Putting Theory-Ladenness to the Test

Ioannis Votsis (ioannis.votsis@nchlondon.ac.uk / i.votsis@lse.ac.uk)
Philosophy Faculty, New College of the Humanities, 19 Bedford Square, London WC1B 3HH, UK
Department of Philosophy, London School of Economics, Houghton Street, London WC2A 2AE, UK

Abstract

This paper explores two experiment designs that seek to determine the extent to which, if at all, observation can be free from theory. The two designs are compared and found to be similar in certain ways. One particular feature critical to both is that they seek to create conditions that compel test subjects with diverse theoretical backgrounds to resort to bare observational skills. If judgments made on the basis of these skills converge, such convergence would provide support for the view that theory-neutral observations can be had.

Keywords: theory-ladenness; cognitive penetrability of perception; observation reports; perceptual beliefs.

Introduction

Genuine tests of theories crucially rest upon the veridicality of observation reports. Such reports can only be said to confirm or disconfirm a theory if they truthfully represent certain things about the world as it is independently of our conceptions of it. This is precisely the kind of claim that several advocates of the theory-ladenness thesis deny. Put simply, the thesis holds that theory influences the content of such reports to the point of distortion. In this paper, we compare two experiment designs that seek to determine the extent to which, if at all, observation can be free from theory.

The paper is structured as follows. In the ensuing section, we consider different versions of the theory-ladenness thesis. Immediately after that, we explore an experiment design proposed by Schurz (2015). The aim of this design is to determine the extent to which, if at all, observation can be free from theory. We then turn to my own proposal for an experiment design: the stimulus exchange procedure. Finally, in the last section the two designs are compared and found to share some vital features. One such feature is that they both attempt to force test subjects in a position where they have to rely on their bare observational skills. If judgments made on the basis of these skills converge, such convergence would provide support for the view that theory-neutral observations can be had.

The Many Guises of Theory-Ladenness

There isn't just one theory-ladenness thesis but many. Whenever we speak of theory-ladenness in the indefinite we mean something general like a schema that applies to a whole family of such theses. Understood thus, theory-ladenness has two variables. One takes as values those things that effect the change. We can call this 'the input'. And the other takes as values those things that absorb the change. We can call this 'the output'. Up to now, we have been using theory as a value to the first and observation reports as a value to the second. Clearly, this approach does not do justice to the various versions of the theory-ladenness thesis out there. Besides good old fashioned theories, the input has been variably interpreted to include one or more of the following: linguistic frameworks, conceptual schemes, prior beliefs, factors relevant to sensory physiology and even environmental cues. And besides observation reports, the output has been similarly interpreted in a variety of ways so as to include one or more of the following: sense-data, perception, experience, observational judgment and empirical data.[1] Whether the resulting theses are indeed substantively different is not immediately obvious but depends on the specific interpretation of the relevant concepts. Even so, let me at least try to demonstrate some of these differences with a couple of examples.[2]

Take a theory-ladenness thesis whose input is linguistic frameworks and whose output is observation reports. This is sometimes called 'the language-relativity of observation'. The linguistic framework one uses affects the way they report what they observe. Clearly, reports can only be as detailed as our language allows. So, if one has a very poor language, say a language that contains only two terms for colours, then any observational report concerning the colour of an object will have to be accordingly confined. Similar remarks apply if we opted for a theory-ladenness thesis that takes as input conceptual schemes. This would result in the conceptual-relativity of observation. Poor conceptual schemes presumably affect the content of observation reports as much as linguistic frameworks do. Whether these two versions of the theory-ladenness thesis, the linguistic and the conceptual, are truly distinct depends on how we understand the relation between language and thought. If the two are inseparable, e.g. if linguistic frameworks just mirror conceptual schemes, then the two theses reduce to one. If, however, there is some divergence between the two, then the two theses preserve their autonomy.

Now take factors relevant to sensory physiology as the input and perception as the output. Consider what would happen to perception if the channels through which perceptual processing is made were substantially different, either from birth or because of subsequent changes. We are all familiar with cases of colour-blindness – see, for example, Zeki (1990). The cause of this condition may be either genetic or acquired. Affected areas vary and may include one or more of the following: cone cells, the optic nerve and parts of the brain, e.g. the *ventromedial occipital*

[1] Items like observational reports and empirical data need not be sourced from lone individuals but may instead be sourced from scientific groups or instruments.

[2] For an in-depth survey and classification of the various kinds of theory-ladenness, see Brewer (2012).

lobe. Take two individuals, one with colour-blindness and one without. Some objects will appear identical in colour to the colour-blind individual but will appear distinct to the individual without the condition. Beyond colour-blindness, there are also less publicised conditions where factors in sensory physiology play a big role in determining our perceptual content. Take prosopagnosia, the neurological disorder that impairs our ability to recognise faces – see, for example, Towler, Fisher and Eimer (2016). Like colour-blindness, it is either congenital or acquired, e.g. through injury. One of the affected areas of the brain appears to be the *fusiform gyrus*. This area helps to coordinate, among other things, facial perception and memory. Both colour-blindness and prosopagnosia involve what are typically characterised as 'impairments' to the normal functioning of perception. But one can easily imagine individuals with 'enhanced', as opposed to 'impaired', colour- or face-detection abilities. Such individuals would also deviate from the aforesaid norm. Since the veridicality of perception is a point of contention in the philosophical literature, we can put aside any judgments that such individuals are either impaired or enhanced and merely note that sensory physiology differences affect their perceptions.[3] That's precisely what is needed to understand what a version of the theory-ladenness thesis that takes sensory physiology as the input and perception as the output involves.

It may be argued that the sheer diversity of all of the inputs and outputs makes the use of one umbrella term, i.e. 'theory-ladenness', to capture their interactions unwise. The reason why I want to resist such an argument is that, their diversity notwithstanding, all of these interactions have the potential to undermine the neutrality of scientific theory testing. For if we assume that the outputs, to the extent that they are distinct kinds, are related stages on the path from stimulus to observational reports, then it is not unreasonable to maintain that any change effected early on in that path is likely to be preserved downstream. To give a toy example, if the presence of a prior belief can somehow distort what we perceive, then it is not likely that our observational judgments or reports are going to nullify that distortion.

It is not hard to show how the debate over the cognitive penetrability of perception can be related to the current discussion. Roughly speaking, those advocating the cognitive penetrability thesis claim that cognitive states, e.g. beliefs, affect perceptual states. The bringing together of these two debates, theory-ladenness and cognitive penetrability, goes back some time and has been explored in a number of works – see, for example, the introductory chapter in Zeimbekis and Raftopoulos (2015). As is well-known, even before the advent of the cognitive penetrability debate, results in the psychological study of perception were harnessed to promote specific viewpoints within the philosophy of science. Thus, Feyerabend, Hanson and Kuhn made use of studies from Gestalt and New Look psychologies to argue for the claim that observations in science are tainted by theory.

The connection between the psychology of perception and the philosophy of science was reinforced in the early 1980s, as the emergence of the cognitive penetrability debate coincided with a renewed discussion about the possibility of theory-neutral observation. Indeed, Fodor (1984), who is one of the founders of the cognitive penetrability debate, is unequivocal about this connection. He argues that were perception to be cognitively penetrable, observation would not be able to occupy the role of neutral adjudicator in science. In his own words: "The main contention of this paper is that there is a theory-neutral observation/inference distinction; that the boundary between what can be observed and what must be inferred is largely determined by fixed, architectural features of an organism's sensory/perceptual psychology" (1984, p. 25). The reasons for Fodor's defence of the objectivity of science are well-documented so I will not dwell on them here. Suffice it to say that he endorses the view that various brain systems, including perception, are modular and hence are impervious to outside influences. Being modular, the integrity of perceptual processing is thus safeguarded. Another way of expressing roughly the same thought is that top-down cognitive processes have little to no effect on bottom-up perceptual processes.[4]

Fodor targets those who have pushed Feyerabend's, Hanson's and Kuhn's claims to their social constructivist extreme. That is, he targets claims to the effect that observation states (or states denoted by cognate notions) are cognitively malleable so much so that even in cases of convergent judgments, the convergence can be explained away as nothing more than the result of social negotiation. The implication of course being that observation cannot reflect any aspect of the world as it is independently of us. Although Kuhn appears to want to deny such radical constructivist interpretations (see the 'Postscript' in Kuhn 1970), a number of his pronouncements can't help but fuel them. When comparing experts to non-experts, for example, he asserts that "… viewing a cloud chamber [the expert] sees (here literally) not droplets but the tracks of electrons, alpha particles, and so on..." (1970, p. 197). In other words, the expert sees the world differently to the non-expert. And the same point is made in relation to experts belonging to different paradigms. Moreover, Kuhn seems to suggest that there is no such thing as one world being observed when he asserts that "[p]racticing in different worlds, the two groups of scientists see different things when they look from the same point in the same direction" (1970, p. 150).

[3] One referee rightly noted that psychologists and philosophers have different reactions on this matter. Moreover, they noted that psychologists commonly take perception to be partially constructive – think of perceptual constancy. Here I merely wish to add that such construction is not incompatible with veridicality.

[4] Those who argue against Fodor sometimes point out that there are top-down perceptual processes that penetrate low-level perception – think of the memory colour effect. Though not engaging with Fodor directly – after all, Fodor restricts his claims to top-down cognitive processes – the threat that such cases pose to veridicality is still palpable.

Returning to the topic of importing experimental results from psychology into the philosophy of science, it would be foolish to deny that these results teach us something about the limits of cognition and perception. Isn't this fact a ringing endorsement that cognitive effects on perception are widespread and hence that some version of the theory-ladenness thesis holds? No, not exactly. One potential explanation why so many experiments turn up this way is that they make for sensational news. Allow me to explain. Those who conduct such experiments are acutely aware that it is more headline-grabbing to find cases where cognitive differences lead to divergent rather than convergent perceptual judgments. To be clear, the worry is not that the studies are somehow fraudulently manipulated to produce the desired results but rather that there is a kind of bias that favours the performing (and publication) of the former instead of the latter kind of studies. After all, we must not forget that there is a strong incentive for journals, especially leading ones, to publish studies with dramatic and/or unexpected results – see Young, Ioannidis and Al-Ubaydli (2008). Less cynical explanations may be more appropriate here and I would certainly not want to exclude them. The key point in all of this is that we should at the very least be careful in what we conclude from such studies.

It is also important not to exaggerate the scope of these studies. Many experiments done in labs tend to impose conditions, e.g. short time-intervals between priming a subject with a specific belief and asking them to make a perceptual judgment, that cannot be said to faithfully reflect conditions present in the real world. As Brewer and Lambert note, stimuli in such experiments are ''either ambiguous, degraded, or requir[e] a difficult perceptual judgment'' (2001, p. 179). Indeed, it's not at all easy to design the kinds of ambiguous shapes, e.g. Leeper's famous young/old woman figure, found in many studies. Otherwise put, if the kinds of shapes and, more generally, the conditions under which such studies are conducted are uncommon outside of the psychology lab then there is less reason to fret about outputs like perceptual judgments being distorted.

The cognitive penetrability debate, as it is conducted today, is largely concerned with the level at which such cognitive effects take place. There are those, like MacPherson (2012), who argue that cognition affects perception itself, not just perceptual judgment. On this view, cognition doesn't just show up at the level of interpreting what we have experienced but penetrates deep into perceptual processing. But there are also those, like Lyons (2011), who suggest that cognition's effects are typically more shallow, e.g. targeting perceptual judgment alone. Finally, there are those who are getting exasperated with the lack of progress in the debate. Machery (2015) expresses this sentiment by arguing that the various experiments utilised on either side are unable to fix the location of cognitive penetration. Though it would clearly be invaluable to know how deep cognition penetrates, it does not really matter for the purposes of this paper. So long as the effects are likely to be preserved downstream, it makes no difference to the theoretical neutrality of observation reports if they appear early or later. For even if such effects penetrate all the way to early vision but happen very infrequently and do not distort the incoming structure of the stimuli substantially, then scientists employing observation reports downstream have nothing specific to worry about. Conversely, if such effects do not penetrate early vision – see, for example, Raftopoulos (2014) who makes a compelling case for this view – but nonetheless happen regularly and with severity, then scientists have something specific to be concerned about. For these reasons, and unless otherwise noted, this paper will put the otherwise very important issue of the locus of penetration to one side.

The Ostensive Learnability Criterion

In the remaining sections, we examine two experiment designs that seek to determine the extent to which, if at all, observation can be free from theory. We first turn to a design that originates in Schurz (2015). Some preliminary remarks are in order. Schurz concedes that various observations are theory-laden (or as he calls them 'theory-dependent'). Even so, he indicates that "the existence of observations that are weakly theory-neutral in the sense that they don't depend on acquired background knowledge" may still be possible (2015, p. 139). To find out whether this is the case, he proposes a criterion whose purpose is to decide whether a given concept is theory-neutral or theory-laden. His focus on concepts is deliberate. Concepts are the basic constituents of propositions. If what we are after are observational propositions that are theory-neutral and therefore apt for the purposes of adjudicating between rival theories then such propositions must surely have as constituents theory-neutral concepts.

In his search for a criterion that would enable us to discriminate between theory-laden and theory-neutral concepts, Schurz imposes a number of conditions. Any such criterion must: (a) distinguish theory-neutral from theory-laden concepts along the lines of human sensorial capacities, (b) itself be empirically testable and (c) not rely on culturally specific verbal behaviour. He then goes on to propose a criterion that he claims satisfies these conditions. He calls it 'the ostensive learnability criterion' and provides an experimental framework within which this criterion can be put to the test. The experiment has two phases: a training and a testing phase. In the training phase, a number N of made-up terms t_i (where $i \in N$) each denoting some distinct concept C_i is introduced to the test subjects. The concepts are made-up so as to not evoke any meaning associations. Each time a new such term is introduced, the experimenter presents the test subjects with a small number of positive and negative instances of the corresponding concept using ostension and simple expressions like 'This is a t_i!' or 'This is not a t_i!'. An instance being positive or negative is predetermined by the experimenters. The instances may be concrete objects, videos or photos. Normal observation conditions must hold across all the instances. At the end of the training phase the test subjects are expected to have

extracted a concept. In the testing phase, a new set of positive and negative instances is presented to the test subjects. This time no identification is made about which instances can or cannot be classed under the given C_i. Instead, the question is asked: Is this an instance of t_i? The test subjects reply with a 'Yes' or 'No'. The variables measured include the individual success rate, i.e. the number of instances classified correctly by each test subject divided by the total number of instances, as well as the learning curve, i.e. how quickly (if at all) the test subjects reach a success rate threshold – say 90%.

What are we meant to gather from these variables? Schurz reasons that concepts learned swiftly by "(almost) all" the test subjects, regardless of their cultural, linguistic and theoretical background, are theory-neutral. Thus, an implicit assumption is that the test subject population must be varied. Conversely, concepts that take time to be learned or cannot be learned at all are deemed theory-laden. The rationale is quite simple. The training phase utilises a teaching method, namely ostension, which does not involve the *communication* of theoretical prejudices but rather the mere sensorial exposure of test subjects to a small number of positive and negative instances of concepts. Their ability to learn these concepts is thus a testament to those sensorial abilities. If the majority of the test subjects reach some high success rate threshold quickly, it is not unreasonable to suggest that cultural, linguistic or theoretical baggage does not interfere with their perceptual judgments. Hence, the concepts learned are not likely to be theory-laden. To put this into context, the theory-ladenness thesis targeted by this criterion is one where cultural, linguistic and theoretical backgrounds affect the content of concepts.

As Schurz stresses, his view does not imply that all cultures have the same observation concepts. Rather, it at best implies that cultures can acquire all theory-neutral observation concepts through ostensive learning. This sentiment is captured in the two definitions he proposes:

"Definition 1: A concept φ is a *theory-neutral* observation concept (or an observation concept i.n.s. [in the narrow sense]) iff almost all humans can acquire this concept in an ostensive learning experiment, under normal observation conditions, independently of their background information, language and culture" (p. 151) [original emphasis].

"Definition 2: A concept is *the less theory-dependent* (or the more theory-independent), the more humans of a representative sample with mixed cultural background can acquire φ in an ostensive learning experiment, and the faster they can acquire φ" (p. 152) [original emphasis].

As the second definition makes clear, any given concept may be more or less theory-laden. Thus, on this account, theory-ladenness comes in degrees and should reveal itself through the extent to which the learning of concepts is delayed (or even grinds to a halt) in test populations with subjects from various backgrounds.

The last thing we need to consider in this section is how exactly the ostensive learnability criterion is meant to meet the foregoing adequacy conditions. Take the first condition. Is this criterion able to distinguish theory-neutral from theory-laden concepts along the lines of human sensorial capacities? Presumably yes, as learning how to make correct classification judgments through ostension involves using one's senses. What about the second condition? Is the criterion empirically testable? Once again, there is good reason to answer in the affirmative as judgment convergence it not determined a-priori. For example, there is no guarantee that almost all test subjects will learn what we intuitively deem as theory-neutral concepts fast. Finally, the third condition asks that the criterion not rely on culturally specific verbal behaviour. This is presumably achieved by putting almost all the weight of the experiment on ostension. Moreover, language use is restricted to simple expressions like 'This is a t_i!', 'Is this an instance of t_i?' and 'Yes/No'. Presumably these are the kinds of expressions that are likely to be found in all cultures. This concludes our introduction to the ostensive learnability criterion.

The Stimulus Exchange Procedure

In this section, I propose the design of an experiment whose aim is to determine whether differences in the observational judgments of experts vs. laypersons, where these exist, disappear under controlled conditions. Clearly, if that were the case at least sometimes, theory-ladenness of this sort, i.e. where theoretical beliefs affect observational judgments, would pose less of a threat to the objectivity of those judgments and the corresponding reports.

Consider an image of what are presumably cellular details of organic matter taken with a scanning electron microscope (SEM). Were an expert to produce an observational report of this image, they would identify several rich features, e.g. the structure of the nucleus, the mitochondrion and the endoplasmic reticulum. Their report would thus be laden with theoretical descriptions from the field of cellular biology. A layperson or non-expert, by contrast, would have no such theory to fall back on, though they may certainly infuse their observational reports with some theoretical descriptions of their own. No assumption need be made here that the relevant theoretical beliefs of the expert, or indeed the non-expert, distort the content of the observation reports or judgments, though they may very well do. All that matters for our purposes is that those reports or judgments differ on account of the distinct theoretical backgrounds possessed by experts and non-experts.

Is there some layer of content in those reports or judgments that is impervious to theory? By design, the proposed experiment is intended to bring about a set of primitive or basic observation conditions that allows experts and non-experts to leave their theory behind; as much as such theory can be left behind of course. If that's possible, then, under such conditions, we should expect to find agreement between the resulting 'raw' observation reports or judgments of the two groups. This would be tantamount

to a demonstration that at least some theory-laden effects can be removed and hence that the specific theory-ladenness thesis may not be as menacing a threat as first thought.

Let us describe the proposed experiment. Take a number of experts from the same scientific field and an equal number of laypersons who meet the following preconditions: they possess (i) normal visual perception and (ii) decent drawing skills. Say there are ten such experts and hence ten non-experts. These will make up our test subjects. Ask the experts to select twelve instrument-produced images, e.g. SEM images, from their field. Of these images, the experts should deem four of them as strongly dissimilar (call this 'collection A'), four as moderately dissimilar (call this 'collection B') and four as weakly dissimilar (call this 'collection C'). Allow me to explain. Any one image in collection A should be quite easy for an expert to discriminate from another image in the same collection. Similarly, any one image in collection B should be moderately difficult for an expert to discriminate from another image in the same collection. And, finally, any one image in collection C should be quite difficult (yet still feasible) for an expert to discriminate from another image in the same collection. Now, ask both the experts and laypersons to each draw a faithful, i.e. no detail spared, reproduction of all twelve images by hand. Gather all the resulting drawings together, digitise them using a high-resolution scanner and present the digitised images of the drawings to each individual in a random order on a computer screen. Ask each individual to judge (in isolation) which digitised images of the drawings are similar to which original images. According to the numbers assumed above, there should be 240 digital images of drawings in total. Each test subject should thus match each of the original images with twenty digital images of drawings. Their choices will then be recorded and the data statistically analysed.

Let us call this experiment design the 'stimulus exchange procedure' on account of the fact that the stimuli associated with the drawings get exchanged between the test subjects. What could such an experiment show? To the extent that the classification judgments of experts and non-experts are highly convergent across one or more collections, it is reasonable to conclude that the two groups recognise the same patterns of features and hence make the same observational judgments. For how could their observational judgments turn out to be substantially different if each one matches images to drawings made by 19 others – some experts and some non-experts – in at least approximately the same way as those others do? The stimulus exchange procedure is designed to create conditions that permit test subjects to decouple their observational judgments from their theoretical background. If successful, such decoupling would mean that the latter cannot distort the former and hence the former's veridicality can be safeguarded.

Comparing the Two Designs

Let us begin with features that are not shared. One of them is concept-dependence. The ostensive learnability design asks test subjects to make judgments on the basis of concepts learned in a training phase. This dependence is absent from the stimulus exchange design as no concepts are expected to be extracted from the assigned task. The latter design may thus be seen as having an advantage over the former. That's because the possibility of constructing a concept that (either deliberately or inadvertently) is not sufficiently detached from concepts already known to the test subjects does not even arise in the case of the stimulus exchange design. Take the term 'meran'. Suppose the experimenters decide that it denotes the concept *blood orange*. In Indonesian as well as Malaysian 'merah' means *red*. A test group composed of some subjects with knowledge of either of those two languages is more likely to gravitate towards the intended concept in an accelerated manner. After all, blood oranges have a red-ish hue. To be clear, I don't think that this is an insurmountable problem. The point, rather, is that the very possibility of this problem does not even come up in the stimulus exchange design.

Another feature that the two designs differ on is the supposition of correctness. In the ostensive learnability design, one must assume that the made-up concept has certain positive and negative instances and hence that test subject judgments in relation to these instances are correct or incorrect. This requirement is absent from the stimulus exchange design as the test subjects are only evaluated from the perspective of whether or not their matching of drawings to original images converges, not whether the matching is correct.[5] Once again, I take this as an advantage of the latter design. That's because by assuming that some judgments are correct we open ourselves up to accusations of potential experimenter bias. Generally speaking, the fewer assumptions made, the better.

Yet another feature that is not shared is scope. The ostensive learnability design can be applied to individuals from different cultures, languages and theoretical backgrounds. By contrast, the stimulus exchange design is more restricted in scope as it targets differences between experts and non-experts. Here I find in favour of the former design as it is more flexible and can potentially determine neutrality in a greater variety of contexts. This finding is, however, preliminary. That's because there is a way to liberalise our understanding of expertise that allows the scope of the stimulus exchange design to be expanded. If by expertise we mean simply the possession of any particular theoretical background, then any individual not in possession of that background counts as a non-expert and hence can be compared to the individual who is indeed in possession of it. To illustrate: Take two individuals α, β with different religious beliefs, say B_α and B_β respectively. Individual β is a non-expert in relation to α's theoretical background, namely B_α, and individual α is a non-expert in relation to β's theoretical background, namely B_β. Understood thus, the distinction between experts and non-

[5] Having said this, it is reasonable to assume that convergence is less likely to occur without correct matching.

experts enables a broad range of comparisons to be made; perhaps as broad as those that can be performed in the ostensive learnability design.

What about features shared by the two designs? Let's start with the obvious. Both make use of visual stimuli. In the ostensive learnability design, these are photos, videos or concrete objects. In the stimulus exchange design, they are instrumentally-produced images from a scientific field and drawings thereof. Another shared feature is that the two designs approach the problem with a classification task. In the ostensive learnability design, test subjects are asked to decide whether a photo, video or concrete object is an instance of a concept. In the stimulus exchange design, test subjects are asked to match drawings to the original images. Indeed the classification task in both designs crucially depends on the ability to make similar judgments. For example, to determine whether an object given in the test phase of the ostensive learnability experiment instantiates a concept, one needs to compare how similar that object is to those presented in the training phase. Likewise, to match a drawing to one of the original images in the stimulus exchange experiment a similarity judgment is necessary.

The most critical feature shared by both designs is by far the eliciting of observational judgments under a set of primitive observation conditions. We have already seen how this manifests in the stimulus exchange procedure. The ostensive learnability criterion does something similar. It attempts to put test subjects in a situation where they leave theory behind, e.g. by presenting them with made-up terms and by asking for simple 'Yes' and 'No' answers to visual tasks. In both designs, the hope is that the said conditions will compel test subjects with diverse theoretical backgrounds to fall back on their bare observational skills. Convergence of judgment, or lack thereof, under such conditions would thus be most telling. A positive evaluation to the question whether theory-neutral observations can be had is borne out of such judgment convergence. In the ostensive learnability design, this takes the form of how swiftly (if at all) the test subjects reach a certain success rate threshold in the acquisition of a novel concept. In the stimulus exchange design, this takes the form of the closeness of expert and non-expert perceptual judgments.

Let me bring this paper to a close by considering an objection to the very idea that underwrites the aspirations of these experiments, viz. that we can potentially learn something about theory-ladenness theses through careful design. It may be argued that if observations are theory-laden then we cannot use the observations of a proposed experiment to support (or refute) the theory at hand – in this case, the theory that theory-laden effects are ever-present and severe. Although clever-sounding at first, this kind of objection is ill-conceived and, indeed, self-undermining. That's because the main motivation for the view that observation is theory-laden to the extent that its objectivity flies out of the window is precisely the aforementioned experimental results in psychology. If we were to start doubting the veridicality of observations wholesale, then we would have to deny the validity of those experimental results and, as a consequence, remove the most powerful motivation we have for placing our trust in theory-ladenness theses. Put another way, to employ such results in support of theory-ladenness theses one needs to endorse their veridicality. Moreover, note that there is more than a whiff of unfalsifiability to the claim that no experiment can conceivably chip away at theory-ladenness. If the thesis is not open to the *possibility* of empirical refutation, how are we to choose between it and the countless other alternatives, including the extreme opposite (and in my view also mistaken) position that all observations are veridical. Logic alone is clearly not the answer. That's why the current paper explores the two experiment designs. We need to find the most unobjectionable way to determine the extent to which veridicality can be preserved.

References

Brewer, W. F. (2012). The theory ladenness of the mental processes used in the scientific enterprise. In R. W. Proctor & E. J. Capaldi (Eds.), *Psychology of science: Implicit and explicit processes.* Oxford: Oxford University Press.

Kuhn, T. S. (1970). *The structure of scientific revolutions.* 2nd ed., Chicago: University of Chicago Press.

Lyons, J. C. (2011). Circularity, reliability, and the cognitive penetrability of perception. *Philosophical Issues*, 21, 289–311.

Machery, E. (2015). In J. Zeimbekis & A. Raftopoulos (Eds.), *The cognitive penetrability of perception: New philosophical perspectives.* Oxford: Oxford University Press.

MacPherson, F. (2012). Cognitive penetration of colour experience: Rethinking the issue in light of an indirect mechanism. *Philosophy and Phenomenological Research*, 84(1), 24–62.

Raftopoulos, A. (2014). The cognitive impenetrability of the content of early vision is a necessary and sufficient condition for purely nonconceptual content. *Philosophical Psychology*, 27(5), 601–20.

Schurz, G. (2015). Ostensive learnability as a test criterion for theory-neutral observation concepts. *Journal for General Philosophy of Science*, 46, 139-153.

Towler, J., Fisher, K. & Eimer, M. (2016). The cognitive and neural basis of developmental prosopagnosia. *The Quarterly Journal of Experimental Psychology*, 70:2, 316-344.

Young, N. S., Ioannidis, J. P. A., & Al-Ubaydli, O. (2008). Why current publication practices may distort science. *PLoS Med*, 5(10), e201.

Zeimbekis, J., & Raftopoulos, A. (Eds.) (2015). *The cognitive penetrability of perception: New philosophical perspectives.* Oxford: Oxford University Press.

Zeki, S. (1990). A century of cerebral achromatopsia. *Brain*, 113(6), 1721-1777.