however that they are based on totally different physical assumptions, a fact that renders the possibility of unveiling common plausible sources of errors explaining their convergence toward a high value very unlikely.

This makes the fourth result published during the Santa Barbara conference even more disturbing. Indeed, the Carnegie-Chicago-Hubble program, led by W. Freedman, presented their Hubble value during this exact same conference, based on a new version of the cosmic ladder using the Tip of the Red Giant Branch stars (TRGBs) as a standard candle. Their announcement, far from solving the issue, created a new puzzle, since their value lies right in the middle between the low value based on the CMB and the high value of the local universe methods, at $H_0 = 69.8 \pm 0.8\,\mathrm{km\,s^1\,Mpc^{-1}}$.

3.3.2 Houston, We have a Rogue Measure

TRGBs offer an excellent standard candle to calculate the distance of supernovae. This phase of the evolution of red giants corresponds to the moment when a star of around 1 to $2M_\odot$ has exhausted the hydrogen of its core and started the fusion in the outer shell. Unlike what happens for stars of higher mass though, the core does not contract but is entirely sustained by the electron degeneracy pressure. As a result, the temperature increases in the core as the helium piles up there, without any corresponding dilatation of the star, until the temperature reaches 100,000 K. At this point a triple α reaction is triggered. Under the combined effects of the

temperature and the pressure exerted in the core, the fusion of helium into beryllium and into carbon turns into a runaway reaction and creates an extremely violent flash of helium, with a release of energy superior to the entire output of a whole galaxy. The infrared luminosity of stars going through an helium flash is independent from their mass and composition, which makes them eligible for the title of standard candles.

Probably one of the most interesting features of this technique is that it constitutes the perfect counterpart to the Cepheids technique. This means that, even though we currently have no way to decide whether one of these results, that obtained based on the Cepheids or that based on TRGBs is correct, a comparison between them is both very telling and very informative, as each one fills the lacuna of the other. Whereas Cepheids are short lived-stars that live in dusty environments, TRGBs are old stars living in isolation, in the outskirts of galaxies. As such, they are not as exposed to extinction and crowding effects as Cepheids are. Likewise, while the period-luminosity relationship of the latter depends on their metallicity, that of TRGBs can be accurately accounted for in two different ways: first, the infrared I-passband is almost not affected by their metallicity. Second, this metallicity manifests itself in the color of the star, by a widening of the RGB color that has been well-studied and calibrated empirically. Finally, TRGBs are not variable stars and do not require multiple exposures. As we will see in the next section, these complementary features constitute an ideal investigation path and offer a fecund scenario not only to test the robustness of the Hubble value, but also to discover new sources of uncertainties not necessarily accounted for in the report of the accuracy of these measurements. The question is thus the following: how can we explain the fact that, TRGBs excluded, the different methods based on the local universe agree on a high value and the methods based on the early universe on a low value, whereas at the same time the two methods that are the closest to each other, the more complementary and the more likely to agree fail to do so? How can we account for this success on one hand and this failure on the other?

3.4 Should We Call it a Crisis?

How should we thus react to these diverging measures? For many cosmologists, as reflected in the number of papers attempting to resolve the Hubble problem published since 2019, the convergence of the SH0ES, the H0LiCOW, the MIRAs and the Megamaser techniques towards a high value of H_0 is taken to indicate that the standard model of cosmology is undergoing a crisis. Although it is difficult to reconstruct the exact argument supporting a crisis, the papers that endorse the idea of a 'Hubble tension' or a 'Hubble crisis' tend to agree on the following statements: first, the discrepancy is much higher than the error estimates associated with each of these results. Put differently, the error bar is sufficiently small for a meaningful assessment of the discrepancy. Second, the independence of the techniques converging toward a high value for H_0 excludes that one single, shared,

source of systematic errors could waive the tension. Hence, the latter will likely resist an improvement of the accuracy of these techniques and a subsequent decrease in the size of their error bar.[17] The quote below summarizes this view:

> Given the size of the discrepancy and the independence of routes seeing it, a single systematic error cannot be the explanation. [...] Moreover, a suite of low redshift, different, truly independent measurements, affected by completely different possible systematics, agree with each other; it seems improbable that completely independent systematic errors affect all these measurements by shifting them all by about the same amount and in the same direction" (Verde et al. 2019, 7).[18]

3.4.1 From Robustness to Reliability

Although the word is never explicitly stated in the discussion, this line of argument captures what constitutes the core of robustness analysis as theorized within the tradition starting with Levins (1966), Levins (1993) and Wimsatt (2012).[19] Robustness analysis has been famously suggested by Levins as a way to assess the trustworthiness of models in the absence of a background theory providing analytically solvable equations. Since models have to be simplified to get predictions suseptible to being measured against nature, a method must be developed in order to evaluate the impact of these simplifications on the predictions of the model and to determine "whether a result depends on the essentials of the model or on the details of the simplifying assumptions" (Levins 1966, 423). One way to do so, to Levins' eyes, is to compare different models M_1, M_2, ..., M_N of the same target system, where each model is conceived of as the intersection of a common, plausible core C and of an unshared, variable part V_1, V_2, ..., V_N, and to look for a connection between C and a predicted property R (Levins 1993). The plausible core includes the biological or physical assumptions that are undergoing the test, while the unshared part corresponds to different idealizations or simplifications used to make the problem tractable. If one can show that the intersection of C with the union of the V_i implies R, then one can establish under certain conditions[20] that C alone implies R-put differently, that R does not depend on the V_i, but on the common core of all models whose adequacy is under test:

[17] See notably Verde et al. (2019), Efstathiou (2020), and Riess (2020).

[18] See also (Riess 2020, p. 2).

[19] See also Weisberg (2006) and Weisberg (2012) for a more recent take on this version of robustness analysis. Of course, many schools of thoughts have arisen since Levins' and Wimsatt's accounts of robustness analysis. The goal here, however, is not to address whether robustness is a sound tool, but whether the kind of robustness allegedly displayed in the Hubble context supports the existence of a crisis. Therefore, the paper focuses on their version of robustness, which seems to capture the line of argument defended by cosmologists. For a more up-to-date account of how robustness is used in the actual practice of scientists, see among many others (Soler et al. 2012).

[20] For instance, the condition that the set of V_i exhausts the space of admissible possibilities.

> (...) Thus the search for robustness as understood here is a valid strategy for separating conclusions that depend on the common [...] core of a model from the simplifications, distortions and omissions introduced to facilitate the analysis, and for arriving at the implications of partial truths. (Levins 1993, 554).

Levins, however, remained rather careful about what can be learned through robustness; at least in his 1993 piece:

> (...) the more inclusive the set of V_i's, the more we can have confidence that C implies R. If we feel that the set of V_i's spans a wide enough range of possibilities, then we may generalize to claim that C usually implies R, a result that is not very exciting as a mathematical theorem but may be good biology. (Levins 1993, 554)

It was Wimsatt in 1981 who tightened the bond between robustness and reliability. According to him, robustness analysis is defined through the following three principles: first, it is a procedure aiming at distinguishing the "reliable from the unreliable"; second, it requires one to show the invariance of that which reliability is scrutinized over independent[21] processes or models, in order to build confidence in their independence from these; and finally it requires determining the scope of this invariance. Hence, within this framework, establishing reliability is no longer a mere possible goal of robustness, but one of its core and definitional tenets–robustness and reliability go hand in hand, and where one is to be found, the other is expected:

> [A]ll the variants and uses of robustness have a common theme in the distinguishing of the real from the illusory; the reliable from the unreliable; the objective from the subjective; the object of focus from artifacts of perspective; and, in general, that which is regarded as ontologically and epistemologically trustworthy and valuable from that which is unreliable, ungeneralizable, worthless, and fleeting. (46)

Since Wimsatt, robustness has been generally accepted as an indicator of reliability, that is, as evidence that a prediction is not an artifact of specific modelling assumptions.[22] Orzack and Sober have however convincingly undermined this claim, in a beautiful paper that seeded many of the questions about robustness that have been debated ever since. As they see it, there are three possible scenarios resulting from a Levins-like robustness reasoning:

- Scenario 1: We already know that one of the M's among M_1, M_2, ..., M_N is true.[23] In that case, if for all i, M_i implies R, then R must be true. As emphasized in Justus (2012, 797), this inference is unproblematic but also relatively uninteresting, given that robustness is precisely needed in those cases

[21] For space reasons, I will not dwell on how to characterize what 'independence' means here, given that models have to be of the same target and thus presumably share some core assumptions to be even comparable. This term is present in both Wimsatt's and Levins work, but never fully elucidated and is subject to controversy. See for instance Schupbach (2015).

[22] See for instance Weisberg (2006), Weisberg and Reisman (2008) or Soler et al. (2012), but also Parker (2011) for a criticism.

[23] Although we do not endorse the terminology in terms of 'truth' that Orzack and Sober use, we will keep it here in order to remain faithful to the authors.

where no observations or no analytic solutions are available that could establish the truth of one of the M_i.

- Scenario 2: We know that all the models are false. In this case, we have no reasons to believe that the fact that each M_i implies R is evidence that R is true. Their simple example illustrates this point beautifully: if all models we compare in population biology admit natural selection as the only force acting on the size of the population, then all models will predict populations with infinite size. That they agree does not say anything about the truth of the prediction, but only about the convenience of the assumption (Orzack and Sober 1993, 538).
- Scenario 3: We do not know whether one of the models is true. This is the most common situation in astrophysics and cosmology, and robustness is precisely used in these contexts to help to establish the reliability of predictions converging across different models. According to Orzack and Sober, we have no more reasons yet in this situation to infer that R is true than we have in the second scenario: "if we do not know whether one of the models is true, then it is again unclear why a joint prediction should be regarded as true (ibid., 538–539).

From a purely inferential point of view, I think that Orzack and Sober make a valid logical point here in emphasizing that we have no reason to consider R as true in any of the last two scenarios. Nonetheless, I am still willing to grant that the last scenario can correspond to very different epistemic situations, and that for each of them the degree of confidence possibly supported by the robustness of R or its value as a heuristic guide could vary a lot. A comparison where recently developed techniques, with many shared assumptions, and from which little is understood, are compared, does not support a high degree of confidence in R. But it seems reasonable to say that a much higher degree of confidence in R would be justified–in the words of Levins, would be "good astrophysics"– if the comparison is performed across mature and independent models, that have been rigorously examined and studied such that systematics and statistical errors have been identified and reduced to the best of our knowledge. The question is thus the following: which one of these situations corresponds to the Hubble constant crisis's scenario? Are we in the epistemic position to apply robustness and infer conclusions with a high degree of confidence on this basis? Or are we putting the cart before the horse?

3.4.2 Temporary Discrepancy vs. Residual Discrepancy

When are we justified in thinking that a discrepancy is symptomatic of a crisis? My –presumably uncontroversial– answer to this question would be: when the discrepancy is not a temporary one, but a *residual* discrepancy. That is, when we find ourselves in the case where known sources of systematic and statistical errors have been quantified and sufficiently reduced for a comparison between the two values thus obtained and their error bars to be significant, and when enough efforts have been done to chase unknown unknowns, i.e., new sources of yet undiscovered

errors. A *residual* discrepancy, in other words, is the discrepancy that remains after adequate efforts have been deployed to identify, quantify and reduce all possible sources of uncertainties that could explain away the disagreement.

Of course, this does not mean that we need to be absolutely certain that all sources of errors have been excluded. It does mean, however, that the significance of the tension is directly related to the assumption that all sources of uncertainty have been identified and accounted for. If we have strong evidence that there are still errors, which are not accounted for but which could resolve the tension, the robustness of the tension does not have sufficient epistemic strength to justify a call for a crisis.

Up to the Santa Barbara conference, one can see that the discrepancy between the early and the late universe measures was interpreted as a temporary one, thus merely seen as a tension rather than as a crisis. What changed after the Santa Barbara conference is that the robustness of the values resulting from the local universe measurements[24] was interpreted as an indication that the discrepancy had gone from temporary to residual. Now, do we have good reasons on the basis of the new results to think that the epistemic situation switched from temporary to residual? And is robustness the appropriate tool to decide whether this is the case, that is, if systematic and statistical errors have been sufficiently purged from our measurements?

In the remainder of the paper, I content that we have clear evidence that not only the discrepancy is not residual, but that robustness has so far been unsuccessful in detecting unknown systematic errors in our case study, notably because of the emphasis on comparing measurements *as independent as possible*. I illustrate the latter point by showing that the robustness of the high late-universe value is blind to the systematics since acknowledged (notably) in time-delay cosmography. Next, I show that the tool that could diagnose these errors is actually in competition with robustness and thus often neglected despite its epistemic priority.

3.4.2.1 The Example of Time-Delay Cosmography

Announced in 2019, the H0LiCOW result was considered as one of the most important evidence of a Hubble crisis. As mentioned above, a measure based on gravitational lensing is a *direct* measure of the Hubble constant, in that it does not require appealing to the ladder technique. Furthermore, the H0LiCOW and the Cepheids' measurements are as independent as two measures of the Hubble constant can be. They rely on completely different objects and different physics, whereas the Mira project is a version of the cosmic ladder which might suffer from the same

[24] This robustness holds only at the price of excluding the TRGBs' result of course. The SH0ES team (Yuan et al. 2019) has justified such an exclusion on the basis of, as they argued, a calibration error on the TRGBs side. This claim has now been debunked several times (Freedman et al. 2020; Freedman 2021; Mortsell et al. 2021).

issues as the Cepheids or the TRGB technique (due to common anchors for instance) and the Megamaser Project involves the maser located in NGC4258, which is a common zero-point for the Cepheids and the TRGB measures. Thus, an agreement between the two independent measures released by H0LiCOW and by SH0ES was considered as particularly exciting and telling, and a major reason to interpret the Hubble tension as a Hubble crisis.

But this technique is a really young one, and a lot of work still needs to be done to understand how the different assumptions that enter this measurement might distort the result. Note that this is not pure speculation about possible future developments for time-delay measurements of gravitational lensing. The effects of relaxing assumptions about the mass density profile of the deflector have already been carefully studied, with surprising results. Indeed, one of the most important sources of uncertainties in time-delay lensing is the mass profile of the deflector. If no assumption is made about it, the precision of the measurement, based on the 7 H0LiCOW lenses, drops from 2% precision to 8%. Such an error budget is far too important to resolve the Hubble tension. So where do the mass assumptions used in this context come from?

A lens model should ideally be able to reproduce the observables associated with the lens with as few unconstrained parameters as possible. With respect to quasar astronomy however, there are too few observational constraints to reach this standard, which means that different models can reproduce the same set of observations but give different $H_0\delta t$ product and thus different values for H_0. This degeneracy can be broken by relying on stellar kinematics–from which the above 8% uncertainties estimate is obtained– or, for an improved precision, by further constraining the mass distribution of the lensing system. Traditionally, the two main solutions adopted are that of a power-law, or of a constant mass-to-light ratio plus the so-called Navarro-Frenk-White dark matter halo density profile[25] inferred from simulations (Navarro et al. 1997). These mass assumptions are not however chosen on theoretical grounds or observational constraints. To be sure, surveys did show that the mean slope of the density profile of lenses is nearly isothermal. But this slope is an average, and thus need not be adequately described by a power law (Schneider and Sluse 2013). On the contrary, there are good reasons to think that the central regions of the lens would significantly depart from a perfect power law. Yet, the studies published by Birrer et al. (2020) and Birrer and Treu (2021) have shown that, with a sample of lenses increased from 7 to the 33 lenses from the SLAC-TDCOSMO collaboration, if the mass modelling assumptions are relaxed to be maximally degenerated, then the value obtained for the Hubble constant is no longer of $H_0 = 73.3 \pm 1.7 \pm 1.8\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$, but of $H_0 = 67.4 \pm 4.1 \pm 3.2\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$, no longer in significant tension with the CMB measure. Such a result demonstrates how much we need to further improve our understanding of the systematics for time-delay cosmography before we can claim a 2% precision measurement that does not

[25] Note that the NFW profile is challenged, as it fails to reproduce the observations especially for low-surface-brigtness galaxies. See for instance Bullock and Boylan-Kolchin (2017).

involve unjustified assumptions when it comes to gravitational lensing. Clearly, the high Hubble value with a small error bar highly depends on an assumption that has no strong physical justification. Acknowledging what we still do not know, while not taking any bets, leads to a low value with a much larger error bar. But more importantly, it shows that the robustness of the high Hubble value is no guarantee that this high value is not an artefact from systematic errors. Had the robustness argument for a crisis been taken at face value, the track for these unknown facts would have stopped and the importance of the mass distribution assumption not been properly understood.

3.4.2.2 Systematic Replication and Unknown Unknowns

Now, what about the Cepheids's claim of 1% precision measurement, grounded in more than four decades of refining the ladder technique? Can we legitimately believe that new systematics could remove the discrepancy at this stage of the scientific investigation?

Before addressing this question, a short detour through another toolbox, that of replication, is necessary. The Replication Crisis,[26] according to which many findings in social, behavioral and biomedical sciences have failed to replicate at alarming rates, has led to many interesting developments when it comes to understand what is a replication and what purpose it can serve. I will briefly present a typology of replication adapted from Schmidt and Oh (2016), Schmidt (2016), Zwaan et al. (2018), and Fletcher (2021) before going back to our current issue. The typology I suggest orders replication along four categories. 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:[27]

- *Direct replication*: direct replication is an attempt to reproduce exactly the original study, on a different statistical set. It is especially useful to exclude errors related to the statistical sample or to contextual factors. In the case of the Cepheids' technique for instance, a successful direct replication would consist of obtaining the same Hubble constant value based on different Cepheids calibrators.
- *Methodological replication*: this kind of replication is a simple re-analysis of an experiment, ideally by another team. As Fletcher puts it, methodological replication "ensures that the results of a scientific study are not due to data-entry, programming or other suchlike technical errors", or in general to human errors.

[26] An introduction to the replication crisis and to its importance can be found in Romero (2019).

[27] One way to think about this spectrum is that the typology aims to capture situations going from: nothing is changed in the replication, only one thing is changed, to several if not all are changed. How you quantify the number of variables that vary also depends on how fine-grained your perspective is.

The re-analysis of the SH0ES data found in Javanmardi et al. (2021) is such an example, among many ones, of such methodological replication.

- *Systematic replication*: it consists in systematically varying one of the variables of the experiment or measure while maintaining the others fixed. The goal of systematic replication is to help to identify which variables causally contribute to the final outcome, but more importantly to better understand and circumscribe the causal contribution of a given variable to the result. Contrary to the other three, systematic replication is more about understanding a protocol, measurement, experiment or model than about assessing the reliability of its prediction.
- *Conceptual replication*: here, the goal is to measure the same phenomenon or test the same hypothesis as in the initial study, but by using different methods, techniques or models. Conceptual replication subsumes robustness analysis: an agreement between different models or different types of measurements amount to a successful replication. But given that the more independent the measurements under comparison are, the stronger the link between robustness and reliability is, robustness is on the far-end of the spectrum–it is a form of conceptual replication that insists on the fact that the original and the replicated study ideally have nothing in common but the targeted value. In other words, the most extreme form of conceptual replication is needed to warrant strong robustness-based conclusions. The alleged agreement between the four local measures detailed above is supposedly of such a nature: it is a successful conceptual replication, allegedly between fully independent measures, inasmuch as the TRGB result is excluded.

The comparison between the TRGB and the Cepheids results exemplifies how the replication spectrum goes from systematic replication to conceptual replication. As the reader may remember, TRGB and Cepheids have complementary weaknesses and strengths, as well as common and independent zero-point anchors: on the Cepheids side, the zero-point calibration is based on Milky Way parallax, on the LMC or on NGC4258, and the sample of Type I SNe used is the Supercal sample. On the TRGB side, the zero-point calibration has been done on the basis of the distance modulus to the LMC based on DEBs + Hubble Space Telescope parallax calibration in 2019, and on the basis of the LMC, NGC 4258 and the Milky Way globular clusters in 2020. The sample of supernovae can be that of the Carnegie-Chicago Program or the Supercal SNIa sample used by the SH0ES team. Galaxies where the calibrators (either TRBG or Cepheids) can be found as well as SNe include 18 host galaxies on the TRGB side, 19 on the Cepheids's side, 11 of which are common to the two groups. In other words, the comparison between the two results can be constructed such as to maximally overlap and leave only the choice of standard candle as the variable explored–which amounts to the perfect picture of a systematic replication, or to be fully independent, which would amount in the case of an agreement to a perfect conceptual replication, inasmuch as the standard candle used is no longer considered a mere variable but a method. As it turns out, neither the systematic replication nor the conceptual replication are successful replications in this context: recently updated TRGB and Cepheids measurements result in differing

values of $H_0 = 69.6 \pm 1.9\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ (Freedman et al. 2019) for the TRGB and 73.04 ± 1.4 (Riess 2020) for the Cepheids. But the conceptual replication does not teach us anything about how to locate the problem, as differences in several variables do not allow us to pin down the most probable culprit. Differences in one variable only, as it the case with systematic replication, do not fall prey to this problem. If the two measurements only differ in the sample of supernovae chosen, then the calibration or a possible bias in the sample of supernovae is most likely responsible for the disagreement. It is only because the Carnegie Chicago Hubble Program led by W. Freedman proceeded to a detailed systematic comparison between TRGB stars and Cepheids that we now have a better idea about where to look for possible unidentified unknown unknowns. While the two methods show excellent agreement on the distance modulus to 28 galaxies for instance, the study shows that this agreement no longer holds when comparing the distance to the 10 SNIa host galaxies that the two have in common. Future observational campaigns with much higher resolution, notably thanks to the JWST telescope, might be in a position to elucidate this disagreement.

The failure of systematic replication not only indicates with no possible doubts that new systematics are yet to be discovered, but informs us about where to look for them: if one wants to test an hypothesis about a possible source of systematics (e.g., the distance modulus to the LMC), the overlap between these two techniques easily allows one to design a crucial test that permits one to verify such an hypothesis–for instance by comparing the TRGB result to the value inferred from Cepheids only on the basis of the Milky Way and of NGC 4258. Likewise, the fact that the metallicity can be constrained for TRGB stars allows one to decouple the problem of metallicity and of extinction, given that the TRGB I-band is not affected by metallicity effects. Metallicity and extinction can thus be individually solved, and the measure of the extinction obtained from TRGB stars can inform the calibration of Cepheids for common objects.[28] Hence the epistemic superiority of systematic replication in contexts of highly uncertain measurements, and the need to wait for successful systematic replication before applying robustness arguments:

- Given that the focus of robustness is on comparing models or measurements that are as independent as possible, arguments drawn on its basis remain mute and offer no explanations and no guide to locate the problem when the robustness analysis fails. A good example of this is a comparison between the success of the robustness strategy when excluding the TRGB result and its failure when including it. Once the claim of poor calibration on the TRGB side is excluded, how ought we to account for this failure? How can scientists decide where to start to explain it? Systematic replication, on the other hand, is maximally informative and is in a position to identify possible sources of failure. Conceptual replication does not have in principle to be mute about such possibles sources of errors, but

[28] For a detailed comparison between TRGB stars and Cepheids for the different replications performed, see section 4 of Freedman et al. (2019).

the part of the conceptual replication spectrum that corresponds to robustness does, inasmuch as it necessitates the maximization of the independence of different measurements.

- As we have seen above, a successful robustness analysis or conceptual replication does not logically establish the reliability of a given prediction. But the failure of systematic replication demonstrates unequivocally that new systematics have yet to be identified. Hence, if we grant the claim made in 3.4.1 that a high degree of confidence is better supported when the comparison holds between mature and independent models, free as much as possible of unknown systematic and statistical errors, then systematic replication does not only have the epistemic priority to assess a discrepancy, but also the chronological priority. Indeed, it is the role of systematic replication to diagnose whether such unknown unknowns have still to be identified. It is only when successful systematic replications are performed that one can be confident that the most important sources of systematic and statistical errors have been accounted for. In other words, if systematic replication cannot be performed successfully, it demonstrates that the conditions for applying conceptual replication understood as robustness are not yet met, at least in the conditions that would allow for a high degree of confidence in conclusions drawn from it.[29] In the case of the Hubble constant, the failure of the systematic replication performed on the Cepheids and TRGB results show that the precision of these measurements, though by far the most mature techniques for determining the value of the Hubble constant, has not reached a sufficient level for robustness arguments to be telling and/or trusted.

Two remarks are needed to qualify the claim made here. First, one does not have to reject robustness analysis altogether on these grounds. Robustness analysis and systematic replication are complementary tools and can work together very well to address different problems. They cannot however be deployed at the same stage of the inquiry, as the latter indicates when the conditions for justified inferences based on robustness are met. The appeal to robustness has no epistemic grounds if systematic replication is not successfully achieved. If systematic replication fails, no meaningful robustness-based conclusions about the reliability of the measure can be drawn. Second, when used too early, robustness analysis is actually an obstacle to its companion and leads to neglect of the chase for unknown unknowns. This happens because robustness, as an extreme form of conceptual replication, focuses on developing independent techniques that have, ideally, absolutely nothing to do with each other–e.g., time-delay cosmography and Cepheids-based distance ladder. On the other hand, systematic replication requires measurement techniques sufficiently close to each other to be mutually informative, as Cepheids and TRGB

[29] Although space reasons prevent me to expand on this point, it would be interesting to analyze the different scenarios that can arise: failure of CR vs success of SR, failure of both, and so on and so forth. Here we only address the claim that a specific kind of conceptual replication, that usually referred to as robustness analysis, is sufficient to warrant the reliability of the late-universe measurements and thus the problematic nature of the discrepancy.

can be. Robustness deployed too early leads to developing techniques that are too independent from each other to offer the grounds needed for systematic replication. Hence the need to understand their roles and places in the scientific investigation, so as to not let robustness becomes the crack in the track of the Hubble constant. The failure of systematic replication offered by the TRGB and the Cepheids' measurements tells us that the discrepancy between early universe and local universe measurements is not a residual discrepancy, contrary to what the defenders of a crisis would like to see on the basis of their robustness strategy. Moreover, it informs us about how measurements can be further improved and where to start doing so. Robustness cannot be used to justify such as crisis if it is not established that the track for unknown unknowns has gone far enough, and that we are indeed comparing mature and well-understood techniques. There might be a cosmological crisis to come, but such a crisis is not justified by the current epistemic situation, and certainly not on the basis of robustness arguments made at this stage of the scientific investigation.

3.5 Conclusion

Recent developments in astrophysics have seen the community working on the Hubble constant shaken by the robustness of the high value found by local universe measurements, and subsequently by the significant discrepancy between this value and the one obtained from early universe measurements. Some have gone as far as claiming that these developments prove that the standard model of cosmology is undergoing a crisis. The robustness of the high-values, they contend, shows that the discrepancy between the early and late universe will hold, and thus is the long-wished for evidence that new physics is needed to amend the standard model. I hope to have shown here that we do not have good reasons to interpret the discrepancy as residual on the basis of robustness arguments, and to think that we are currently facing a crisis. The Hubble debate does not offer the conditions that would warrant a strong degree of confidence in robustness-based inferences. The fact that we seem to have reached 1% precision measurements is not yet a sign that systematic errors have been almost eliminated, but that the unknown systematics are getting harder and harder to track. The blossoming of new techniques is an opportunity not to establish the robustness of the discrepancy, but to use their overlap to deploy systematic replication and refine our understanding of where the skeletons could still be hiding.

Acknowledgments The author is very grateful to Chris Smeenk, Barry Madore and Wendy Freedman for their helpful feedback. They would also like to thank the Bonn History and Philosophy of Physics Seminar, the APC colloquium (Paris 7), and the audience of the "What can Astrophysics teach us about Replicability?" workshop organized by the CSH center (Bern) for insightful discussions. Finally, the author would like to thank the two anonymous referees for their helpful feedback. This paper is based on work done while funded under the John Templeton Foundation grant: "New Directions in Philosophy of Cosmology" (grant number 61048).

References

Birrer, S., A. Shajib, A. Galan, M. Millon, T. Treu, A. Agnello, M. Auger, G.-F. Chen, L. Christensen, T. Collett, et al. 2020. TDCOSMO-IV. Hierarchical time-delay cosmography–joint inference of the Hubble constant and galaxy density profiles. *Astronomy & Astrophysics* 643: A165.

Birrer, S., Treu, T. 2021. TDCOSMO-V. Strategies for precise and accurate measurements of the Hubble constant with strong lensing. *Astronomy & Astrophysics*, 649: A61.

Bullock, J.S., M. Boylan-Kolchin. 2017. Small-scale challenges to the ΛCDM paradigm. arXiv:1707.04256.

Collaboration, P. 2020. Planck 2018 results. VI. Cosmological parameters. *Astronomy and Astrophysics* 641(A6): 1–56.

Di Valentino, E., O. Mena, S. Pan, L. Visinelli, W. Yang, A. Melchiorri, D.F. Mota, A.G. Riess, J. Silk. 2021. In the realm of the hubble tension—a review of solutions. *Classical and Quantum Gravity* 38(15): 153001.

Efstathiou, G. 2020. A Lockdown Perspective on the Hubble Tension (with comments from the SH0ES team). Manuscript. arXiv preprint arXiv:2007.10716. Accessed on 12 May 2023.

Fletcher, S.C. 2021. The role of replication in psychological science. *European Journal for Philosophy of Science* 11(1): 1–19.

Fong, Y.C. 2011. Measuring the hubble constant. https://www.imperial.ac.uk/media/imperial-college/research-centres-and-groups/theoretical-physics/msc/dissertations/2011/Yick-Chee-Fong-Dissertation.pdf

Freedman, W.L. 2021. Measurements of the hubble constant: Tensions in perspective. *The Astrophysical Journal* 919(1): 16.

Freedman, W.L., B.F. Madore, D. Hatt, T.J. Hoyt, I.S. Jang, R.L. Beaton, C.R. Burns, M.G. Lee, A.J. Monson, J.R. Neeley, 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(1): 34.

Freedman, W.L., B.F. Madore, T. Hoyt, I.S. Jang, R. Beaton, M.G. Lee, A. Monson, J. Neeley, J. Rich. 2020. Calibration of the tip of the red giant branch. *The Astrophysical Journal* 891(1): 57.

Guralp, G. 2020. Calibrating the universe: The beginning and end of the hubble wars. *Standardization in Measurement. Philosophical, Historical and Sociological Issues* 125–138.

Huang, C.D., A.G. Riess, W. Yuan, L.M. Macri, N.L. Zakamska, S. Casertano, P.A. Whitelock, S.L. Hoffmann, A.V. Filippenko, D. Scolnic. 2020. Hubble space telescope observations of Mira variables in the SNIa host NGC 1559: An alternative candle to measure the hubble constant. *The Astrophysical Journal* 889(1): 5.

Javanmardi, B., A. Mérand, P. Kervella, L. Breuval, A. Gallenne, N. Nardetto, W. Gieren, G. Pietrzyński, V. Hocdé, S. Borgniet. 2021. Inspecting the Cepheid distance ladder: The hubble space telescope distance to the SN Ia host galaxy NGC 5584. *The Astrophysical Journal* 911(1): 12.

Justus, J. 2012. The elusive basis of inferential robustness. *Philosophy of Science* 79(5): 795–807.

Leavitt, H.S. 1908. 1777 variables in the magellanic clouds. *Annals of Harvard College Observatory* 60: 87–108.

Levins, R. 1966. The strategy of model building in population biology. *American Scientist* 54(4): 421–431.

Levins, R. 1993. A response to Orzack and Sober: Formal analysis and the fluidity of science. *The Quarterly Review of Biology* 68(4): 547–555.

Mortsell, E., Goobar, A., Johansson, J., Dhawan, S. 2021. Sensitivity of the Hubble Constant Determination to Cepheid Calibration. Manuscript. arXiv preprint arXiv:2105.11461. Accessed on 12 May 2023.

Navarro, J.F., C.S. Frenk, S.D. White. 1997. A universal density profile from hierarchical clustering. *The Astrophysical Journal* 490(2): 493.

Orzack, S.H., E. Sober. 1993. A critical assessment of Levins's the strategy of model building in population biology (1966). *The Quarterly Review of Biology* 68(4): 533–546.

Parker, W.S. 2011. When climate models agree: The significance of robust model predictions. *Philosophy of Science* 78(4): 579–600.

Pesce, D.W., Braatz, J.A., Reid, M.J., Condon, J.J., Gao, F., Henkel, C., Kuo, C.Y., Lo, K.Y., Zhao, W. 2020. The Megamaser Cosmology Project. XI. A Geometric Distance to CGCG 074-064. *The Astrophysical Journal*, 890, pp. 340–398.

Potter, C., J.B. Jensen, J. Blakeslee, P. Milne, P.M. Garnavich, P. Brown. 2018. Calibrating the type Ia supernova distance scale using S brightness fluctuations. In *American astronomical society meeting abstracts# 232*, vol. 232, 319–02.

Riess, A.G. 2020. The expansion of the universe is faster than expected. *Nature Reviews Physics* 2(1): 10–12.

Riess, A.G., S. Casertano, W. Yuan, L.M. Macri, D. Scolnic. 2019. Large magellanic cloud cepheid standards provide a 1% foundation for the determination of the Hubble constant and stronger evidence for physics beyond λCDM. *The Astrophysical Journal* 876(1): 85.

Riess, A.G., W. Yuan, L.M. Macri, D. Scolnic, D. Brout, S. Casertano, D.O. Jones, Y. Murakami, G.S. Anand, L. Breuval, et al. 2022. A comprehensive measurement of the local value of the Hubble constant with 1 km s^{-1} mpc^{-1} uncertainty from the Hubble space telesope and the SH0ES team. *The Astrophysical Journal Letters* 934(1): L7.

Romero, F. 2019. Philosophy of science and the replicability crisis. *Philosophy Compass* 14(11): e12633.

Sandage, A.R. 1970. Cosmology: A search for two numbers. *Physics Today* 23(2): 34–41.

Schmidt, F.L., I.-S. Oh. 2016. The crisis of confidence in research findings in psychology: Is lack of replication the real problem? Or is it something else? *Archives of Scientific Psychology* 4(1): 32.

Schmidt, S. 2016. Shall we really do it again? The powerful concept of replication is neglected in the social sciences. In *Methodological issues and strategies in clinical research*, ed. A.E. Kazdin, 581–596. Washington: American Psychological Association.

Schneider, P., D. Sluse. 2013. Mass-sheet degeneracy, power-law models and external convergence: Impact on the determination of the Hubble constant from gravitational lensing. *Astronomy & Astrophysics* 559: A37.

Schupbach, J.N. 2015. Robustness, diversity of evidence, and probabilistic independence. In *Recent developments in the philosophy of science: EPSA13 Helsinki*, 305–316. Berlin: Springer.

Soler, L., E. Trizio, T. Nickles, W. Wimsatt. 2012. *Characterizing the robustness of science: After the practice turn in philosophy of science*, vol. 292. Berlin: Springer.

Verde, L., T. Treu, A.G. Riess. 2019. Tensions between the early and late Universe. *Nature Astronomy* 3: 891–895.

Weisberg, M. 2006. Robustness analysis. *Philosophy of Science* 73(5): 730–742.

Weisberg, M. 2012. *Simulation and similarity: Using models to understand the world*. Oxford: Oxford University Press.

Weisberg, M., K. Reisman. 2008. The robust Volterra principle. *Philosophy of Science* 75(1): 106–131.

Wimsatt, W.C. 2012. Robustness, reliability, and overdetermination (1981). In *Characterizing the robustness of science*, 61–87. Berlin: Springer.

Wong, K.C., S.H. Suyu, G.C. Chen, C.E. Rusu, M. Millon, D. Sluse, V. Bonvin, C.D. Fassnacht, S. Taubenberger, M.W. Auger, et al. 2020. $H0LiCOW$–XIII. A 2.4% measurement of H_0 from lensed quasars: 5.3 σ tension between early-and late-universe probes. *Monthly Notices of the Royal Astronomical Society* 498(1): 1420–1439.

Yuan, W., A.G. Riess, L.M. Macri, S. Casertano, D.M. Scolnic. 2019. Consistent calibration of the tip of the red giant branch in the large magellanic cloud on the hubble space telescope photometric system and a redetermination of the hubble constant. *The Astrophysical Journal* 886(1): 61.

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