PLACEBO TRIALS WITHOUT MECHANISMS: HOW FAR CAN THEY GO?

David Teira
Dpto. de Lógica, Historia y Filosofía de la ciencia, UNED
Senda del rey 7 | 28040, Madrid (Spain)
dteira@fsof.uned.es

Published in *Studies in History and Philosophy of Biol & Biomed Sci*,
https://doi.org/10.1016/j.shpsc.2019.101177

ABSTRACT

In this paper, I suggest that placebo effects, as we know them today, should be understood as experimental *phenomena*, low-level regularities whose causal structure is grasped through particular experimental designs with little theoretical guidance. Focusing on placebo interventions with needles for pain reduction -one of the few placebo regularities that seems to arise in meta-analytical studies- I discuss the extent to which it is possible to decompose the different factors at play through more fine-grained randomized clinical trials. My sceptical argument is twofold. On the one hand, I argue that experiments alone are not enough to standardize interventions, and that it is necessary to include theories. On the other hand, I argue that the social interactions that seem to be part of placebo effects are difficult, if not impossible, to blind. Therefore, the measurement biases arising from the participants' reactivity to the experimental setup cannot be controlled for. Further decomposition of placebo effects requires a theoretical account of the existing experimental regularities that may guide further tests.

Keywords: placebo; clinical trials; reactivity; needling; mechanisms; standardization

1. Decomposing placebo interventions

Over the last fifty years, the experimental social sciences have created a catalogue of *phenomena*, regularities that arise from the data accumulated in a family of experiments (Bogen & Woodward, 1988; Guala, 2005, pp. 41-44). These tests aim at capturing and describing the details of a *prima facie* causal structure, with or without the support of a pre-existing theory. Experimentalists may look for phenomena guided by a theory that may explain them, but phenomena can be also freestanding regularities awaiting an explanation or further analysis[1]. For instance, the regularities emerging from the popular *dictator game* in experimental economics (a test about how people share resources) can be appraised as a test of the predictions of economic theory about

[1] (Gervais & Weber, 2015) discusses how these regularities often emerge from what they aptly call *orientation experiments*, in which experimenters, often without a guiding theory, offer a rough, qualitative characterization of the mechanism responsible for a salient causal capacity.

individual decisions, but also as an independent experimental setup for investigating all sort of sharing behaviours (Jiménez-Buedo, 2015).

The placebo effect in medicine often appears as one such phenomenon. It originated in informal observations that led to a collection of systematic experiments, all of which showed how treatments without, e.g., a pharmacologically active principle had nonetheless measurable therapeutic benefits. Placebo effects have a freestanding status in medicine: they have been documented almost without theoretical guidance and still today there is no consensual account explaining why placebos work. In this regard, as I will argue in section 2, the status of placebo effects as phenomena has been linked to an experimental method to capture them, in particular the randomized clinical trial (RCT). During the second half of the 20th century, many different interventions were tested in RCTs for placebo effects with often positive outcomes.

All these outcomes have been challenged from the 1980s onwards (Kienle & Kiene, 1997) when a statistical reanalysis of the raw data showed how the purported phenomenon was most likely a statistical artefact: a simple regression to the mean (McDonald, Mazzuca, & McCabe, 1983). Under certain statistical assumptions, it can be shown that, on average, patients exhibiting abnormal levels of an outcome variable (such as pain) will regress to the mean level, whatever the treatment they receive: if the patients' pain was abnormally high, it will decrease; if it was abnormally low, it will increase. A recent Cochrane review (Hrobjartsson & Gotzsche, 2010) found that placebo interventions had no relevant clinical effects, except for patient reported outcomes, especially when using physical devices (such as needles) for pain reduction –*pace* (Howick, 2017). Interventions without a clear causal pathway such as acupuncture can relieve the self-reported degree of pain in patients, according to a standardized scale, by a clinically relevant magnitude.

In other words, the placebo effect would only constitute a real experimental phenomenon for a particular class of interventions on just an outcome variable: roughly speaking, placebo physical devices for pain reduction. Taking this as my starting point, I discuss the possibility of decomposing the only placebo effect so far documented into further components through new experimental designs. The general question I want to address is whether more clinical trials can lead to the discovery of new regularities about placebos or it is rather time for basic research to pave the way for further experimentation.

The social and biomedical sciences have witnessed throughout the last five years how many experimental phenomena have not resisted the test of systematic replication (Fanelli, Costas, & Ioannidis, 2017). In this context, this paper is a reflection on the limits of the experimental decomposition of a complex causal structure, the placebo effect, involving social and biomedical factors. Are clinical trials enough to grasp the factors at play in placebo effects?

My case in point is a particularly insightful trial design: (Kaptchuk, et al., 2008) investigated whether a placebo effect could be triggered by three different interventions that had been merged in previous tests. The interventions were assessment and observation, "a therapeutic ritual", and a supportive patient-practitioner relationship. The trial targeted patients with irritable bowel syndrome (IBS)[2]: in the first arm, the participants were put on a waiting list (observation); in the second arm, the participants received sham acupuncture administered by a therapist with a sober interaction script; in the third arm, sham acupuncture was administered with an engaging interaction script in which the therapist showed active support for the patient. The effects of these interventions were assessed with four patient-reported outcomes, and the statistical analysis showed that sham acupuncture either with a limited or an augmented interaction produced statistically and clinically significant outcomes. The three interventions, the authors concluded, "can be progressively combined in a manner resembling a graded dose escalation of component parts" (Kaptchuk, et al., 2008).

My question in this paper is whether these factors can be further decomposed in subsequent trials. In order to answer it, I will proceed as follows. In the following section, I will explain in what sense RCTs capture a regularity in placebo effects and what sort of phenomenon we are witnessing in the IBS trial. In sections 3 and 4, I will discuss the two interventions in the trial, needling and interaction scripts. I will defend that experiments alone do not allow us to standardize these interventions to the degree that it would be necessary for a more fine-grained study of placebo analgesia. This lack of standardization, I will contend, is the source of the high heterogeneity and poor outcomes in so many experiments on needles and social scripts. Only basic research on the underlying mechanisms of placebo will allow trials to obtain further regularities about placebos. This may not be a popular conclusion in a field that has built until recently on the accumulation

[2] From now on, I will refer to (Kaptchuk, et al., 2008) as the IBS trial.

of simple empirical regularities, but, as I will argue in the conclusion, it is the better option to overcome the crisis of confidence in blind experimental research.

2. Measuring placebos in RCTs

The first step in my argument will be to explain in what sense the placebo effect has been, at most, an experimental phenomenon – one in which the causal structure did not arise from any theory but rather from the very design of the experiment. Historians of the placebo effect usually refer to the tests conducted by the Franklin commission in 1784 as the first experimental study of the placebo effect (Teira, 2016). At the request of the French monarchy, the commission had to assess the efficacy of J.B. Mesmer's magnetic therapy. Among its members there was Antoine Lavoisier, the great chemist, who systematized the use of blinding for the causal analysis of medical treatments. Lavoisier suspected that there was no real connection between the treatment and the effects observed in patients. He disentangled both following the standard laboratory practice of chemist: analysis and synthesis:

> There is only one way to achieve it: demagnetize the more sensitive persons without their awareness, and persuade them of being magnetized when they actually aren't. Gathering these two types of experiments, we will obtain separately the effects of magnetism and those of imagination, and we will be able to conclude what we should attribute to each one of them. (Lavoisier, 1865)

Lavoisier installed a screen between the therapist and the patient, so that this latter could not tell whether an intervention was being performed. Mesmer was discredited when, blinded, patients did not exhibit any of the effects of the therapy administered. From then on, blinding became a standard control in clinical trials, probably in response to an already widespread intuition among physicians: the mental states of patients (Lavoisier's "imagination") played some sort of role in their recovery. Blinding was therefore the device that allowed experimenters to decompose the different factors contributing to a treatment effect, without an actual understanding of their causal mechanisms.

Lavoisier showed that the therapeutic role of "imagination" could be isolated in an experimental setup through blinding. But it took 150 years more to quantify the *placebo effect* and transform it into a *phenomenon* (Miller, Colloca, & Kaptchuk, 2013) For this, we need to measure the differences between treatments in order to ascertain their degrees of efficacy. RCTs made this measurement possible. RCTs became a standardized

experimental design for testing treatments in the 1940s and the 1950s, articulating the statistical foundations laid by Ronald Fisher the previous decade with a long-standing tradition of controls in medical tests. RCTs are comparative experiments: at least two interventions are administered to two groups of patients, measuring the respective effects. If any difference in the outcome is observed, the design of the experiment will allow us to assess whether it is statistically significant and, if so, whether it is big enough for patients to benefit.

RCTs implement John Stuart Mill's *method of difference* in causal analysis: the two groups should be entirely alike except for the intervention administered. Hence, if there is a difference between the treatment effects, we will be entitled to attribute it to the causal power of each treatment. However, in every other respect, RCTs operate as a *causal black box*: we focus on the size of the difference between treatment outcomes, abstracting away the causal mechanisms by which such effects occur. Throughout the drug development process, researchers might have learnt about these mechanisms reasonably well. But success in a RCT depends only on the difference between treatment outcomes, independently of how good our mechanistic understanding of these outcomes is. Still in the 1950s and 1960s, treatments reached the pharmaceutical markets without such causal understanding of their action (Gonzalez-Moreno, Saborido, & Teira, 2015).

The placebo effect was quantified through RCTs in this black-boxed manner: without any assumption about the mental states involved, it was possible to quantify the effect of a placebo intervention in an arm of the trial and compare it to the outcome of the other treatments administered (or lack of them). We speak of the placebo effect when there is statistically significant difference between the outcome of a placebo intervention and the outcome of a no treatment group. However, in order to reach clinical significance (i.e., actual benefit for a patient), the placebo should not be worse than the standard treatment. As in every other RCT, treatments are not assessed on themselves but always regarding a clinical consensus on the appropriate intervention (or a placebo, if there is none).

A great deal of controversy on the placebo effect hinges on the diversity of experimental approaches to capture it. The last seven decades have witnessed all sorts of tests documenting placebo effects: apart from RCTs, there are experiments on animals, behavioural analysis, studios of brain images and physiological responses, etc. –see

(Miller, et al., 2013) for a survey. This wealth of evidence would suggest that placebo effects are a well-consolidated phenomenon. Therefore the paradox in the results of (Hrobjartsson & Gotzsche, 2010) meta-analysis: if only physical interventions on pain measured via subjective reports cause statistically significant effects, how shall we interpret all the other evidence about placebos? Here is where the concept of *phenomenon* plays a clarificatory role. Until a theory emerges to provide a unified explanation of all the placebo phenomena, what we actually have is a collection of experimental regularities, each one of them emerging from a particular setup. In these setups, the phenomenon will only exist to the extent that it is replicable: experimenters can reproduce it at will. Although placebo experiments often target similar variables (pain, expectations, etc), there is no reason to consider all the emerging regularities as instantiations of the same phenomenon. This is something only theories do, and there is still no consensual theoretical account of placebos (Miller, et al., 2013).

Why then focusing on RCTs as the *locus* of placebo effects? A more or less implicit assumption in placebo literature is that placebos are significant phenomena to the extent that they have significant clinical effects. The benchmark for what counts as one, for the last seventy years, has been the RCT: only to the extent that placebo interventions show efficacy under the same conditions as standard treatments will they count as conventional therapies –instead of, e.g., alternative medicine. If RCTs successfully track a regularity connecting placebo interventions with clinically relevant effect, this phenomenon is enough to consider placebo a medicine on equal grounds with other successful drugs that were accepted as such without a full understanding of their causal mechanisms.

From all this I draw two assumptions for the rest of my analysis. On the one hand, the placebo effect as an experimental phenomenon has its most interesting expression for therapeutic purposes in RCTs. On the other hand, the only solid placebo phenomenon documented RCTs occurs with physical devices to treat pain with a patient reported outcome. The available evidence suggests that there is something special to the placebo effects caused by, e.g,, needling and the question I want to address is whether more sophisticated RCT designs are enough to disentangle the active principle behind such interventions. The beauty of the IBS trial (Kaptchuk, et al., 2008) is that it takes exactly

this approach and uses a brilliant design in order to tear apart the components of the placebo effect in an RCT setup[3].

The treatment arms address three potential ingredients of this effect: the patients' response to observation and assessment (being in a waiting list); the patients' response to a perceived physical intervention (sham acupuncture); and the patients' response to this latter augmented with a positive interaction with the therapist (sham acupuncture plus a scripted conversation). The trial's causal box remains black: we may safely ignore the mechanism underlying the three interventions. But the size of the box is reduced: the placebo effect is disentangled, thanks to the patients' blinding. To the extent that they actually ignore the goal of the trial and the actual intervention being performed, their differential responses will show the size of each component of the placebo effect. The trial showed that these three components add up progressively, reaching a maximum effect with the augmented placebo, reaching a clinically significant effect in the treatment of the condition. The outcome variable was a patient reported outcome (the IBS Global improvement scale) which asked patients about the improvement of their symptoms. These are usually abdominal pain and discomfort associated with altered bowel habits.

The authors of the IBS trial are clearly aware that the three ingredients under study are big enough to encompass many different factors. As we will later discuss in section 4, it is impossible to rule out in this trial setup that the outcome owes much to the patients' reactivity: the patients are behaving in the test in a way that owes more to the interaction with the experimenter than to their own spontaneous reactions to the interventions. Following Kaptchuk's approach (Kaptchuk, 2002; Kerr, et al., 2011), the placebo treatment would capture the patient's response to the administration of a therapeutic ritual, for which the list of sub-ingredients is not short: the characteristics of both patient and practitioner, their interaction, the nature of the illness treated and the very treatment and treatment setting. The latter two arms will pick up on two of these sub-ingredients: the patient-practitioner interaction and the (purported) treatment.

The question the IBS trial opens is: how much deeper can we go into the decomposition of these different ingredients in the placebo effect through more fine-grained RCTs? Is it possible to unbundle the different elements in each arm into more

[3] I will assume, for the sake of the argument, that the IBS is indeed a replicable regularity. My argument does not presuppose it though: lack of reproducibility could be easily explained precisely for the causal roughness of the interventions tested, as discussed in sections 3 and 4.

specific interventions so that the different placebo components are more clearly identified? Although these are empirical questions that only further experiments will settle, I want to argue that experiments alone are not enough to standardize the two interventions under study at the IBS trial. At least, not to the degree required to obtain a solid replicable regularity. Let me start with needles in section 3 and then proceed to interaction scripts in section 4.

3. On the necessity of treatment standardization: needles

Testing a treatment in a clinical trial requires a certain degree of standardization. RCTs test a hypothesis about the comparative efficacy of, at least, two interventions. The statistical design of most RCTs assume a frequentist interpretation of probability. A p-value is the probability of observing a range of trial outcomes (under the hypothesis of no difference between treatments) if the RCT was repeated time and again. In a frequentist approach, the probabilistic assessment of a trial outcome is then tied to a particular experimental design: if the experiment is impossible to repeat, probabilities like p-values stop making sense. If an RCT can be repeated and outcomes appear with the initially predicted frequency, it is a sign of experimental control of the intervention under study (Spanos & Mayo, 2015). Under certain assumptions (Norton, 2015), the replicability of an experiment suggests that the intervention has a causal structure that does not depend on the particular characteristics of the experimenter: anybody following the protocol can obtain the same outcomes.

Therefore, if RCTs are adopted as the yardstick test of safety and efficacy, therapeutic interventions should be standardized enough to allow for direct replication. Pharmacological interventions easily meet this degree of standardization. On the one hand, they are industrially produced with strict quality control procedures. On the other hand, the administration instructions for most drugs can be easily summarised in a simple brochure. RCTs have been particularly successful at testing pharmaceutical compounds, but not so much in the assessment of less standardized interventions, such as surgical procedures. The development of surgical techniques involves a continuous refining process (Wartolowska, et al., 2016) in which a procedure is improved through gradual and constant modifications. Surgical degrees rarely remain the same and RCTs are rare and often have no significant impact on surgical praxis (Wartolowska, et al., 2014).

Drawing on these premises, my argument in this section will be as follows. So far, the only placebo effects with clinical and statistical significance have been obtained with needling techniques for treating subjectively reported pain. The IBS trial provides a decomposition of the factors at play in these interventions. I am going to argue next that any further decomposition requires a clinical trial design that controls for the two sources of measurement error that most often appear in assessing pain relief interventions. This trial design requires a degree of standardization in the interventions that experiments alone on neither needling nor scripting can yield. In this section, I will make the case for needling, leaving the analysis of interactions scripts for the following one.

Since standardization is often a matter of degree, let me first explore which particular degree is necessary for testing any intervention for pain relief, following (Dworkin, McDermott, Farrar, O'Connor, & Senn, 2014). First, regarding pain our species is not a homogenous biological population: not all patients are equally responsive to treatment, be it because of genetic differences in pain perception or in analgesic responses (Lotsch & Geisslinger, 2006; Mogil, 2012). Therefore, depending on the patient, the treatment effect might be systematically bigger or smaller, i.e. not random variations around the mean. Focusing on the mean treatment effect averaged over all patient potentially misses clinically relevant differences in response between patients. Measuring this potential *treatment-by-patient interaction* in an RCT requires a multiple-period-cross over design, in which each patient receives each treatment tested in the trial in at least two different periods. This would be, for instance, a 4-period cross-over trial with 2 periods of active treatment and 2 periods of placebo (Dworkin, et al., 2014). Taking up again the IBS trial interventions, in order to measure the treatment-by-patient interaction regarding the interaction script, each patient should receive both sham-acupuncture *without* the positive script and sham-acupuncture *augmented with* the positive script, in two different rounds. This would allow us to detect whether the sensitivity to the script varies *between patients*. At the same time, this multiple-period-cross-over design would allow us to grasp the potential *within-patient* variation over time: the measured outcome may also vary depending on the different responses a patient may give to the same treatment in different treatment periods. For instance, a potential source of variation is the lack of consistency between a single patient's responses, when she is asked to report her pain reduction in response to the same treatment. Even when patients use a standardized score to grade the analgesic effects of the treatment, they often assess

it differently in different exposures to the same intervention. This is a source of error to be controlled for whenever subject-reported outcomes are used, as it often happens in placebo trials on pain (Hrobjartsson & Gotzsche, 2010).

Multiple period cross-over designs allow us to measure these two sources of variability but they are, of course, complicated to implement and there are not many available in the pain literature (Dworkin, et al., 2014, p. 458). Yet, from a methodological standpoint, if we are to advance in the decomposition of the placebo effect through clinical trials, this particular cross-over design seems the most adequate option. On the one hand, there are clear signs of placebo treatment per patient interactions: although there is no precise identification of the *placebo responder* in terms of personality traits (Kaptchuk, 2002), there is evidence enough about placebo responses varying with individual patient's expectations (Kirsch & Rosadino, 1993; Price, et al., 1999). The formation of these expectations is a clear target for experimental decomposition. On the other hand, as we already saw, placebo effects have been shown to originate in simple statistical variation unrelated to the treatment, the so-called regression to the mean (McDonald, et al., 1983). To prevent a spurious statistical artefacts, experimenters should control for errors in measuring the outcome variable. Therefore in placebo pain trials with subjective reported outcomes it is crucial to check for within patient variation.

Now, if we are to decompose further the placebo factors featuring in the IBS trial, and the cross-over design proposed by (Dworkin, et al., 2014) is the best option, any further refinement of the interventions at play in the trial (needling + interaction script) should be standardized enough not just to allow replications of the same experiment, but even to perform the same intervention more than once during the experiment. I am going to discuss now to what extent experimentation with patients alone allows us to reach this degree of standardization. Let us first examine needling.

In the IBS trial, there were two weekly sessions of sham acupuncture, in which six to eight dummy needles were placed for 20 minutes over pre-determined non-acupuncture points on the arms, legs and abdomen. Acupuncture was then treated as it were already a standardized procedure, but a closer look at the literature on needling reveals that, if we wish to decompose further the analgesic effects of acupuncture through trials, we need a more rigorous standardization of the technique that experiments alone have not yielded.

Indeed, needling techniques have been notoriously difficult to standardize in RCTs on pain. For a start, there are different theoretical approaches to needling: e.g., whereas the old tradition of acupuncture is primarily based on a traditional understanding of body energies, contemporary techniques such as dry needling are based on a physiological theory about how pain originates in the muscles (Zhou, Ma, & Brogan, 2015). Dry needling was originally presented as the mechanical use of a hypodermic needle (without injecting a solution) thick enough to puncture contraction knots in the muscles, the so-called *trigger points*, the cause of actual pain (Simons, Travell, Simons, & Travell, 1999). Having a candidate mechanism to explain the effects of the intervention provides, in principle, a more solid basis to standardize the technique (as compared to other approaches to needling). Yet, there has been a significant gap between the purported physiological mechanisms of pain targeted by dry needling and the diagnostic criteria therapists use to identify them by physical examination (e.g., palpation) (Lucas, Macaskill, Irwig, Moran, & Bogduk, 2009; Tough, White, Richards, & Campbell, 2007). In blinded tests, examiners have obtained a poor inter-examiner diagnostic reliability, not succeeding at locating consensually the trigger points to be then punctured (Myburgh, Larsen, & Hartvigsen, 2008). This fundamental ambiguity pervades many other elements of the intervention: there is no solid consensus as to how many points to needle in each patient, how deep and for how long should the needle be inserted and how many times should the treatment be repeated (Quintner, Bove, & Cohen, 2015; Tough, White, Cummings, Richards, & Campbell, 2009). Dry needling trials have not succeeded either at identifying a clear standard between all the candidate techniques, in terms of its superior therapeutic effects. To the contrary, the specific effects are absent or small (Cummings & White, 2001; Espejo-Antunez, et al., 2017; Tough, et al., 2009). Furthermore, lack of standardization makes comparisons between trials difficult: e.g.,(Tough, et al., 2009) examined 1517 studies and found only seven that were of high enough quality (according to standard quality scores) for meaningful analysis.

RCTs on needling therapies are indeed not very conclusive: they usually lack methodological quality (measured with a standard scale) and the aggregate results in meta-analyses are poor –see (Madsen, Gotzsche, & Hrobjartsson, 2009). A charitable interpretation of this apparent failure would be as follows. If replicability signals a solid experimental control of the intervention, this lack of conclusive results would be a sign of a still preliminary understanding of the causal structure of the phenomenon under

study, the mechanisms underlying pain relief through needling. Only a better grasp of the physiological processes by which needles trigger the analgesic response will allow a precise standardization of the technique. It should establish precise enough thresholds for the identification and selection of the piercing points, depth and duration of the needling, etc.

In other words, and this is the relevant point for my argument: experimenting with needles on patients alone has not been enough to standardize the technique. Unlike in other experimental fields, in needling experimenting based on loose theoretical concepts has not yielded regularities that can be exploited for further theorizing on placebos. Without a clear standard, it is impossible to advance in the experimental decomposition of the needling factor in the placebo effect through multiple-period cross-over trials. Unless therapists can perform the needling with a high enough degree of agreement (measured in the usual reliability tests), the measurement error will grow unnoticed in the multiple treatment rounds necessary to implement the cross-over design. We will not be able to grasp the individual variability of treatment responses and introduce further refinements in how needles achieve their analgesic effects.

We should now turn our attention to the second placebo factor revealed in the IBS experiment, the scripted interaction, and discuss whether it is possible to standardize it enough to decompose it further.

4. Standardizing interaction scripts

The patient-practitioner relationship was implemented in two forms in the IBS trial. In the limited interaction arm, the script was described as follows

> The limited patient-practitioner relationship was established at the initial visit (duration <5 minutes) during which practitioners introduced themselves and stated they had reviewed the patient's questionnaire and "knew what to do." They then explained that this was "a scientific study" for which they had been "instructed not to converse with patients." The placebo needles were then placed, and the patient left alone in a quiet room for 20 minutes, a common acupuncture practice, after which the practitioner returned to remove the "needles."

In the augmented interaction arm, the description reads:

> Unlike participants in group 2 (limited), however, [the participants] received an augmented patient-practitioner relationship that began at the initial visit (45 minutes duration) and was structured with respect to both content (four primary

> discussions) and style (five primary points). Content included questions concerning symptoms, how irritable bowel syndrome related to relationships and lifestyle, possible non-gastrointestinal symptoms, and how the patient understood the "cause" and "meaning" of his or her condition. The interviewer incorporated at least five primary behaviours including: a warm, friendly manner; active listening (such as repeating patient's words, asking for clarifications); empathy (such as saying "I can understand how difficult IBS must be for you"); 20 seconds of thoughtful silence while feeling the pulse or pondering the treatment plan; and communication of confidence and positive expectation ("I have had much positive experience treating IBS and look forward to demonstrating that acupuncture is a valuable treatment in this trial"). We based this intervention model on research concerning an optimal patient-practitioner relationship. Only after completing this nine item agenda did the acupuncturist place the placebo needles and leave the participant in a quiet room for 20 minutes. On returning, the practitioner "removed" the placebo needles and exchanged a few words of encouragement.

The participants were not told that the study included two different degrees of interaction until the trial ended, so their blinding regarding this particular goal of the experiment was not tested. The augmented interaction script incorporated a bundle of micro-interventions targeting the patients, for which "future investigations will have to determine the relative importance". My question is whether RCTs alone can unbundle this script and establish the relative contribution of its sub-components. My conjecture is again pessimist, although for a different reason. Whereas the standardization of needling treatments might be achieved through basic research, the standardization of interaction scripts seems intrinsically more difficult. On the one hand, as of today, there is no fundamental discipline providing the fine grain causal structure of these sort of interactions. As I mentioned in the introduction, most experimental social sciences do not rely on a fundamental theory to grasp the causal variables, but they are grasped and refined through their experimental setups. On the other hand, the more complicated the script, the more difficult it is to blind it and fend off the participants' reactivity. Let me start with the former.

Decomposing the interaction scripts in the IBS so they can be further developed into a multiple-period cross-over trial poses the same sort of problems that we already witnessed in needling. The causal structure of the intervention is rough: just as the insertion technique of the needle was defined with relatively broad margins of precision, the interaction script is general enough to admit countless variations. Think for instance of the experimenter's characteristics (gender, age, etc.): without a theory guiding the

selection of the relevant factors, a further decomposition of the script would imply sampling any number of experimenter characteristics and randomizing the allocation in search of potential placebo components. This is, of course, unfeasible and, just as it happened with needling techniques, in most social sciences experimenting with interaction scripts, the alternative is to try to intervene with this degree of causal roughness in search of phenomena. Some fields have indeed accumulated a catalogue of relatively robust experimental regularities (e.g., experimental economics (Camerer, et al., 2016)). Yet some other fields targeting loose interactions, like psychology (and, in particular, social psychology), have experimented the same proliferation of non-replicable findings that we already witnessed in acupuncture (Collaboration, 2015).

For instance, unconscious thought theory is a recent approach in psychology that vindicates the superiority of unconscious processes in solving different tasks. Although it gained a certain momentum during the last decade, a recent meta-analysis and replication study found that the experimental effects observed "concern nothing but spurious effects obtained with an unreliable paradigm" (Nieuwenstein, et al., 2015). The interesting point is that the authors conducting the replications identified in the literature, at least, 12 different general factors (each of which with different sub-specifications) that might contribute to the outcome, not all of them implemented in the designs examined (Strick, et al., 2011). This lack of agreement between the relevant factors contributing to the experimental outcome is probably behind the high statistical heterogeneity detected in the meta-analyses of experimental psychology: the variation observed between treatment effects across studies suggests that there is something other than chance causing it (Stanley, Carter, & Doucouliagos, 2018)

The 45 minutes interaction in the IBS is a bundle of interventions rather than a single distinctive script: its nine items (four sets of questions plus five styles of engaging behaviour) are surely candidates for separate decomposition in further trials, so that the contribution of each sub-factor in the exchange can be assessed separately. The challenge is, of course, to formulate each one of them in a standardized enough manner as to secure repeatability within the experiment (in the multiple period cross-over design discussed before) and replicability of the trial itself. Short of a theory to rely on for the standardization of these items, we may wonder how placebo research would succeed in standardizing these scripts where many other psychological disciplines seem to have failed.

An immediate rejoinder would be denying the necessity of any further standardization: if the 45 minutes interaction triggers the effect, maybe there is nothing else to decompose: if this long script is enough to augment the therapeutic effect of the needles in a replicable manner, that is enough for consolidating the placebo effect as a phenomenon. Yet interactions this long and so lightly standardized risk to trigger a well-known source of bias in the experimental social sciences: the participants' reactivity. A fundamental assumption in the IBS trial is that the interaction script under study will operate in the test just as it would work outside the experiment: the patient's response should not be different in both setups. Yet, the reactivity of the patients to the experimental setup has been amply documented in various fields since the 1940s (Morawski, 2015). In the 1950s, psychologists became increasingly concerned about the "authenticity" of the participants' responses and developed tests to detect their ability to fake them: e.g., giving socially desirable answers instead of manifesting their real beliefs or desires. In addition, researchers increasingly adopted precautions not to influence the subject's performance themselves, testing whether the experimenters' characteristics had any influence on the test outcome (Morawski, 2015, pp. 576-581). Standardizing the experiment's tasks in the experiment became a warrant of objectivity: researchers should be entirely interchangeable in order to make the intervention the only difference-making factor in the experimental setup. Yet, Robert Rosenthal's studies in the 1950s soon alerted about the limits of standardization: the experimenter may be reading his participants identical instructions and still treat the comparison groups differently "in subtle ways" according to his own expectations about the test outcome (Morawski, 2015, p. 588).

From the 1970s onwards, blinding both the experimenters and the participants regarding the goals of the experiment became the default solution to control for biases they may introduce in the outcome. The most radical form of blinding was deception: lying to the participants about the true goals of an experiment (ultimately, about the very fact that they were part of an experiment) (Korn, 1997). Blinding therefore makes up for a lack of standardization.

Can blinding control for the participants' reactivity in a long slightly standardized script such as the augmented interaction in the IBS trial? In clinical trials of pharmacological treatments, blinding is best achieved with, precisely, a placebo externally as similar as possible to the true treatment, but without the active principle. The factor to control for are the participants' preferences about the treatments on trial and

how they relate to the treatment outcomes: e.g., if they don't want the intervention they are actually receiving, these preferences should not have any systematic correlation with the outcome –(Zizzo, 2010); see, for a trace of this effect, (Luparello, Leist, Lourie, & Sweet, 1970). In placebo tests like the IBS trial, the question is whether there is a way to blind whatever preferences' the therapists or the participants may have about the interaction script. As the authors themselves admit, they could not separate the effects of observation and assessment (Kaptchuk, et al., 2008, p. 1003). Let me elaborate on this point.

There are three possible scenarios. First, neither the therapists nor the participants know which interaction script they are having. This would be the perfect analogue of the placebo pill, but this is obviously not possible. The second option would be to deceive the participants about the existence of alternatives, but there are strong arguments for banning deception in the experimental social sciences, and not just for ethical reasons: if the participants suspect they are being deceived, the inferences from the trial outcome might be easily challenged (Ortmann & Hertwig, 2002) (Hersch, 2015). The third option is to do away with the blinding and explicitly disclose the real nature of the scripted interaction, as in the so-called open label placebos, and check for whatever impact this disclosure may have on the outcome[4]. Yet, the only outcome variable in which solid placebo effects have been detected is a subject reported pain score: patients assess in a scale the relief caused by the treatment. This is a variable that is highly sensitive to a number of reactivity-linked biases: e.g., the patient's score may reflect more a desire to please an engaging therapist than an actual reduction in pain (Gracely, Dubner, Deeter, & Wolskee, 1985). A meta-analysis on trials with subject reported outcomes has documented how, if the assessor taking the score is not blinded regarding the treatment administered, the score differs from the one obtained in blinded assessments (Savovic, et al., 2012).

In other words, if instead of decomposing and standardizing the interaction script, placebo researchers opt for keeping it long and complex, given the outcome variable under study, the measurement process may be contaminated by the participants' expectations. It is surely possible to implement additional checks to control for this possibility, but, at this point, we should be aware that the sources of error we are trying

[4] As (Kaptchuk, et al., 2010) admit, in the trial of open label placebos, report biases derived from the therapist-patient interaction are, in principle, impossible to eliminate.

to control for in the decomposition of the IBS trial keep piling up (standardization of the needle treatment and the scripted interventions plus reactivity) which is in itself a reason for pessimism. It is time to take stock and consider some final objections.

5. Concluding remarks

Let us review the whole argument now. I have proposed to take placebo effects as experimental regularities, *phenomena*, until a consensual unified theory explains them all. Of all these potential phenomena, I take that placebo effects measured in RCTs are the most interesting for medical purposes, since those placebos achieving clinical and statistical significance would count as legitimate treatments. Even if its full causal structure is not yet understood, there would be evidence enough about their benefits.

The solid placebo phenomenon acknowledged in the most stringent meta-analysis (Hrobjartsson & Gotzsche, 2010) is the treatment with medical devices of pain measured with patient-reported outcomes. How should placebo researchers invest their resources now? Should they run more trials in search of additional placebo regularities or should they instead invest in basic research before engaging in more experimentation? I have tried to provide an argument for this second alternative, raising some doubts about the possibility of further decomposing the causal structure of placebo interventions in trials. The IBS trial provides an insightful decomposition of the factors contributing to placebo effects, precisely with the sort of intervention and outcome in which it appears as a phenomenon.

If multiple-period cross-over trials are the best methodology for investigating pain treatments, it is dubious that the two factors studied in the IBS trial (needle treatments and interaction scripts) can be standardized with enough precision as to allow for solid reproductions. On the one hand, even the needling with more solid causal foundations have found sustained trouble to standardize the basic points in the intervention, leading to RCTs with high heterogeneity and inconclusive results. On the other hand, the standardization of interaction scripts has proven problematic even in experimental disciplines that exclusively focus on such scripts. The source of the problem is probably in the lack of agreement on the causal factors at play. Blinding, the default debiasing procedure to deal with interfering causes, cannot be properly implemented when the intervention under analysis (the interaction script) needs to be run in the open.

My conclusion is that placebo research should focus on the basic physiological mechanisms underlying placebo in order to find guidance for further experimentation, so that solid phenomena emerge from the tests. This conclusion, like my entire argument, can be proven wrong with just a single (but replicable) experiment documenting a placebo effect with lowly standardized interventions. Philosophy is, obviously, not conclusive.

Let me close with a brief discussion of two global objections to my argument. First, some placebo researchers will radically argue that there is no causal structure to grasp. The anthropologist Dan Moerman has famously objected that experimental interventions with the same structure implemented in two different countries yield different response rates (Moerman, 2002, p. 125), suggesting a contextual dimension that can only be grasped locally investigating the meaning of treatments for each particular group of patients. My argument suggests instead that the variability may arise from the lack of standardization of the interventions. As a matter of fact, in order to find out the part of the variance that emerges from cultural factors, it is a prerequisite to standardize the interventions.

A second objection may be that I adopt unjustifiably high experimental standards: most research on pain is not conducted with multiple-period cross-over designs; most research on needling and interaction scripts in different fields proceeds with rough methodological approaches that nonetheless yield valuable outcomes. RCTs, in particular, may be poorly equipped to deal with the nuances of placebo interventions so it might be worth trying other experimental designs. All this may be the case, but I would not say that any of the fields I have examined has suffered from lack of methodological pluralism –see again (Miller, et al., 2013) about diversity in placebo research. And yet, solid, replicable experimental regularities are not easy to find. As I see it, if placebo testing is guided by more solid hypotheses on the underlying mechanisms and more strict methodological standards, the outcome is more likely to persuade an already sceptical audience (Norton, 2015) than a myriad of experiments of the sort we already see.

Acknowledgements

My most sincere thanks to Phil Hutchinson and the participants in the Placebo Workshop held at Manchester Metropolitan University in April 2016, where this paper was first presented, and to Karina Makhnev, Leen De Vreese, Jeroen van Bouwel, Erik Weber and the audience at the CLPS seminar in Ghent for their objections and suggestions. I am also

grateful to María Jiménez Buedo, José Antonio Pérez Escobar for their comments. Mario Santos Sousa and two reviewers helped me to improve the text substantially. This work was funded by the Spanish Ministry of Science research grant FFI2014-57258-P.

References

Bogen, J., & Woodward, J. (1988). Saving the phenomena. *Philosophical Review, 97*, 303-352.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., & Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science, 351*, 1433-1436.

Collaboration, O. S. (2015). Estimating the reproducibility of psychological science. *Science, 349*.

Cummings, T. M., & White, A. R. (2001). Needling therapies in the management of myofascial trigger point pain: a systematic review. *Arch Phys Med Rehabil, 82*, 986-992.

Dworkin, R. H., McDermott, M. P., Farrar, J. T., O'Connor, A. B., & Senn, S. (2014). Interpreting patient treatment response in analgesic clinical trials: implications for genotyping, phenotyping, and personalized pain treatment. *Pain, 155*, 457-460.

Espejo-Antunez, L., Tejeda, J. F., Albornoz-Cabello, M., Rodriguez-Mansilla, J., de la Cruz-Torres, B., Ribeiro, F., & Silva, A. G. (2017). Dry needling in the management of myofascial trigger points: A systematic review of randomized controlled trials. *Complement Ther Med, 33*, 46-57.

Fanelli, D., Costas, R., & Ioannidis, J. P. (2017). Meta-assessment of bias in science. *Proc Natl Acad Sci U S A, 114*, 3714-3719.

Gervais, R., & Weber, E. (2015). The role of orientation experiments in discovering mechanisms. *Stud Hist Philos Sci, 54*, 46-55.

Gonzalez-Moreno, M., Saborido, C., & Teira, D. (2015). Disease-mongering through clinical trials. *Stud Hist Philos Biol Biomed Sci, 51*, 11-18.

Gracely, R. H., Dubner, R., Deeter, W. R., & Wolskee, P. J. (1985). Clinicians' expectations influence placebo analgesia. *Lancet, 1*, 43.

Guala, F. (2005). *The methodology of experimental economics*. Cambridge; New York: Cambridge University Press.

Hersch, G. (2015). Experimental economics' inconsistent ban on deception. *Studies in History and Philosophy of Science Part A, 52*, 13-19.

Howick, J. (2017). Measuring placebo effects. In M. Solomon, J. Simon & H. Kincaid (Eds.), *The Routledge Companion to Philosophy of Medicine* (pp. 134-143). London: Routledge.

Hrobjartsson, A., & Gotzsche, P. C. (2010). Placebo interventions for all clinical conditions. *Cochrane Database Syst Rev*, Cd003974.

Jiménez-Buedo, M. (2015). The Last Dictator Game? Dominance, Reactivity, and the Methodological Artefact in Experimental Economics. *International Studies in the Philosophy of Science, 29*, 295-310.

Kaptchuk, T. J. (2002). The placebo effect in alternative medicine: can the performance of a healing ritual have clinical significance? *Ann Intern Med, 136*, 817-825.

Kaptchuk, T. J., Friedlander, E., Kelley, J. M., Sanchez, M. N., Kokkotou, E., Singer, J. P., Kowalczykowski, M., Miller, F. G., Kirsch, I., & Lembo, A. J. (2010). Placebos without deception: a randomized controlled trial in irritable bowel syndrome. *PLoS One, 5*, e15591.

Kaptchuk, T. J., Kelley, J. M., Conboy, L. A., Davis, R. B., Kerr, C. E., Jacobson, E. E., Kirsch, I., Schyner, R. N., Nam, B. H., Nguyen, L. T., Park, M., Rivers, A. L., McManus, C., Kokkotou, E., Drossman, D. A., Goldman, P., & Lembo, A. J. (2008). Components of placebo effect:

randomised controlled trial in patients with irritable bowel syndrome. *Bmj, 336*, 999-1003.

Kerr, C. E., Shaw, J. R., Conboy, L. A., Kelley, J. M., Jacobson, E., & Kaptchuk, T. J. (2011). Placebo acupuncture as a form of ritual touch healing: a neurophenomenological model. *Conscious Cogn, 20*, 784-791.

Kienle, G. S., & Kiene, H. (1997). The powerful placebo effect: fact or fiction? *J Clin Epidemiol, 50*, 1311-1318.

Kirsch, I., & Rosadino, M. J. (1993). Do double-blind studies with informed consent yield externally valid results? An empirical test. *Psychopharmacology (Berl), 110*, 437-442.

Korn, J. H. (1997). *Illusions of reality : a history of deception in social psychology*. Albany: State University of New York Press.

Lavoisier, A. (1865). "Sur le magnétisme animal.". In A. Lavoisier (Ed.), *Oeuvres de Lavoisier, vol. II* (pp. 499-527). Paris: Imprimerie Impériale.

Lotsch, J., & Geisslinger, G. (2006). Current evidence for a genetic modulation of the response to analgesics. *Pain, 121*, 1-5.

Lucas, N., Macaskill, P., Irwig, L., Moran, R., & Bogduk, N. (2009). Reliability of physical examination for diagnosis of myofascial trigger points: a systematic review of the literature. *Clin J Pain, 25*, 80-89.

Luparello, T. J., Leist, N., Lourie, C. H., & Sweet, P. (1970). The interaction of psychologic stimuli and pharmacologic agents on airway reactivity in asthmatic subjects. *Psychosom Med, 32*, 509-513.

Madsen, M. V., Gotzsche, P. C., & Hrobjartsson, A. (2009). Acupuncture treatment for pain: systematic review of randomised clinical trials with acupuncture, placebo acupuncture, and no acupuncture groups. *Bmj, 338*, a3115.

McDonald, C. J., Mazzuca, S. A., & McCabe, G. P., Jr. (1983). How much of the placebo 'effect' is really statistical regression? *Stat Med, 2*, 417-427.

Miller, F. G., Colloca, L., & Kaptchuk, T. J. (2013). *The Placebo: A Reader*: Johns Hopkins University Press.

Moerman, D. E. (2002). *Meaning, medicine, and the "placebo effect"*. Cambridge ; New York: Cambridge University Press.

Mogil, J. S. (2012). Pain genetics: past, present and future. *Trends Genet, 28*, 258-266.

Morawski, J. (2015). Epistemological Dizziness in the Psychology Laboratory: Lively Subjects, Anxious Experimenters, and Experimental Relations, 1950–1970. *Isis, 106*, 567-597.

Myburgh, C., Larsen, A. H., & Hartvigsen, J. (2008). A systematic, critical review of manual palpation for identifying myofascial trigger points: evidence and clinical significance. *Arch Phys Med Rehabil, 89*, 1169-1176.

Nieuwenstein, M. R., Wierenga, T., Morey, R. D., Wicherts, J. M., Blom, T. N., Wagenmakers, E.-J., & van Rijn, H. (2015). On making the right choice: a meta-analysis and large-scale replication attempt of the unconscious thought advantage. *Judgment and Decision Making, 10*, 1.

Norton, J. D. (2015). Replicability of Experiment. *Theoria: Revista de Teoría, Historia y Fundamentos de la Ciencia, 30*, 229-248.

Ortmann, A., & Hertwig, R. (2002). The Costs of Deception: Evidence from Psychology. *Experimental Economics, 5*, 111-131.

Price, D. D., Milling, L. S., Kirsch, I., Duff, A., Montgomery, G. H., & Nicholls, S. S. (1999). An analysis of factors that contribute to the magnitude of placebo analgesia in an experimental paradigm. *Pain, 83*, 147-156.

Quintner, J. L., Bove, G. M., & Cohen, M. L. (2015). A critical evaluation of the trigger point phenomenon. *Rheumatology (Oxford), 54*, 392-399.

Savovic, J., Jones, H. E., Altman, D. G., Harris, R. J., Juni, P., Pildal, J., Als-Nielsen, B., Balk, E. M., Gluud, C., Gluud, L. L., Ioannidis, J. P., Schulz, K. F., Beynon, R., Welton, N. J., Wood, L.,

Moher, D., Deeks, J. J., & Sterne, J. A. (2012). Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials. *Ann Intern Med, 157*, 429-438.

Simons, D. G., Travell, J. G., Simons, L. S., & Travell, J. G. (1999). *Travell & Simons' myofascial pain and dysfunction : the trigger point manual* (2nd ed.). Baltimore: Williams & Wilkins.

Spanos, A., & Mayo, D. (2015). Error statistical modeling and inference: Where methodology meets ontology. *Synthese, 192*, 3533-3555.

Stanley, T. D., Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychol Bull, 144*, 1325-1346.

Strick, M., Dijksterhuis, A., Bos, M. W., Sjoerdsma, A., van Baaren, R. B., & Nordgren, L. F. (2011). A Meta-Analysis on Unconscious Thought Effects. *Social Cognition, 29*, 738-762.

Teira, D. (2016). Debiasing Methods and the Acceptability of Experimental Outcomes. *Perspectives on Science, 24*, 722-743.

Tough, E. A., White, A. R., Cummings, T. M., Richards, S. H., & Campbell, J. L. (2009). Acupuncture and dry needling in the management of myofascial trigger point pain: a systematic review and meta-analysis of randomised controlled trials. *Eur J Pain, 13*, 3-10.

Tough, E. A., White, A. R., Richards, S., & Campbell, J. (2007). Variability of criteria used to diagnose myofascial trigger point pain syndrome--evidence from a review of the literature. *Clin J Pain, 23*, 278-286.

Wartolowska, K., Collins, G. S., Hopewell, S., Judge, A., Dean, B. J., Rombach, I., Beard, D. J., & Carr, A. J. (2016). Feasibility of surgical randomised controlled trials with a placebo arm: a systematic review. *BMJ Open, 6*, e010194.

Wartolowska, K., Judge, A., Hopewell, S., Collins, G. S., Dean, B. J., Rombach, I., Brindley, D., Savulescu, J., Beard, D. J., & Carr, A. J. (2014). Use of placebo controls in the evaluation of surgery: systematic review. *Bmj, 348*, g3253.

Zhou, K., Ma, Y., & Brogan, M. S. (2015). Dry needling versus acupuncture: the ongoing debate. *Acupunct Med, 33*, 485-490.

Zizzo, D. J. (2010). Experimenter demand effects in economic experiments. *Experimental Economics, 13*, 75-98.