Accepted version of article: Stephan Guttinger, "A New Account of Replication in the Experimental Life Sciences," *Philosophy of Science* 86, no. 3 (July 2019): 453-471.
https://www.journals.uchicago.edu/doi/abs/10.1086/703555?af=R

A New Account of Replication in the Experimental Life Sciences

Stephan Guttinger
Centre for Philosophy of Natural and Social Science, London School of Economics, Lakatos Building, Houghton Street, London, WC2A 2AE, UK

Abstract:

The life sciences are said to be in the midst of a replication crisis because 1) a majority of published results are irreproducible, and 2) scientists rarely replicate existing data. Here I will argue that point 2) of this assessment is flawed because there is a hitherto unidentified form of replication in the experimental life sciences, which I will call 'micro-replications' (MRs). Using a case study from biochemistry I will illustrate how MRs depend on a key element of experimentation, namely experimental controls. I will end by reflecting on what MRs mean for the broader debate about the replication crisis.

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. 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.

1. Introduction

The experimental life sciences have been a success story in many ways. From insights into the functioning of our cells to the development of new cancer treatments, findings from fields such as biochemistry, molecular biology, immunology and genetics have not only fundamentally re-shaped our understanding of biological systems but also translated into advances in the clinical sciences.

However, in recent years several reports raised alarm over the trustworthiness and usability of the data produced in the life sciences (Prinz et al. 2011; Begley and Ellis 2012). The reports in particular claimed that the vast majority (75%-90%) of the studies produced in (academic) wet-lab research are not reproducible.[1] These numbers acquired explosive power in light of the fact that researchers rarely set up dedicated replication studies to test existing data. Scientists, therefore, seem to move ahead blindly (or even recklessly), as they use potentially flawed data without testing it first. As I will discuss in more detail in section 2, this assessment of the status quo fuelled talk of a replication crisis in the biological sciences and led to calls for fundamental changes to the way research is being conducted and funded.[2]

[1] Note that the issue of replication has also been hotly debated in other fields such as the psychological sciences. I will return to this wider debate in sections 5 and 6.

[2] Daniele Fanelli shows that since 2013 there has been a rapid increase in talk of a 'replication crisis' in the sciences (Fanelli 2018). There clearly is a correlation between

Here I will claim that the above assessment of the status quo is flawed and that more replication is taking place in the experimental life sciences than it is usually assumed. More specifically, I will claim that a key element of the experimental process – namely experimental controls – provides replications of existing data. I will refer to these replications as 'micro-replications' (MRs). Using the case study of the in vitro binding assay (introduced in section 3), I will show that controls in the experimental life sciences can establish links between different experiments by embodying elements from previous experiments in new experimental settings (section 4). 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 (section 5). I will end the paper by reflecting on what the concept of MRs could mean for the broader debate about replication in the experimental sciences (section 6).

2. Replication in the Experimental Life Sciences

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; Schmidt 2009) and

---

the publication of the studies by (Prinz et al. 2011) and (Begley and Ellis 2012) and the crisis narrative picking up steam.

to ensure the reliability and/or robustness of experimental output (Soler et al. 2012).[3] However, despite this theoretical consensus, in practice replication seems to be more an idea than a reality, in particular as replications are rarely performed. As I will show in the next section, this and two other claims have led to the idea that the experimental life sciences are in the midst of a replication crisis.

2.1 The Three Claims Fuelling Talk of a Replication Crisis

The first claim that underlies the current narrative of a replication crisis in the life sciences is an existence claim, namely the simple acknowledgment that there are published (and usually peer-reviewed) data that are irreproducible. This is probably the least controversial part of the debate as most scientists and commentators seem to agree that research is difficult and that scientists won't always get it right. Most commentators therefore seem to accept that some level of failure is to be expected (see (Firestein 2015) and (Redish et al. 2018) on the importance of failure in the experimental life sciences). The simple existence of irreproducible data is therefore unlikely to cause alarm on its own.

In addition to this existence claim, however, a more specific claim about the *extent* of failure in the biological sciences has emerged in recent years. In 2011 and

[3] 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)).

2012, two papers presented specific numbers for the percentage of irreproducible studies that are published in the experimental life sciences (with a focus on pre-clinical cancer research). The numbers presented were truly staggering, ranging from 75% to 90% (Prinz et al. 2011; Begley and Ellis 2012).

The realisation that failure rates might be significantly higher than anyone expected quickly led to calls for reform, in particular regarding questions of research procedure and conduct. There is broad agreement amongst scientists that research practice can and has to be improved in order to reduce the risk of failure (see, e.g., Collins and Tabak 2014; Begley et al. 2015; Munafò et al. 2017). A key part of this debate focuses on the importance of quality control, in particular the quality of the materials used, such as cell lines or antibodies (Baker 2016a). The question of how best to report data is another part that has gained significant attention in this context (see, e.g., Landis et al. 2012).

But again, it could be argued that these numbers on their own would not have been sufficient to trigger talk of a crisis. Rather, they seem to have gained their explosive power in light of a third claim about the status quo in the experimental life sciences, namely the idea that researchers usually don't test existing data. There is a broad consensus, also among scientists, that replications are rarely performed in the experimental sciences (Collins 1985; Baker 2016b; Goodman et al. 2016).[4] There is

[4] This not only applies to the biomedical sciences but also to research in psychology (Makel et al. 2012).

little empirical data on the exact extent of the problem, but current estimates place the percentage of replication studies (which might vary from field to field and also over time) somewhere between 1-4% (Ioannidis 2012; Makel et al. 2012; Iqbal 2016).

This third claim matters because it makes all the difference. With no replication studies being performed scientists seem to be moving forward blindly, taking existing data at face value and even willingly ignoring potential problems. If replication studies were routinely performed then even a high percentage of irreproducible studies would be less of a problem because scientists would be able to identify the problematic studies before using them. Science would of course be a highly inefficient enterprise, but it would not necessarily risk its trustworthiness.

It is this risk of losing trustworthiness that ultimately led to calls for a radical reform of the way in which research is being conducted and funded. Daniel Sarewitz, for instance, called for basic research to be cut back in favour of research that is tied to practical problem-solving (by tying it more closely to what he calls the 'national innovation complex' of which the military-industrial complex was a precursor). This should ensure that science becomes more accountable and reliable again (Sarewitz 2016a).

2.2 The Problem of Incentive Structures

The above assessment of the status quo raises several questions. One is why so many studies fail. Another is why researchers rarely perform replication studies. Here I will focus on the second question.

A key culprit that has been identified in response to this question is the prevailing publish-or-perish culture in the sciences (and the existing incentive structures more generally). The idea is that replications are not performed because there is no reward to be had from doing so: replications are not only expensive and time-consuming to do but they also don't translate into high-impact publications (if they can be published at all). Researchers might be aware of the potential problems with published data, but they simply don't have the time and money to perform replication studies. To solve this problem, so the current thinking, the incentive structures in the sciences have to be changed (see, e.g., Alberts et al. 2014; Sarewitz 2016a; 2016b; Rosenblatt 2016; Romero 2017).

However, this explanation of why there are so few replication studies being performed in the life sciences is problematic for several reasons. One is that there seems to be little correlation between the emergence of the current incentive structures and the number of replication studies performed. If the recent changes in the incentive structures were indeed what keeps scientists from performing replication studies, one would expect to observe a decline in the number of replication studies produced over the last decade or two. However, from the longitudinal data we have on the prevalence of replication studies there is no indication that there has been such a decline. In fact, a study on the situation in the psychological sciences (which compared the time period from 1950-1999 with the period from 2000-2012) showed that if anything there was a slight increase in the number of published replication studies over time (Makel et al.

2012).[5] This data is in line with the fact that the general absence of replications in science has already been reported by observers in the 1980s who were looking back at the research done in the 1960s and 70s (see, for instance, (Collins 1985)). It is also in line with a recent study by Daniele Fanelli who claims that new analyses of meta-studies show no indication of an increase in issues related to reproducibility (and research integrity more generally) over recent years (Fanelli 2018).

But if growing financial and career-related pressures don't explain the obvious lack of published replication studies, how can we make sense of the way in which scientists proceed? As Jim Bogen pointed out, the fact that scientists seem to accept non-replicated data as a trustworthy basis for further research represents a puzzle in the context of the prevailing analytic framework, which assumes that only replicated work is epistemically sound (Bogen 2001).

Bogen proposes an interesting explanation of the phenomenon: based on an analysis of the clinical and pathological sciences in the 19th century, he claims that some data simply don't need to be replicated in order to be judged by scientists as trustworthy. Researchers, at least in the observational disciplines he analysed, have

[5] The only study I am aware of that focuses on the biomedical sciences (Iqbal et al. 2016) analysed papers published between 2000 and 2014. The percentage of replications it found (1.5%) matched the numbers found in the psychological sciences. Unfortunately, this study did not look at the long-term changes in the number of replication studies.

other ways of establishing trust in existing data. Bogen therefore suggests that the existing doctrine of the importance of replications is too narrow.

Bogen's assessment allows us to make sense of the way some scientists proceed with confidence in the absence of replication studies. But it is not clear how generalisable his position is, in particular because he focuses on observational data from the clinical and pathological sciences. In other fields, such as the experimental life sciences, evidence is based less on observational reports and more on the manipulation of specially prepared specimens in particular experimental setups. This context will likely pose different challenges to the cases that Bogen analysed.

Here I want to propose a different explanation of why some scientists move forward without performing dedicated replication studies that applies more directly to the situation in the experimental life sciences. Like Bogen, I will claim that the analytic framework which guides the debate about replication is flawed. Unlike Bogen, I will focus on the different forms replications can take on and not on whether replications are needed at all. I will claim that researchers move forward without performing dedicated replication studies because they can rely on a special form of replication that is built into regular research practice and which is not recognised by the current analytic framework. To develop this account of what I will call 'micro-replications' (MRs) I will make use of recent work in philosophy of experimentation to analyse how researchers in the life sciences build on existing knowledge without testing it first.

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).[6] 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. In section 3, I will introduce an experimental system that can be used for both EE and TDE, namely the so-called 'in vitro binding assay'. This experimental system is widely used in the life sciences to study protein-protein interactions. Comparing its different uses will allow us to gain more insight into how researchers move forward with confidence even in the absence of dedicated replication studies.

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.

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

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.[7]

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

[7] Note that the in vitro binding assay can also be used to study the interactions between other entities, such as DNA, RNA or small molecules such as hormones.

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[8] coupled to Western blot analysis or mass spectrometry, two of the main methods used in molecular biology to identify specific proteins.

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 molecule, in this case 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

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

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 (Steinle 1997).

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 of 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 first 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 that arises in every laboratory-based research setup (and elsewhere, for instance when making measurements). When the entity or process of interest is placed in a context that is different from its native environment (in the case of biological entities or processes this is usually the cell or the organism) there is a chance that behaviours are detected that are only specific to the new but not to the native context (or that native behaviours are completely lost). Such artefacts negatively affect the trust a researcher can put in the results obtained as they might lead to false positive or false negative outcomes.

In the context of protein studies a key problem 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 a 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 for the in vitro binding assay 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 (this applies to any use of this or similar assays).

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 consists a) 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.[9]

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.

And like the negative control, 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).

[9] 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 4.2.

4. The Different Dimensions of Experimental Controls

The analysis above has shown that negative and positive controls 1) serve as calibration devices and 2) can be used to exclude artefacts. The controls allow the researcher to put 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.

But as I will show below, this ability to create trust and readability does not simply stem from the intra-experimental role a control plays but also from the inter-experimental links they establish.

4.1. 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 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, I will argue below, not only applies to molecular interaction studies but also to any experiment that recreates biological events in a non-native setting.

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).[10]

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

[10] 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.

factors involved are in principle active and allow the researcher to make reliable statements 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 work, which I will refer to as the *inter*-experimental role of controls. This mode, I claim, is a crucial part of what makes controls tools for establishing trust when building on the work of others.

4.2. The Inter-Experimental Role of Controls

The two setups of the in vitro binding assay described in section 3 have shown that 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.

The description of the guided experiment (section 3.3) 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. 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 specific 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. This means that in addition to the intra-experimental role described above there is also an *inter*-experimental role controls can play.

The inter-experimental mode of controls is significant in the context of this paper because it entails the replication of earlier results. What the case study shows is that previous experiments are brought into the experiment at hand through the controls. The results from previous studies are re-produced in control samples to prove that the system works as expected. They are therefore also 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.

4.3. Replication in the Experimental Life Sciences – The General Importance of Controls

This way of moving forward in experimentation is, I claim, a general feature of research in the experimental life sciences. The setup of the case study discussed here was not determined by the fact that an interaction between proteins is analysed. The same principles for the use of positive and negative controls would also apply if interactions between RNA, DNA or, for instance, membrane vesicles were studied. There is also nothing in this general setup that depends on the fact that it is an *interaction* study we are looking at (elsewhere I illustrate the power of MRs using an example from plant biology (Guttinger 2018)).

What calls for a positive control is rather the fact that a particular entity or phenomenon is analysed in a setting that is not native to it. Specific biological entities or processes are transferred into a new context where they are combined with different

materials (for instance synthetic beads or buffer solutions that would not be encountered in a cell or organism). The researcher therefore needs to make sure that despite all of these changes the entity or phenomenon of interest still behaves as expected. As discussed above, the main aim of the controls used is to check for artefacts *and* to make sure researchers can get a readable output.

The power of the positive control in particular is to demonstrate continuity and accuracy – what a functioning positive control shows is that the current setup is in line with earlier settings and that an accurate representation of earlier effects is possible in this new setup. This uniformity and accuracy requirement is also something that Bogen highlighted when he discussed why some data is accepted by scientists without replicating it first (Bogen 2001). Bogen states that if the uniformity of the object of interest and the accuracy of the observation report are established researchers might have no need to perform additional replications of the existing data. It is simply accepted that the results are in line with what is already known. As my case study illustrates, in cases in which uniformity and accuracy cannot simply be assumed (for instance because of the extensive manipulations needed to isolate and purify a protein) researchers will use specific controls to ensure a reliable and readable output. As I will explain in more detail in the next section, these controls represent a new form of replication that the prevailing analytic framework in the replication crisis debate is not accounting for.

5. Replication-via-controls vs. Replication-as-add-on

An interesting aspect of the whole debate about the replication crisis in the experimental sciences is that it is exclusively based on the idea that replication studies are add-ons to regular experimental practice: even though there is little consensus in the literature (both within the sciences and philosophy of science) on the exact definition of replications, everyone seems to agree that replications are something that has to be done *on top* of what researchers normally do. Replications are seen as add-ons that cost money and time. In this framework it is little surprise that researchers don't seem willing to perform replications.

What the analysis of the in vitro binding assay has shown, however, is that this picture is too simple. Replications of earlier results happen as part of regular experimentation and not just in what is explicitly designed and labelled as a replication of earlier results; they are a built-in part of standard research practice. These replications-via-controls don't necessarily aim to repeat a whole study or a particular figure from earlier work. They rather pick out one aspect that is crucial in guiding the experiment at hand and make it part of the current setup in order to establish its readability and trustworthiness (the two being intertwined). Because of the small (but important) role they play in the new study they are part of I will refer to these replications as 'micro-replications' (MRs).[11]

[11] Note that positive controls can reproduce crucial aspects of existing studies and that these reproductions can be time- and resource-consuming. In this latter sense there is nothing 'micro' about MRs. See also section 5.1 on this point.

These MRs offer a different explanation of why scientists are often happy to move ahead without setting up dedicated replication studies first. Researcher trust the particular data they are relying on because they *are* replicating it through the positive controls they are using. Elements of previous work thereby become part of the current experimental setting. This also implies that scientists are not simply moving ahead blindly or recklessly (at least if they implement the appropriate controls).

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 repeat it in their own laboratory to see for themselves. This is part of what allows them to navigate a realm that can be filled with potentially problematic data. Unsurprisingly perhaps, Begley (2013) has identified the absence of adequate experimental controls as one of six red flags for suspect work.

Once we realize that (micro-)replications happen as part of normal experimentation, the picture of a crisis in science changes. What the analysis provided here suggests is that scientists do more (successful) replications than current analyses suggest. Because of the controls scientists use they not only trust the output of their own experiments but they also know when to trust the data published by others.

5.1. Open Questions

There are of course a range of questions or objections that the idea of MRs raises. The aim of this paper was to introduce the idea of MRs, but more work (both empirical and philosophical) will be needed to understand their role and structure in more detail. It is, for instance, not clear yet how much power MRs ultimately have. MRs might be present in experimental science, but they are unlikely to pick up and therefore cover all of the existing data in a field. This would mean that MRs could not make up for the obvious lack of replication studies.

This surely is a valid worry, however, it is also important to point out that the suggestion here is not that MRs are a complete substitute for full-blown replication studies. Dedicated replication studies are certainly an important (but relatively rare) part of experimental reality. What the MR account proposes is that MRs are an additional level of replication that is (potentially) wide-spread and which has so far been overlooked in the debate about the replication crisis. Further research will have to establish the exact prevalence and power of MRs in the experimental life sciences and elsewhere.

Another issue that could limit the power of MRs is that researchers might simply pick the low-hanging fruit when setting up controls.[12] Researchers might, for instance, choose a positive control that has little biological relevance or which has already been tested many times before. This could mean that MRs – whilst a real thing – might not represent an important or informative type of replication.

[12] I thank one of the anonymous reviewers for raising this point.

There certainly are cases in which researchers will choose a safe, easy, and/or well-known option when selecting a suitable control. But even if this is the case in some instances, it does not mean it is the norm.

To understand why this is so it is important to consider some of the differences between EE and TDE. As the case study discussed in this paper has illustrated, in the case of EE any known interactor of factor X will do as a positive control. This means that the researcher is free to choose a protein that is well-studied and/or easily accessible. All that the researcher needs to show is that system in principle works.

However, in the case of guided experiments (in which a specific effect or phenomenon is further investigated) researchers will usually be much more restricted in their choice of controls as they will have to demonstrate that they can observe the original effect in their own setup (no matter whether this is easy or difficult to achieve). This also restricts the choice of positive control(s) that is open to them. In the example discussed in section 3.3, for instance, the researchers had to use members of a particular class of proteins (containing a specific type of signal peptide) as a positive control; simply picking any convenient or well-established control is not an option.

More often than not, the specific phenomenon or finding of interest in TDE will be novel and hence not yet well-tested. Given that a significant amount of research in the life sciences is guided in the sense of TDE, it is reasonable to assume that a significant amount of the positive controls used are micro-replications of interesting and novel pieces of data, rather than bland repetitions of well-established findings.

Another open question the MR account faces is how broadly it applies to science more generally. As mentioned in footnote 1, talk of a replication crisis not only affected the biological sciences but also fields such as the psychological sciences. It is not clear yet to what extent MRs are present in these fields and what work they might do there. My initial sense is that they are present in the psychological sciences but that they are less abundant than in the biological sciences because the experimental setups used are very different. However, a more detailed answer to this issue will depend on an in-depth comparison of the experimental setups used in the different fields and of the roles controls play within them.

6. Conclusions: Replications and the Dark Matter of the Experimental Sciences

There are (at least) two questions the reproducibility crisis in the biological sciences raises: 1) Why is so much data irreproducible and 2) why do scientists not perform more replications of previous data? It is usually assumed that the answer to the second question is found in the prevailing incentive structures in science (scientists don't want to/can't afford to invest the time and money needed for replications because there is little reward for doing so).

Here I have claimed that there is another reason why dedicated replication studies are rarely performed. Based on the analysis of a case study from the experimental life sciences, I claim that there is a form of replication that has so far been overlooked by commentators on the issue, namely what I have called 'micro-replications' (MRs). This form of replication is part of everyday research practice, as it is built into normal

experimentation through the inter-experimental use of controls. It allows researchers to have a readable and trustworthy output of their particular experiment and it also gives them a tool to judge the quality of the work of others. The presence of MRs suggests that the extent of the reproducibility crisis might be less dramatic than some of the ongoing discussions imply, as more replications are performed than it is usually assumed.

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 research process that is still poorly understood, namely the experimental controls. Whilst controls have gained significant attention in philosophy of statistics, this is not necessarily the case when it comes to the use of controls in the experimental life sciences.

This 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.[13] Controls are also crucial elements of the peer review process, another element of

[13] An exception that confirms the rule is (Schickore 2017) who presents crucial insight into the history of controlled experiments through her in-depth analysis of snake venom research.

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 they have 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 in the experimental life sciences. 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.

References

Alberts, Bruce, Marc W. Kirschner, Shirley Tilghman, and Harold Varmus. 2014. "Rescuing US biomedical research from its systemic flaws." *Proceedings of the National Academy of Sciences* 111 (16): 5773–77.

Baker, Monya. 2016a. "How quality control could save your science." *Nature* 529 (7587): 456-58.

——. 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. 2013. "Reproducibility: six red flags for suspect work." *Nature* 497 (7450): 433–34.

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

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

Bogen, Jim. 2001. "'Two as good as a hundred': poorly replicated evidence in some nineteenth-century neuroscientific research." *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 32 (3): 491-533.

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*. Chicago: University of Chicago Press.

Collins, Francis S., and Lawrence A. Tabak. 2014. "NIH plans to enhance reproducibility." *Nature* 505 (7485): 612.

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): 313–36.

Fanelli, Daniele. 2018. “Opinion: Is science really facing a reproducibility crisis, and do we need it to?” *Proceedings of the National Academy of Sciences* 115 (11): 2628–31.

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

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

Goodman, Steven N., Daniele Fanelli, and John P. A. Ioannidis. 2016. “What does research reproducibility mean?” *Science Translational Medicine* 8 (341): 341ps12.

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

Guttinger, Stephan. 2018. "Replications Everywhere: Why the replication crisis might be less severe than it seems at first." *BioEssays* 40 (7): 1800055.

Ioannidis, John P. A. 2012. "Why science is not necessarily self-correcting." *Perspectives on Psychological Science* 7 (6): 645-54.

Iqbal, Shareen A., Joshua D. Wallach, Muin J. Khoury, Sheri D. Schully, and John P. A. Ioannidis. 2016. "Reproducible research practices and transparency across the biomedical literature." *PLoS biology* 14 (1): e1002333.

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.

Landis, Story C., Susan G. Amara, Khusru Asadullah, Chris P. Austin, Robi Blumenstein, Eileen W. Bradley, Ronald G. Crystal, Robert B. Darnell, Robert J. Ferrante, Howard Fillit, Robert Finkelstein, Marc Fisher, Howard E. Gendelman, Robert M. Golub, John L. Goudreau, Robert A. Gross, Amelie K. Gubitz, Sharon E. Hesterlee, David W. Howells, John Huguenard, Katrina Kelner, Walter Koroshetz, Dimitri Krainc, Stanley E. Lazic, Michael S. Levine, Malcolm R.

Macleod, John M. McCall, Richard T. Moxley III, Kalyani Narasimhan, Linda J. Noble, Steve Perrin, John D. Porter, Oswald Steward, Ellis Unger, Ursula Utz, and Shai D. Silberberg. 2012. "A call for transparent reporting to optimize the predictive value of preclinical research." *Nature* 490 (7419): 187–91.

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

Munafò, Marcus R., Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware, and John P. A. Ioannidis. 2017. "A manifesto for reproducible science." *Nature Human Behaviour* 1: 0021.

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.

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): 712–13.

Redish, A. David, Erich Kummerfeld, Rebecca Lea Morris, and Alan C. Love. 2018. "Opinion: Reproducibility failures are essential to scientific inquiry." *Proceedings of the National Academy of Sciences* 115 (20): 5042–46.

Romero, Felipe. 2017. "Novelty versus Replicability: Virtues and Vices in the Reward System of Science." *Philosophy of Science* 84 (5): 1031-43.

Rosenblatt, Michael. 2016. "An incentive-based approach for improving data reproducibility." *Science Translational Medicine* 8: 336ed5.

Sarewitz, Daniel. 2016a. "Saving science." *The New Atlantis* 49:4–40.

——. 2016b. "The pressure to publish pushes down quality." *Nature* 533 (7602): 147.

Schickore, Jutta. 2017. *About Method: Experimenters, Snake Venom, and the History of Writing Scientifically*. Chicago: University of Chicago Press.

Schmidt, Stefan. 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, Léna, Emiliano Trizio, Thomas Nickles, and William Wimsatt, eds. 2012. *Characterizing the Robustness of Science: After the Practice Turn in Philosophy of Science. Boston Studies in the Philosophy and History of Science*. Vol. 292. Berlin: 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.

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.