The role of replication studies in theory building

Abstract:

At least since Meehl's (in)famous (1978) paper, the state of theorizing in psychology has often been lamented. Replication studies have been presented as a way of directly supporting theory development by enabling researchers to more confidently and precisely test and update theoretical claims. In this paper I use contemporary work from philosophy of science to make explicit and emphasise just how much theory development is required before 'good' replication studies can be carried out, and show just how little theoretical pay-off even good conceptual replications offer. I suggest that in many areas of psychology aiming at replication is misplaced, and that instead replication attempts are better seen as exploratory studies that can be used in the cumulative development of theory and measurement procedures.

1. Introduction

At least since Meehl's (in)famous (1978) paper, the state of theorizing in psychology has often been lamented. Replication studies have been presented as a way of directly supporting theory development by providing more robust data sets and promoting the use of diverse ways of testing theories. This is supposed to enable researchers to more confidently and precisely test theoretical claims, and suggests how to revise theoretical frameworks where necessary.

In this paper I use contemporary work from philosophy of science to make explicit and emphasise just how much theory development is required before 'good' replication studies can be carried out, and show just how little theoretical pay-off even good conceptual replications offer. I suggest that in many areas of psychology aiming at replication is misplaced, and that instead replication attempts are better seen as exploratory studies that can be used in the cumulative development of theory and measurement procedures.

Replication studies are in fact often used in a more exploratory way, and the core themes of the discussion here will probably not be news to psychologists. What is analysed here are the implications of taking these themes seriously. Call a replication study that appears to test the core claim from an original study, and does so in a fairly convincing way (i.e. cannot be easily dismissed), a 'good' replication. While conducting good replication studies is generally seen to be fairly difficult, it still seems to be generally maintained that aiming at replication is a reasonable goal. Instead, if it is accepted that conducting good replications is currently not possible in some areas of psychology (which seems to fit some of rhetoric in

the literature), then aiming at replication, at least in the short term, is clearly not a reasonable thing to do. In addition, if it is accepted that a single round of replication is not sufficient by itself to confirm or disconfirm the existence of an effect, then many of the current practices and language around replication need to change.

The paper proceeds as follows. In Section 2 I briefly outline the basic reasoning behind the use of robustness analyses in science, of which replication studies are an example. In Section 3 I identify broad areas of theory that need to be fairly well developed in order for 'good' replications to be carried out. In Section 4 I analyse how little theoretical pay-off is generated by both direct and conceptual replications. In Section 5 I outline the contemporary view on the nature of theory development and cumulative progress from philosophy of science, and use this to sketch an alternative role for 'replication' studies in psychology, in which replication itself largely drops out of the picture.

2: Robustness and replication

To set up the rest of the paper, I briefly outline the nature of robustness analyses, of which replication studies are an example. Robustness analyses are ways of testing how sensitive theoretical estimates, inferences, measurement outcomes, models, phenomena, and more, are to differences in the way we generate or investigate them. If there are multiple ways of finding things out about a target phenomenon, either theoretically or experimentally, and they all generate similar outcomes, we have a result that is *robust* to variations in theory and experimental procedures (see e.g. Weisberg, 2006; Wimsatt, 2007; Woodward, 2006).

Replication studies are a type of robustness analysis: measurement outcomes are compared across studies in order to make claims about how robust results are across differences in measurement procedures. Direct replications are where the (relevant aspects of) experimental procedures of a selected study are reproduced as closely as possible. Successful direct replications help to rule out false positives and possible experimenter effects. Conceptual replications retain the basic theoretical reasoning in the selected study but use different procedures or operationalisations of variables to test an experimental hypothesis. Successful conceptual replications provide information about the underlying theory's 'generalizability'.

Conceptual replications in particular are usually the focus in philosophy of science. Woodward's (2006) explanation of the reasoning around conceptual replications is recognisable from similar discussions of replication strategies in psychology (e.g. see discussions in Schmidt, 2009, and Stroebe and Strack, 2014). Independent measurement procedures use different causal pathways to access the value of the target, through, for example, using different instrumentation and/or different ways of experimentally intervening on the target. If the outcomes from a range of independent procedures are coherent, that is, all the various measures of the value of X are roughly the same, then we infer that we have fairly accurately measured the value of X. This is because it is more likely that measurement procedures that validly and accurately measure the value of a target will generate the same outcome, compare to a set of measurement procedures that measure the target inaccurately (in different ways), or perhaps do not measure it at all. So, if it turns out that a set of measurement procedures produce similar outcomes, then we have "grounds for increasing our confidence that the quantity has been measured accurately"

(Woodward 2006, p. 234). Similarly, Schmidt (2009) states that "With every difference that is introduced [to the replicating study] the confirmatory power of the replication increases" (p. 93).

3. Theoretical inputs into replication studies

3.1 Direct vs conceptual replications

First I accept the claim from e.g. (Stroebe and Strack 2014) that all replications in psychology are conceptual. As Fabrigar and Wegener (2016) point out (see also Schmidt, 2009; Stroebe & Strack, 2014), the original materials or operationalisations used in a study may be developed for a specific population or context, so they may not generate the same psychological phenomenon when used with other populations or in other contexts. Instead, Fabrigar and Wegener recommend focusing on psychometric invariance: that 'direct' replications should try, as closely as possible (bearing in mind context/population effects) to recreate the same psychological conditions as the original study. In effect, this recognizes that the causal pathways across replications will never be identical: due to the subject matter of psychology, all replications are conceptual. However, it also recognizes that there are ways to minimize the causal independence of studies where this is valuable. However, for brevity, I continue to use the terminology of direct and conceptual replications. Below these should be read as replications that are less ('direct') or more ('conceptual') causally independent from an original study, but all of which are essentially conceptual.

The fact that all replications are conceptual does not make things easy. Indeed, conceptual replications have been criticised as drivers of theory development and scientific progress because failed conceptual replications are easy to dismiss on the grounds that they were simply not appropriate replications in the first place (e.g. Nosek et al., 2012; Pashler & Harris, 2012). If all replications are conceptual to at least some degree, this makes even the more 'direct' replications open to this criticism too.

3.2 Three demands for theoretical input

Call a replication study that appears to test the core claim from an original study, and does so in a fairly convincing way (i.e. cannot be easily dismissed), a 'good' replication. Good replications might succeed or fail to find evidence that supports the original core claim.

There is not, as far as I can tell, a huge amount of discussion about the level of theoretical development required to design a good replication. There is (perhaps understandably) more discussion on how to identify successful and failed replications. For example, a number of authors have recommended that rather than simply counting up apparent successes and failures, replication study by replication study, (where success is identified as finding a statistically significant result in the same direction as the original), that multiple studies are routinely combined in meta-analyses (Braver et al., 2014; Fabrigar & Wegener, 2016; Gelman, 2018; Maxwell et al., 2015; Schmidt & Oh, 2016; Stanley & Spence, 2014). However, this suggestion relies on replications being good ones that do genuinely target the core claim from the original study. The computer simulations used in Stanley and Spence (2014) and Braver et al. (2014), and the examples discussed in Maxwell et al. (2015), build

'goodness' in, and the discussion of direct replications in Fabrigar and Wegener (2016) is prefaced with the phrase "Presuming that the replication experiments have achieved psychometric invariance…" (p. 75) (i.e. are good replications).

What counts as a good replication will be determined by local factors, but there are some very general areas of theory that are required to build good replications. To be clear, these areas of theory do not have to be finalised or entirely correct, but they do at least need to be fairly plausible and fairly well worked out. I outline these areas here, and briefly discuss their likely availability.

1) To do a good direct replication, and preserve psychometric invariance, one needs to have a pretty good idea of how differences in e.g. social or historical context, or demographic, educational or cultural differences across test populations, might affect whether an original set of stimuli will trigger the same psychological response in a new population, and if so, how to update the study. Again, the idea here is not that replication studies have to be criticism-proof from this point of view. Instead, the idea is that *if* the aim is to try to replicate a finding, such that others will take the replication results seriously, the replication needs to be 'good' enough that it, in a fairly defensible way, seems to target the core claim from the original study.

Adapting studies in this way might sometimes be straightforward, but as suggested by responses to the Many Labs replication projects (Ebersole et al., 2016; Klein et al., 2014b; Klein et al., 2018), sometimes it is not. Some of the adaptations or other factors questioned in these commentaries include changing associations to national symbols, the degree to

which certain ethnic/religious groups are treated as outgroups across cultures, ambient lab temperature, participants doing multiple studies in one sitting, whether "lots of work-related travel" is now rated as a neutral feature of a parent, and more (e.g. Crisp et al., 2014; Ferguson et al., 2014; IJzerman et al., 2016; Petty & Cacioppo, 2016; Wilson, 2016). The wealth of possible factors that might affect whether psychometric invariance is achieved, particularly in social psychology, is vast. As a result of this, there is often no well developed and widely accepted set of background theory that can be easily referred to in order to confidently and precisely inform the design of good direct replications.

2) To do a good conceptual replication, one that is deliberately (more) causally independent of the original study, one needs to know even more. Using new measurement procedures and operationalisations of key variables requires having a pretty good idea of how to generate different ways of intervening on the target phenomenon, and of assessing its impacts on behaviour. This demands having a fairly fleshed out theory of the general causal profile of the target phenomenon: what reliably causes it and what it reliably causes in turn, relative to key situational factors.

The (apparent) general aversion to carrying out conceptual replications suggests that researchers are even less confident of being able to do these than direct replications. The worry that failed conceptual replications can be easily dismissed suggests that researchers often have to go out on a theoretical limb in order to construct a conceptual replication, filling in areas of theoretical detail, which can be easily rejected by the original authors. These areas of theoretical detail could be in both relevant background theory, as is required for direct replications, but also in terms of the core theory being tested. Conducting good

conceptual replications is therefore particularly demanding in terms of the kind of theory development required.

3) To know what kind of replication one is looking at in the first place, and so what kind of inferences can be drawn from it, one also needs to know how causally independent the replication is from the original study. Strong inferences about the accuracy and validity of measurements, and so the adequacy of underlying theory, can only be based on procedures that are fairly causally independent from each other (i.e. more conceptual). However, knowing the degree of causal independence between studies again demands a fairly comprehensive understanding of the causal profiles of the measurement procedures and target phenomena involved. Philosophers of science have argued that this is often very difficult to assess (e.g. Stengenga, 2009).

To be clear, this problem of identifying degrees of causal independence is not obviously solved by the more complex taxonomies of replication studies offered by, for example, Hüffmeier et al. (2016) and LeBel et al. (2017). LeBel et al. describe how replications can be more or less similar to an original study based on counting how many of seven 'design facets' are the same or different (e.g. operationalizations and stimuli of dependent and independent variables). This certainly tracks one notion of similarity, but it may be have little to do with causal independence, and this is the crucial feature. In principle, a procedure that incorporates changes to all of the design facets of the original study may still be fairly similar in terms of basic causal structure. Indeed, implementing all these changes may sometimes be necessary for preserving psychometric invariance if working with a radically different population/context to the original study, but intending to do a more

'direct' replication. In contrast, a replication study that changes only one design facet, but does so in a way that changes the causal structure of the experimental intervention in a major way may be substantially causally independent of the original study. Assessing causal independence is therefore messier than counting mere changes in experimental design.

In sum then, to do a good direct replication, one needs to have a fairly well developed and widely accepted set of background theory about the extended causal profile of the target phenomenon: enough to adapt experimental stimuli to new populations and contexts. In order to do a good conceptual replication, one needs to have a more comprehensive theory of the causal profile of the target phenomenon: how to reliably intervene on the target and how to capture its causal effects in a significantly different way to the original study. And in order to know whether you're looking at a more direct or a more conceptual replication, and so what kind of inferences once can draw from it, one again needs a pretty good idea of the causal profile of the target. Without this information, a researcher cannot claim to have performed a good replication. And without this, the apparent success or failure of the replication is not obviously informative on whether or not original findings can, in fact, be replicated.

The concern that follows directly from this is that in areas of psychology where theory development is most needed, and so where replication studies might, on the face of it, be most useful, informative replication studies are much harder to perform. Indeed, in some areas they may currently be impossible. That is, in some areas of psychology, relevant areas of theory might not be sufficiently developed to enable good (i.e. generally defensible) replications to be carried out. In the next section I argue that the power of replication

studies to inform theory is also (sometimes) overstated in the psychological literature. That is, even where one can perform good replications, they do not offer a huge theoretical pay-off.

4: The theoretical pay-off from replication studies.

4.1 Direct replications

In philosophical circles, direct replications are deemed to be largely uninformative about the theoretical claims an experiment is aimed at testing. At most, what successful direct replications show is that the measurement procedure is repeatable: when you apply the procedure under relevantly similar conditions, it generates the same outcome. Mere direct replication shows nothing though, by itself, about the accuracy or validity of the measurement procedures or the truth of the underlying theory being tested. Very similar measurement procedures that produce repeatable outcomes might measure a target in similarly inaccurate ways, or might not measure the target at all. So, while repeatable procedures clearly do *something* in a reliable way, it is not clear from the fact of repeatability alone what it is.

A useful example to illustrate this is from Chang's (2004) work on the history of thermometry. In the 1840s, amid ongoing debates about how to define temperature, and what type of thermometer should be used as a laboratory standard, experimentalist Henri Victor Regnault sought to establish which type of thermometer (mercury or air) was more

repeatable: which type of thermometer best agreed with itself under similar conditions. He found that air thermometers were in fact the most repeatable. Mercury, air and the glass used to make the thermometers all expand when hot, and when glass expands, it affects thermometer readings. However, air expanded so much over the temperature range tested that the effects of the expansion of the glass used to hold it were (in relation) fairly minimal. That is, the physical features of the air thermometer made it more robust to a factor that significantly affects the way that thermometers operate. As Chang notes though, Regnault "never strayed from the recognition that comparability [repeatability] did not imply truth" (p. 83, *op cit*), that is, the repeatability of air thermometers was no guarantee that they were capable of accurately measuring temperature. As above, they were definitely doing something reliably, but with the background theory available at the time, it wasn't clear what it was.

This fits with some of the claims made about the power of replication studies in the psychological literature. For example, Crandall and Sherman (2016) argue that conceptual replications "can contribute more to theoretical development and scientific advance" (p. 94) essentially by offering ways to falsify the theories by confronting them with findings that are likely to be accurate. However, these authors state that direct replications, still offer *some* increase in "Confidence in methods of [original] study" and a "modest improvement" in the confidence in related theory (Crandall and Sherman, 2016, see Table 1, p. 95). One possibility is that these claims are limited to the specific operationalisations used. That is, insofar as our operationalisations go, we can use successful direct replications to increase our confidence that they are getting at something, and so we can marginally increase our confidence in the theory they are based on. The findings might still of course be artifacts of

poor operationalisations.

Zwaan et al. (2018) in their defence of direct replications seem to say something rather stronger though: that direct replications can be informative at a theoretical level. Failed direct replications can show that "a theoretically predicted effect is not empirically supported" (p. 9, *op cit*), or they might lead to further work investigating why a specific replication failed (e.g. perhaps it failed to maintain psychometric invariance). That is, direct replications really can tell you about the truth of the theoretical claims being tested. This however is based on the assumption that there is some pre-existing confidence that the measurement procedures used are valid and reasonably accurate. For example, Zwaan et al. (2018) write that "It goes without saying that scientific judgment should be used to assess the validity and importance of a study before deciding whether it is worth replicating" (p. 9) and that a published measurement procedure is likely to be at least somewhat valid "because its authors and the reviewers and editors who evaluated it endorsed the method as a reasonable test of the underlying theory" (p. 8). Say then that we are fairly confident that a measurement procedure does validly test a theoretical prediction, but there is a lack of clarity about the statistical power of the study (for whatever reason). In this case, direct replication can help to resolve this problem. Here then, being fairly confident already about the validity of measurement procedure, we can confront our theoretical predictions with more statistically robust data.

However, this does not seem to capture the relevant state of play in psychology, at least not in the areas where there are stronger concerns about the state of theorizing and cumulative progress. It is rarely the case here that there are a battery of measurement procedures that

are known to be largely valid prior to any form of conceptual replication, and that provide neat and clear tests of theoretical predictions, but which (for whatever reason) are generally not applied with sufficient statistical power. Instead, there are often questions about just how valid the measurement procedures actually are. Clearly, some measurement procedures (even published ones!) are more or less obviously valid than others. However, without conducting conceptual replications and so testing the validity of measurement procedures in that way, it is not clear where a sufficient degree of confidence in the validity of measurement procedures could come from such that direct replications alone could support theoretical claims. Again, direct replications, in the sense of generating appropriate statistical power, are important. But alone, they can't do much.

4.2 Conceptual replications

Conceptual replications, where measurement procedures are more causally independent, are usually assumed to support inferences about the validity and accuracy of measurement procedures and measurement outcomes. With more replications, and more replications that are significantly causally independent of each other, one can be more and more confident in these inferences. In turn, one can be more confident in the implications they have for the underlying theory being tested, in particular how accurate it is, and where and how well it generalises across different domains of application.

A potentially bigger concern then is whether good conceptual replications are all that powerful. In fact, there is an emerging consensus in philosophy of measurement that even good conceptual replications do not offer a significant theoretical pay-off. In particular,

Hudson (2014) has argued that they are rarely used across the sciences in the way standardly depicted, and he provides an alternative historical analysis of (among others) the often cited case of estimating Avogadro's number. One of the reasons for this shift in view is that the 'causal independence' requirement on conceptual replications is vague, and in any case hardly ever met to a significant degree. That is, measurement procedures used across conceptual replications are often (unsurprisingly) based on broadly the same theoretical assumptions, and so show little causal independence in the way they intervene on the target phenomenon. In this case though, even good conceptual replications are not particularly informative. As they are based on interacting with the target in broadly similar causal ways, the fact that replication studies can generate the same outcome is not particularly surprising, and therefore cannot support very strong claims about the accuracy of the measurement outcomes or the validity of the measurement procedure. In turn, even good conceptual replication studies do not provide particularly strong tests of underlying theory.

One reaction to this has been to acknowledge some of the problems associated with conceptual replication studies, but at the same time deny that they have no value (e.g. Basso, 2017; Eronen, 2015; Soler, 2014). There may be a range of fairly significant constraints on the successful application of replication studies, and in isolation (i.e. without well-developed theory) they are not capable of driving theory development. However, these authors argue that suitably contextualised within relevant theory, it is possible that conceptual replication studies can be informative and productive.

These suggestions are in fact roughly in line with Meehl's original (1978) recommendations about how to use consistency tests. Meehl's consistency tests assess a different kind of robustness to the one tested in replication studies, but the same idea applies. Meehl suggests that comparing two results, in a meaningful way, requires "that methods of setting permissible tolerances exist" (1978, p. 829). That is, comparing two outcomes requires having already established roughly how close you expect them to be if they do in fact both measure the same target variable. He claims that precise statistical tests are not used to compare outcomes in the hard sciences, but that on the basis of background theory researchers have a good idea of what counts as outcomes being " 'reasonably close'". He implies that this is the only sensible way of doing things in psychology too.

Descriptions of 'coherent calibration' (Tal, 2017a) or 'measurement assessment via robustness' (Basso, 2017) from philosophy of science develop this further. Calibration relies on the same basic idea that drives simple replication studies, but has an added twist that allows for procedures to be largely causally dependent. The method requires systematically comparing the outcomes from a set of measurement procedures with each other, *and* against theoretical predictions about the kinds of errors or uncertainties associated with each procedure, to assess how well the procedures successfully measure the target phenomenon as it is theoretically defined. Where the outcomes cohere, in light of expectations about measurement errors or uncertainties, then this is taken as evidence that the procedures do fairly accurately measure the target phenomenon, and that the background theory is largely correct. If they do not cohere, in light of these expectations, then this is used to further investigate and revise the procedures used, or the theoretical expectations about them, in order to measure the target more accurately.

To illustrate this method, Basso (2017, pp. 63-64) gives an example of different measurement procedures for assessing poverty. These are based on somewhat different definitions of poverty, but it is generally accepted that they mostly track the same thing. According to $poverty_1$ an individual is in poverty if they have an income below a certain threshold. According to $poverty_2$ an individual lives in poverty if they have insufficient income to maintain a lifestyle that is customary in the society they belong to (e.g. including diet, living conditions, activities). To illustrate how measurement calibration works in these cases, Basso reports one study (Hick, 2015) which compared the groups identified by measurement procedures based on these different definitions as being 'at risk' of poverty. It is widely thought that measurement procedures based on $poverty_2$ systematically underestimate poverty in older populations. However, once this is taken into account, the outcomes from these two sets of measurement procedures cohere, and identify similar groups as being 'at risk'.

This comparison shows two things. First, it shows that the expectation of systematic error in one set of procedures is accurate, that is, procedures based on $poverty_2$ do indeed underestimate poverty in older populations. Second, it shows that since the outcomes of these two procedures are coherent, once systematic errors are taken into account, that both groups of procedures fairly accurately identify groups of people who are genuinely at risk of poverty.

This analysis of the inferential power of conceptual replications adds to the case made in Section 3 that replication studies demand a fairly rich set of background theoretical

knowledge. In addition to the types of theory development detailed above, one also needs (again, with a reasonable degree of confidence, but not absolute certainty), support for the general validity and accuracy of the measurement procedures used, and the ability to estimate, in advance, likely measurement errors across different studies.

This is clearly demanding stuff. In 1978 at least, Meehl was sceptical that this theoretical richness was available across social psychology (and some degree of scepticism seems appropriate now too):

"For example, [say] Meehl's Mental measure correlates .50 with SES in Duluth junior high school students, as predicted from Frisbee's theory of sociability. When Jones tries to replicate the finding on [Mexican-American] seniors in Tucson, he gets *r* = .34. Who can say anything theoretically cogent about this difference? Does any sane psychologist believe that one can do much more than shrug?" (Meehl, 1978, p. 814)

This analysis also shows that even conceptual replications, done with a reasonable amount of statistical power and against a reasonable set of background theory, do not support very strong or novel claims about the accuracy and validity of measurement procedures. At best, they make it possible to make local updates to the specific ways in which procedures are judged to be valid and accurate, where researchers have a reasonable degree of confidence in the general adequacy of measurement procedures and theory already. In turn, even good conceptual replications, in themselves, do not support very strong claims about the adequacy of relevant theory.

5. Theory development and the idea of progress

In this section I use the analyses from Sections 3-4 above to present an alternative view of the role of replication studies in theory development in psychology. To do this I first briefly present the contemporary view of scientific progress from philosophy of science. I then argue that in many areas of psychology aiming at replication *per se* is misplaced, and also misleads about the way that 'replications' can inform theory development and drive cumulative progress. To be clear, similar points to those I make here have been made in the psychology literature. For example, it has been well recognised that 'replication' attempts should ideally inform theory in an iterative way, with theory informing replication design, replication results informing theory, and on again (e.g. Earp & Trafimow, 2015; Ebersole et al., 2017; Klein et al., 2014a). However these points are (unsurprisingly) rarely seen as undermining the aim and practice of replication in general.

To start, there is a very basic problem that, on the face of it, makes cumulative progress in science impossible: the problem of co-ordination. This problem is that identifying a valid measurement procedure for a target phenomenon cannot be achieved without a developed theory about the properties of that target, and vice versa. In other words, you cannot know if you have a good way of 'getting at' something unless you have a good idea what it is, and you cannot know what something is without having a good way of 'getting at' it. Various moves have been made in philosophy of measurement to get around this problem, many very familiar in psychology (e.g. operationism), but most of which have met a sorry end (for review see Tal, 2017b).

The current view in philosophy of science is that scientific progress occurs in a coherentist manner: advances in theory and experiment go hand in hand and slowly self-correct (Chang (2004) and van Fraassen (2010)). One starts from somewhere, and iteratively and slowly makes local improvements to get somewhere better: "What we have is a process in which we throw very imperfect ingredients together and manufacture something just a bit less imperfect" (Chang 2004, p. 226). To illustrate this, Chang offers a simple but useful taxonomy of progress as enrichment and self-correction. Enrichment includes adding precision to theory or experimental techniques, and expanding the scope of theoretical claims and experimental techniques. Self-correction includes making more accurate theoretical predictions and measurements, based on the current state of empirical knowledge. At all points though, progress is only made possible by building on existing theoretical knowledge and existing experimental procedures: *both* are required.

The analyses above essentially suggested that conducting good replication studies across at least some areas of psychology is not currently possible. That is, in some areas of psychology there is not yet enough well developed theory to conduct replication studies that would be widely agreed to test the core claim from an original study. In addition, even good conceptual replications do not, afterall, come with a big theoretical pay-off.

While this might sound like a problem, in the context of this newer conception of scientific progress, it is not. Instead, the problem is with the expectations that it is possible to generate good replications in the absence of well developed theory, and that a single round

of replication studies should be able to confirm or disconfirm the existence of an effect. These expectations are wholly unrealistic in any science, and should be dropped.

Instead, replication studies conducted in the absence of well developed theory (and often in the presence of well developed theory too) are much better seen as *exploratory* studies that support different stages of theory development concerning a target phenomenon. Generating exploratory studies, both of direct and conceptual kinds, often requires constructing theory where there was none before, making assumptions about what factors might or might not be causally relevant, and using these to (re)test a claim. Often, the results are different to what was predicted, at which point more theory is constructed to explain the differences. This in turn is used to update the overall theory, and/or better experimental controls are identified, and testing begins over again. Parts of this process can involve enrichment, for example by making theoretical claims more precise, and/or by identifying the range of conditions under which an effect occurs. This process may also require taking a step back from testing the main theory to test sub-claims: self-correction can involve systematically identifying the underlying assumptions of a theory or measurement procedure, testing and exploring them to see whether they are warranted, and updating them where necessary. If and when theoretical work degenerates, the focus might temporarily shift (as it did for Regnault) onto questions that can be tackled in a fairly direct experimental way with relatively limited theoretical backing (the downside being that this comes with limited or no theoretical pay-off).

Importantly, this cycle of iteratively testing, updating and repeating, might go on for a while before anything like a good replication, as defined in Section 3, can be carried out. By this

point of course, there will already be a reasonable degree of confidence in the validity of measurement procedures used and the general adequacy of relevant theory. And, as the current state of knowledge informs what counts as a 'good' replication, what counts as a 'good' replication can change. The limited theoretical pay-off that conceptual replications offer is perfectly in line with this view of cumulative scientific process.

As an illustration of this model, Luttrell et al. (2017) identified possible factors that had led to a null result in a replication attempt concerning a core findings from the Elaboration Likelihood Model of persuasion (from Ebersole et al., 2016) and explicitly manipulated these factors experimentally to investigate whether they were indeed the cause of the null result. Ebersole et al. (2017) praised this as a model of good science and extended it by conducting direct replications of Luttrell et al.'s study across 9 locations. Although Luttrell et al.'s findings did not all replicate (or replicate very closely), the general strategy is praised by Ebersole et al.:

"With this observe-hypothesize-test sequence, [the authors] treated the different outcomes of [an original study and replication attempt] as worthy of study rather than simply hypothesizing about the failure to replicate in defense of the original results. In this regard, Luttrell, Petty, and Xu have provided a model of productive scientific critique worth emulating." (Ebersole et al. 2017, p. 186)

From a philosophy of science perspective, the fact that this model is in some way seen as novel, or worth pointing out as a model to emulate is a little worrying. It is essentially just a description of how cumulative progress in science is made. So, something else seems to be

going on here. One possibility is that the language and practices around replication just don't invite engagement. If replication is treated as something that is done in a single round, with either the success or failure left to stand, then obviously further engagement is not obviously required or beneficial. Treating replication attempts as exploratory studies, which is often what they actually are, shifts this: exploratory studies invite engagement and further exploration.

One possible complaint with the claims made here is that going through this slow and iterative process for investigating psychological phenomena is hard, because these phenomena are so complex. The very idea of having a reasonably well mapped out theory of what factors might affect a specific psychological phenomenon and how to causally intervene on it in multiple ways might strike some as implausible. But if that is the case, then aiming at replication is deeply misplaced anyway, because it will *always* be possible to come up with an alternative explanation of why a particular replication succeeded or failed: replication attempts would just never be informative. Indeed, the fact that replication is seen as a plausible aim in psychology, even if it is actually only possible in the long-term, suggests that researchers do in fact view the slow, plodding, iterative exploratory research detailed above as a do-able enterprise. Of course, pursuing a science of complex phenomena in not straightforward, but it is possible.

In sum then, treating replication attempts as exploratory studies better recognises the role that they do and should play in research, and drops unreasonable expectations about the roles that they (in many cases) cannot yet play. As exploratory studies, replication attempts function as one of a battery of empirical and theoretical practices used to slowly and locally

improve the veracity and scope of theory, and the validity and accuracy of measurement procedures, that together drives cumulative scientific progress.

References

Basso, A. (2017). The appeal to robustness in measurement practice. *Studies in History and Philosophy of Science Part A*, *65*, 57–66.

Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science*, *9*(3), 333–342.

Chang, H. (2004). *Inventing temperature: Measurement and scientific progress*. Oxford University Press. https://books.google.co.uk/books?hl=en&lr=&id=JHtPAwAAQBAJ&oi=fnd&pg=PR11&dq=inventing+temperature+chang&ots=O_wGClKaQY&sig=-iSssMYbwq93AyEsVWMcSNnBRdM

Crisp, R. J., Miles, E., & Husnu, S. (2014). Support for the replicability of imagined contact effects. *Social Psychology*, *45*(4), 303–304.

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00621

Ebersole, C. R., Alaei, R., Atherton, O. E., Bernstein, M. J., Brown, M., Chartier, C. R., Chung, L. Y., Hermann, A. D., Joy-Gaba, J. A., Line, M. J., Rule, N. O., Sacco, D. F., Vaughn, L. A., & Nosek, B. A. (2017). Observe, hypothesize, test, repeat: Luttrell, Petty and Xu (2017) demonstrate good science. *Journal of Experimental Social Psychology*, *69*, 184–186. https://doi.org/10.1016/j.jesp.2016.12.005

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., Baranski, E., Bernstein, M. J., Bonfiglio, D. B. V., Boucher, L., Brown, E. R., Budiman, N. I., Cairo, A. H., Capaldi, C. A., Chartier, C. R., Chung, J. M., Cicero, D. C., Coleman, J.

A., Conway, J. G., … Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68–82. https://doi.org/10.1016/j.jesp.2015.10.012

Eronen, M. I. (2015). Robustness and reality. *Synthese*, *192*(12), 3961–3977.

Fabrigar, L. R., & Wegener, D. T. (2016). Conceptualizing and evaluating the replication of research results. *Journal of Experimental Social Psychology*, *66*, 68–80.

Ferguson, M. J., Carter, T. J., & Hassin, R. R. (2014). Commentary on the attempt to replicate the effect of the American flag on increased Republican attitudes. *Social Psychology*, *45*(4), 301–302.

Gelman, A. (2018). Don't characterize replications as successes or failures. *Behavioral and Brain Sciences*, *41*, 19–20.

Hick, R. (2015). Three perspectives on the mismatch between measures of material poverty. *The British Journal of Sociology*, *66*(1), 163–172.

Hudson, R. (2014). *Seeing things: The philosophy of reliable observation*. Oxford University Press.

Hüffmeier, J., Mazei, J., & Schultze, T. (2016). Reconceptualizing replication as a sequence of different studies: A replication typology. *Journal of Experimental Social Psychology*, *66*, 81–92.

IJzerman, H., Szymkow, A., & Parzuchowski, M. (2016). Warmer hearts, and warmer, but noisier rooms: Communality does elicit warmth, but only for those in colder ambient temperatures — Commentary on Ebersole et al. (2016). *Journal of Experimental Social Psychology*, *67*, 88–90. https://doi.org/10.1016/j.jesp.2015.12.004

Klein, R., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W.,

Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., … Nosek, B. A. (2014a). Theory building through replication response to commentaries on the “many labs”replication project. *Social Psychology*, *45*(4), 307–310. https://doi.org/10.1027/1864-9335/a000202

Klein, R., Ratliff, K., Vianello, M., Adams, R., Bahník, Š., Bernstein, M., Bocian, K., Brandt, M., Brooks, B., Brumbaugh, C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E., … Nosek, B. (2014b). Investigating Variation in Replicability: A “Many Labs” Replication Project. *Social Psychology*, *45*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178

Klein, R., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., & Bahník, Š. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490.

LeBel, E. P., Berger, D., Campbell, L., & Loving, T. J. (2017). Falsifiability is not optional. *Journal of Personality and Social Psychology*, *113*(2), 254–261. https://doi.org/10.1037/pspi0000106

Luttrell, A., Petty, R. E., & Xu, M. (2017). Replicating and fixing failed replications: The case of need for cognition and argument quality. *Journal of Experimental Social Psychology*, *69*, 178–183. https://doi.org/10.1016/j.jesp.2016.09.006

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does “failure to replicate” really mean? *American Psychologist*, *70*(6), 487.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*(4), 806.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*(6), 615–631.

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, *7*(6), 531–536.

Petty, R. E., & Cacioppo, J. T. (2016). Methodological choices have predictable consequences in replicating studies on motivation to think: Commentary on Ebersole et al.(2016). *Journal of Experimental Social Psychology*, *67*, 86–87.

Schmidt, F. L., & Oh, I.-S. (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. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, *13*(2), 90–100.

Soler, L. (2014). Against robustness? Strategies to support the reliability of scientific results. *International Studies in the Philosophy of Science*, *28*(2), 203–215.

Stanley, D. J., & Spence, J. R. (2014). Expectations for replications: Are yours realistic? *Perspectives on Psychological Science*, *9*(3), 305–318.

Stegenga, J. (2009). Robustness, discordance, and relevance. *Philosophy of Science*, *76*(5), 650–661.

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science*, *9*(1), 59–71.

Tal, E. (2017a). Calibration: Modelling the measurement process. *Studies in History and Philosophy of Science Part A*, *65*, 33–45.

Tal, E. (2017b). *Measurement in Science*. The Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/archives/fall2017/entries/measurement-science/

Van Fraassen, B. C. (2010). *Scientific representation: Paradoxes of perspective*. Oxford University Press.

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

Wilson, A. E. (2016). Exact replications in an inexact context: Commentary on Ebersole et al. (2016). *Journal of Experimental Social Psychology*, *67*, 84–85. https://doi.org/10.1016/j.jesp.2015.12.008

Wimsatt, W. C. (2007). *Re-engineering philosophy for limited beings: Piecewise approximations to reality*. Harvard University Press.

Woodward, J. (2006). Some varieties of robustness. *Journal of Economic Methodology*, *13*(2), 219–240.

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