A New Account of Replication in the Experimental Life Sciences

Stephan Guttinger

-Egenis, Centre for the Study of Life Sciences, University of Exeter, Byrne House, St German's Road, Exeter, EX4 4PJ, UK

-CPNSS, London School of Economics, Lakatos Building, Houghton Street, London, WC2A 2AE, UK

E-mail: s.m.guettinger@lse.ac.uk

Abstract:

The natural sciences are in the midst of a reproducibility crisis. Scientists don't replicate existing data and when they attempt to do so they often fail. Nevertheless, survey data shows that scientists largely trust the non-replicated data they are using. The question of why this is so has not been raised in the debate about the reproducibility crisis. Here I will claim that one reason for this trust is a hitherto unidentified form of replication, which I will call 'micro-replications' (MRs). Using a case study from the experimental life sciences I will illustrate how MRs depend on a crucial part of the experimental sciences that is poorly understood, namely experimental controls. The existence of MRs suggests that more replication is taking place in the life sciences than current analyses imply.

1. Introduction

It is widely accepted among both scientists and philosophers of science that the replication of previous experiments is a key element of the scientific process. Experiments are replicated to confirm earlier findings (Collins 1985) and to ensure the reliability and/or robustness of experimental output (Soler et al. 2012).

However, two problems have been raised regarding replication in science: first, whilst in theory replication might be crucial, in practice scientists rarely reproduce existing data (this phenomenon has already been observed more than 30 years ago by Harry Collins (1985)).[1] Second, when scientists actually make the effort to replicate previous results they are often unable to do so. This has led to talk of a 'reproducibility crisis' in science (Prinz et al. 2011; Begley and Ellis 2012; Mobley et al. 2013). This crisis is thought to not only affect the natural sciences but also fields such as psychology (Makel et al. 2012).

Whilst the lack of reproducible data certainly is an alarming state of affairs, it is at the same time fascinating to note that scientists –in particular in the life sciences, on which this paper will focus – have been marching on for decades without being fazed by

[1] Note that I will use the terms 'reproduction' and 'replication' interchangeably here even though there are debates on potential differences between the two terms (see, e.g., (Drummond 2009) or (Casadeval and Fang 2010)).

the obvious lack of attempts to replicate data.[2] Moreover, not only are scientists continuing their experimental work in the absence of reproductions, survey data also indicates that they still trust the (non-replicated) data they are using (Baker 2016b).

This raises a series of fascinating and important questions that are not addressed in the current debate on the reproducibility crisis: Where does this basic trust in the work of others come from? Is this a case of blind trust and ignorance or is there more to it? And if the latter is the case then what is it that makes scientists so confident in non-replicated data?

Here I will claim that this trust is built on a type of replication that current analyses of the problem have overlooked, namely what I will call 'micro-replications' (MRs). This type of replication has several key features (which also explain why MRs have so far evaded the attention of those involved in the debate). First, MRs are not done as experiments whose main goal is to recreate (in some form or other) existing results. Rather, they are part and parcel of everyday experimental practice. This goes against the current consensus in the scientific community that replications are an add-on

[2] Those sounding the alarms in recent years were mainly researchers at industry giants who heavily depend on findings from pre-clinical research performed in academia (Baker 2016a). Life scientists working at universities have now joined the discussion and the question of replication has turned into its own field of investigation with several studies under way to establish the extent of the problem (see, e.g., Errington et al. 2014; Open Science Collaboration 2015; Nosek and Errington 2017).

to regular experimentation (discussed in section 2). Second, MRs rely on elements of the experimental process that are not well understood, namely experimental controls. As I will show in section 3 using the case study of the in vitro binding assay, controls do more than just check for artefacts. I will introduce a distinction between an intra- and an inter-experimental role of controls to highlight that controls can also form inter-experimental links by embodying elements from previous experiments in new experimental settings (section 4). As I will show using my case study, these links are not only crucial for the researcher to have an interpretable experimental output but they also serve as built-in replications (MRs) of earlier findings.

Identifying these MRs is important as it changes the debate about the reproducibility crisis (section 5). The dominant view of replication as an add-on can give the appearance that there is hardly any replication happening in the life sciences. But once we have a more in-depth understanding of the experimental process and the role controls play in it we see that – because of MRs – more (successful) replication is taking place in everyday research than is generally assumed. This additional level of replication also explains (in part at least) the trust researchers put in existing data.

2. Replications as Add-ons

Even though there is little consensus in the literature (both within the sciences and philosophy of science) on the exact definition of replications there is a clear consensus that replications are rarely performed in the sciences (Baker 2016b; Goodman et al.

2016). This point has already been made by Harry Collins in the 1980s (Collins 1985) and has been corroborated by more recent studies (see, e.g., (Makel et al. 2012)).

The low amount of actual replication attempts in the experimental sciences is baffling if we take into account the widely reported problems with irreproducible data: if indeed 50% to 80% of existing data cannot be replicated, as some claim (Begley and Ellis 2012; Vasilevsky et al. 2013), then shouldn't the first priority of any researcher be to test all the data she is building on?

In the debate within the sciences it is usually assumed that replications are not performed because there is no reward to be had from doing so: replications are not only expensive and time-consuming but they usually also don't translate into high-impact publications (if they can be published at all). A large part of the debate has therefore focused on how the incentive structures of science could be changed (see, e.g., Alberts et al. 2014; Begley et al. 2015; Begley and Ioannidis 2015; Sarewitz 2016a; 2016b).

But whilst it is certainly the case that the current incentive structures are highly problematic and do not encourage the replication of earlier work, the problem with the above explanation is that it cannot make sense of the trust researchers clearly put in existing data. Not only are researchers continuing to build on existing data without replicating it first, survey data also confirms that they largely trust the data they are using (Baker 2016b). This raises the intriguing question of where this trust comes from.

A key assumption that underlies the whole debate about the reproducibility crisis – and which is rarely questioned – is the idea that replications are something that has to be done *on top* of what researchers normally do. Replications are seen as add-ons to

regular experimental practice that cost money and time. In this framework, it is little surprise that scientists don't perform replications.

But what if replications, these creators of trust, come in forms that the current analytic frameworks don't account for? In his book 'Failure. Why Science is so Successful' the biologist Stuart Firestein makes an interesting observation about the replication of positive results in science (Firestein 2015). First, Firestein states that scientists are right to refrain from reproducing existing results as this would be a waste of time and money. But in a somewhat contradictory twist he then also claims that researchers *do* replicate existing work, namely by building on previous work. As Firestein puts it: "[E]xperiments get replicated because people from other labs use the published results and the methods in their own experiments" (ibid, p. 151).

There is a different sense of replication implied (but not explained) here: somehow replication is part of everyday science as researchers 'use' previous results and methods in their own work. The challenge, then, is to explain what such re-use looks like. What exactly does it mean to use existing data *in* an experiment? How exactly is existing data incorporated, thus becoming part of the current setup? And how is this related to replication and trust?

To find answers to these questions it will be helpful to compare a case of experimental work that builds on previous knowledge with a case where there is no or little use of existing data. Such a comparative study will allow us to identify factors that are present, or more prominent, when researcher build on previous work. These factors

might then give us a better picture of what this form of replication could look like and how it might work.

Interestingly, philosophers of science have identified two types of experimentation that differ in the extent to which they build on existing knowledge, namely 'exploratory experimentation' (EE) and 'theory-driven experimentation' (TDE) (Steinle 1997).[3] In the case of the former there is usually little information available on the system or phenomenon of interest and researchers have very little or nothing to build their new experiments on. In the case of TDE, there is usually a wealth of previous knowledge available that is used to inform the setup, execution and interpretation of the experiment.

These different forms of practice can be of great help to an investigation into experimentation and replication: if we have an experimental system that can be used for both EE and TDE, an analysis of how it is used in practice should allow us to identify how researchers replicate existing results by using them in their own experiments, as Firestein suggests.

In the next section, I will turn to a case study that allows us to do exactly that, namely the so-called 'in vitro binding assay'. This experimental system is widely used in the life sciences to study protein-protein interactions and can be used for both EE and TDE. The analysis of the different applications of the assay will identify a somewhat

[3] On the topic of EE and TDE see also (Burian 1997; Steinle 2002; Franklin 2005; Elliott 2007; O'Malley 2007; Burian 2007; Waters 2007; Karaca 2013).

surprising element that is crucial for the re-use and hence the replication of existing experimental data, namely experimental controls.

3. The in vitro Binding Assay

Proteins are key players in almost all biological systems as they fulfil a variety of roles, such as signal propagation, structural support or the catalysis of chemical reactions. In order to fulfil these roles proteins must not only be able to interact with other elements of the cell (such as DNA molecules or lipids) but also with each other. The analysis of protein-protein interactions is therefore a central part of the research conducted in the molecular life sciences.

To perform interaction studies scientists make use of the fact that proteins can be extracted from cells, either in a purified form or as part of a whole-cell extract (i.e. an extract of all the soluble proteins of a particular cell type). These isolated proteins or protein mixtures can then be used to study protein-protein interactions in vitro. One of the key assays used for this purpose is the so-called in vitro binding assay.

3.1. The General Setup of the in vitro Binding Assay

The basic idea behind the in vitro protein binding assay is relatively simple: a protein of interest is isolated from its original cellular context and incubated in a test tube with another protein (or a mixture of proteins) in a suitable buffer solution. This incubation period (usually in the range of one to several hours) allows for the formation of protein-protein complexes. After incubation, the protein of interest is retrieved from the reaction

mixture using a specific retrieval system (see next paragraph). If any of the other proteins present in the reaction mixture are able to bind to the protein of interest they will be co-retrieved with the protein of interest and can subsequently be identified.

A modified version of the protein of interest has to be used in this assay in order to be able to retrieve it from the reaction mixture. The modification usually consists of what is referred to as a 'tag', often a short polypeptide that is fused to one end of the protein of interest. The tag has a specific binding target (either a small molecule or another polypeptide), which can be chemically coupled to synthetic microbeads. The modification of the beads with a target and of the protein of interest with a tag provides the researcher with a powerful and specific retrieval system: adding the modified beads to the reaction mixture will result in the recruitment of the tagged protein of interest (and everything that is bound to it). The beads can then be separated from the reaction mixture by centrifugation and, following a washing step, all proteins bound to them can be eluted using high salt or denaturing conditions (which interrupt regular protein-protein interactions). These eluted proteins can then be analysed by gel electrophoresis[4] coupled to Western blot analysis or mass spectrometry, two of the main methods used in molecular biology to identify specific proteins.

[4] Gel electrophoresis allows to separate proteins according to their size. Proteins of different size will appear on the gel as distinct bands.

3.2. Using the in vitro Binding Assay for Exploratory Purposes: Mapping Protein Interactions

A key application of the in vitro binding assay is to map the interaction space of a protein X. Such mapping usually represents a form of exploratory research, in particular if there is no data available on the interaction partners of X and if there are no known binding domains or signal peptides present in X. In such a case the researcher is unlikely to have a clear idea about the possible intracellular interactions X might engage in. An in vitro binding assay using tagged X and a cell extract can be used to screen for potential interaction partners of X.

The exploratory use of the in vitro binding assay has several characteristic features. The readout of the mapping experiment will, for instance, consist of a general detection of proteins of all sizes using gel electrophoresis and/or mass spectrometry as the point of the experiment is to explore the whole space of possible protein-protein interactions for factor X. There is therefore no restriction on what proteins the researchers are looking for.

The openness of the mapping experiment is also reflected in the variation of parameters that the researchers are likely to make use of. They might, for instance, use a range of different cell extracts derived from different cell types or organisms to explore a protein space that is as large as possible. Other parameters that the researchers might alter are the salt concentration or the pH of the buffer(s) used (as these parameters can directly affect the ability of proteins to interact with each other) or also the duration of the incubation period.

This variation of parameters and the openness of the readout are needed because the exploratory in vitro binding assay does not build in any strong way on existing data; there simply is very little specific information that could inform the setup, execution or interpretation of this exploratory assay.

3.3. Using the in vitro Binding Assay for Guided Experimentation

Besides the exploratory setup the in vitro binding assay can also be used to test hypotheses about the interaction between two particular proteins. This is a case of guided experimentation, meaning it builds directly on existing data (which formed the basis for the hypothesis being tested).

To illustrate this application of the assay I will use the following example: assume a) that researchers have previously identified two proteins X and Y which form a stable complex and b) that X contains a signal peptide known to mediate binding to proteins of class 'Z'. Further assume c) that Y is a member of Z. The presence of the signal peptide in X would imply that X and Y can interact directly with each other (hypothesis 1) and that this interaction is mediated by the signal peptide (hypothesis 2). Both of these hypotheses could be tested using the in vitro binding assay.

To test hypothesis 1, the researcher would isolate both X and Y and use them in a binding assay (with either of them modified with a tag) to check whether retrieving one protein from the reaction mixture will co-retrieve the other. As both proteins have been isolated from their cellular context the researcher can assume that there are no other

proteins present in the reaction mixture. Therefore, if an interaction is observed it can be concluded that the interaction is direct and not mediated by another factor.

To test hypothesis 2, the researcher would not only have to test the direct interaction between X and Y but also check for an interaction between the two proteins in the absence of a functional signal peptide in X. One way to create such a context would be to remove the signal peptide altogether, for instance by creating a mutant of X that lacks the signal peptide. If this mutant form of X does not show any binding to Y whilst the full-length version of X does, hypothesis 2 would be supported.

In contrast to the exploratory use of the assay the readout of the guided experiment would focus exclusively on the specific detection X and Y, as it is only these two factors the researcher is interested in. This also means that the researchers are unlikely to engage in an extensive variation of experimental parameters as they know what they are looking for (and how to look for it). They would simply use the settings that have worked before when X and Y were found to form a stable complex. All these different features are in line with what Steinle describes as guided experimentation or TDE (Steinle 1997).

3.4. Artefacts and Controls

An important issue that affects both the exploratory and guided uses of the in vitro binding assay is the possibility of artefacts. This is a crucial issue as artefacts negatively affect the trust a researcher can put in the results obtained.

A key factor in this context is that proteins can, in principle at least, interact with a great range of surfaces. Depending on parameters such as pH, temperature, and salt concentration the protein will display particular features on its surface (such as charged or hydrophobic patches). These features will allow the protein to interact with any matching surface, including that of synthetic beads.

This is a problem as everything that is bound to the beads after the retrieval and washing steps will be defined as a potential interaction partner of the protein of interest. The researcher therefore needs to be able to identify such unspecific binding events (often referred to as 'background binding'). If there is no system in place to do so the researcher will not be able to judge whether the marks on the gel represent true binding events or whether the experimental system is misfiring, i.e. producing false positives. To exclude such artefacts the researcher will therefore usually include a negative control in the experiment.

3.4.1. The Negative Control

In an in vitro binding assay there are three potential sources of background binding: 1) the surface of the beads, 2) the target with which the beads are modified, and 3) the tag that is fused to the protein of interest. The proteins present in the reaction mixture could bind to any of these sites.

To control for all three sources of background binding the researcher will prepare a separate sample that a) consists of beads that are b) modified with a target and c) pre-loaded with the tag that was used to modify the protein of interest. The only difference

between this sample and the others used in the assay is the absence of protein X (as only an empty tag is used). This control can be used to exclude background binding as any signal that appears in this sample cannot be due to the presence of X. Any signal that is equally strong in the negative control and the actual sample will therefore be classified as a false positive. If the signal appears in both the negative control and the sample containing X but is stronger in the latter this indicates that there could be a real interaction taking place (as the signal is above background binding). This illustrates another important role controls can play, namely as calibration devices that set the baseline signal of the retrieval system (Grinnell 1992).

3.4.2. The Positive Control

Performing an in vitro binding assay means to manipulate the protein of interest (as it has to be modified, isolated and then immobilised on the beads). All of these interventions risk deactivating the protein of interest, as changes in salt concentration, pH or temperature can lead to the unfolding or lysis (disintegration) of its polypeptide chain. If this happens the basic setup of the assay becomes faulty and it might no longer be able to produce positive results. If this fault is not detected the system could produce false negative results.

To exclude such false negatives the researcher will include a positive control which verifies that the protein of interest is active under the conditions chosen (Baker and Dunbar 2000). The positive control will usually contain a known binding partner of the protein of interest that is tested in parallel to the other samples of the binding assay.

By including this control the researcher will be able to interpret negative results: if the positive control shows an interaction with factor X but all the other samples don't show any interaction, the researcher knows that she is dealing with a true negative result. If the positive control does not show any signal she knows that factor X has become inactivated at some point and that negative results might be an artefact.[5]

As in the case of negative controls, the positive control has to do with the interpretation of the marks obtained in the experiment: if the positive control is missing or not working the researcher cannot exclude that negative results are due to the inactivity of the protein of interest, meaning she will not be able to obtain an interpretable readout.

The positive control can also be used as a calibration device. If, for instance, different mutants of an enzyme are tested for activity (and if it is known that the full-length protein is active), then the signal provided by the full-length sample can serve as a measuring stick for the other samples and give the researcher an idea of the signal strength that can potentially be reached under the conditions used (Grinnell 1992).

In summary, we see that both the negative and positive controls 1) serve as calibration devices and 2) can be used to exclude artefacts. Controls allow the researcher to put

[5] Note that in this case the researcher will also perform a positive control on the positive control to make sure it is not the source of the problem. Controls ultimately only work as part of a complex network, a point I will return to in section 4.

trust in the system they are using, the manipulations they are performing and the results they obtain. Because of this they help to obtain a meaningful, i.e. interpretable output of the experiment. Without controls the researcher cannot read the marks she obtains.

3.5. The Intra-Experimental Role of Controls

In section 3.4.1 we have seen how the negative control is used to separate the bands that appear on a gel into meaningful sets: by having a negative control that was performed in parallel to the other samples (and which is analysed as part of the same gel) the researcher is able to partition the bands visible on the gel into two classes (‘potential interactors’ and ‘background binding’).

This means that an initial interpretation of the raw data provided on the gel (all the bands that appear) is done *in situ* when looking at the gel, comparing the different lanes with each other. Crucially, the controls serve as an ‘other’, i.e. as a difference maker (not in a causal but a semiotic sense); only by including a negative control is it possible for the researcher to create sets of marks that can be compared in a fruitful manner, i.e. to have a meaningful readout for the experiment. Its presence creates the context in which researchers can talk about facts and artefacts.

This particular use of the negative control is an example of what I will refer to as the *intra*-experimental mode in which controls can function: by creating a crucial difference between the samples of the same experiment the use of a negative control opens up a space in which meaningful output can be created. This space is created

through the juxtaposition of two samples that have been processed in parallel and which are present on the same output (a gel in this case).[6]

A positive control can play a similar intra-experimental role as it is again the differential space it creates within the *same* experiment that is important for its function. A sample in which no bands become visible (for instance in the above-described assay that looks at the interaction between X and Y) can be compared to the positive control (which, if it works, confirms that both X and Y are active under the conditions chosen). This comparison between the marks obtained for each sample confirms that all the factors involved are in principle active and allow the researcher to make a reliable statement about the interaction (or absence of interaction) between X and Y.

This intra-experimental use of controls, which can be part of both guided and unguided experiments, corresponds to the more traditional role of controls, i.e. their function to check for artefacts. However, as I will explain in the next section, the examples discussed here allow us to identify an additional mode in which controls can function, which I will refer to as the *inter*-experimental role of controls. This mode, I

[6] If the controls were loaded and analysed on different gels the comparison that is essential to the use of controls would no longer work. If, for instance, there were differences in the intensity of the signals obtained the researcher could not exclude that the two gels display a different staining behaviour, which could mean that one shows a weaker signal than the other even though the same amount of protein is present.

claim, is a crucial part of what makes controls tools for establishing trust when building on the work of others.

4. Building on Previous Work

The two setups of the in vitro binding assay described in section 3 are ideal in the sense that the assay, as used in daily practice, will contain adjustments to account for particularities of the proteins of interest or the specific question that is being addressed. Nevertheless, the examples help to illustrate some of the key elements that are required to make use of the assay for both guided and unguided uses. Interestingly, even though basic positive and negative controls are used in both cases, there are crucial differences in how the controls are employed in each case. These differences come to the fore when we look at the inter-experimental use of controls.

4.1. The Inter-Experimental Role of Controls

The above description of the guided experiment has highlighted several ways in which the researchers might make use of existing knowledge about the entities and processes analysed. They already know, for instance, the sequence and the behaviour of the signal peptide in X ('The type of signal peptide present in X mediates the interaction with proteins of class Z'). They also have information about the behaviour of X and Y as they know that these two proteins form a stable complex with each other (see the starting assumptions made in section 3.3). It is this and other previously established knowledge that lead to the formulation of the two hypotheses that are tested, namely

that proteins X and Y interact directly and that they do so via the signal peptide present in X.

This knowledge is the result of specific experiments and sequence analyses that have gone before: the sequence of the signal peptide will have been defined using functional assays performed with one or several other proteins containing that specific peptide. In the course of such experiments it will also have turned out that the peptide mediates the direct interaction with proteins of class Z. This knowledge is therefore the outcome of particular experiments that have been performed earlier and/or elsewhere using the same class of proteins that is also used in the current experiment. This knowledge not only guides the questions being asked but also informs the setup and the execution of the assay.

This can also be seen in the way controls are being used. If we compare the positive controls used in the guided and unguided experiments described in section 3 we discover interesting differences. For instance, if a positive control is used at all in the unguided case it will be a random protein, in the sense that any protein that is known to interact with X can be used to verify that X is active. This also means that the experimental conditions used for the positive control (e.g. pH, salt concentration, etc.) are not necessarily binding for the actual exploration performed – other proteins might require very different conditions in order to interact with protein X and the researcher might therefore use a range of salt concentrations and different pH values.

The guided experiment, however, is building on specific experimental findings and specific events happening between two known factors. The controls used therefore

have to be specific as well: the point is not simply to show that factors X and Y are active but that they are capable of undergoing the activities that have been ascribed to them in earlier experiments. Factor X, for instance, has to be able to bind to proteins of class Z (to which factor Y belongs). The aim is to show that the signal peptide in X is accessible and hence functional, as it was found to be in past experiments. To show this the researcher will have to reproduce this past event (the same has to be done for Y, i.e. it has to be shown that Y can, in principle at least, bind to signal peptide-containing proteins).

In the guided experiment the positive controls will therefore consist of a specific protein belonging to class Z (controlling for the activity of X) and a protein that contains a signal peptide (controlling for the activity of Y). Specific positive controls are used because it is a particular type of event that needs to be verified in order for the researcher to trust the output of the experiment. This also means that the experimental conditions used will have to be the same as those used for the positive control (and by extension that of the previous experiments), since the positive control is of the same class as the proteins analysed and all samples have to be directly comparable.

The controls therefore create a close link with previous experiments, meaning they establish a continuity between the experiment at hand and the earlier work on which it builds. With this continuity also come expectations, experimental conditions and trust. So in addition to the intra-experimental role described above there is also an inter-experimental role controls can play, at least in the experimental life sciences.

4.3. Creating Trust through Experimentation

The inter-experimental mode of controls is significant in the context of this paper because it entails the replication of earlier results. Particular previous experiments are brought into the experiment at hand through the inter-experimental controls used. The previous results are re-produced in control samples and are part of what makes the data of the current experiment readable and trustworthy. Only if such a local network with guiding and interpretative power is established do researchers have a well-defined experimental outcome to work with.

We start to see here, then, a possible interpretation of the statement Firestein made, namely that scientists reproduce earlier experiments by using them in their own experiments. Replications of earlier results happen as part of regular experiments and not just in what is explicitly designed and labelled as a replication of earlier results. These replications-via-controls don't attempt to repeat a whole study or figure from earlier work. They rather pick one aspect that is crucial in guiding the experiment at hand and make it part of the current setup to establish readability and trustworthiness (the two being intertwined). Because of their small-scale character I will refer to these replications as 'micro-replications' (MRs).

Importantly, scientists not only *use* MRs as part of their regular experimentation but they are also able to *read* them when they encounter work by others. They know when controls are missing and this will often make them question the data they are presented with. Scientists are likely to ignore data that is poorly controlled or to set out to repeat it in their own laboratory to see for themselves.

Identifying MRs is not only important for practising scientists but also for those analysing the reproducibility crisis. A lot has been made of recent claims that up to 80% of existing data cannot be reproduced. However, not only are these claims based on small-scale studies (a problem current projects in meta-science aim to address, see footnote 2) but they are also based on a particular picture of what replications are and how they work.

Once we realize that (micro-)replications happen as part of normal experimentation, the picture starts to change. What the analysis provided here suggests is that scientists do more (successful) replications than current analyses are able to identify. Because of the controls they use scientists not only put trust in the output of their own experiments but they also build trust in the data published by others. The presence of MRs explains (in part at least) why researchers have been so confident in pushing ahead without setting up replication experiments, as defined by existing accounts.

5. Conclusions: Trust and the Dark Matter of the Experimental Sciences

There are (at least) two questions the reproducibility crisis raises: 1) Why is so much data irreproducible? and 2) why do scientists not make more replications of previous data? These two questions lead to further questions such as 3) how could reproducibility be increased.

The greater part of the literature on replication in science focuses on questions 1) and 3). Hardly anyone asks why scientists are not more pro-active, constantly checking

previous results. Part of the reason no one asks question 2) is because it is usually assumed that we already have the answer (scientists don't want to/can't afford to invest the time and money needed for replications because there are no benefits to be had from doing so). But this answer is unsatisfactory in light of the trust researchers clearly put in the data they are using.

Here I have claimed that this trust is – in part at least – based on a form of replication that has so far not been picked up by commentators on the issue, namely what I have called 'micro-replications' (MRs). This form of replication is part of everyday research, as it is built into normal experimentation through the inter-experimental use of controls. This suggests that the extent of the reproducibility crisis – at least in the experimental life sciences – might be less dramatic than some of the ongoing discussions imply, as crucial forms of replications are overlooked due to a flawed conceptual framework within which the analysis of the crisis is taking place.

An interesting question the analysis provided here raises is why MRs have evaded our attention for so long. A key reason for the invisibility of MRs, I think, is the fact that they depend on a part of the experimental process that is still poorly understood, namely the experimental controls. The invisibility of controls might be explained by the fact that their use is not something that is discussed in review articles, original research articles or textbooks. How to use a control and what controls to use are questions that come up in the Q&A section of talks or in informal laboratory meetings, making it an element of scientific practice that can be difficult to track for philosophers and historians of science. Controls are also crucial elements of the peer review process,

another element of science that is largely hidden from sight and difficult to access and assess (asking for different/additional controls is probably one of the key parts of the review process in the experimental sciences). Controls therefore represent something like the dark matter of experimentation, at least from the viewpoint of philosophy: they are a central part of what holds the (experimental) universe together but they are almost invisible to the researcher who is trying to understand that universe.

But despite these challenges, if controls indeed have the importance for the progress and the reliability of the experimental sciences that I propose here then it will be crucial for philosophers and historians of science to develop a more detailed understanding of how they shape the research process and the thinking of researchers. If we do so we will also be in a better position to develop an understanding of more general issues, such as the reproducibility crisis in science.

Acknowledgments:
I would like to thank John Dupré, Roman Frigg, Sabina Leonelli, Jutta Schikore and Nicolas Wüthrich for critical input on this and/or an earlier version of this paper. I would also like to thank the audiences at the &HPS6 2016 conference in Edinburgh, UK and the Sixth Biennial SPSP 2016 conference in Glassboro, USA, where elements of this paper have been presented. The research leading to this paper has received funding from the Swiss National Science Foundation (grant nr. PA00P1_134166) and the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement nr. 324186.

References

Alberts, B., Kirschner, M.W., Tilghman, S. and Varmus, H., 2014. Rescuing US biomedical research from its systemic flaws. *Proceedings of the National Academy of Sciences* 111 (16): 5773–77.

Baker, Monya. 2016a. Biotech giant publishes failures to confirm high-profile science. *Nature* 530 (7589): 141.

——. 2016b. Is there a reproducibility crisis? *Nature* 533 (7604): 452–55.

Baker, Lisa M., and Kevin Dunbar. 2000. "Experimental design heuristics for scientific discovery: the use of baseline and known standard controls." *International Journal of Human-Computer Studies* 52 doi:10.1006/ijhc.2000.0393.

Begley, C. Glenn, and Lee M. Ellis. 2012. Drug development: Raise standards for preclinical cancer research. *Nature* 483 (7391): 531–33.

Begley, C. Glenn, and John P. Ioannidis. 2015. Reproducibility in science. *Circulation Research* 116 (1): 116–126.

Begley, C. Glenn, Alastair M. Buchan, and Ulrich Dirnagl. 2015. Robust research: Institutions must do their part for reproducibility. *Nature* 525 (7567): 25–27.

Burian, Richard M. 1997. “Exploratory Experimentation and the Role of Histochemical Techniques in the Work of Jean Brachet, 1938-1952.” *History and Philosophy of the Life Sciences* 19 (1): 27–45.

——. 2007. “On MicroRNA and the Need for Exploratory Experimentation in Post-Genomic Molecular Biology.” *History and Philosophy of the Life Sciences* 29 (3): 285–312.

Casadevall, Arturo and Ferric C. Fang. 2010. “Reproducible Science.” *Infection and Immunity* 78 (12): 4972–75.

Collins, Harry. 1985. *Changing order: Replication and induction in scientific practice*. University of Chicago Press.

Drummond, Chris. 2009. “Replicability is not reproducibility: nor is it good science.” *Proc. Eval. Methods Mach. Learn*. Workshop 26th ICML, Montreal, Quebec, Canada. http://www.csi.uottawa.ca/cdrummon/pubs/ICMLws09.pdf.

Elliott, Kevin C. 2007. “Varieties of Exploratory Experimentation in Nanotoxicology.” *History and Philosophy of the Life Sciences* 29 (3): 313–336.

Errington, T.M., Iorns, E., Gunn, W., Tan, F.E., Lomax, J. and Nosek, B.A., 2014. An open investigation of the reproducibility of cancer biology research. *Elife* 3: p.e04333.

Firestein, Stuart. 2015. *Failure: Why science is so successful*. Oxford University Press.

Franklin, Laura R. 2005. "Exploratory Experiments." *Philosophy of Science* 72 (5): 888–99.

Goodman, S.N., Fanelli, D. and Ioannidis, J.P., 2016. What does research reproducibility mean? *Science Translational Medicine* 8 (341): 341ps12.

Grinnell, Frederick. 1992. *The Scientific Attitude*. New York: The Guildford Press.

Karaca, Koray. 2013. The strong and weak senses of theory-ladenness of experimentation: Theory-driven versus exploratory experiments in the history of high-energy particle physics. *Science in Context 26* (1): 93–136.

Makel, M.C., Plucker, J.A. and Hegarty, B., 2012. Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science* 7 (6): 537–42.

Mobley, A., Linder, S.K., Braeuer, R., Ellis, L.M. and Zwelling, L., 2013. A survey on data reproducibility in cancer research provides insights into our limited ability to translate findings from the laboratory to the clinic. *PLoS One* 8 (5): p.e63221.

Nosek, Brian, and Timothy Errington. 2017. Reproducibility in cancer biology: making sense of replications. *Elife* 6: p.e23383.

O'Malley, Maureen A. 2007. "Exploratory Experimentation and Scientific Practice: Metagenomics and the Proteorhodopsin Case." *History and Philosophy of the Life Sciences* 29 (3): 337–58.

Open Science Collaboration, 2015. Estimating the reproducibility of psychological science. *Science* 349 (6251): p.aac4716.

Prinz, Florian, Thomas Schlange, and Khusru Asadullah. 2011. Believe it or not: how much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery* 10 (9): 712.

Sarewitz, Daniel. 2016a. The pressure to publish pushes down quality. *Nature* 533 (7602): 147.

——. 2016b. Saving science. *The New Atlantis* 49: 4–40.

Soler, Léna, Emiliano Trizio, Thomas Nickles, and William Wimsatt, eds. 2012. *Characterizing the Robustness of Science: After the Practice Turn in Philosophy of Science* (Vol. 292). Springer Science & Business Media.

Steinle, Friedrich. 1997. "Entering New Fields: Exploratory Uses of Experimentation." *Philosophy of Science* 64 (Proceedings): S65–S74.

——. 2002. "Experiments in History and Philosophy of Science." *Perspectives on Science* 10 (4): 408–32.

Vasilevsky N. A., Brush M. H., Paddock H., Ponting L., Tripathy S. J., LaRocca G. M., Haendel M. A. 2013. On the reproducibility of science: unique identification of research resources in the biomedical literature. *PeerJ* 1:e148 https://doi.org/10.7717/peerj.148

Waters, C. Kenneth. 2007. "The Nature and Context of Exploratory Experimentation: An Introduction to Three Case Studies of Exploratory Research." *History and Philosophy of the Life Sciences* 29 (3): 275–84.