Synthese
https://doi.org/10.1007/s11229-018-01992-y

S.I. : LOGPERSCICOG

Theory-ladenness: testing the 'untestable'

Ioannis Votsis[1,2]

Received: 20 February 2018 / Accepted: 16 October 2018

Abstract
In this paper, I investigate two potential ways to experimentally test the thesis that observation is theory-laden. One is a proposal due to Schurz (J Gen Philos Sci 46:139–153, 2015) and the other my own. The two are compared and found to have some features in common. One such feature is that both proposals seek to create conditions that compel test subjects with diverse theoretical backgrounds to resort to bare (or at least as bare as possible) observational judgments. Thus, if judgments made under those conditions are convergent across test subjects, the said convergence would lend credibility to the view that theory-neutral observations are feasible. This still leaves the question of why any such convergence exists unanswered. Towards the end of the paper, it is argued that the best explanation for observational judgment convergence is the veridicality of those judgments.

Keywords Theory-ladenness · Cognitive penetrability · Observation reports · Observational judgments · Perceptual beliefs · Perceptual judgments · Social constructivism

1 Introduction

Observation reports are used throughout science to test clusters of theories plus auxiliaries—hereafter just 'theories' for expedience. But for tests to carry real weight, the observation reports must be veridical. That is, they must not just stand in the right inferential relations to theories, e.g. be deductive consequences of theories, but must also truthfully represent certain things about the world. Consider what would happen if we endorsed the view that such reports need not be veridical. The ensuing notion of testability would become severely enfeebled. That's because we would no longer

✉ Ioannis Votsis
ioannis.votsis@nchlondon.ac.uk; i.votsis@lse.ac.uk

[1] Faculty of Philosophy, New College of the Humanities, London WC1B 3HH, UK

[2] Department of Philosophy, Logic and Scientific Method, London School of Economics, London WC2A 2AE, UK

Published online: 28 November 2018

be able to justifiably assert, with any degree of definitiveness, that one theory is better supported than another. Take an observation report O_1 that supports (in this enfeebled manner) a theory T_1 but opposes a rival theory T_2. Suppose that O_1 is not veridical or at the very least we do not know, with any degree of definitiveness, whether it is veridical. How could we justifiably decide in favour of T_1 if a competing observation report O_2 that is also not veridical (or that we do not know with any degree of definitiveness that it is) supports T_2 in the same enfeebled manner but not T_1? There are those who deny that observation reports need to be veridical and, as a consequence, are happy to accept such an impoverished version of the notion of testability. One major motivation for this approach is the thesis that observation reports are theory-laden. In simple terms, theories distort the content of observation reports and hence such content cannot truthfully represent things about the world.[1]

Implicit in the theory-ladenness thesis is the assumption that distinct theories distort the content of observation reports in distinct ways. As a result, the observation reports of individuals endorsing distinct theories diverge in content and hence such reports cannot form a neutral adjudication basis on which theory choices can be justifiably made. Proponents of the thesis thus predict that the observation judgments of such individuals must be divergent. This paper has two aims. The first is to compare two experiment designs that seek to determine the extent to which, if at all, such divergence is present. In other words, the two designs seek to help us ascertain whether observation reports are, or can be, free from theory. Note that even if observation turned out to be strongly theory-free, veridicality would not be guaranteed. For, although verdicality implies observational judgment convergence—i.e. two individuals who both correctly adjudge the same observational situation could not be in disagreement—the same is not true the other way around.[2] In a nutshell, convergence in observational judgments is necessary but not sufficient for their veridicality. The second aim of this paper is to bridge as much of this gap (between veridicality and observational judgment convergence) as possible. This is achieved by providing some general arguments for the view that the most likely explanation for observational judgment convergence is the veridicality of the corresponding observation reports. Alternative explanations in the form of constructivist views are considered and dismissed as inadequate.

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). This design endeavours to determine the extent to which, if at all, observation can be free from theory. We then turn to my own design which seeks to do something similar. In the section that follows, the two designs are compared and evaluated. One shared feature vital to both is that the designs attempt to force test subjects in a position where they have to rely on their bare observational judgments or at least on observational judgments that are

[1] Not all advocates of theory-ladenness are anti-veridicalists. Indeed, some support the view that, even when theory-laden, observation reports are often veridical—see, for example, Maxwell (1962). On this view, theory is thought of as capable of playing a corrective role, making observation reports more (not less) truthful. The theory-ladenness thesis endorsed by such scholars is thus clearly distinct from the one(s) considered in this paper.

[2] This is, of course, a well-known and respected point that has many different manifestations. Here's another: inter-subjective agreement does not imply truth.

as bare as possible. If such judgments converge across test subjects with distinct theoretical commitments, then that convergence would provide support for the view that theory-neutral observations are feasible. In the penultimate section, a case is also made for the claim that observational judgment convergence is best explained by the veridicality of observational judgments and not by constructivist alternatives. The paper concludes with a summary of the main points.

2 Theory-ladenness: the landscape

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 an input value and observation reports or judgments as output values. Clearly, this approach does not do justice to the various versions of the theory-ladenness thesis out there. Besides good old fashioned scientific 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 and judgments, the output has been similarly interpreted in a variety of ways to include one or more of the following: sense-data, percepts, perceptions, experiences and empirical data. 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 the potential differences with a few examples.[3]

Take a theory-ladenness thesis whose input is linguistic frameworks and whose output is observation reports. This is sometimes called 'the linguistic relativity' or 'the language relativity' thesis. That is, 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. This idea is underwritten by *The World Color Survey* (Kay et al. 2009), according to which some colours have corresponding terms in all languages, e.g. terms for black/dark and white/light, but others, e.g. red, can only be found in some.[4] Since the veridicality of observation reports is disputed in the philosophical literature, there is no need to assume here that a rich linguistic framework is better (in some truth-apt way) than a poor one. The colour example can merely be used to demonstrate the broader point, namely that differences in linguistic frameworks influence observation reports. Similar remarks apply if we opted for a theory-ladenness thesis that takes as input conceptual schemes. This would result in

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

[4] The survey also reveals that the order in which such terms populate the vocabulary of a language is partly fixed. For example, if a language has a third colour term, i.e. after black and white, it is for the colour red. The data in the survey was collected from 110 unwritten languages. For a critique of this view on the basis of one such language, namely Candoshi, see Surrallés (2016).

the conceptual-relativity of observation. Poor conceptual schemes presumably affect the content of observation reports as much as poor 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 physiological 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—for more details see 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 cortex*. 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 et al. (2016). Like colour-blindness, it is either inherited 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. As before, 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 differences in sensory physiology have an impact on perceptions.

It may be argued that the sheer diversity of all the inputs and outputs makes the use of one umbrella term, i.e. 'theory-ladenness', unwise. The reason why I want to resist such an argument is that, their diversity notwithstanding, all of these input–output relations threaten the neutrality and truth-probing ability of scientific theory testing. Moreover, note that the threat is even more unified than perhaps first imagined in that the aforesaid relations feed off each other. For if we assume that the outputs, to the extent that they are real and distinct kinds, are linked 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 not likely to be undone downstream. To give an example, if the presence of a prior belief can somehow distort what we perceive, then it is not likely that by the time we form the corresponding observational judgment or report that distortion is going to vanish. Thus, the fact that these linked input–output relations threaten the neutrality and truth-probing ability of scientific theory testing warrants the use of the same umbrella term to describe them.

The debate over theory-ladenness is, and has been for quite some time now, joined at the hip with the debate over the cognitive penetrability of perception—roughly speaking, those advocating the cognitive penetrability thesis claim that cognitive states, e.g. beliefs, affect perceptual states. The crosspollination between these debates has been explored in a number of works—see, for example, the introductory chapter in Zeimbekis and Raftopoulos (2015). Indeed, 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, among others, 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, was unequivocal about this connection. He argued 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" (p. 25). The reasons for Fodor's defence of the objectivity of science are well-documented so we 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 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.[5]

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, indeed 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…" (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" (p. 150).[6] The spirit of this constructivist approach is still

[5] 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.

[6] It is often argued that Kuhn toned down his constructivist views significantly after the publication of the first edition of *The Structure of Scientific Revolutions*. Although some concessions were indeed made—see,

very much alive and kicking today, as demonstrated by the following quotation, which appears in a highly cited constructivist work: "Within social constructionism there can be no such thing as an objective fact… In fact, the very word 'discover' presupposes an existing, stable reality that can be revealed by observation and analysis, an idea quite opposed to social constructionism" (Burr 2015, pp. 9–14).

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? Not quite. We should be careful in what we conclude or how we generalise from such 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 are arguably atypical in the real world. To be clear, it's not that such effects cannot be found in the real world but rather that they may be the exception rather than the rule. As cognitive psychologists 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 all that easy to design the kinds of ambiguous shapes, e.g. Rubin's vase, found in experimental 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 think that outputs like perceptual judgments are always or even frequently distorted.[7]

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 can affect 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 can penetrate deep into perceptual processing. But there are also those, like Zeimbekis (2013), who suggest that cognition's effects are typically more shallow, e.g. targeting perceptual judgment alone.[8] 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 the depth at which cognition penetrates, it does not really matter for the purposes of this paper. So long as the effects are not likely to be undone downstream, it makes no difference to the theoretical neutrality of a final product like an observation report if they appear early or later. Moreover, if such effects penetrate all the way down to early vision but happen very infrequently and do not distort the incoming structure of the stimuli substantially, then scientists employing observation reports downstream

Footnote 6 continued
for example McMullin (1993)—Kuhn remained adamant that science is not in the business of discovering truths.

[7] This is not even to mention that such outputs would have to be distorted in specific ways for the corresponding scientific judgments to go astray.

[8] Firestone and Scholl (2016) survey a number of experiments that claim to have established the cognitive penetrability of perception. They argue that such experiments fall prey to a handful of objections, e.g. the El Greco fallacy. Moreover, they empirically ground their viewpoint by replicating some of the experiments and showing how cognitive penetrability can be explained away.

have nothing specific to worry about. Conversely, if such effects do not penetrate early vision—see, for example, Pylyshyn (1999) and Raftopoulos (2014) who make 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.

3 The ostensive learnability criterion

Schurz (2015) concedes that observations can be theory-laden (or as he calls them 'theory-dependent') in a number of ways. Even so, he indicates that "the existence of observations that are weakly theory-neutral in the sense that they don't depend on acquired [as opposed to innate] background knowledge" may still be possible (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 the said propositions must surely have as constituents theory-neutral concepts.

For a criterion to be suitable, Schurz reasons, it needs to meet three constraints. It must: (a) rely on sensorial capacities inherent to human beings to distinguish theory-neutral from theory-laden concepts, (b) be empirically testable and (c) not rely on culturally specific verbal behaviour. The first constraint is there to make sure that any concepts deemed theory-neutral are indeed concepts whose application is decided purely (or at least predominantly) by sensory abilities we possess qua human beings. Concepts whose application is decided via scientific instruments or indeed theory are thus excluded. For example, concepts like 'magnetic field' should not come out as theory-neutral since we do not possess the ability to (directly) detect such fields with our unaided senses, though other animals may very well be able to do so. On the other hand, concepts like 'round fruit' should come out as theory-neutral since we do possess the ability to (directly) detect such things with our unaided senses. The second constraint is there to ensure that the proposed criterion does not guarantee by fiat or determine a priori the existence of a non-empty set of theory-neutral concepts. If such a set exists, then it should be borne out of the results of an experiment. Finally, the third constraint seeks to avoid language-dependent effects sneaking in through the back door. It does so by requiring that any communication relies as little as possible on language and, where language is unavoidable, on expressions that are universally shared.[9]

Schurz then goes on to propose a criterion, which he calls 'the ostensive learnability criterion', and which he claims satisfies the above constraints. To evaluate whether it does, we need first to understand what the criterion involves. The criterion is embedded within an experimental framework. The proposed experiment has two phases: a training phase 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 made-up concept C_i is introduced to the

[9] Assuming, of course, that such expressions exist. See footnote 11 for some more thoughts on this issue.

test subjects. The terms and 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!’. Whether an instance is positive or negative is something that the experimenters decide in advance and in a coherent fashion. The instances may be concrete objects, videos or photos. Normal observation conditions, e.g. sufficient lighting, are envisaged to hold across all presentations of 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 are or are not classed under the given concept. Instead, the question is asked: Is this an instance of t_i? The test subjects reply with a ‘Yes’ or ‘No’. The dependent variables measured include the individual success rates, i.e. the number of instances classified correctly by each test subject divided by the total number of instances, as well as the learning curves, i.e. how quickly (if at all) the test subjects reach a success rate threshold (e.g. 90%).[10] Both play a very important role in determining a concept’s degree of theory-neutrality.

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, a tacit assumption is that the test subject population must be fittingly varied. Conversely, concepts that take time to be learned or cannot be learned at all are deemed theory-laden. 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. In short, 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 makes sure to stress, 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 italics].

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 italics].

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

[10] A missing assumption in Schurz’s proposal is the minimum number of instances that need to be presented before we start checking whether the success rate threshold is reached.

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 more swiftly a concept is learned, the more theory-neutral that concept is deemed to be.

Does the ostensive learnability criterion meet the three aforementioned constraints? Take the first one. Is this criterion able to distinguish theory-neutral from theory-laden concepts on the basis of sensorial capacities inherent to human beings? Presumably yes, as learning how to make correct classification judgments through ostension involves appealing to output from our in-built sensory organs and nothing else besides. What about the second constraint? Is the criterion empirically testable? Once again, there is good reason to think that the ostensive learnability criterion is indeed empirically testable in the sense outlined above. This is because whether or not the results of the experiment indicate judgment convergence in the test subjects is not something that is guaranteed by fiat or 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 or indeed that any of the concepts will be learned fast. Finally, the third constraint 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’.[11] The rationale behind this approach is that no theoretical prejudices could conceivably be communicated with such expressions. This concludes our outline of the ostensive learnability criterion.

4 The stimulus exchange procedure

In this section, I propose the design of an experiment whose aim is to determine whether differences in theoretical background, specifically those found between experts versus laypersons, are inconsequential to observational judgments under some well-specified conditions.[12] Clearly, if that were the case at least sometimes, the prevalence of theory-ladenness of this sort, i.e. where theoretical beliefs affect observational judgments, would be undermined and hence theory-neutrality bolstered.

Take a scanning electron microscope (SEM) image of what an expert would presumably identify as cellular details of organic material. Were that 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 could infuse their observational reports with some theoretical descriptions of their own. Is there some layer of content in the reports of experts and non-experts 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 theories behind; as much as such theories can be left behind of

[11] Presumably these are the kinds of expressions that are likely to be found in all cultures. That some way to express assent and dissent exists in all natural languages is pretty uncontroversial in empirical linguistics—see, for example, Dryer (2005).

[12] An early version of this proposal can be found in Votsis (2015).

course. If that's possible, then, under those conditions, we should expect to find agreement between the 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. Suppose individuals in both groups meet the following preconditions: they possess (i) normal visual perception, i.e. no visual impairment, and (ii) decent drawing skills.[13] Say there are twenty such experts and hence twenty non-experts.[14] These will make up our test subjects. Ask the experts to collectively 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*). That is to say, 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 as similar as possible to which original images. According to the numbers assumed above, there should be 480 digital images of drawings in total. Each test subject should thus match each of the original images with forty 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 in the images, the drawings and hence the world or at least that any theoretical prejudices related to expertise are kept at bay.[15] For how could their observational judgments turn out to be substantially different if each one matches images to drawings made by 39 others—some experts and some non-experts—in at least approximately

[13] What counts as decent drawing skills is explored in the section that follows.

[14] The sample size chosen is hypothetical.

[15] What about the prospect that drawing itself is a theory-laden activity? After all, being able to draw requires some sort of expertise. In reply, it can be pointed out that one can guard against this kind of theory-ladenness affecting the results of the proposed experiment by including test subjects without any or with little drawing experience but who are still tasked with matching the images and drawings of the other test subjects. If the convergence across all test subjects (including the ones with little to no drawing experience) were to persist, then it would be reasonable to conclude that the said theory-ladenness is either negligible or non-existent.

the same way as those others do? The design of the stimulus exchange procedure creates the kinds of conditions that, if successful, facilitate test subjects to decouple their observational judgments from their theoretical background. This in effect means that, under such conditions, the latter cannot distort the former and hence that one strong reason against the veridicality of observational judgments can be foiled.

Why are the original images split into three collections (strongly dissimilar, moderately dissimilar and weakly dissimilar)? This is to enable us to make more nuanced claims about the extent to which the classification judgments of experts and non-experts converge or diverge. It may be the case, for example, that the expert and non-expert classification judgments align perfectly in relation to collections A and B but diverge in relation to collection C. Recall that the images in collection C are deemed by the experts to be quite difficult to discriminate. If the more difficult it is for experts to discriminate such images, the more they rely on theoretical considerations to make those discriminations, then one might expect that (in the absence of those theoretical considerations) the non-experts are less likely to be able to discriminate those images in the same way as the experts and hence less likely to, for example, match the same images to the same expert drawings as the experts do. More concretely, suppose $c1$ to $c4$ are the four images in collection C. Suppose, moreover, d_{c1}^{En} to d_{c4}^{En} are the corresponding drawings of a given expert, where superscript En stands for the nth expert and the subscript stands for the image used by that expert to produce the specific drawing. Finally, suppose that d_{c1}^{Lm} to d_{c4}^{Lm} are the corresponding drawings of the given layperson, where superscript Lm stands for the mth layperson and the subscript, again, stands for the image used by that layperson to produce the specific drawing. It may turn out that the only significant divergence in the classification judgments of experts and non-experts is in relation to images $c1$ to $c4$. More concretely, experts may match each other's C images and drawings in the same way—e.g. $E1$ may judge that d_{c1}^{E2} is an image of $c1$, that d_{c2}^{E2} is an image of $c2$ and so on—but laypersons may do considerably less well when matching each individual expert's C images and drawings—e.g. $L1$ may judge that d_{c1}^{E2} is an image of $c4$, d_{c2}^{E2} is an image of $c3$ and so on. I hope this explanation demonstrates why there is reason to split the images into three collections.[16]

The above example was not chosen randomly. To the extent that any significant divergence across the observational judgments of the test subjects is measured, it is my conjecture that it is only likely to appear in relation to that collection, i.e. C. Such restricted divergence would support the claim that theory-ladenness of the sort described above permeates some observational judgments but not all. Moreover, depending on the magnitude of this divergence it would provide some indication as to the strength of the theory-ladenness in the case under test. For now, of course, this is just a conjecture. What will in fact happen is not determinable or guaranteed a priori. That's why an actual experiment needs to be conducted.

Before we draw this section to a close, it is worth considering what a convergence between the judgments of experts and non-experts does and does not tell us about the two groups of individuals. Well, it does not tell us that the non-experts can be made

[16] One could, of course, opt for a more fine-grained approach thereby increasing the number of collections. Running such a study would, as a result, become increasingly more demanding as, provided we keep the other variables (i.e. the number of test subjects and the number of images per collection) fixed, each test subject would have to make 160 additional judgments per new collection.

to see what the experts can see. This, doubtless, is what training is for! That is to say, nobody in this debate will deny that non-experts can, through appropriate training, be turned into experts and hence that they can be made to see what the other experts can see.[17] What the convergence does in fact reveal is exactly the opposite. To be exact, the experts can (still) see what the non-experts can see. This ability could not appear out of thin air. The more sensible explanation is that, even after years of training, experts retain the ability to observe the world in ways that are untainted by their theoretical background and indeed shared with the non-experts. This is something that advocates of the theory-ladenness thesis deny.

5 Comparing the two designs

Let us begin the comparison 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 artificial 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 the experimenters 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 artificial term 'meran'. Suppose the experimenters decide that it denotes the concept *blood-coloured and stringy*. In Indonesian as well as Malaysian 'merah' means *red*. A test group composed of some subjects with knowledge of either of those two languages may thus be more likely to gravitate towards the intended concept in an accelerated manner. After all, the two terms, meran and merah, are virtually indistinguishable and blood-coloured things have a red-ish hue. To be clear, I don't think that this is an insurmountable problem for the ostensive learnability design. Even so, this problem does not even come up in the case of the stimulus exchange design since no concept extraction is involved.

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 or not the matching is correct.[18] Once again, I take this to be an advantage of the stimulus exchange design. By assuming that some judgments are correct we open ourselves up to accusations of potential bias, e.g. experimenter bias. Other things being equal, the fewer assumptions made in an experiment, the better.

[17] Having said this, it is mystifying how the theory-ladenness advocate can expect non-experts to become experts without presupposing the veridicality of observation in training. After all, successfully issuing and receiving instructions does not involve mind-reading but observation. See Sect. 6 below for a general argument along these lines.

[18] Still, I am of the view—as I will argue in the section below—that convergence is less likely to occur without correct matching.

A final unshared feature—actually more like a practical consideration—worth considering is the ease with which a design may be implemented. The ostensive learnability design has, as we have already seen, some difficulties in relation to the selection of appropriate test concepts. Beyond these difficulties, however, implementing it should be plain sailing as the test merely involves verbal assent to visual stimuli. The stimulus exchange design, by contrast, imposes the requirement of possessing decent—notice: not good—drawing skills on the test subjects. How exactly this requirement must be fulfilled is not a straightforward matter. One idea is to use some standardised drawing task as a pre-test in order to select test subjects for the main study, making sure that they are all at approximately the same level. For example, one can check hand-drawing steadiness and co-ordination through a tracing task. The task requires test subjects to try to draw a line right in between two versions (one smaller than the other) of the same shape. Any deviation from the middle can then be measured. On the assumption that the distribution is normal, one can then select for participation in the main study only those test subjects whose drawing falls within one standard deviation either side of the mean.[19] This would reflect the expectation that most of us have decent drawing skills. Another idea is to swap the drawing requirement (in both the pre-test and the main study) with a requirement that's effectively equivalent. For example, one could use software to automatically create several morphed versions of the original images and then ask the test subjects to decide which morphed images resemble which original ones the most.[20] Convergence of judgment would again be telling. The more similar the judgments of experts and non-experts, the more likely that they recognise the same patterns of features in the world. To summarise, although the stimulus exchange procedure has some non-trivial implementation hurdles, there are several promising ways to overcome them.

Let us now turn to features shared by the two designs. One feature that at first glance appears to be different, but is in fact shared, is scope. The ostensive learnability design can be applied to individuals of different cultures, languages and theoretical backgrounds. By contrast, the stimulus exchange design seems to be more restricted in scope as it targets differences between experts and non-experts. This being said, the operative notion of expertise can be stretched to accommodate all sorts of differences between individuals. Understood thus, any disparity in scope between the two experiment designs soon disappears. We can illustrate this point with an example of 'cultural expertise'. Take two individuals α, β with different cultural, e.g. religious, beliefs, say sets B_α and B_β respectively. Individual β is a non-expert in relation to α's B_α beliefs and individual α is a non-expert in relation to β's B_β beliefs. Similar examples can be given in relation to linguistic expertise. Thus, the scope of the stimulus exchange design appears to be as broad as that of the ostensive learnability design.

Three more shared features are worth noting. One is pretty obvious and concerns the use of visual stimuli. In the ostensive learnability design, these take the form of photos, videos or concrete objects. In the stimulus exchange design, they take the form of instrumentally-produced images from a scientific field and drawings thereof. The

[19] Needless to say, the choice of test subjects falling within one standard deviation of the mean should be balanced across both groups, i.e. experts and non-experts.

[20] This was suggested to me by experimental psychologist Fintan Nagle.

second shared feature is that the two designs essentially 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 similarity judgments. For an object given in the test phase of the ostensive learnability experiment to be deemed as instantiating a concept, a determination must be made about 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-based assessment is necessary.

The third and most critical feature shared by both designs is the attempt to elicit judgments under a set of primitive or bare observation conditions. We have already seen how this attempt is meant to materialise in the stimulus exchange design. To remind the reader, the test subjects are just asked to make visual similarity judgments. In the absence of a demand to describe what they see, the test subjects may thus be able to resort to something like pure (or as pure as can be) observational judgments. The ostensive learnability criterion does something similar. It attempts to put test subjects in a situation where they leave theory behind 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 judgments. Under these conditions, convergence of judgments, or lack thereof, would thus be telling.

At this point, it is worth pausing to reflect on a worry that takes the form of a dilemma. On the one hand, if the test subjects are given no guidance on what counts as similar and dissimilar then any experimental results run the risk of becoming meaningless. That's because, as the well-known dictum goes, 'anything is similar to anything else in some sense or other'. Let us call this the 'meaningless' horn of the dilemma. On the other hand, if some guidance is given to the test subjects, then the suspicion will arise that any discovery of observational judgment convergence is manufactured by the guidance itself. Let us call this the 'circularity' horn of the dilemma. In what follows, we consider each of these horns in relation to the two designs.

Take the meaningless horn first. Neither of the two designs appears to be afflicted by this horn of the dilemma as both provide some guidance to the test subjects. In the stimulus exchange design, the test subjects are told to produce *faithful* drawings—indeed, *no details spared* versions—of the images. They are then told to visually match the drawings that are most similar *overall* to the images. In the ostensive learnability design, guidance is provided in the training phase through the researchers' assent and dissent to positive and negative instances of the given concepts. One potential spanner in the works for this design concerns the issue of concept reference fixing. As Quine (1960) famously argued, merely pointing to objects and establishing assent in relation to instances of (and dissent in relation to counter-instances of) those objects is not sufficient to fix reference as alternatives are always possible. Somewhat similarly, Devitt (1981) argues that unless one knows the specific *kind* of object they are naming they cannot succeed in naming it.[21] Both problems raise worries in relation to the concept extraction envisaged by the ostensive learnability design. Even so, I am disinclined to consider these as weighty reasons to reject the design. That's because, practically speaking, reference-fixing doesn't need to be perfect in order for successful communication to take place.[22] Were this not the case, the point of writing and reading academic papers such as the current one would be void. Thus, unless objectors have specific reasons to doubt that a satisfactory level of reference-fixing can be achieved through the methods employed by the ostensive learnability design, I am disposed to leave these worries behind.

Now take the circularity horn. Neither of the two designs appears to be afflicted by this horn of the dilemma as both provide guidance that is considerably aloof. In the stimulus exchange design, the guidance would be circular if it made mention to specific patterns in the images or drawings. No such mention is made. All that is requested from the test subjects, to repeat the previous paragraph, is the production of faithful drawings of the images as well as the subsequent visual matching of the drawings that overall resemble images the most. Similarly, in the ostensive learnability design, so long as the conditions under which test subjects are exposed to concepts remain invariant, e.g. training for each concept involves the same number of instances (both negative and positive), there is no suggestion of any theoretical bias creeping in.[23] To summarise, the guidance given by both designs does not appear to theoretically steer the test subjects towards observational judgment convergence.

Readers will have noticed that, though the stimulus exchange design is presented somewhat more favourably in relation to the ostensive learnability design, my overt stance is that both designs are promising as tests for theory-neutrality. There is thus no absolute preference for one over the other. Indeed, I want to leave the door open that a third design may be better than both. After all, the designs compared here may in fact be flawed in some unforeseen way that is much more pernicious than the deficiencies identified in the current section. But where these designs fail, a different one will hopefully succeed. The only constraint I can specify in advance is that any other design must ensure that, like in the two designs compared here, judgments are elicited under the most primitive observational conditions attainable.

6 General arguments for veridicality

It is now time to turn our attention to the second aim of this paper. In this section, some general reasons in support of the view that the veridicality of observation reports can be safeguarded are put forth. Should these reasons hold, they would go some way toward

[21] I thank one of the referees for bringing this problem to my attention. Devitt calls it the 'qua-problem' because, knowing what kind an object is, say an animal, means being able to classify it as such, namely *qua* an animal. The problem is utilised to argue against causal theories of reference, which, like Schurz's proposal, rely heavily on ostension.

[22] Some of my thoughts on the matter can be found in Votsis (2015).

[23] That's not to say, of course, that theoretical bias cannot be introduced in other ways, i.e. not through the instructions test subjects receive. The example of the syntactic and semantic similarity between the terms 'meran' and 'merah' given above illustrates one such way, namely though the inappropriate choice of made-up terms.

explaining any convergent observational judgments found in the carrying out of the aforementioned experiments or experiments very much like them. Conversely, should the experiments not yield the requisite convergence, the potency of those reasons would be considerably undermined. Among other things, this section compares the veridicality view with some of its main competitors. The said competitors are strong forms of constructivism—henceforth, just constructivism.[24] As will be made clear below, these are either social or neural in nature.

We begin 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 premised on the idea that existing experimental results in psychology are undeniable or nearly so. 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, lose the most powerful motivation that we can potentially have for placing our trust in theory-ladenness theses. Note, moreover, that there is more than a whiff of blind intransigence in the fallback claim that no experiment can conceivably chip away at theory-ladenness, i.e. in the claim that theory-ladenness is untestable. If such theses are not open to the *conceptual possibility* of refutation, then there is no reason to choose between them and the countless other alternatives, including the extreme opposite (and in my view also mistaken) position that all observations are theory-neutral.

Consider next how the constructivist would explain observational judgment convergence. According to the social variety of constructivism, the reason why different individuals converge in their observational judgments has nothing to do with their sensory organs faithfully tracking some feature of the world and everything to do with the process of coming to an agreement. In more detail, what turns out to be an observation report in a scientific setting like a lab is merely the result of discussion and negotiation between members of the relevant scientific community. Thus, whether or not they all endorse a report such as 'The needle points to the middle of the scale' is not a matter of their sensory organs producing truth-tracking output but only a matter of a social mediation process whose aim is to settle on a given observational report. A similar picture is painted by the neural variety of constructivism.[25] On this view, the reason why different individuals converge in their observational judgments has to do with the way our brains are structured. If brains are structured in a similar way then it's no wonder that they yield the same perceptual outputs and indeed the same observational judgments when presented with the same stimuli. The only thing left

[24] I take it that advocates of genuinely weak forms of constructivism would be happy to concede that some observation reports are veridical.

[25] This is a view suggested in outline by one of the referees. I have not found any publications associated with it. As a consequence, I have tried to fill in the details myself.

for neural constructivists to add is that such outputs are constructs and hence are not veridical.

Though alluring, constructivist explanations face some sizable challenges. Let us start with social constructivism. Recall that, on this view, observation reports are the result of a scientific community coming to an agreement through negotiation. I will here explore two challenges against it. First, the conditions envisioned by the aforementioned experiment designs preclude any such negotiation between the test subjects. That's because each of them undergoes the test *independently* of, and *without contact* to, any of the others. There is thus no scope for agreement. The relation between test subjects and researchers is more complex as it does require contact. Even so, both experiment designs are such that contact between the two groups is minimal and, as argued above, likely to be neutral relative to the potential outcomes. Opportunities to reach agreement through negotiation are thus snuffed out.

Second, for individuals to be able to reach genuine agreement they must first engage in successful communication. But the very act of successful communication, presupposes the veridicality of observation. That's because for linguistic expressions to be shared and understood some signalling must be accurately sent and received. Take the expression 'The needle points to 39.817 on the scale' communicated by one individual to another. The sender may convey this expression in any number of ways including writing and speech. My point can be made with either. Suppose it is through speech. This presumably involves the production of certain structured sounds on behalf of the sender. Unless the receiver sensorially detects these sounds, there cannot be communication between the sender and the receiver, let alone agreement. The same point holds in relation to the *correct* discrimination of the structure of these sounds by the receiver as well as its translation into the corresponding content. If any part of this process breaks down, then there is no sense in which the two individuals agreed about anything since any claim to mutual understanding quickly disappears. That is, if the receiver could not correctly discriminate and translate into content the structure of those sounds then the corresponding meaning, namely that the needle points to 39.817 on the scale, could not be agreed upon. That's because, under such circumstances, the receiver could understand any number of different things, including the needle points to 5.012 but also something completely different, e.g. the room is green. Note that if the constructivist concedes that observations are veridical in relation to any one of the stages required for successful communication, e.g. the *correct* discrimination of the structure of sounds by the receiver, then it beggars belief why they are not veridical in relation to needles pointing in a certain direction on a scale.[26]

The neural constructivist faces challenges that are no less worrisome. On this view, convergent observational judgments are the result of our brains producing the same perceptual constructs when confronted with the same stimuli. Two challenges against this view will be explored here. First, the endorsement of the same constructs-same stimuli claim unwittingly leads the neural constructivist to ruin. That's because no

[26] Neural constructivism does not escape this challenge. For even though, on this view, observation reports are not constructed through agreement but somehow by the brain, successful communication presumably still needs to take place between individuals.

matter how constructed the perceptions are, so long as the same constructs follow (or even are likely to follow) the same stimuli, then the resulting observational judgments will preserve some veridicality. After all, the same constructs-same stimuli claim is tantamount to an injective mapping from constructs to stimuli, i.e. distinct elements in the set of perceptual constructs are mapped to distinct elements in the set of stimuli, and hence reveals something about the structure of the stimuli.[27] This point has been known since Locke introduced the inverted spectrum argument—see Votsis (2015) for an in-depth analysis.

Second, the most powerful argument for the veridicality of observational judgments is, as it always was, the success those judgments confer on our ability to predict, and interact with, the world.[28] Short of being a coincidence of cosmic proportions, such continued success cannot be secured without our observational judgments somehow having latched on to the world. Unsurprisingly, this view is bolstered by evolutionary considerations. Species would not be able to stick around long enough if their ability to make life and death choices was persistently hindered by misleading, i.e. false or largely false, observational judgments. Otherwise put, species whose members are equipped with the ability to produce trustworthy, i.e. true or largely true, observational judgments are likely to linger a while longer.

7 Conclusion

It has been argued, I hope compellingly, that the two experiment designs compared in this paper are both promising ways to determine the extent to which, if at all, observation can be theory-neutral. Both seek to create conditions that drive test subjects, even when these come from varied theoretical backgrounds, to something like bare observational judgments. Plans are under way to put versions of these experiments into practice. If the experiments result in judgment convergence, then this will provide support for the view that theory-neutral observations are feasible. In such a case, we may also ask why such convergence occurs. The preceding section provides some forceful reasons in favour of the view that any substantial convergence is best explained by the veridicality of observation reports and not by constructivist alternatives.

Acknowledgements I wish to thank three anonymous referees for their astute remarks on an earlier version of this paper. These remarks have helped shape this paper into its current form. I also wish to thank audiences at the Celebratory Colloquium in Honour of Gerhard Schurz (held at Heinrich-Heine Universitaet Duesseldorf), the 23rd Annual Meeting of the European Society for Philosophy and Psychology (held at the University of Tartu), the Work-in-Progress Seminar in Philosophy at LSE and the Experience and Reality conference (held at the Catholic University in Ružomberok) as well as my colleague Fintan Nagle for invaluable feedback. Moreover, I am grateful to Peter Brössel, Christian Feldbacher-Escamilla, Alexander Gebharter and Markus Werning (the editors of this special issue) for various fruitful discussions on the subject matter. Last but not least, I wish to thank Gerhard Schurz for inspiring some of this work and, more generally, for being a mentor in a crucial period of my philosophical development.

[27] An injective mapping is a function that preserves the distinctness of elements.

[28] This challenge afflicts social constructivism as much as it does neural constructivism.

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* (pp. 289–334). Oxford: Oxford University Press.

Brewer, W. F., & Lambert, B. L. (2001). The theory-ladenness of observation and the theory-ladenness of the rest of the scientific process. *Philosophy of Science, 68*(S3), S176–S186.

Burr, V. (2015). *Social constructionism* (3rd ed.). New York: Routledge.

Devitt, M. (1981). *Designation*. New York: Columbia University Press.

Dryer, M. (2005). Negative morphemes. In M. Haspelmath, et al. (Eds.), *The world atlas of language structures* (pp. 454–457). Oxford: Oxford University Press.

Firestone, C., & Scholl, B. J. (2016). Cognition does not affect perception: Evaluating the evidence for 'top-down' effects. *Brain and Behavioural Sciences, 39*(e229), 1–77.

Fodor, J. (1984). Observation reconsidered. *Philosophy of Science, 51*(1), 23–43.

Kay, P., et al. (2009). *The world color survey*. Stanford: CSLI Publications.

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

Machery, E. (2015). Cognitive penetrability: A no-progress report. In J. Zeimbekis & A. Raftopoulos (Eds.), *The cognitive penetrability of perception: New philosophical perspectives* (pp. 59–74). 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.

Maxwell, G. (1962). The ontological status of theoretical entities. In H. Feigl & G. Maxwell (Eds.), *Scientific explanation, space, and time, Vol. 3, Minnesota studies in the philosophy of science* (pp. 3–15). Minneapolis: University of Minnesota Press.

McMullin, E. (1993). Rationality and paradigm change in science. In P. Horwich (Ed.), *World changes: Thomas Kuhn and the nature of science* (pp. 55–78). Cambridge, Massachusetts: MIT Press.

Pylyshyn, Z. (1999). Is vision continuous with cognition? *Behavioral and Brain Sciences, 22*(3), 341–365.

Quine, W. V. O. (1960). *Word and object*. Cambridge, MA: MIT Press.

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), 601–620.

Schurz, G. (2015). Ostensive learnability as a test criterion for theory-neutral observation concepts. *Journal for General Philosophy of Science, 46*(1), 139–153.

Surrallés, A. (2016). On contrastive perception and ineffability: Assessing sensory experience without colour terms in an Amazonian society. *Journal of the Royal Anthropological Institute, 22*(4), 962–979.

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.

Votsis, I. (2015). Perception and observation unladened. *Philosophical Studies, 172*(3), 563–585.

Zeimbekis, J. (2013). Color and cognitive penetrability. *Philosophical Studies, 165*(1), 167–175.

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.