Uljana Feest. Paper to be presented at PSA2018: The 26th Biennial Meeting of the Philosophy of Science Association, Nov. 1-4, 2018, Seattle

Why Replication is Overrated

Current debates about the replication crisis in psychology take it for granted that direct replication is valuable and focus their attention on questionable research practices in regard to statistical analyses. This paper takes a broader look at the notion of replication as such. It is argued that all experimentation/replication involves individuation judgments and that research in experimental psychology frequently turns on probing the adequacy of such judgments. In this vein, I highlight the ubiquity of conceptual and material questions in research, and I argue that replication is not as central to psychological research as it is sometimes taken to be.

1. Introduction: The "Replication Crisis"

In the current debate about replicability in psychology, we can distinguish between (1) the question of why not more replication studies are done (e.g., Romero 2017) and (2) the question of why a significant portion (more than 60%) of studies, when they *are* done, fail to replicate (I take this number from the Open Science Collaboration, 2015). Debates about these questions have been dominated by two assumptions, namely, first, that it is in general desirable that scientists conduct replication studies that come as close as possible to the original, and second, that the low replication rate can often be attributed to statistical problems with many initial studies, sometimes referred to as "p-hacking" and "data-massaging."[1]

[1] An important player in this regard is the statistician Andrew Gelman who has been using his blog as a public platform to debate methodological problems with mainstream social psychology (http://andrewgelman.com/).

I do not wish to question that close (or "direct") replications can sometimes be epistemically fruitful. Nor do I wish to question the finding that are severe problems in the statistical analyses of many psychological experiments. However, I contend that the focus on formal problems in data analyses has come at the expense of questions about the notion of *replication* as such. In this paper I hope to remedy this situation, highlighting in particular the implications of the fact that psychological experiments in general are infused with conceptual and material presuppositions. I will argue that once we gain a better understanding of what this entails with respect to replication, we get a deeper appreciation of philosophical issues that arise in the investigative practices of psychology. Among other things, I will show that replication is not as central to these practices as it is often made out to be.

The paper has three parts. In part 1 I will briefly review some philosophical arguments as to why there can be no exact replications and, hence, why attempts to replicate always involve individuation judgments. Part 2 will address a distinction that is currently being debated in the literature, i.e., that between direct and conceptual replication, highlighting problems and limitations of both. Part 3, finally, will argue that a significant part of experimental research in psychology is geared toward exploring the shape of specific phenomena or effects, and that the type of experimentation we encounter there is not well described as either direct or conceptual replication.

2. The Replication Crisis and the Ineliminability of Concepts

When scientists and philosophers talk about successfully replicating an experiment, they typically mean that they performed the same experimental operations/interventions. But what does it mean to perform "the same" operations as the ones performed by a previous experiment? With regard to this question, I take it to be trivially true that two experiments cannot be identical: At the very least, the time variable will differ. Replication can therefore at best aim for *similarity* (Shavit & Ellison 2017), as is also recognized by some authors in psychology. In this vein, Lynch et al (2015) write that "[e]xact

replication is impossible" (Lynch et al 2015, 2), arguing that at most advocates of direct replication can aim for is to get "as close as possible," i.e., to conduct an experiment that is similar to the previous one. In the literature, such experiments are also referred to as "direct replications." (e.g., Pashler & Harris 2912).[2]

The notion of similarity is, of course, also notoriously problematic (e.g., Goodman 1955), since any assertion of similarity between A and B has to specify with regard to what they are similar. In the context of experimentation, the relevant kinds of specifications already presuppose conceptual and material assumptions, many of which are not explicated, about the kinds of factors one is going to treat as relevant to the subject matter (see also Collins 1985, chapter 2). Such conceptual decisions will inform what one takes to be the "experimental result" down the line (Feest 2016). For example, If I am interested in whether listening to Mozart has a positive effect on children's IQ, I will design an experiment, which involves a piece by Mozart as the independent variable and the result of a standardized IQ-test at a later point. Now if I get an effect, and if I call it a Mozart effect, I am thereby assuming that the piece of music I used was causally responsible *qua being a piece by Mozart*. Moreover, when I claim that it's an effect on intelligence, I am assuming that the test I used at the end of the experiment *in fact measured intelligence*. These judgments rely on conceptual assumptions already built into the experiment qua choice of independent and dependent variables. In addition, I need *material assumptions* to the effect that potentially confounding variables have been controlled for. I take this example to show that whenever we investigate an effect *under a description*, we cannot avoid making conceptual assumptions when determining whether an experiment has succeeded or failed. This goes for original experiments as well as for replications.

[2] Both advocates and critics of direct replication sometimes contrast such replications with "conceptual" replications" (Lynch et al 2015). We will return to this distinction below.

One obvious rejoinder to this claim might be to say that replication attempts need not investigate effects under a description. They might simply imitate what the original experiment did, with no particular commitment to what is being manipulated or measured. But even if direct replications need not explicitly replicate effects under a description, I argue that they nonetheless have to make what Lena Soler calls "individuation judgments" (Soler 2011). For example, the judgment that experiment 2 is relevantly similar to experiment 1 involves the judgment that experiment 2 does not introduce any confounding factors that were absent in experiment 1. However, such judgments have to rely on some assumptions about what is relevant and what is irrelevant to the experiment, where these assumptions are often unstated auxiliaries. For example, I may (correctly or incorrectly) tacitly assume that temperature in the lab is irrelevant and hence ignore this variable in my replication attempt.

It is important to recognize that the individuation judgments made in experiments have a high degree of epistemic uncertainty. Specifically, I want to highlight what I call the problem of "conceptual scope," which arises from the question of how the respective independent and dependent variables are described. Take, for example, the above case where I play a specific piece by Mozart in a major key at a fast pace. A lot hangs on what I take to be the relevant feature of this stimulus: the fact that it's a piece by Mozart, the fact that it's in a major key, the fact that it's fast? etc. Depending on how I describe the stimulus, I might have different intuitions about possible confounders to pay attention to. For example, if I take the fact that a piece is by Mozart as the relevant feature of the independent variable, I might control for familiarity with Mozart. If I take the relevant feature to be the key, I might control for mood. Crucially, even though scientists make decisions on the basis of (implicit or explicit) assumptions about conceptual scope, their epistemic situation is typically such that they don't know what is the "correct" scope. This highlights a feature of psychological experiments that is rarely discussed in the literature about the replication crisis, i.e., the deep epistemic uncertainty and conceptual openness of much

research. This concerns both the initial and the replication study. Thus, concepts are ineliminable in experimental research, while at the same time being highly indeterminate.

3. Is the dichotomy between direct and less direct replication pragmatically useful?

One way of paraphrasing what was said above is that all experiments involve individuation judgments and that this concerns both original and replication studies. While this serves as a warning against a naïve reliance on direct (qua non-conceptual) replication, it might be objected that direct replications nonetheless make unique epistemic contributions. This is indeed claimed by advocates of both direct and less direct (="conceptual") replication alike. I will now evaluate claims that have aligned the distinction between direct and "conceptual" with some relevant distinctions in scientific practice, such as that between the aim of establishing the existence of a phenomenon and that of generalizing from such an existence claim on the one and that between reliability and validity on the other. I will argue that while these distinctions are heuristically useful, but on closer inspection bring to the fore exactly the epistemological issues just discussed.

3.1 Existence vs. Generalizability

Many scientists take it as given that there cannot be two identical experiments, but nonetheless argue that there is significant epistemic merit in trying to get *close enough*., i.e., to conduct direct replications. In turn, the notion of a direct replication is frequently contrasted with that of a "conceptual" replication. In a nutshell, direct replications essentially try to redo "the same" experiment (or at least something very close), whereas the conceptual replications try to operationalize the same question or concept/effect in a different way. The advantage of direct replications, as viewed by its advocates, is that by being able to redo an experiment faithfully and to create the same effect, one can show that the effect was real: "Exact and very close replications establish the basic existence and stability of a

phenomenon by falsifying the (null) hypothesis that observations simply reflect random noise" (LeBel et al, forthcoming, 7).

Advocates of conceptual replication don't deny this advantage of close replications, but hold that we want more than to establish that a given effect – created under very specific experimental conditions – is real. We want to know whether our findings about it can be generalized to: "When the goal is generalization, we argue that 'imperfect' conceptual replications that stretch the domain of research may be more useful" (Lynch et al 2015, 2). From a strictly Popperian perspective, the idea that non-falsification of the hypothesis of random error can provide proof of stability and existence is questionable , of course. But even if we abandon Popperian ideology here and take the falsification of H0 (that the initial effect was due to random error) to point to the truth of H1 (that there is a stable effect), the question is how to describe the effect. In other words, when claiming to have confirmed an effect, we have to say *what kind of effect* it is. And there we face the following dilemma:

a) Either we describe the effect as highly specific to very local experimental circumstances, involving the choice of a specific independent variable, delivered in a specific way etc.
b) Or we describe it in slightly broader terms, e.g., as a Mozart effect.

Advocates of direct replication might indeed endorse something like a), thereby exhibiting the kind of caution that motivated early operationists, in that no claim is made beyond the confines of a specific experiment. If, on the other hand, psychologists endorsed a description such as b), they would immediately run into the question of conceptual scope, i.e., the question *under what description* the independent variable can be said to have caused an effect. I argue that no amount of direct replication can answer this question, and hence, even if direct replication can confirm the existence of an effect, it cannot say what kind of effect. By asserting this, I am not saying that it's never useful to do a direct replication. My claim is merely that it will tell us relatively little. More pointedly: Direct replication can (perhaps) provide evidence for the existence of something, but it cannot say *existence of what*. Rolf

Zwaan makes a similar point when he states that "replication studies "tell us about the reliability of those findings. They don't tell us much about their validity." (Zwaan 2013).

In a similar vein, I argue that direct replication, with its narrow focus on ruling out random error, is epistemically unproductive, because it has nothing to say about *systematic error*. Systematic error arises if one erroneously attributes an effect to a specific feature of the experiment, when it is in fact due to another feature of the experiment. This can include, but is not limited to, the above-mentioned problem of conceptual scope. Fiedler et al. (2012) make a similar point when they argue that a narrow focus on falsification (with the aim of avoiding false positives) can be detrimental to the research process. Differently put, by privileging direct replication, we are not in a position to inquire about the kind of effect in question. This question, I argue, is best addressed by paying close attention to the possibility of systematic error, and hence by doing conceptual work. In other words, experimentally probing into systematic errors of conceptual scope is a valuable and productive part of the research process as it enables scientists to gradually explore what kind of effect (if any) they are looking at.[3]

3.2 Generality

I have argued that (a) scientists typically produce effects under a description and (b) that it can be epistemically productive to probe the scope of the description and to investigate the possibility of systematic error with regard to experiments that draw on such descriptions. It is epistemically productive, because it forces scientists to explore the nature and boundaries of the effect they are investigating. With this I have argued against a narrow focus on direct replication and I have cautioned against overstating the epistemic merits of such replication. But when we are concerned with effects

[3] I take this to be a contribution to arguments that philosophers of experimentation have made for a long time; e.g., Mayo 1996.

under a description, we are confronted with questions about the adequacy of the description. It is this question that advocates of "conceptual replication" claim to be able to address when they emphasize that their approach can deliver generality (over mere existence).

We have to distinguish between two notions of generality, namely (a) what kinds of descriptions one can generalize or infer to *within the experiment*, and (b) does the effect in question hold *outside the lab* (see Feest & Steinle 2016). These types of generality are also sometimes referred to as internal vs. external validity, respectively (Campbell & Stanley 1966; Guala 2012), where the former refers to the quality of inferences within an experiment and the latter refers to the quality of inferences from a lab to the world. The notion of generalizability raises questions about two kinds of validity. My focus here will be on internal validity, i.e., with the question of whether the effect generated in an experiment really exists as described by the scientist.[4]

Internal validity can fail to hold because of epistemic uncertainties regarding confounding variables both internal and external to experimental subjects. For example, prior musical training might make a difference to how one responds to Mozart music, but the experimenter may not have taken this into consideration in their design. But internal validity can also fail to hold is by virtue of what I have referred to as the problem of conceptual scope (for example, we may refer to the effect as a Mozart effect when it is in fact a Major-key effect). Effectively, when I treat a major-key effect as a Mozart effect, I have misidentified the relevant causal feature of the stimulus. In turn, this means that I will neglect to control for major/minor key as I will regard this as irrelevant, which can result in systematic errors. In both cases, scientists can go wrong in their individuation judgment. What is at stake is not whether there is an effect, but what kind of effect it is. Now, given that those kinds of problems can

[4] In this respect I differ from some advocates of conceptual replication, who have highlighted external validity as a desideratum (E.g., Lynch 1982, 3/4).

occur, we turn to the question of whether "conceptual replication" has an answer. I will now argue that it does not.

To explain this, let me return to the above characterization of conceptual replication, according to which such replication consists in repeating an experiment, using different operationalizations of the same construct. For example, a conceptual replication of an experiment about the Mozart effect might operationalize the concept Mozart effect differently by using a different piece of Mozart music and/or a different measure of spatial reasoning. But there is a major caveat here: If I want to compare the results of two experiments that operationalized the same construct differently, I already have to presuppose that both operationalizations in fact have the same conceptual scope, i.e., that they in fact individuate the same effect. But this would be begging the question, since after all – given the epistemic uncertainty and conceptual openness highlighted above – that's precisely what's at issue. Differently put, experiment 2 might or might not achieve the same result as experiment 1, but the reason for this would be underdetermined by the experimental data. Thus, the problem of conceptual scope prevents us from being able to say whether we have succeeded in our conceptual replication.

Given the uncertainties as to whether one has in fact succeeded in conceptually replicating a given experiment, I am weary of the language of replication here. If anything, I would argue that the method in question should be regarded as a research strategy that is aimed at helping to demarcate and explore the very subject matter under investigation. But as I will argue now, this is perhaps better described as exploration, not as replication.

4. Putting Replication in its Proper Place

The conclusion of the previous paragraphs seems pretty bleak: Direct replication is either extremely narrow in what it can deliver or it runs into the joint problems of confounders and conceptual scope. Conceptual replication, on the other hand, cannot come to the rescue, because it also runs into the

exact same problems. Should we then throw up our hands and conclude that since ultimately neither direct nor conceptual replication are possible the crisis of replication is much more severe than we previously thought? This would be the wrong conclusion, however. This would only follow if replication was in fact as central to research as it is sometimes taken to be. I claim that it is not. My argument for these claims has three parts. The first part holds that exploring (the possibility of) systematic errors is an important part of the investigative process, which is not well described as replication. Second, if we take seriously this process of exploring and delineating the relevant phenomena, we find that there is indeed a great deal of uncertainty in psychological research, but this, in and of itself, does not necessarily constitute a crisis. Lastly, while it is fair to say that there is a crisis of confidence in current psychology, it is not well described as a replication crisis.

Let me begin with the first point. I have argued that direct replication (even where it is successful) is of limited value, because it can at most rule out random error, but completely fails to be able to address systematic error. But if we appreciate (as I have argued we should) that direct replication inevitably involves individuation judgments, it is obvious that there is always a danger of systematic error, because I have to assume that all confounding variables have been controlled for. One important class of confounders follows from what I have referred to as the problem of conceptual scope, i.e., the difficulty of correctly describing both the independent variable responsible for a given effect and the dependent variable.[5] Epistemically productive experimental work, I claim, therefore needs to focus on systematic errors, specifically those brought about by unstated auxiliary assumptions.

Indeed, if we look at the story of the Mozart effect, we find that this is exactly what happened. This example also nicely illustrates my claim about the conceptual openness and epistemic uncertainty

[5] My focus here has been mainly on the former. But of course the problem of conceptual scope concerns both.

in many areas of experimental psychology. The Mozart effect was first posited by Rauscher and colleagues (Rauscher et a. 1993). It can now be regarded as largely debunked. While it is true that several people tried (and failed) to replicate the effect (e.g., Newman et al. 1995; Steele 1997), it is important to look at the details here. It is not the case that the effect was simply abandoned for lack of replicability. Rather, when we look at the back and forth between Rauscher and her critics, we find that the discussion turned on the choices and interpretations of independent and dependent variables. In this vein, Newman et al (1995) and Steele (1997) used different dependent variables, prompting Rauscher to argue that her effect was more narrowly confined to the kind of spatial reasoning measured by the Stanford-Binet. I suggest that we interpret this case as one where Rauscher was forced to confront (and retract) an unstated auxiliary assumption of her initial study, namely that the spatial reasoning subtest of the Stanford-Binet (which she had used as her dependent variable), was representative of spatial reasoning more generally. Likewise, her choice of the Mozart's Sonata for Two Pianos in D-major as the independent variable was put under considerable pressure by critics, who suggested that the relevant feature of the independent variable was not that it was a piece by Mozart, but that it was up-beat and put subjects in a good mood (Chabris 1999). My point here is that the debates surrounding the Mozart effect are best described as conceptual work, exploring consequences of possible errors that might have arisen from the problem of conceptual scope. At issue, I claim, was not primarily whether Rauscher really found an effect, but rather what was the scope of the effect.

I argue that this is a typical case. Rather than, or in addition to, attempting to conduct direct replications of previous experiments, researchers critically probed some hidden assumptions built into the design and interpretation of the initial experiment. My point here is both descriptive and normative. Thus, I argue that this is a productive way to proceed. However, I claim that it is not well described as replication, let alone conceptual replication. Rather, what we see here is a case in which scientists explore the empirical contours of a purported effect in the face of a high degree of epistemic

uncertainty and conceptual openness, and this is precisely why the case is not well described as employing conceptual replication. The reason for this is quite simple: For a conceptual replication to occur, one needs to already be in the possession of some well-formed concepts, such that they can be operationalized in different ways. It also presupposes that in general the domain is well-understood, such that operationalizations can be implemented and confounding variables can be controlled. But this completely misses the point that researchers often investigate effects precisely because they don't have a good understanding (and hence concept) of what it is.

Therefore I argue that while direct replication can only contribute a very small part to the research process, conceptual replication cannot make up for the shortcomings of direct replication. Instead, productive research should (and frequently does) proceed by exploring, and experimentally testing, hypotheses about possible systematic errors in experiment. Such research, I suggest, can contribute to conceptual development by helping to explore and fine-tune the shape and scope of proposed or existing concepts. The fact that this is riddled with problems does not in and of itself constitute a crisis, let alone a replication crisis.

5. Conclusion

The upshot of the above is that when we talk about the importance of replication, we need to be clear on what we mean by replication and why it is so important, precisely.

In this paper I have argued that if by replication we mean either "direct" or "conceptual" replication, we need to first of all be clear that direct replications are not non-conceptual. I then turned to some alleged epistemic merits of direct replication, for example that they can establish the existence of effects or the reliability of procedures that detect effects. I argued that insofar as such replications involve concepts, they run (among other things) into the problem of conceptual scope, i.e., the difficulty of determining, on the basis of independent and dependent variables of experiments what precisely is

the scope of the effect one is trying to replicate. I highlighted that this is a real and pernicious problem in experimental research in psychology, due to the high degree of epistemic uncertainty and conceptual openness of many fields of research.

While my emphasis of the conceptual nature of replication may suggest that I would be more favorably inclined toward conceptual replication, I have argued that conceptual replication runs into the same problems, and for similar reasons: The very judgement that one has successfully performed a conceptual replication of a previous experiment presupposes what is ultimately the aim of the research, namely to arrive at a robust understanding of the relevant area of research. This, I argue that since conceptual replication presupposes a relatively good grasp of the relevant concepts, it is begging the question, and I suggested instead that researchers (should) engage in a process of specifically investigating possible systematic errors in original studies as a means to develop the relevant concepts. This process is not best described as one of replication, however. Summing up, then, I conclude that in general, replications are less useful and important than is widely assumed – at least in the kind of psychological research I have focused on in this paper.

Now, in conclusion let me return to the notion of a crisis in psychology as it is currently discussed in the literature. Obviously, I do not mean to deny that there is a crisis of confidence in (social) psychology (Earp & Trafimov 2015) as well as in other areas of study. However, based on the analysis provided in this paper, I argue that this crisis is not well described as a crisis of replication. Rather, it seems to be to a large degree a crisis that turns on questionable research practices with regard to the use of statistical methods in psychology (see Gelman & Loken 2014). While acknowledging the valuable philosophical and scientific work that is being done in this area, I suggest that a broader focus on the notion of replication provides us with a deeper appreciation of the conceptual dynamics characteristic of experimental practice.

REFERENCES

Campbell, D. T., and Stanley, J. C. (1966), Experimental and Quasi-Experimental Designs for Research (Chicago: Rand McNally).

Chabris, C. (1999): Prelude or requiem for the 'Mozart Effect?' "Scientific Correspondence", Nature, 4000, 826.

Collins, H. (1985). Changing order. Replication and induction in scientific practice. Chicago and London: The University of Chicago Press.

Earp, Brian & Trafimow, David (2015): "Replication, falsification, and the crisis of confidence in social psychology." Front. Psychol, 19 May 2015 | https://doi.org/10.3389/fpsyg.2015.00621

Feest, U., 2016, "The Experimenters' Regress Reconsidered: Tacit Knowledge, Skepticism, and the Dynamics of Knowledge Generation". Studies in History and Philosophy of Science, Part A 58 34-45.

Feest, U. & Steinle, F., 2016, "Experiment." In P. Humphreys (Ed.): Oxford Handbook of Philosophy of Science. Oxford University Press, 274–295.

Fiedler, K.; Kutzner, F. & Krueger, J. (2012): „The Long Way from alpha-error control to validity proper: Problems with a short-sighted false-positive debate." Perspectives on Psychological Science 7(6), 661-669

Gelman, Andrew & Loken, Eric (2014): The Statistical Crisis in Science. Data-dependent analysis—a "garden of forking paths"— explains why many statistically significant comparisons don't hold up. American Scientist 102 (6) 460-464. DOI: 10.1511/2014.111.460

Goodman, Nelson (1983/1955): Fact. Fiction and Forecast. Harvard University Press; 4 Revised edition edition

Guala, F. (2012), "Philosophy of Experimental Economics." In U. Mäki (ed.), Handbook of the philosophy of science. Vol. 13: Philosophy of Economics (Boston: Elsevier/Academic Press), 597–640

LeBel, E.P.; Berger, D., Campbell, L.; Loving, T. (2017): "Falisifiability is not Optional." Journal of Personality and Social Psychology (forthcoming)

Lynch, J. (1982): "On the External Validity of Experiments in Consumer Research. Journal of Consumer Research 9, 225-239. (December)

Lynch, J.; Bradlow, E.; Huber, J.; Lehmann, D. (2015): "Reflections on the replication corner: In praise of conceptual replication." IJRM ???

Mayo, Deborah (1996): Error and the Growth of Experimental Knowledge. University of Chicago Press.

Newman, J., Rosenbach, J., Burns, K.; Latimer, B., Matocha, H., Vogt, E. (1995: An experimental test of the 'Mozart Effect': Does listening to Mozart improve spatial ability? Perceptual and Motor Skills, 81, 1379-1387.

Open Science Collaboration (2015) Estimating the reproducibility of psychological science. Science 349

Pashler, Harold & Harris, Christine (2012): "Is the Replication Crisis Overblown?" Perspectives on Psychological Science 7(6), 531-536.

Rauscher, F., Shaw, G.; Ky, K. (1993). Music and spatial task performance. Nature ,365, 611.

Romero, Felipe (2017): "Novelty vs. Replicability. Virtues and Vices in the Reward System of Science." Philosophy of Science.

Shavit, Ayelet & Ellison, Aaron (eds.) (2017): Stepping in the Same River Twice. Replication in Biological Research. Yale University Press

Soler, Lena (2011): "Tacit Elements of Experimental Practices: analytical tools and epistemological consequences." European Journal for Philosophy of Science 1, 393-433.

Steele, K., (2000). Arousal and mood factors in the 'Mozart effect'. Perceptual and Motor Skills, 91, 188-190.

Zwaan, Rolf (2013): "How Valid are our Replication Attempts?"

https://rolfzwaan.blogspot.de/2013/06/how-valid-are-our-replication-attempts.html