Philos Stud (2021) 178:4119–4149
https://doi.org/10.1007/s11098-021-01639-8

# What theoretical equivalence could not be

Trevor Teitel[1]

Accepted: 5 April 2021 / Published online: 16 April 2021

**Abstract** Formal criteria of theoretical equivalence are mathematical mappings between specific sorts of mathematical objects, notably including those objects used in mathematical physics. Proponents of formal criteria claim that results involving these criteria have implications that extend beyond pure mathematics. For instance, they claim that formal criteria bear on the project of using our best mathematical physics as a guide to what the world is like, and also have deflationary implications for various debates in the metaphysics of physics. In this paper, I investigate whether there is a defensible view according to which formal criteria have significant non-mathematical implications, of these sorts or any other, reaching a chiefly negative verdict. Along the way, I discuss various foundational issues concerning how we use mathematical objects to describe the world when doing physics, and how this practice should inform metaphysics. I diagnose the prominence of formal criteria as stemming from contentious views on these foundational issues, and endeavor to motivate some alternative views in their stead.

**Keywords** Theoretical equivalence · Representation · Logical positivism · Substantivalism · Metaphysical realism

Formal criteria of theoretical equivalence are mathematical mappings between specific sorts of mathematical objects, such as sets of sentences (understood as syntactic strings), or sets of mathematical models, or categories of mathematical models (in the sense of category theory). Philosophers of science working on such

✉ Trevor Teitel
trevor.teitel@utoronto.ca

[1] Department of Philosophy, Faculty of Arts and Science, University of Toronto, Jackman Humanities Building, 4th Floor, 170 St. George Street, Toronto, ON M5R 2M8, Canada

criteria first associate different physical theories with some such mathematical objects. They then use theorems about which of these mathematical objects stand in one of these mathematical mappings to each other in order to draw conclusions about which physical theories are (or fail to be) “theoretically equivalent”.

These formal approaches to theoretical equivalence have been around for a while, but there has been something of an explosion of work on them over the past decade.[1] I take the mathematical fruits of this work to be unassailable (results to the effect that such-and-such mathematical objects do or do not stand in such-and-such mathematical relation). However, those working on formal criteria take their results to have significant implications that go beyond pure mathematics. For instance, they take their results to bear on the project of using our best mathematical physics as a guide to what the world is like, and also to have deflationary implications for various debates in the metaphysics of physics. And, without question, the interest of work on formal criteria would be diminished if formal equivalence results were conceded to lack such non-mathematical implications. But it has yet to be made clear precisely what these non-mathematical implications might be, and how they are supposed to follow from a formal equivalence result. My primary goal here is to argue that the prospects for filling in this story are dim. I will investigate various views one might hold about the non-mathematical significance of these formal criteria, and argue that none is tenable. My tentative conclusion is that formal criteria are of limited non-mathematical interest.

Along the way, I shall discuss various foundational issues concerning how we use mathematical objects to describe the world when doing physics, and how this practice should inform metaphysics. I will suggest that the prominence and allure of formal criteria rests on certain contentious assumptions about these foundational issues, and will endeavor to motivate some alternative views in their stead (see especially Sect. 6). To preview, formal equivalence proofs by their nature consider only the mathematics we use to express our best physics. Yet more than just this mathematics contributes to the conceptions of reality inspired by contemporary physics. It is no surprise that criteria of equivalence that ignore these additional components are inadequate in important respects. The foundational assumptions that I shall challenge are quite prevalent in the metaphysics of physics and the philosophy of science more generally. So even those unconcerned with the topic of theoretical equivalence should still find material of interest in what follows.

[1] For a sampling see Barrett (2015, 2019), Barrett and Halvorson (2016a, 2016b, 2017), Butterfield (2018), Curiel (2014), Coffey (2014), Glymour (2013), Halvorson (2012), Halvorson (2013), Hudetz (2019), North (2009), Rosenstock et al. (2015), Teh and Tsementzis (2017), Tsementzis (2017), and Weatherall (2015). For some older work on the topic see Glymour (1970, 1977), Quine (1975), Sklar (1982), and Putnam (1983). For a helpful overview of the literature see Weatherall (2019). There has also been a burgeoning interest in the related topic of dualities. Much of what I will say also bears on this topic. However, discussing dualities explicitly would require another paper, so I shall confine my attention here to theoretical equivalence.

# 1 Preliminaries

Let us start with a bit of background about theoretical equivalence generally, and and an example of a formal criterion of equivalence.

## 1.1 Semantic equivalence and deflationary strategies

There are several things you might be interested in when asking whether two theories are equivalent, and there is little sense in fighting over which "really" deserves to be labeled 'equivalence'. In particular, someone might just be interested in whether the two theories are empirically equivalent, or are formulated using mathematical structures that stand in a certain purely mathematical relation, and choose to call the theories 'equivalent' as a result. I have no objection to them doing so; I am not interested in fighting over the word.

That being said, the sense of equivalence I have in mind throughout is the one philosophers of science are generally interested in, namely whether two theories say the same thing about the world, or have the same semantic content, or the same interpretation, or express the same proposition. (I will understand all of these glosses to amount to the same thing; more on this in a moment.) The crucial contrast is that a mathematical object, such as a set or category of mathematical models, is on its own just a piece of mathematics, which does not represent the world as being any way whatsoever. However, such objects are amongst the tools we use to represent the world, and in doing so we bring about an association of the objects with some propositional content, depending on what we are using the mathematics to represent. To avoid terminological issues about what in fact suffices for "equivalence," let us henceforth say that two mathematical objects that are being used to say the same thing about the world (or have the same content, and so on) are *semantically equivalent*. Note, we shall soon see that all of these glosses can never be understood absolutely; after all, one and the same sentence or mathematical object can be used to say different things about the world on different occasions. I will ignore this point for the time being, but we will see its importance in due course.

You might immediately worry that there are large debates about the nature and fineness of grain of propositions or contents themselves. For instance, do sentences used to express metaphysically necessarily equivalent propositions say the same or different things about the world? Fortunately, we do not need to get embroiled in such controversies. My goal here is to challenge whether there is a defensible view according to which formal criteria illuminate when two theories say the same thing about the world. And my arguments apply even on very weak or coarse-grained construals of what this requires, which are maximally hospitable to proponents of formal criteria. For instance, my arguments will show that formal criteria fail even to secure metaphysically necessary equivalence. Indeed some of my arguments point to cases where formal criteria do not even imply material equivalence (having the same truth value). Thus those who hold quite fine-grained views about propositions or semantic content can substitute various weaker relations in place of my talk of semantic equivalence throughout without affecting the cogency my arguments: relations like expressing mutually entailing propositions, expressing metaphysically necessarily equivalent propositions, and so on. I shall stick with the imperfect label 'semantic equivalence', though it should be understood in a liberal sense throughout compatible with very coarse-grained accounts of what saying the same thing about the world involves.[2]

Like 'equivalence', the word 'theory' often leads to terminological confusion. At times 'theory' is used to describe certain uninterpreted mathematical objects (for instance, a certain solution space satisfying some equations). At other times, the word describes this mathematics together with an interpretation; that is, the mathematical object plus some associated semantic content encoding a way the world might be. Quite often interpreting certain claims involving the word 'theory' requires one to shift between these two senses. The former sense is the one generally operative in discussions of semantic equivalence; the issue under debate is precisely which uninterpreted mathematical objects have the same interpretation. To avoid terminological confusion, I will generally avoid the word 'theory' unless it is unambiguous what sense I intend. I will use 'representational vehicle' to describe an uninterpreted object that might come to have an interpretation (express a proposition, have content, and so on). So representational vehicles include uninterpreted strings in a formal or natural language, and also uninterpreted mathematical objects like a set or category of mathematical models. In this terminology, formal criteria of equivalence are mappings between certain representational vehicles.[3]

Why have philosophers of science been interested in semantic equivalence? Discoveries about which representational vehicles are semantically equivalent enable one to collapse certain distinctions. If such discoveries are non-obvious, then the result of semantic equivalence may enable one to diagnose some extant debate as misguided. Suppose two philosophers are debating about the fundamental metaphysics of the world. The former expresses her view with representational vehicle $A$, and the latter with vehicle $B$. Then if you could show $A$ and $B$ to be semantically equivalent, you would have thereby deflated the debate, by showing

[2] No label here is perfect, but I have found 'semantic equivalence' to be the least misleading. Another option would be 'worldly equivalence'. Other labels one finds in the literature for the target phenomenon include 'metaphysical equivalence', 'full equivalence', 'interpretational equivalence', and 'representational equivalence'. Readers should feel free to substitute whichever label they prefer throughout.

[3] I should flag that I think there are problems with lumping sentences together with mathematical objects in this way. In particular, I am skeptical of the common practice of treating mathematical objects as things that, like sentences, might be or fail to be semantically equivalent to one another. When we use a sentence to describe the non-mathematical world, we do so by using it to express some proposition or content. By contrast, when we use a mathematical object like a set of mathematical models to describe the non-mathematical world, we do say by saying something *about* that object and the non-mathematical world, usually highlighting some salient respect in which the two are similar. There is little sense in asking, even on some particular occasion of use, what a mathematical model "says about the non-mathematical world"; rather, it is similar in certain respects and different in others. Some of my skepticism about this contrast will crop up below, but I will try to set it aside as much as possible, and acquiesce in the standard practice of treating mathematical objects as things that may be semantically equivalent to one another. Doing so allows me to focus on my concerns about formal criteria of equivalence in particular.

that the philosophers have not succeeded in articulating a contentful difference to be disagreeing over. For instance, suppose $A$ and $B$ are identical sentences except for being written in different fonts. Given standard representational conventions, we do not take our choice of font in writing sentences down to change their semantic content.[4] Hence, $A$ and $B$ are plausibly semantically equivalent on standard occasions of use, and any appearance of a substantive disagreement between $A$-advocates and $B$-advocates is illusory. Notice that this sort of criticism is much stronger than the common complaint that some debate is *epistemically* intractable: deflating a debate via semantic equivalence reveals that the participants have failed to even carve out a meaningful distinction to disagree over in the first place. Now disagreements where the parties employ representational vehicles that differ only over their fonts will never arise in practice, so an equivalence-based deflationary strategy must employ non-obvious results about which representational vehicles are semantically equivalent.

The classic version of such a strategy was logical positivism, which took "empirical equivalence" to suffice for semantic equivalence. Positivists thus purported to deflate debates framed in terms of representational vehicles used to express empirically equivalent contents. Indeed, they regarded such debates as no more sensible than debates framed in terms of representational vehicles differing only over their fonts. Generally the vehicles appeared to be expressing contentful, albeit empirically inaccessible, differences (for instance, differing over whether they imply that space is infinite). However, for the positivists such appearances were illusory. The doctrines at issue in most philosophical debates do not differ over their empirical consequences; as a result, if the positivists were correct, we would have had reason to consign most philosophical debates to the flames.

Nowadays the positivist program is rightly regarded as a dramatic failure, resting on dubious assumptions across metaphysics, epistemology, and the philosophy of language.[5] However, this deflationary strategy illustrates why semantic equivalence is usually the sense of 'equivalence' that philosophers of science are interested in, particularly those set on discrediting metaphysical inquiry (again, always keeping in the mind our weak use of 'semantic equivalence' emphasized above). For semantic equivalence is what must be at issue if some debate is to be deflated via assimilation to the debate between representational vehicles differing only over their fonts. We shall see that proponents of formal criteria see their results as implying neo-positivist deflationary conclusions about certain extant debates amongst metaphysics-oriented philosophers of physics, such as the debate over whether there are spacetime points. And my sense is that many take work on formal criteria to cast

[4] Choice of font might affect the *truth* of certain token sentences given standard representational conventions (consider 'this sentence is written in Times New Roman'). However, the issue in the main text concerns the bearing of font choice on the proposition expressed.

[5] See Soames (2003, ch. 12–13) for an overview of some reasons for the fall of positivism. In the main text I described the standard characterization of the positivist program, and my comments are directed at the program only understood in this way (according to which it is committed to a flat-footed empiricist criterion of semantic content). An anonymous referee points out that some commentators argue that the positivists in fact held more sophisticated and defensible views than the standard characterization would suggest. For discussion, see Friedman (1999) and Creath (2020).

doubt on these more metaphysics-oriented debates. Yet these deflationary conclusions follow only if formal criteria illuminate semantic equivalence, which I shall challenge here.

### 1.2 An example: definitional equivalence

My arguments will generally concern the idea of a formal criterion of theoretical equivalence in the abstract. Hence, for the most part we need not delve into the details of particular criteria on offer. Nonetheless, it will be helpful to give you a feel for one of the criteria, so you have an example in mind moving forward. The criteria can helpfully be grouped into two broad categories: sentential and non-sentential. Sentential criteria relate logical theories, understood as sets of (uninterpreted) sentences of some formal language. Non-sentential criteria relate non-sentential representational vehicles, such as categories of mathematical models.

Let us start with the sentential criterion of *definitional equivalence*, first introduced into the philosophy of science by Glymour (1970, 1977). The criterion relates theories in first-order languages. Any two such theories that are formulated in different signatures (primitive vocabularies) cannot be logically equivalent. Definitional equivalence is meant to capture the intuition that nevertheless such theories might have the same expressive resources. Here is the rough idea.[6] Let $\Sigma$ and $\Sigma^+$ be first-order signatures such that $\Sigma \subseteq \Sigma^+$. Given a first-order theory $T$ in $\Sigma$, we can define the *definitional extension* of $T$ to $\Sigma^+$. This is a first-order theory $T^+$ in $\Sigma^+$ that extends $T$ by adding explicit definitions of all vocabulary in $\Sigma^+ \setminus \Sigma$ in terms of the vocabulary in $\Sigma$. For example, let $F$ and $G$ be monadic predicate constants, and suppose $\Sigma = \{F\}$ and $\Sigma^+ = \{F, G\}$. Then, a definitional extension $T^+$ might extend $T$ by adding the explicit definition $\forall x(Fx \leftrightarrow Gx)$. Now, consider any first-order theories $T_1$ in signature $\Sigma_1$ and $T_2$ in signature $\Sigma_2$. We say $T_1$ and $T_2$ are *definitionally equivalent* iff there is a definitional extension $T_1^+$ of $T_1$ to the signature $\Sigma_1 \cup \Sigma_2$, and a definitional extension $T_2^+$ of $T_2$ also to the signature $\Sigma_1 \cup \Sigma_2$, such that $T_1^+$ and $T_2^+$ are logically equivalent. In a slogan, definitionally equivalent theories have a "common definitional extension".

To side-step having to axiomatize realistic physics in a first-order language, Glymour instead works with a model-theoretic analogue of definitional equivalence. Though, as emphasized by Weatherall (2015, 1079–1980), the analogue employs the notion of elements of one model being "uniquely and covariantly definable" in terms of the elements of the other, and the need for first-order formulations recurs in trying to make this notion precise.[7] This point provides perhaps the central explanation for the recent prominence of non-sentential criteria, in particular a

[6] See Barrett and Halvorson (2016a) for a rigorous presentation.

[7] Notably, as Weatherall appreciates, for Glymour's purposes the need for first-order formulations does not arise. We shall see that he regarded definitional equivalence only as a *necessary* condition for semantic equivalence, and, in the cases he was interested in, the "uniqueness" clause sufficed for his results (which concerned verdicts about only *in*equivalence).

category-theoretic criterion, amongst philosophers of science working on equivalence. We shall discuss this criterion later on.

Let us turn now to our main task of exploring the views one might hold about the non-mathematical significance of a formal criterion of theoretical equivalence.

## 2 Trivial semantic conventionality

Here is a natural first-pass view for the proponent of a formal criterion who wants to argue that it has some bearing on semantic equivalence: the criterion straightforwardly "tells us which representational vehicles are semantically equivalent to which others." But, as adumbrated above, any claim of this sort cannot be correct, because of the familiar platitude that any representational vehicle can in principle be used to represent the world as being just about any way whatsoever (what Putnam 1983, 41) calls *trivial semantic conventionality*). For example, we generally use the (uninterpreted) sentence 'all dogs have fleas' to say that all dogs have fleas, however there is nothing incoherent about a community that uses that very same sentence to instead say that all philosophers have fleas. Similarly, we noted above that generally we use representational vehicles differing only over their fonts to express the same semantic content, which motivated the claim that such vehicles are semantically equivalent. However, there is nothing incoherent about a community that uses English sentences exactly the way we do with the exception that writing a sentence in a particular font is a way of negating it.

The platitude of trivial semantic conventionality shows that it does not make sense to ask what a representational vehicle says about the world simpliciter (or its interpretation simpliciter, or its semantic content simpliciter, and so on). As a result, trivial semantic conventionality shows that it also does not make sense to ask which representational vehicles are semantically equivalent simpliciter. Rather, such questions must be relativized, whether to interpretations or occasions of use (where interpretations are mappings from representational vehicles to contents, and different interpretations can be operative on different occasions of use). Thus formal criteria of equivalence between two representational vehicles $A$ and $B$ cannot tell us something about the semantic properties of $A$ and $B$ simpliciter, absent information about how $A$ and $B$ are being used to represent the world. Moreover, we should not ask whether two representational vehicles are semantically equivalent relative to *every* interpretation: we know from trivial semantic conventionality that *no* representational vehicles are semantically equivalent relative to every interpretation. Similarly, we should not ask merely whether two representational vehicles are semantically equivalent relative to *some* interpretation: again from trivial semantic conventionality, we know that *every* pair of representational vehicles is trivially semantically equivalent relative to some interpretation. Thus, in light of trivial semantic conventionality, the question facing proponents of a formal criterion is whether there is some interesting range of interpretations relative to which the criterion illuminates semantic equivalence.

Notice that the platitude and attendant moral are not peculiar to sentences, but hold true of representational vehicles generally. For instance, a common example that motivates semantic equivalence for non-sentential representational vehicles (analogous to choice of font) involves the choice of signature when writing down a general relativistic theory. Consider two general relativistic solution spaces (understood as sets of uninterpreted mathematical models) that differ only over the choice of a Lorentzian metric of signature (1, 3) rather than of signature (3, 1) in each solution. Analogously to how we generally do not take our choice of font when writing a sentence down to affect its propositional content, the choice between these solution spaces (which are distinct mathematical objects) is universally regarded as a mere sign convention that does not affect semantic content (roughly, the convention of whether to associate time-like distances with positive numbers and space-like distances with negative numbers, or the other way around). A moral one might be tempted to draw is that the two solution spaces "are semantically equivalent," because they "say the same thing about the world," but we now see that these glosses cannot strictly hold without qualification. Just as there is nothing incoherent about a community that allows choice of font to make a contentful difference, there is nothing incoherent about a community that does the same for choice of signature. For example, consider a community, call them the "+sitivists" (bad pun), where it is ingrained in their applied mathematical practice that only positive numbers in a Lorentzian metric correspond to time-like vectors, and only negative numbers in the metric correspond to space-like vectors. Thus, in this community, the solution space where each solution has a metric of signature (1, 3) might be used to represent the world as containing a familiar general relativistic spacetime (with its one time-like and three space-like directions at every point). By contrast, they take the (3, 1) solution space to correspond to the (perhaps metaphysically impossible) proposition that at every spacetime point there are three mutually orthogonal time-like directions and no two orthogonal space-like directions. This community's representational conventions are alien to our own, but they are perfectly coherent. We see then that the moral from trivial semantic conventionality extends to all representational vehicles, including mathematical objects like a solution space. On their own such objects are just mathematics, which do not "say anything about the world," or have any "interpretation," and so on. Hence, any talk about whether such objects are semantically equivalent must be understood relative to some operative interpretation or particular occasions of use. Finally, notice that the reasoning that led to this conclusion applies irrespective of how syntactically or structurally similar or different the mathematical objects at issue may be.

It likely seems as though I am belaboring the obvious, but this moral reveals that most extant glosses on the non-mathematical significance of formal criteria cannot be taken at face-value. For example, a standard gloss is that formal criteria holding between two representational vehicles reveal that the vehicles have "the same capacities to represent physical situations" (this gloss on the non-mathematical significance of the popular category-theoretic criterion of equivalence is repeated by Weatherall (2015, p. 1081, p. 1087) and Rosenstock et al. (2015, 315); compare also Hudetz (2019, 52–53)). But, on the most straightforward reading of these glosses, all representational vehicles considered on their own have the *same* "capacities to represent physical situations," simply due to trivial semantic conventionality:

namely, the capacity to represent just about any physical situation whatsoever. So such glosses must not be meant at face-value. But we shall see in Sect. 5 that it is unclear what precisification of such glosses might serve the purposes of proponents of formal criteria. Similarly, Rosenstock et al. (2015, 315–316) claim that because their category-theoretic criterion holds between two representational vehicles, the vehicles "encode precisely the same physical facts about the world, in somewhat different languages." But this gloss does not do any better. As I have emphasized, no representational vehicle, whether a sentence or a mathematical object, encodes *any* facts about the world simpliciter. One more: Barrett (2019, pp. 1188–1192) argues for a connection between which of our theories satisfy some formal criterion of equivalence and which "features of our theories are significant or contentful." But the "theories" at issue in his discussion are uninterpreted mathematical objects drawn from mathematical physics; and again, *no* features of such uninterpreted theories are significant or contentful full-stop.

The moral in this section is a different route towards the moral emphasized by all extant criticisms of formal criteria. Here I have in mind the investigations of Sklar (1982), Coffey (2014), Nguyen (2017), and Butterfield (2018) (see also Putnam, 1983, 38 and van Fraassen, 2014). These criticisms rightly point out, from different directions and using different examples, that we can, and often do, use one and the same mathematical object in different ways on different occasions. For example, in the literature on the metaphysics of non-relativistic quantum mechanics, flash and matter-density conceptions of the world are presented with the aid of one and the same stochastic collapse mathematical formulation of a quantum theory, such as the GRW theory.[8] Because of this point, these critiques rightly conclude that no purely formal relation can illuminate semantic equivalence absolutely; rather, a relation can do so only if it is sensitive to the interpretation or semantic content being associated with the representational vehicles at issue.

I am very sympathetic with all of these critiques as far as they go. However, I think we can go considerably further. Indeed, despite these criticisms work on formal criteria of equivalence has not let up, and I think there are a few reasons for this. First, taking on board the need for relativization does not scotch the attempt to provide a rationale for the non-mathematical significance of formal criteria; as noted above, for all we have said so far the criteria may correlate with semantic equivalence relative to some important but circumscribed range of interpretations or occasions of use. I think we can also cast doubt on such scaled-back ambitions for the non-mathematical significance of formal criteria, as I shall attempt to do in the rest of the paper. Second, as Weatherall (2019) emphasizes when responding to the critiques just mentioned, what I have been calling 'formal criteria of equivalence' thus far are often presented as being sufficient for equivalence only when conjoined

[8] Compare also van Fraassen (2014, 279): "If the same diffusion equation is presented to describe gas diffusion and, elsewhere, temperature distribution over time, would anyone think that one and only one theory was being presented? [...] A representation has content. A representation of gas diffusion is not the same thing as a representation of temperature distribution, even if the math is the same." Though because this example involves empirically inequivalent contents, it will likely not worry proponents of formal criteria, for reasons I outline below in the main text.

with empirical equivalence.[9] And as I have just been emphasizing, it makes no sense to talk about the "empirical content" or "observational content" of a representational vehicle like a set or category of models in the abstract, despite the prominence of this way of speaking. Such vehicles on their own have no content whatsoever, whether empirical or extra-empirical. Thus arguably proponents of formal criteria have never meant to be propounding *purely* formal criteria for equivalence, which apply to representational vehicles in the abstract, but rather criteria which apply only to such vehicles together with an interpretation. Taking this moral on board, let us ask whether there is a tenable scaled-back view according to which these criteria have non-mathematical significance. (Note, I shall continue to use the label 'formal criteria of equivalence', though I will be explicit about the role of empirical equivalence when it is relevant.)

## 3 Sentential criteria

Let me start by discussing sentential criteria. I think there are clear counterexamples to any view that regards such criteria as illuminating semantic equivalence, even relative to some circumscribed range of interesting interpretations. The reason is that such criteria are manifestly extensionally inadequate relative to any interpretations we in fact employ: indeed the criteria fail even to imply *material equivalence* relative these interpretations. Yet these interpretations include the ones operative when philosophers say things like 'there are spacetime points', or engage in other metaphysical speculation. So I take this result to cast doubt on there being any defensible and interesting view according to which sentential criteria have non-mathematical significance. After defending these claims, I shall devote the rest of the paper to non-sentential (in particular category-theoretic) criteria.

The kind of counterexample I have in mind has been forcefully presented by Sklar (and bracket the point just mentioned about empirical equivalence for a moment):

> Let the two theories be 'All lions have stripes', and 'All tigers have stripes', with all the words in both theories taking on their usual meanings. The theories are inter-translatable in the purely formal sense. They are exactly alike in logical form and one can be obtained from the other by a simple term for term substitution. But they are most assuredly not equivalent [...] mere commonality of logical form, even of a total theory when compared with another total theory, is certainly not by itself sufficient for theoretical equivalence. The meanings of the terms in the theories, however construed, are crucial to questions of equivalence. (Sklar, 1982, 93)

The natural regimentations of Sklar's single-sentence theories into first-order logic are deemed equivalent by every extant sentential criterion that I am aware of. In

[9] Those who have explicitly conceived of formal criteria of equivalence as ways to strengthen empirical equivalence include Quine (1975, 319), Sklar (1982), Glymour (2013, 289), Rosenstock et al. (2015), Hudetz (2019), and Weatherall (2015, 2019).

particular, the theories are deemed equivalent according to (i) Glymour's (1970, 1977) criterion of definitional equivalence, which I outlined above, (ii) a recent generalization of definitional equivalence called "Morita equivalence" due to Barrett and Halvorson (2016b), (iii) Quine's (1975) criterion in terms of inter-translatability, and (iv) the generalization of Quine's criterion spelled out by Barrett and Halvorson (2016a), which like Morita equivalence turns out to be implied by definitional equivalence. Yet the two sentences relative to their operative interpretation are not semantically equivalent; indeed they are also not necessarily equivalent, nor even materially equivalent. And it is easy to multiply examples of this sort indefinitely. If these counterexamples succeed, they reveal sentential criteria to be woefully extensionally inadequate relative to the interpretations we in fact employ. In the rest of this section, I shall argue that this deceptively simple challenge stands up to scrutiny: all replies on behalf of proponents of formal criteria are problematic.

A first reply to these counterexamples appeals to the point mentioned above, that formal criteria are generally intended to strengthen empirical equivalence (recall footnote 9). As applied to Sklar's example, the idea would be that although the regimentations of 'all lions have stripes' and 'all tigers have stripes' satisfy the various sentential criteria, these criteria are sufficient for semantic equivalence only relative to interpretations where the sentences express empirically equivalent contents. Yet on the relevant interpretations the two sentences fail this test.

An initial challenge for this reply are the familiar issues that arise for all views that place considerable theoretical significance on the distinction between what is and is not observable, of the sort that plagued positivists. For instance, what is observable, and hence what is empirically equivalent to what, seems vague and to vary as our experimental capacities advance (for classic discussions see Maxwell, 1962; van Fraassen, 1980). Any view tied to the distinction will then seem to inherit these features. That being said, I do not want to delve into these large and thorny issues here; the present reply fails even setting such issues aside.

The central problem with this reply is that it either threatens to collapse into the discredited positivist criteria for semantic equivalence, or else does not address the issue that the counterexamples bring out. Proponents of formal criteria should (and generally explicitly do) allow for some contentful distinctions that cut finer than empirical equivalence.[10] Doing so allows them to avoid dubious claims to the effect that there is no intelligible distinction between, say, Lorentzian and Minkowskian conceptions of special relativity, or between ascribing to the world a Newtonian versus neo-Newtonian spacetime structure (which differ over whether there is a standard of absolute velocity). Yet once proponents of formal criteria allow for some intelligible distinctions that cannot be teased apart empirically, we can resuscitate Sklar-style counterexamples. For we can now find sentences which, on the operative interpretation, (i) are not semantically equivalent, (ii) have the requisite syntactic similarity to satisfy every extant sentential criteria, yet which (iii) are also empirically equivalent. For example, consider a world with a Newtonian

[10] For a recent example see Barrett (2019, p. 1191). Compare also Putnam (1983, 30).

spacetime structure. Suppose on some occasion one person in this world says ‘the centre of mass of the universe is moving at an absolute speed of 1m/s’ and a second says ‘the centre of mass of the universe is moving at an absolute speed of 2m/s’, where both parties intend to be using standard English representational conventions. The two sentences are used to say different things about the Newtonian world despite being empirically equivalent. Yet, exactly like Sklar’s original example, the sentences’ regimentations into first-order logic will satisfy every extant sentential criterion of equivalence. Thus it looks like supplementing sentential criteria with empirical equivalence simply fails to get to the heart of the problem posed by the counterexamples.

A second potential reply to the counterexamples appeals to semantic holism, claiming that we cannot consider single-sentence theories like Sklar’s, but must instead consider the speaker’s, or perhaps the entire linguistic community’s, total “background theory” of the expressions that figure in the relevant single-sentence theories (expressions like ‘lion’, ‘tiger’, or ‘moving at an absolute speed of 1m/s’). For example, in the original case such a “background theory” might include the sentence ‘nothing is both a lion and a tiger’. Yet the theory of this sentence and ‘all lions have stripes’ fails to be definitionally equivalent to the theory consisting of the sentence and ‘all tigers have stripes’ (because the theories have the same signature yet are not logically equivalent).

This reply is untenable, however. A first issue is that holistic metasemantic theories are now widely rejected (see see Soames, 2003, ch. 17 for discussion of some of the worries these holistic theories face). But even if we set that point aside, there is a straightforward problem with this reply: no extant formal equivalence proofs consider total theories of this sort, and it is dubious that such a theory could ever be written down in practice. Rather, such proofs generally consider standard mathematical formulations of our best physical theories. And as I hinted at in the introduction, and will expand on in Sect. 6, these standard formulations are plausibly further embellished by the user’s or community’s “background theory” of concepts like space, time, or mass.[11] Thus I doubt proponents of formal equivalence proofs would opt for this holistic reply, on pain of having to abandon their entire project.

A third reply would be to claim that the formal criteria (perhaps supplemented with empirical equivalence) are intended only as *necessary* conditions for semantic equivalence. Indeed, Glymour himself originally put forward definitional equivalence only as a necessary condition, given that for his purposes he sought a verdict only about which representational vehicles *fail* to be semantically equivalent. However, it is hard to see how this reply can be accepted by contemporary proponents of formal criteria. Those in this literature spend much of their time proving positive results. Moreover, these positive results are what is needed in order to implement an equivalence-based deflationary strategy. Relegating formal criteria

[11] And the same would be true even of candidates for what physicists sometimes describe as a “total,” “complete,” or “final” theory, such as string theory or some other candidate theory of quantum gravity.

to mere necessary conditions makes it mysterious why such considerable energy has been exerted on these positive results.

Proponents of formal criteria will likely respond by extending the third reply as follows: although formal criteria (perhaps supplemented with empirical equivalence) are merely necessary conditions for semantic equivalence, they together with empirical equivalence form a non-redundant component of some informative sufficient condition for semantic equivalence, which rationalizes the extensive focus on proving positive results. The tenability of this reply depends on what exactly is taken to be sufficient for semantic equivalence only when conjoined with a formal equivalence result plus empirical equivalence. In the abstract, my objection is that either this extra ingredient will be objectionable on independent grounds, or else render the formal equivalence result a redundant idle-wheel. Let us see how this dilemma plays out with some particular instances of this strategy.

For example, proponents of formal criteria cannot just declare some formal criterion sufficient for semantic equivalence relative to some interpretation when the theories at issue also have the same content relative to that interpretation. That amounts to saying that a formal equivalence result conjoined with semantic equivalence is sufficient for semantic equivalence: the formal equivalence result is patently redundant, rather than offering some independent handle on semantic equivalence. Yet various other candidates for the extra ingredient arguably face the same problem, only in a less direct manner. For instance, Putnam (1983) suggests that a formal equivalence result is sufficient for semantic equivalence when conjoined with the non-formal requirement that the result "preserves the relation of *explanation* and that *the same phenomena are explained by both*" (39). But the explanations provided by some representational vehicle as interpreted on some occasion of use depend, of course, on the vehicle's content on that occasion. So Putnam's proposal avoids the charge of rendering the formal criterion at issue redundant only in a circumscribed range of occasions of use: namely, those in which we know enough about the representational vehicles' contents on the occasion to know that the vehicles explain the same phenomena, yet are still unsure whether the vehicles are semantically equivalent (that is, have the same contents full-stop) on the occasion. How prevalent will such occasions be? Answering this question would require going through the candidate contemporary accounts of explanation. So I will instead lean on a second worry. But let me still note that the advocate of semantic *in*equivalence in some of the disputed cases in the debate takes the representational vehicles at issue to describe, what at least purport to be, altogether different conceptions of reality. For this reason, in such cases she can also be expected to take the interpreted vehicles to offer explanations with radically different underlying structures so as to render them inequivalent on Putnam's proposal. For this reason Putnam's proposal may offer little solace to those seeking to carve out a distinctive role for formal equivalence results to play in adjudicating cases of semantic equivalence.

A second and more pressing worry for Putnam's proposal turns on the familiar issue of what the "phenomena" are.[12] I shall rehearse the dialectic here in more detail in Sect. 5, but let me briefly spell out the issue. First, the phenomena must encompass more than just our experiences themselves, for familiar reasons from the failure of the positivist program (for example, Putnam's proposal so-understood would render our familiar scientific accounts of the world semantically equivalent to various skeptical or idealist accounts). And as mentioned above, proponents of formal criteria generally explicitly allow for distinctions that cut finer than empirical equivalence. The trouble is that proponents of semantic *in*equivalence in the cases of interest to proponents of formal criteria might take their differing metaphysical pictures of the world to engender differences in the phenomena. For example, recall again the flash and matter-density interpretations of some stochastic collapse formulation of a quantum theory. For the flash-theorist, the phenomena might include a short-lived pointer-shaped object momentarily appearing in space. For the matter-density theorist, the phenomena might instead include a certain field taking on high values across some pointer-shaped spatial region. What proponents of formal criteria need is some principled (even if vague) intermediate level of content that includes and extends beyond our experiences yet not far enough to also encompass these sorts of underlying metaphysical differences that they wish to jettison. One such account generates what I call the *physics deference* proposal, which I shall discuss in Sect. 5. For now, let us just grant that such an intermediate level of content can be carved out. The issue is that proponents of formal criteria would still need independent motivation for the non-mathematical premise that a formal equivalence proof that also preserves this intermediate level of content on some occasion suffices for semantic equivalence on that occasion. Without such motivation, the strategy under discussion would just amount to declaring the sought-after non-mathematical conclusions true by fiat. Yet now this extra (and to my mind dubious) non-mathematical premise—the bridge between a formal equivalence result that preserves the still amorphous intermediate level of content and semantic equivalence—is doing the heavy-lifting in securing the non-mathematical conclusions about semantic equivalence. So, although this proposal does not render formal criteria redundant, it does leave them with a subsidiary role. And most importantly, I am not aware of any attempt by proponents of formal criteria to defend the critical non-mathematical premise. So if some form of this proposal indeed undergirds the substantial non-mathematical import that has been claimed on behalf of formal criteria, such claims are premature.

The dialectic in the previous two paragraphs applies in general to any version of the reply under discussion, which recall claims that formal equivalence results, plus

[12] Putnam (1983, 39) is aware of this challenge. He offers a list of some candidate phenomena in the context of different Lorentz frames in Special Relativity. Still, one wants some precise characterization of what counts as the phenomena in general, otherwise we still would lack a general proposal for when a formal equivalence proof plus empirical equivalence licenses a substantial non-mathematical conclusion like a claim of semantic equivalence. Moreover, the dialectic I rehearse in this paragraph applies to Putnam's specific examples (in particular, we still lack justification for the non-mathematical premise that a formal equivalence result plus empirical equivalence plus explaining exactly these specific candidate phenomena suffices for semantic equivalence).

empirical equivalence, plus some extra ingredient suffice for semantic equivalence. The closer the extra ingredient comes to semantic equivalence itself, the more the charge of redundancy becomes stark. Yet the closer the extra ingredient comes to simply empirical equivalence, the more dubious the inference from formal equivalence plus empirical equivalence plus the extra ingredient to semantic equivalence becomes. As a result, the demand for some independent justification for the inference begins to seem all the more urgent.[13]

A final potential reply to Sklar-style counterexamples for proponents of formal criteria is that the sentential criteria are meant to apply only to first-order theories formulated in some privileged vocabulary, of the sort we encounter when doing physics, rather than ordinary natural language expressions like ‘lion’ or ‘tiger’. And as concerns such first-order theories, we cannot straightforwardly appeal to the interpretations we in fact employ to generate counterexamples to sentential criteria (most ordinary speakers never use the relevant vocabulary).

But this reply is also untenable. We must ask what the relevant interpretation of the vocabulary at issue is, relative to which sentential criteria are supposed to bear on semantic equivalence. The natural answer here is the interpretations employed by practicing physicists using the relevant vocabulary. And now two problems arise. First, even relative to these interpretations, it is dubious that the sentential criteria, even conjoined with empirical equivalence, will imply semantic equivalence. Consider my example above concerning the absolute speed of the centre of mass of the universe at a Newtonian world: those single-sentence theories arguably pass the privileged vocabulary restriction we are considering, and remain straightforward counterexamples to any extant sentential criterion plus empirical equivalence implying semantic equivalence on the relevant interpretations. Second, even bracketing that point, I will argue below against a similar proposal for non-sentential criteria (what I call the *physics deference* proposal). Analogues of the points I will make there can be made against the present attempt to restrict sentential criteria.

The upshot of this section is that the original counterexamples cannot be easily dismissed. I conclude that sentential criteria seem to be straightforwardly bad guides to semantic equivalence. I take this moral to cast doubt on any plausible and interesting view according to which such criteria have non-mathematical significance, let alone bear on semantic equivalence. For, any interpretations of the sentential theories at issue on which such a claim might be true will be far-fetched and patently unrelated to the interpretations we in fact employ and care about.

[13] For example, the dialectic applies to the version of the reply sketched by Hudetz (2019, 48), that formal criteria are sufficient for semantic equivalence when conjoined both with empirical equivalence and equivalence of “theoretical content beyond the empirical (if there is any)” (48). Hudetz is admirably upfront that this sketch must be fleshed out, but already we can see how the dialectic might go. If ‘theoretical content’ is just non-empirical content, then the proposal amounts to declaring semantic equivalence sufficient for semantic equivalence, and the formal criteria are rendered redundant. So ‘theoretical content’ plus ’empirical content’ must amount to the sort of intermediate-level of content described above in the main text. One precisification of such content makes the strategy exactly akin to the ‘physics deference proposal’ that I shall argue against in Sect. 5. Still, however the notion is made precise, the inference to semantic equivalence must be defended, not just assumed.

Some proponents of formal criteria will not be too fazed by this upshot: as mentioned above, some have switched to the non-sentential category-theoretic criterion. Perhaps only non-sentential criteria are meant to bear on semantic equivalence or have other significant non-mathematical implications? And unlike with sentences, it is less clear what might be meant by ‘the interpretations we in fact employ’ for the non-sentential representational vehicles often at issue in scientific practice. Indeed, I suspect some working on sentential criteria never viewed themselves as doing anything but pure mathematics or logic.[14] As I emphasized at the outset, I have no criticisms of the purely mathematical upshots of work on formal criteria (results claiming that such-and-such mathematical objects do or do not stand in so-and-so formal relation). Still, some work on sentential criteria is premised on such criteria having significant non-mathematical implications, in particular implying semantic equivalence (for example, the conclusions drawn by Barrett and Halvorson, 2017, 1060–1061). My conclusion in this section reveals this position to be untenable.

## 4 Two formulations of general relativity

I now turn to non-sentential criteria, focusing on the popular category-theoretic criterion. This criterion will occupy us for the rest of the paper. In this section I will walk through a central application of this criterion from the recent literature. We will use this test-case to explore whether there is a defensible view according to which non-sentential criteria like the category-theoretic criterion bear on semantic equivalence relative to some relevant interpretations. We shall see that the most plausible option here requires adopting what I will call the *physics deference* proposal. I will then argue against this proposal.

But let us start with the central example. The example concerns two mathematical formalisms in which one can couch General Relativity (hereafter GR). It is a paradigm success-case in the eyes of proponents of formal criteria, and moreover meant to cast doubt on the intelligibility of the venerable metaphysical debate about whether there are spacetime points. It thus presents an ideal example to use for our investigation of how one might vindicate non-sentential criteria having significant non-mathematical implications.

The first formulation is the textbook treatment in terms of differential geometry. Here one begins with a set of mathematical models, each of which contains a four-dimensional smooth manifold of points, and various mathematical structures, called tensor fields, defined on this manifold (including a Lorentzian metric field), all satisfying some equations. The second formulation involves an algebraic structure

[14] This diagnosis strikes me as a plausible reading of Tsementzis (2017), and is suggested by the prominence of examples from pure mathematics in Halvorson (2012) and Barrett and Halvorson (2016a, 2016b). (Though, as I will note shortly in the main text, in other places these latter authors make claims that presuppose more than purely mathematical ambitions for sentential criteria.) The diagnosis is also suggested by work applying sentential criteria to different logics, such as Wigglesworth (2017), Dewar (2018), and Woods (2018).

called in Einstein algebra, due to Geroch (1972). Roughly, Einstein algebras begin with every smooth real-valued function on some manifold of points, and then kick away the manifold and understand these functions as algebraic objects in their own right. Geroch then showed how to transform any tensor field from the textbook formulation in terms of differential geometry into an operation on this algebra of functions.[15] Do not worry if this mathematics is unfamiliar. The important point is that, because Geroch explicitly rigged up Einstein algebras to reproduce any general relativistic solution space couched in the textbook formalism of differential geometry, unsurprisingly there is a strong structural resemblance between analogous general relativistic solution spaces couched in the different formalisms. (Here by 'analogous' I mean the solution spaces employ the same sorts of matter fields and impose the same constraints on these fields.) This hunch has been made precise using the category-theoretic formal criterion by Rosenstock, Barrett, and Weatherall (hereafter RBW), in their (2015).

Spelling out RBW's mathematical result would require going through requisite background in algebra and category theory. Fortunately, the informal idea behind what they show will suffice for our purposes. A category consists of a collection of objects and a collection of arrows, which are mappings from one object to another required to satisfy various axioms (see Mac Lane, 1998, ch. 1 for the basic background). The first category at issue in RBW's proof is the category of every model in a solution space couched in the textbook formalism, each of which contains a manifold of points with a metric defined on that manifold (and the arrows of the category are isometries). The second category is the category of every model in the analogous solution space couched in the Einstein algebra formalism, each of which contains an algebra of functions with the operation that is the analogue of a metric defined on that algebra (and the arrows of the category are algebra homomorphisms). Given the structural analogies between the formalisms, these two categories resemble one another at some low-level of abstraction, and this is precisely what RBW (2015) show (in particular that the categories are dual). Moreover, the methods they use extend to most general relativistic solution spaces.

How is this mathematical result supposed to bear on whether there are spacetime points? The reason is that some have claimed to be using the different formalisms to express conceptions of the world that disagree over this question. The Einstein algebra formalism was first introduced into foundational discussions of spacetime theories by Earman (1979, 1986, 1989), under the heading of "Leibniz algebras". Earman took himself to be using the Einstein algebra formalism to express a metaphysics of spacetime that "eschews substantivalism in the form of spacetime points" (1989, 193), by contrast with how he was using the standard formalism. So Earman at least took himself to have associated analogous solution spaces couched in the different formalisms with different propositional contents: that is, to have

[15] See Rynasiewicz (1992) and Rosenstock et al. (2015) for clear expositions of the formal details. These expositions differ somewhat, but the differences need not concern us here.

effected an interpretation where these representational vehicles are not semantically equivalent.[16] I shall explore exactly how this process might work below.

How is learning about RBW's formal proof meant to bear on Earman's proposal? Is there a tenable view according to which the proof has non-mathematical significance, and moreover somehow casts doubt on the intelligibility of the question of whether there are spacetime points? RBW (2015) seem to think so; in their conclusion they write:

> [Our result] establishes a sense in which the Einstein algebra formalism is equivalent to the standard formalism for general relativity. This sense of equivalence captures the idea that, on a natural standard of comparison, the two theories have precisely the same mathematical structure—and thus, we claim, the same capacities to represent physical situations ... Insofar as one wants to associate these two formalisms with "substantivalist" and "relationist"—or at least, non-substantivalist—approaches to spacetime, it seems that we have a kind of equivalence between different metaphysical views about spatiotemporal structure. (315)

Let us now turn to how one might try to make good on this conclusion, and the more general doctrine that the category-theoretic formal criterion has significant non-mathematical implications. As mentioned, for concreteness I shall stick with this one example throughout, and the attendant debate over whether there are spacetime points. The example has the nice features of being reasonably familiar, and perhaps avoiding more high-powered ideology at issue in other debates in metaphysics (such as fundamentality).[17] However, I want to emphasize that the points I make are not

[16] Earman hoped that the Einstein algebra inspired metaphysics would offer a metaphysics of spacetime that addresses the hole argument (Earman & Norton, 1987 for the classic statement of this argument, and Pooley (2013, Section 7) and Norton (2015) for overviews of the many replies the argument has provoked). This motivation is widely taken to have been undermined by Rynasiewicz (1992), who constructed an analogue of the hole argument in terms of Einstein algebras. However, this issue, and the motivations for Earman's position generally, will not bear on my arguments. Similarly, Earman used 'relationism' to encompass more than just the negation of substantivalism. Hence, he took his Einstein algebra inspired metaphysics to offer a novel third view, that vindicates certain aspects of both substantivalism and relationism, rather than a relationist view. But nothing in what follows turns on the terminological question of which views we label 'relationist'.

[17] This latter issue is contentious. Some metaphysicians (for example Fine, 2001; Schaffer, 2009) argue that most existence questions, whether 'are there spacetime points?' or 'are there numbers?', are trivially answered in the affirmative. They then employ some additional ideology to carve what they see as more interesting questions, such as 'are there numbers at the *fundamental* level?', or 'are there *really* numbers?', and so on. Some have pushed this general line about the substantivalism/relationism debate in particular, proposing that the debate cannot concern merely the *existence* of spacetime points, which even relationists can grant (for different versions of this line, see Field, 1984; North, 2018). I shall ignore this wrinkle in the main text, but incorporating it would not challenge my arguments. The viability of the purely existential framing may also undermine some of RBW's skepticism about the debate; at one point they concede "of course, it remains open to the person who wants to give [the two formalisms] a metaphysical significance to say that one of them is more fundamental than the other" (315), yet find their deflationary conclusion "far more philosophically interesting" (316). Notice though that even skeptics about 'fundamentality' talk can pose the question of whether there are spacetime points. Moreover, if my arguments succeed then their deflationary conclusion is either ill-posed (see Sect. 2) or else garners no support from their formal proof.

wedded to this example. For instance, Weatherall (2015) proves a category-theoretic equivalence result between the solution space of classical electromagnetism framed in terms of the Faraday tensor and the analogous solution space that employs the vector potential. I assume he would take his result to have non-mathematical, and likely deflationary, implications for a debate over whether there is a fundamental property corresponding to the vector potential. And I could make analogues of my points below using this debate. Similarly for the related debate at issue in his discussion concerning whether evidence for Newtonian theories supports believing spacetime to be flat or curved. Readers with independent reasons to dislike the Einstein algebra example may prefer to substitute one of these alternatives. Of course you may have independent reasons to be skeptical of high-powered metaphysical ideology like fundamentality. But such skepticism that stems from considerations other than a formal equivalence proof is irrelevant to the present dialectic. Similarly for other reasons one may have for being skeptical of some metaphysical debate that are independent of any formal equivalence proof.

## 5 Physics deference

The passage from RBW just quoted above is one of the few places where we are given a hint as to how a formal equivalence proof is meant bear on semantic equivalence. But notice that it is not clear what is being claimed. The comment about "precisely the same mathematical structure" is a property of the representational vehicles at issue understood as uninterpreted mathematical objects. How is this purely mathematical fact supposed to have implications for the non-mathematical world? One might be tempted to put the moral of RBW's formal proof as showing that "the metaphysical view" associated with the textbook formalism is the same as "the metaphysical view" associated with the Einstein algebra formalism; or similarly that "what is truly represented" by the textbook formalism is also "what is truly represented" by the Einstein algebra formalism. But we already saw the problems with such glosses in Sect. 2. No formal object on its own, like a solution space in *any* formalism, has some metaphysical view baked-in, or represents the world as being any way whatsoever. In that section we also saw that the "same capacities to represent physical situations" gloss that RBW offer in the passage cannot be taken at face-value. Let me discuss this gloss further, given its prevalence.

As we saw, on the most straightforward reading of the gloss all representational vehicles considered on their own have the *same* "capacities to represent physical situations" because of trivial semantic conventionality: namely, the capacity to represent just about any physical situation whatsoever. How are we supposed to instead read the gloss so that it might apply non-trivially? The works of those who use the gloss do not tell us. Moreover, it is unclear what precise characterization of the gloss could serve the purposes of proponents of formal criteria: namely, render plausible both the inference from a category-theoretic formal equivalence proof to 'same capacities to represent physical situations', properly understood, and also the further inference from this latter property to semantic equivalence relative to the interpretations operative in whatever metaphysical debate is at issue. Some common qualifications of the gloss that I have encountered state that a categorical equivalence proof reveals the vehicles at issue to have the same capacities to represent physical situations *faithfully* or *aptly* or *for some specific purpose*. For any manner of making such glosses precise, two features of the resulting relation must be defended: (i) that the vehicles at issue in the disputed cases (which have been shown to be equivalent relative to some formal category-theoretic criterion) thereby also stand in the relation, and (ii) that if two vehicles stand in the relation, they are plausibly thereby also semantically equivalent relative to the interpretations operative in whatever metaphysical debate is at issue. For every precisification of the common qualifications of which I am aware, either (i) or (ii) becomes a non-starter.

To see a rough example, suppose we have some grip on the idea of an interpretation that assigns contents to non-sentential representational vehicles in a manner where any content represented must be in some sense mirrored in, or isomorphic to, the structural or syntactic properties of the vehicle itself.[18] Call such interpretations ‘picture-theory interpretations’; they are meant to capture the intuition that a map, for example, might be in some sense intrinsically better suited to represent certain properties of certain regions rather than contents of any other sort. I am skeptical that this idea can be made precise in a tenable manner, but we can bring out the problem even conceding its cogency. With this notion in hand, here is one attempted precisification of the claim that two non-sentential representational vehicles have “the same capacities to represent physical situations” (perhaps intended by the ‘faithfully’ or ‘aptly’ qualifiers): there is no picture-theory interpretation relative to which the vehicles have different contents. The trouble is that this precisification flouts requirement (ii) from the previous paragraph, and hence cannot fulfill the ambitions of proponents of formal criteria. For there is no reason to believe that picture-theory interpretations, which can assign only very weak semantic contents, are those operative when philosophers have asked the metaphysical questions at issue, such as whether there are spacetime points.[19] We shall see more on this theme in Sect. 6 as well, and I suspect the same issue would arise for other proposals of this kind.

[18] Thanks to an anonymous referee for suggesting that I address this kind of proposal. There are various attempts to spell out the very rough idea in a more plausible and precise manner in the vast literature on scientific modeling. For some helpful surveys of the lay of the land here, see Suarez (2010) and Frigg and Nguyen (2016).

[19] For instance, even given my tenuous handle on the notion of a picture-theory interpretation, arguably such interpretations can assign only contents that are *purely qualitative* (not about any particular objects). If so, such interpretations can at best assign only contents like *there are some spacetime points or other standing in such-and-such pattern of field values*, rather than contents describing which particular spacetime points have which field values. Yet the latter non-qualitative contents are the ones required to even formulate the hole argument, which is perhaps the central argument that animates the contemporary substantivalism/relationism debate. (For references to some overviews of the hole argument, see footnote 16.) In Sect. 6 we’ll see that arguably even purely qualitative yet topic-specific contents of the sort just described (such as purely qualitative contents about spacetime points) require going beyond the representational resources of anything like a picture-theory interpretation.

What about the ‘for some specific purpose’ qualifier on “same capacities to represent physical situations”? This brings us to the most initially plausible and common qualification that I have encountered. In particular, here is what I suspect proponents of formal criteria will say at this juncture. The qualifier ‘physical’ in the “same capacities to represent physical situations” gloss is doing important work. RBW’s category-theoretic proof reveals the standard solution space and the Einstein algebra solution space to be equally adequate *for the purposes of doing physics*; that is, the proof about the two solution spaces’ mathematical structure somehow reveals that physicists could use the solution spaces interchangeably. Let us say that two representational vehicles are *physically equivalent* relative to some interpretations just in case the vehicles say the same thing about any subject matter relevant to physics relative to those interpretations. As I shall expand on shortly, it is far from clear what precisely this gloss might amount to; still, uncontroversially (i) physical equivalence at least implies empirical equivalence, and (ii) things like the values of fields or distribution of matter across spacetime will fall under the subject matter relevant to physics. Assuming we have some handle on the notion of physical equivalence for now, the claim under consideration is that RBW’s proof reveals the two solution spaces to be physically equivalent relative to the interpretations at issue when practicing physicists use these representational vehicles. The second half of the passage quoted above, which moves from physical equivalence to collapsing the substantivalism/relationism debate, embodies the further step to semantic equivalence. Let us call the *physics deference proposal* the claim that, for every interpretation *I*, if two representational vehicles are physically equivalent relative to *I* then they are also semantically equivalent relative to *I*. The physics deference proposal replaces the positivist’s empirical equivalence with physical equivalence as sufficient for collapsing some distinction. Weatherall adopts something like the proposal when he summarizes the vision underlying his interest in the category-theoretic criterion as follows: “one allows that the distinctions that one can sensibly draw depends on the structure of the world. And the best guide to understanding what those distinctions are will be to study the properties of and relationships between our best physical theories” (Weatherall, 2015, 1088). And something in the vicinity of the physics deference proposal is arguably implicit in much of the work on formal criteria.[20] In this section and the next I shall argue that the physics deference proposal is untenable. Yet this is the only remaining view that I am aware of that may vindicate the striking non-mathematical conclusions drawn from extant

[20] Additional evidence for this claim comes from the common practice amongst philosophers of science, especially in the literature on dualities, of using ‘physical equivalence’ as a label for what I am calling ‘semantic equivalence’. If my arguments against the physics deference proposal are successful then this terminology is highly misleading. For relevant citations and discussion, see Butterfield (2018, 34). Compare also Putnam’s remark—when arguing for the semantic equivalence of traditional continuous conceptions of spacetime and gunky conceptions (on which there are no measure-zero points)—that “it can make no difference to *physical explanation* whether we treat space-time points as ‘real’ or as mere logical constructions” (Putnam, 1983, 43, emphasis original). The difficulties writing down physical laws in gunky spacetimes suggests otherwise (see, for instance, Arntzenius & Hawthorne, 2005 and Arntzenius, 2008, 2012, ch. 4).

formal equivalence proofs. I will conclude that non-sentential criteria, like their sentential kin, are of limited non-mathematical significance.

Let me stress first, though, that it is far from clear what ‘physical equivalence’, as it appears in the proposal, even amounts to. The idea relies on the notion of saying the same thing about any subject matter relevant to physics. We should allow that this subject matter need not be simply what actual physicists would claim to be relevant for their purposes, so the ‘deference’ at issue involves some idealization. But what precisely does this subject matter consist in? It better encompass more than just our observations, otherwise the physics deference proposal will be as implausible as the discredited positivist criteria for semantic equivalence itself. And as mentioned in Sect. 1.1, proponents of formal criteria explicitly want to allow that some intelligible distinctions cut finer than empirical equivalence. However, what exactly is this notion of physical equivalence that strengthens empirical equivalence yet falls short of encompassing the sorts of distinctions metaphysics-oriented philosophers of physics debate about? It must not encompass the latter distinctions, otherwise differences like those embodied in substantivalism and relationism will engender physical *in*equivalence, rendering the physics deference proposal powerless to support deflationary morals about such debates. Yet it is not clear that an intermediate line, even a vague one, can be drawn that somehow siphons off the unwanted “metaphysical” content (recall the discussion of “intermediate” levels of content from Sect. 3). The reason is that familiar issues about the theory-ladenness of observation for empirical equivalence may recur at any such intermediate level of “physical” or “theoretical” content that strengthens empirical equivalence but falls short of semantic equivalence itself. For example, some argue that the very notion of a field must be understood as a property or relation distributed over a substantival spacetime (see Field, 1984). If that view is correct then the distribution of field values across spacetime—a seemingly uncontroversial example of the subject matter relevant to physics—would itself presuppose a stance on the “metaphysical” debate between substantivalism and relationism. Appealing to some notion of an ‘O-term’ or ‘non-structural term’ as providing the basis for an account of physical equivalence does not seem to offer any guidance here. It seems purely stipulative to deem, say, ‘spacetime point’ a T-term rather than O-term (or non-structural term), and then to argue on that basis that the substantivalism/relationism debate is ill-posed, provided we are already allowing for some O-terms that go beyond our experiences themselves.

Fortunately, I think we can bracket these concerns about whether proponents of formal criteria can flesh out the physics deference proposal, and argue against the proposal even granting that some workable notion of physical equivalence can be found. Similarly, because my main arguments target the inference embodied in the physics deference proposal itself—from physical equivalence to semantic equivalence—I am content to bracket another sort of concern one might have. Notice that even granting the physics deference proposal, there still remains the prior question of whether the category-theoretic criterion is in fact a good guide to physical equivalence relative to the interpretations employed by practicing physicists. And one might complain that the category-theoretic criterion itself allows one to generate different verdicts in most cases, depending on how one couches the non-sentential representational vehicles at issue as categories (in particular, which arrows one includes in the categories). Yet only one such verdict can in fact track which vehicles practicing physicists use interchangeably. However, let us also set concerns stemming from this front aside, despite my skepticism that there could be some metaphysically-neutral way to determine which arrows to include in each category at issue. So let us grant that extant formal criteria between non-sentential representational vehicles perfectly track some rigorously characterized notion of physical equivalence relative to the interpretations at issue amongst practicing physicists. What would follow?

Now extant formal equivalence proofs would have *some* non-mathematical implications: they would illuminate physical equivalence, which however ultimately spelled out will concern the non-mathematical world. Still, I do not think that proponents of formal criteria should take much solace in this result, for two reasons.

First, notice that even if the physics deference proposal were true (that is, even if we could infer semantic equivalence from physical equivalence), this non-mathematical premise, rather than any formal equivalence proofs, would arguably be doing the central work in delivering the significant non-mathematical conclusions about semantic equivalence. Given some putative metaphysical debate, the question we must ask would be whether the distinction at issue concerns subject matter relevant to physics. If not, the physics deference proposal dictates that we commit the debate to the flames. Perhaps that is the view of proponents of formal criteria. But if it is, then they should focus their energies on defending the controversial physics deference proposal itself. All of the work proving formal equivalence results would seem to be dialectically far less central than the integral philosophical premise: it is the tendentious non-mathematical inference embodied in the physics deference proposal, not any formal equivalence result, that opponents of the striking non-mathematical conclusions about semantic equivalence will be inclined to challenge.

However, proponents of formal criteria will likely argue that our best guide to physical equivalence are formal equivalence proofs (perhaps conjoined with empirical equivalence), and hence that such proofs are still of central importance under the physics deference proposal. The trouble is that, even if we were to view vindication of the physics deference proposal as a vindication of formal equivalence results, we can nevertheless argue against the physics deference proposal directly. For the proposal seems to deliver false verdicts in various cases. Let me walk through one case, though many could be given. I will then argue in the next section that the cases in fact at issue in the literature (including our central example of the two formalisms for GR) are not importantly disanalogous from this one.

Philosophers of mind interested in panpsychism debate over whether microscopic objects are phenomenally conscious.[21] Now imagine the standard mathematical formalism for doing particle physics embellished in two different ways with a new

[21] Or alternatively instantiate a "proto-consciousness" intrinsic property that grounds facts about which macroscopic objects are phenomenally conscious. I shall ignore this wrinkle in the main text.

expression referring to the property of phenomenal consciousness: one formalism adds the claim that all fundamental particles are conscious, and the other adds the claim that no fundamental particles are conscious.[22] To dramatize this, you might imagine a physics community where writing your claims down in one font indicates the panpsychist option, and in some other font the anti-panpsychist option, yet there is no neutral formalism that physicists could use.[23] These formalisms will be structurally analogous, and can easily be contrived to be deemed equivalent by any formal criterion. And it is plausible that the formalisms will be physically equivalent relative to the interpretations at issue when practicing physicists in this hypothetical community use these formalisms. For, however we cash out the notion of physical equivalence, arguably the subject matter of physics will encompass things like the trajectories taken by fundamental particles, their masses, and so on, not the further issue of whether the particles also happen to be conscious. Indeed, whether there is something it is like to be the fundamental particles seems orthogonal to the concerns of physicists.[24] Yet notice that this result of physical equivalence on its own is of limited non-mathematical significance: it merely reflects what subject matter happens to concern physicists. The physics deference proposal now recommends drawing the further inference from physical equivalence to semantic equivalence, which would be of considerably more non-mathematical significance. Yet this further inference is problematic: nobody should take the fact that physicists need not concern themselves with which things are phenomenally conscious to challenge either the intelligibility of debates over which things are conscious, or the overwhelmingly plausible claim that there is a contentful distinction to be drawn between what is and is not phenomenally conscious. Thus the physics deference proposal delivers the wrong result: we should all reject a wholesale deference to physical equivalence as concerns the limits of intelligibility, on pain of having to collapse the distinction between what is and is not phenomenally conscious (and much more besides).

Proponents of formal criteria will likely reply that the cases they are interested in are importantly disanalogous from this case and other analogous problem cases we might have employed instead.[25] That is, they will likely concede that the physics

[22] Field-theoretically, this could be phrased in terms of whether certain excitations in various quantum fields (which license our talk about a particle being present) also bring about phenomenal consciousness.

[23] We can also imagine a difference that affects the mathematical models the physicists use to express their views. The argument in the main text does not turn on the precise details of how the difference gets formally expressed: many options will still render the resulting vehicles equivalent relative to all extant formal criteria.

[24] In the main text I am bracketing certain fringe views where the truth of panpsychism would percolate up to spoil physical equivalence; for instance, understandings of quantum mechanics where consciousness triggers wave-function collapse. If even the truth of panpsychism were claimed to fall under the subject matter of physics, we could employ numerous other examples instead (such as debates about whether there are moral properties, abstract objects, and so on).

[25] If they were instead to bite the bullet about this case and every other analogous case, I would then lean more on the concerns that I raised but set aside above: (i) the lack of a principled and metaphysically-neutral characterization of physical equivalence, and (ii) the lack of a principled and metaphysically-neutral procedure for couching whatever non-sentential representational vehicles are at issue as categories.

deference proposal recommends the wrong verdict about phenomenal consciousness, but claim that this point does not challenge the application of the proposal to the debates they are concerned with. To discuss this reply, I will continue to focus on our central example of the substantivalism/relationism debate and the two formulations of GR. The task facing proponents of formal criteria is the following: isolate some salient difference between the debate over whether there are spacetime points, on the one hand, and the debate over panpsychism, on the other, which renders plausible the claim that we should follow the verdicts of the physics deference proposal concerning the former debate, despite the proposal delivering false verdicts concerning the latter debate. Successfully completing this task would block my argument. However, I will now argue that these cases are in fact not disanalogous in any respect that bears on the applicability of the physics deference proposal.

## 6 Natural language glosses in mathematical physics

Notice that in the panpsychism case we appealed to a concept we understand independently of the representational vehicles at issue in order to engender the vehicles' semantic inequivalence, despite their physical equivalence and any formal resemblance between them. In particular, given that we antecedently understand the concept of phenomenal consciousness, we could straightforwardly use it to ensure that the two representational vehicles fail to be semantically equivalent even relative to the interpretations adopted by practicing physicists, by stating that one but not the other is being used to describe all fundamental particles as being phenomenally conscious.

Thus the disanalogy proponents of formal criteria must press is that we lack an analogous independent understanding of the concept of a spacetime point. Otherwise, we could straightforwardly use it to express the distinction between a world containing spacetime points and a world containing none, and thereby effect an interpretation of the different solution spaces according to which one but not the other represents each nomic possibility as containing spacetime points. We would then have a failure of semantic equivalence, exactly as in the panpsychism case. This may be what Earman took himself to be doing when he said that he was using the Einstein algebra formalism to represent a metaphysics of spacetime that "eschews substantivalism in the form of spacetime points" (1989, 193). From this dialectical position, the physical equivalence of the two representational vehicles, and hence RBW's formal proof of their structural resemblance, would seem completely beside the point as concerns their semantic equivalence.

How might proponents of formal criteria argue that our understanding of the concept of a spacetime point is importantly different from our understanding of phenomenal consciousness? If RBW's proof is to be relevant, the position must somehow tether our understanding of the concept of a spacetime point to how physicists use the textbook differential geometry formalism (with its manifold of points) rather than other formalisms (like the Einstein algebra formalism). The idea would then continue that our evidence for the intelligibility of the substantivalism/

relationism debate is therefore premised on practicing physicists drawing a distinction between the differential geometry solution space and solution spaces in other formalisms. If that were right, then once we learn that these representational vehicles are physically equivalent (relative to the interpretations practicing physicists employ), we plausibly ought to jettison our belief that we understand what it is to be a spacetime point, and with it our belief in a contentful substantivalism/relationism debate. The general proposal here is as follows: restrict the physics deference proposal to interpretations of representational vehicles that are effected using only concepts analogous to our concept of a spacetime point if the claims in this paragraph are correct. That is, the general proposal is that we should infer semantic equivalence from physical equivalence only in cases where we have appealed to no antecedently understood concepts. This *restricted physics deference proposal* excludes the panpsychism case and the other analogous cases we might have employed, thereby dodging my arguments.

But does the restricted physics deference proposal allow proponents of formal criteria to continue to maintain that their proofs support deflationary positions about certain debates in the metaphysics of physics, or have other substantial non-mathematical implications? Let me confess that I have a tenuous grip on what exactly is being claimed here on behalf of the concept of a spacetime point. We are all familiar with talk about concepts “given by their role in a theory.” Yet notice that in the present setting ‘theory’ talk must somehow encompass both a particular mathematical formalism as well as some semantic content, otherwise an equivalence proof between analogous solution spaces in different formalisms cannot purport to undermine our claim to understand the relevant concepts. Thus the position being articulated concerning our concept of a spacetime point cannot be modeled straightforwardly on the standard Ramsey sentence method of Lewis (1970), which is not tethered to any mathematical formalism. Notice also that uncontroversially the claim that must be made here on behalf of the concept of a spacetime point would be a very particular semantic doctrine. Yet proponents of formal criteria have never tried to argue for any such doctrine. Thus, even bracketing my arguments to come, I regard unearthing this potential presupposition of the view that formal criteria have non-mathematical implications as an important result in its own right, and one which may help to focus the debate over the implications of formal criteria on the central philosophical issues moving forward.

Nevertheless, even given a tenuous understanding of what the restricted physics deference proposal is claiming, we can argue that the proposal does not in fact deliver the desired non-mathematical implications. The reason is that the debates in the metaphysics of physics at issue generally involve concepts that are importantly analogous to those we antecedently understand, like phenomenal consciousness. Hence, this restricted physics deference proposal will not recommend an inference to semantic equivalence even in the desired cases, but will instead exclude them along with the problematic panpsychism case. I shall offer three arguments in defense of this claim, continuing to focus on our central example of the concept of a spacetime point and the substantivalism/relationism debate. Each argument purports to show that our understanding of the concept of a spacetime point never depended on deference to whether practicing physicists draw a certain distinction with a particular formalism. I should stress, though, that analogues of my arguments carry over to the other debates in the metaphysics of physics we might have employed instead.

First, notice that an analogue of the substantivalism/relationism debate in the context of GR goes back at least to Newton and Leibniz, who debated about whether there are spatial and temporal points in pre-relativistic physics well before the advent of differential geometry. Arguably a close analogue of the debate can even be found in pre-Socratic discussions about the reality of unoccupied space (“the void”); and these discussions of course occurred significantly before the advent of contemporary physics or mathematics. These historical precedents should make us very uneasy about claims that our concept of a spacetime point is somehow tethered to physicists marking out certain distinctions using the textbook formulation of GR. These precedents for the contemporary debate suggest that whether evidence for some physical phenomena supports believing that there are things like spacetime points is not some esoteric concern of recent metaphysics.

Second, the concept of a spacetime point arguably has a well-defined role independent of any particular physics or mathematics, and is tied to some of our core concepts like that of an object. For instance, spacetime points might be understood as those objects that material objects (like tables, chairs, or particles) are located at, or those objects that are the bearers of field values, or those objects that instantiate fundamental spatiotemporal properties and relations (which may in turn be a primitive concept, or defined via a connection to dynamical laws and the concept of an inertial trajectory). The details are up for debate, but the point again is that our understanding of the concept does not seem wedded to which distinctions practicing physicists draw or somehow to the textbook differential geometry formalism for GR.

Finally, my main argument stems from the foundational issues mentioned towards the outset. As I have emphasized throughout, a mathematical object, like some general relativistic solution space in the textbook differential geometry formalism, is on its own just a piece of mathematics, which does not represent the world as being any way whatsoever. Investigate the properties of some such object as much as one likes; explore all of the mathematical mappings it does or does not stand in to other uninterpreted mathematical objects; and still no proposition or semantic content will somehow come out, let alone the contentful distinction between a world having or lacking spacetime points. Rather, even physicists must somehow use other antecedently interpreted representational vehicles to endow uninterpreted mathematical objects with semantic content in the first place; for instance, natural language glosses on the mathematics, where these glosses employ concepts the physicists antecedently understand. Notice that this was also exactly our diagnosis of how the failure of semantic equivalence arose in the panpsychism case.

This method of proceeding is apparent in any physics textbook. For example, notice that no textbook on GR just displays a general relativistic solution space (understood as an uninterpreted mathematical object) and then expects the reader to arrive at some semantic content encoding what a world where GR is true happens to be like. Rather, the mathematics at issue when presenting GR (or any other physics) is qualified with natural language glosses, like ‘representing spacetime’, ‘representing mass density’, and so on, which serve to characterize what the mathematics is being used to represent. Crucially, if this method is to succeed then these natural language concepts must be at least partially understood independently of the mathematics at issue. We see, then, that even physicists must take for granted some antecedent understanding of the concept of a spacetime point, given that they employ this concept in characterizing what they are using the formalism of GR to represent. This general mode of proceeding is also apparent throughout the metaphysics of physics: diverging views in some debate often arise by qualifying one and the same mathematical object with different natural language glosses (recall the flash and matter-density example). It is generally assumed that we have an antecedent understanding of the distinctions involved in these glosses (such as between an object and a property, a law and a non-law, and so on).[26]

It is tempting, albeit mistaken, to regard the conceptions of reality inspired by our best physics as somehow arising entirely from how physicists use certain mathematical formalisms, with no need for contributions from the rest of our conceptual repertoire, including those embodied in natural language. And I think this mistaken thought plausibly underlies the guiding vision of proponents of formal criteria of equivalence, where a purely formal relation (even conjoined with empirical or physical equivalence) could have significant non-mathematical implications, and in particular illuminate semantic equivalence. Once we appreciate the importance of antecedently interpreted representational vehicles, like natural language glosses, in arriving at a conception of reality from our best physics, the failure of formal criteria to bear on semantic equivalence is unsurprising. For such criteria by their nature ignore these integral components in the presentations of our best physics, instead considering only the mathematics we use to express our best physics. Supplementing a formal criterion with empirical or physical equivalence does not remedy the problem, given that the distinctions we can draw with these additional representational resources, like natural language glosses, can and generally do cut finer than empirical and even physical equivalence.[27] Of course

[26] Precedents for views in the spirit of the one I am sketching here can be found in Sklar (1980)—who emphasizes the importance of semantic connections to ordinary concepts (like that of an object) via analogies when doing science—and also in Maudlin (2018)—who emphasizes the importance of what he calls a “commentary” to supplement any given mathematical formalism in order to arrive at some semantic content. A similar moral has also been drawn in the vast literature on scientific modeling, where it is now widely recognized that features beyond a mathematical model itself—such as the intentions and natural language glosses of the scientist using the model—are integral to effecting an interpretation of the model. For some helpful overviews of this literature see the works cited in footnote 18. Nguyen (2017) applies this moral about scientific modeling to the debate over theoretical equivalence, supporting the extant critiques mentioned in Sect. 2 (in particular, the critiques of Sklar, 1982; Coffey, 2014).

[27] Category-theoretic criteria allow some additional freedom, stemming from the choice of arrows when couching some mathematical physics in category-theoretic terms. However, this point also does not challenge the moral in the main text. The reason is that the distinctions we can draw with additional representational resources also extend beyond those we can draw by different choices of arrows when deciding on a category-theoretic representation, granting the standard contentful significance of such arrows as erasing distinctions between possibilities. This fact also diminishes the interest of the claims made on behalf of the non-mathematical significance of category-theoretic criteria by Barrett (2019) and Weatherall (2019), to the effect that we can use different choices of arrows to diagnose ambiguities in

these brief remarks are only the beginning of a complete story about how mathematical physics works: we still face the question of how some of these antecedently understood concepts come to have content in the first place, whether our concept of space, time, object, and so on. However, this is just the familiar and perennial problem of metasemantics, which is everyone's problem, and the subject of considerable ongoing investigation. Given that mathematical physics works, and natural language glosses involving these concepts play an important role in its operation, the problem must have some solution.[28]

## 7 Taking stock

We have been searching for a defensible view according to which formal criteria of equivalence, whether sentential or non-sentential, might have significant non-mathematical implications, in particular illuminating semantic equivalence. Our investigation suggests that there is no such view to be found, thereby supporting the moral that formal criteria do not bear on which distinctions are intelligible, nor somehow impugn the kinds of debates that occupy metaphysics-oriented philosophers of physics. Perhaps there is some such view that I have missed, though our investigation does not recommend optimism on this score. Moreover, I think we can confidently claim that, if there is such a view, spelling it out precisely and defending it will be a highly non-trivial task, one bound up with controversial issues in other branches of philosophy. Proponents of formal criteria who want to maintain that their results have significant implications beyond pure mathematics should make explicit which philosophical doctrines they take to support this stance, and try to defend such controversial doctrines directly. If nothing else, I hope my discussion spurs proponents of formal criteria to shift some of their focus from proving formal equivalence results to this foundational task, which is integral to the philosophical interest of formal criteria yet has been relatively neglected.

**Acknowledgements** Thanks most of all to Cian Dorr and Ben Holguín, for extremely helpful comments and discussion at every stage of the paper's development. Many thanks also to David Albert, Thomas Barrett, Dave Chalmers, Hartry Field, and Tim Maudlin, for comments and discussion that led to significant improvements. Finally, for very helpful comments and discussion about this material, thanks to Kevin Coffey, Michael Strevens, Brad Weslake, the anonymous referees, as well as the audiences at NYU's work in progress seminar and Caltech's philosophy of physics reading group.

Footnote 27 continued
how some mathematical physics can be used to represent the world: restricting this claim to any natural class of relevant interpretations (as we must, recall Sect. 2), these ambiguities can involve distinctions that also cut finer than choosing which differences between possibilities wash out when testing for equivalence, which is what a category's arrows are used to represent.

[28] For related remarks about the importance of more general metasemantical issues in the philosophy of language and mind to questions about how we use mathematical objects to represent the world when doing science, see Callender and Cohen (2006).

## References

Arntzenius, F. (2008). Gunk, topology, and measure. *Oxford Studies in Metaphysics, 4,* 225–247.

Arntzenius, F. (2012). *Space, time, and stuff*. Oxford: Oxford University Press.

Arntzenius, F., & Hawthorne, J. (2005). Gunk and continuous variation. *The Monist, 88*(4), 441–465.

Barrett, T. W. (2015). On the structure of classical mechanics. *The British Journal for the Philosophy of Science, 66*(4), 801–828.

Barrett, T. W. (2019). Equivalent and inequivalent formulations of classical mechanics. *The British Journal for the Philosophy of Science, 70*(4), 1167–1199.

Barrett, T. W., & Halvorson, H. (2016a). Glymour and quine on theoretical equivalence. *Journal of Philosophical Logic, 45*(5), 467–483.

Barrett, T. W., & Halvorson, H. (2016b). Morita equivalence. *The Review of Symbolic Logic, 9*(3), 556–582.

Barrett, T. W., & Halvorson, H. (2017). From geometry to conceptual relativity. *Erkenntnis, 82*(5), 1043–1063.

Butterfield, J. (2018). *On dualities and equivalences between physical theories*. Manuscript .

Callender, C., & Cohen, J. (2006). There is no special problem about scientific representation. *Theoria, 55,* 67–85.

Coffey, K. (2014). Theoretical equivalence as interpretative equivalence. *The British Journal for the Philosophy of Science, 65*(4), 821–844.

Creath, R. (2020). Logical empiricism. In E. N. Zalta (Ed.), the stanford encyclopedia of philosophy, Metaphysics Research Lab, Stanford University summer 2020 edn.

Curiel, E. (2014). Classical mechanics is Lagrangian: It is not Hamiltonian. *The British Journal for the Philosophy of Science, 65*(2), 269–321.

Dewar, N. (2018). On translating between logics. *Analysis, 78*(4), 622–630.

Earman, J. (1979). Was Leibniz a relationist? *Midwest Studies In Philosophy, 4*(1), 263–276.

Earman, J. (1986). Why space is not a substance (at least not to first degree). *Pacific Philosophical Quarterly, 67*(4), 225–244.

Earman, J. (1989). *World enough and space-time*. Cambridge: MIT Press.

Earman, J., & Norton, J. (1987). What price spacetime substantivalism? The hole story. *British Journal for the Philosophy of Science, 38,* 515–525.

Field, H. (1984). Can we dispense with space-time? In *PSA: Proceedings of the biennial meeting of the philosophy of science association* (Vol. 2, pp. 33–90). Basil Blackwell. Reprinted in Field. (1989). Realism, Mathematics, and Modality, 171–226

Fine, K. (2001). The question of realism. *Philosophers' Imprint, 1*(1), 1–30.

Friedman, M. (1999). *Reconsidering logical positivism.* Cambridge: Cambridge University Press.

Frigg, R., & Nguyen, J. (2016). Scientific representation. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*, Metaphysics Research Lab, Stanford University winter 2016 edn.

Geroch, R. (1972). Einstein algebras. *Communications in Mathematical Physics, 26*(4), 271–275.

Glymour, C. (1970). Theoretical realism and theoretical equivalence. In *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* (pp. 275–288).

Glymour, C. (1977). The epistemology of geometry. *Nous, 11*(3), 227–251.

Glymour, C. (2013). Theoretical equivalence and the semantic view of theories. *Philosophy of Science, 80*(2), 286–297.

Halvorson, H. (2012). What scientific theories could not be. *Philosophy of Science, 79*(2), 183–206.

Halvorson, H. (2013). The semantic view, if plausible, is syntactic. *Philosophy of Science, 80*(3), 475–478.

Hudetz, L. (2019). Definable categorical equivalence. *Philosophy of Science, 86*(1), 47–75.

Lewis, D. (1970). How to define theoretical terms. *Journal of Philosophy, 67*(13), 427–446.

Maudlin, T. (2018). Ontological clarity via canonical presentation: Electromagnetism and the Aharonov–Bohm effect. *Entropy, 20*(6), 465.

Maxwell, G. (1962). The ontological status of theoretical entities. In H. Feigl, & G. Maxwell (Eds.), *Scientific explanation, space, and time: Minnesota studies in the philosophy of science* (pp. 3–27). University of Minnesota Press.

Nguyen, J. (2017). Scientific representation and theoretical equivalence. *Philosophy of Science, 84*(5), 982–995.

North, J. (2009). The "Structure" of physics: A case study. *The Journal of Philosophy, 106*(2), 57–88.

North, J. (2018). A new approach to the relational-substantival debate. *Oxford Studies in Metaphysics, 11,* 3–43.

Norton, J. D. (2015). The hole argument. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy, Metaphysics Research Lab, Stanford University.

Pooley, O. (2013). Substantivalist and relationalist approaches to spacetime. In R. Batterman (Ed.), *The Oxford handbook of philosophy of physics* (pp. 522–586). Oxford University Press.

Putnam, H. (1983). Equivalence. In *Realism and reason: Philosophical papers* (Vol. 3, pp. 26–45). Cambridge University Press.

Quine, W. V. (1975). On empirically equivalent systems of the world. *Erkenntnis, 9*(3), 313–328.

Rosenstock, S., Barrett, T. W., & Weatherall, J. O. (2015). On Einstein algebras and relativistic spacetimes. *Studies in History and Philosophy of Modern Physics, 52,* 309–316.

Rynasiewicz, R. (1992). Rings, holes and substantivalism: On the program of Leibniz algebras. *Philosophy of Science, 59*(4), 572–589.

Saunders, M. L. (1998). *Categories for the working mathematician*. New York: Springer.

Schaffer, J. (2009). On what grounds what. In D. Manley, D. J. Chalmers, & R. Wasserman (Eds.), *Metametaphysics: New essays on the foundations of ontology, 347–383*. Oxford University Press.

Sklar, L. (1980). Semantic analogy. *Philosophical Studies, 38,* 217–234.

Sklar, L. (1982). Saving the noumena. *Philosophical Topics, 13*(1), 89–110.

Soames, S. (2003). *Philosophical analysis in the twentieth century, Volume 1: The dawn of analysis*. Princeton University Press.

Suarez, M. (2010). Scientific representation. *Philosophy Compass, 5*(1), 91–101.

Teh, N. J., & Tsementzis, D. (2017). Theoretical equivalence in classical mechanics and its relationship to duality. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics, 59,* 44–54.

Tsementzis, D. (2017). A syntactic characterization of morita equivalence. *The Journal of Symbolic Logic, 82*(4), 1181–1198.

van Fraassen, B. C. (1980). *The scientific image*. Oxford University Press.

van Fraassen, B. C. (2014). One or two gentle remarks about Hans Halvorson's critique of the semantic view. *Philosophy of Science, 81*(2), 276–283.

Weatherall, J. O. (2015). Are Newtonian gravitation and geometrized newtonian gravitation theoretically equivalent? *Erkenntnis, 81*(5), 1073–1091.

Weatherall, J. O. (2019). Theoretical equivalence in physics (Parts 1 and 2). *Philosophy Compass, 14*(5), 66.

Wigglesworth, J. (2017). Logical anti-exceptionalism and theoretical equivalence. *Analysis, 77*(4), 759–767.

Woods, J. (2018). Intertranslatability, theoretical equivalence, and perversion. *Thought: A Journal of Philosophy, 7*(1), 58–68.