Mutual Translatability, Equivalence, and the Structure of Theories

Abstract

This paper presents a simple pair of first-order theories that are not definitionally (nor Morita) equivalent, yet are mutually conservatively translatable and mutually 'surjectively' translatable. We use these results to clarify the overall geography of standards of equivalence and to show that the structural commitments that theories make behave in a more subtle manner than has been recognized.

1 Introduction

There is a long tradition in philosophy of comparing theories in terms of their ontological commitments. More recent literature has started to compare them in terms of their structural commitments too. The most famous case of structural comparison comes from the history of classical spacetime theories. It is standard to claim that the Galilean theory of spacetime posits less structure than the Newtonian theory of spacetime. The Newtonian theory comes equipped with the structure required to single out a preferred inertial frame as the rest frame, while the Galilean theory does not.[1]

We can appeal to this kind of structural comparison between theories to inform our judgments about equivalence of theories (North, 2009; Barrett, 2019).[2] If two theories disagree in terms of their structural commitments, then we can infer that they are inequivalent. This is a natural thought. If two theories posit different structure — like the Galilean and Newtonian theories of spacetime do — then they cannot 'have the same content' or 'say the same thing about the world'. It is uncontroversial that equivalent theories must agree in terms of their structural commitments. But there are some further natural principles relating structure and equivalence that are not as obvious. In this paper we will consider the following two.

Principle 1. If T posits all of the structure of T' and T' posits all of the structure of T, then T and T' are equivalent.

[1] See Geroch (1978), Friedman (1983), Earman (1989), Maudlin (2012), and Barrett (2015b) for discussion.

[2] See Weatherall (2019b) for a review of the recent debate on equivalence of theories.

Principle 2. If T can be embedded in T' and T' can be embedded in T, then T and T' are equivalent.

Principle 1 captures a natural idea that one might have about structure and equivalence. If one theory posits all of the structure of the other and the other posits all of the structure of the one, then one might expect them to posit the same structure and to thereby be equivalent. This kind of idea is widespread, especially in the recent literature on structure and equivalence. It has been put nicely by Dewar (2021, p. 6), who suggests the following 'slogan' about structure and equivalence: "for two representations to be equivalent is for them to posit the same structure, and the structure of a representation is that which it has in common with equivalent representations". North (2009, p. 66–7) comes close to endorsing this same idea when she says that "[i]t seems pretty clear why modern physics is so interested in structure. Structure consists in just the kind of things we take to be candidates for objective features of the world, for features of reality. [...] Reality has to do with *structure*." If so, then it is natural to think that two theories that posit the same structure describe the world in precisely the same manner and will thereby be equivalent.

The idea behind Principle 1 also comes up in debates about structural realism. Many structural realists think that the content of a theory is exhausted by its 'structural content'. If so, then two theories that posit the same structure should be equivalent. For example, consider the famous idea that the structural content of a theory is captured by its Ramsey sentence. Many of the problems with this proposal reduce to the following core issue. If one identifies the structural content of a theory with its Ramsey sentence, then one is forced to adopt an unreasonable standard of theoretical equivalence (Dewar, 2019); 'having the same Ramsey sentence' is not a satisfactory criterion of equivalence. The idea lurking behind the scenes here is close to Principle 1. Theories that posit the same structure — which, on this structural realist proposal, amounts to 'having the same Ramsey sentence' in some precise sense — should be considered equivalent.[3]

On the face of it, Principle 2 is perhaps more intuitive than Principle 1. Suppose that one theory can be 'embedded' in the other and the other can be 'embedded' in the one. For many mathematical objects (like, for example, sets) this mutual embeddability would imply that the two objects are 'the same' in the sense that they are isomorphic. One would expect two mutually embeddable theories to posit the same structure and to be equivalent. We will argue here, however, that neither Principle 1 nor Principle 2 is true. For reasons that we will discuss below, it is natural to call Principle 2 the 'Cantor-Bernstein property' of theories and Principle 1 the 'co-Cantor-Bernstein' property of theories.

The aim of this paper is to present a simple example that demonstrates that first-order theories lack both the Cantor-Bernstein and co-Cantor-Bernstein properties. These two results are philosophically significant in their own right,

[3]There is, of course, a small gap between Principle 1 and this idea; it may be that T posits all the structure of T' and vice versa but T and T' do not posit the *same* structure. It is, in a sense, exactly this gap that the example in this paper will exploit.

but we will use them to draw out two further philosophical payoffs that relate to recent discussions of structure and equivalence. First, these results allow us clarify the overall geography of standards of equivalence. And second, they show that the structural commitments that theories make behave in a more subtle manner than has been recognized. In particular, there is a sense in which, contrary to the way that philosophers often speak, structure is not the kind of thing that can be genuinely 'counted'.[4]

2 Two theories

Consider the following two theories.[5]

The theory $\mathbf{T_1}$. T_1 is formulated in the signature $\Sigma_1 = \{p_0, p_1, p_2, \ldots\}$ where each of the p_i is a unary predicate symbol. We define the theory as follows:

$$T_1 = \{\exists_{=1} x(x = x)\}$$

T_1 says that there is exactly one thing, but it is silent on whether or not that thing is p_i.

The theory $\mathbf{T_2}$. T_2 is formulated in the signature $\Sigma_2 = \{q_0, q_1, q_2, \ldots\}$ where each of the q_i is a unary predicate symbol. We define the theory as follows:

$$T_2 = \{\exists_{=1} x(x = x), \forall x(q_0(x) \to q_1(x)), \forall x(q_0(x) \to q_2(x)), \ldots\}$$

T_2 says that there is exactly one thing, and that if that thing is q_0, then it is q_i for all of the other i too. One can think of the predicate q_0 as a kind of 'light switch' that turns on all of the other predicates q_i.

Our first aim is to examine these theories and catalogue the relationships between them. In particular, we will demonstrate the following five claims, which correspond to the five theorems that we prove in sections 3 and 4.

1. T_1 and T_2 are not definitionally equivalent.
2. T_1 and T_2 are not Morita equivalent.
3. There are conservative translations $F : T_1 \to T_2$ and $G : T_2 \to T_1$.
4. There are essentially surjective translations $H : T_1 \to T_2$ and $K : T_2 \to T_1$.

[4] There is a sense in which these results are not surprising. Principle 1 and 2 are strong sufficient conditions on equivalence of theories, and moreover, the Cantor-Bernstein and co-Cantor-Bernstein properties fail for many kinds of mathematical objects — indeed, possibly for more than they hold of — so one might expect them to fail for theories too. The two payoffs about structure and equivalence, however, are surprising. This indicates that, even if one is not surprised by the falsity of Principles 1 and 2, one may not have fully realized the consequences that this has for recent debates about structure and equivalence.

[5] For preliminaries on model theory the reader is encouraged to consult Hodges (2008). The apparatus and terminology that we use will follow Halvorson (2019). For additional discussion of these two theories see Halvorson (2012) and Barrett and Halvorson (2016b).

5. There is no essentially surjective and conservative translation from T_1 to T_2, or vice versa.

It is natural to separate these claims into two kinds. The first two claims are about what kinds of extensions exist for T_1 and T_2 (i.e. they have no common definitional nor Morita extension), while the last three are about what kinds of translations exist between T_1 and T_2. After proving these claims, we will return to the general philosophical issues mentioned above and discuss how these results demonstrate the falsity of Principles 1 and 2.

3 Extension

Let $\Sigma \subset \Sigma^+$ be signatures with $p \in \Sigma^+ - \Sigma$ an n-ary predicate symbol. Recall that an **explicit definition** of p in terms of Σ is a Σ^+-sentence of the form

$$\forall x_1 \dots \forall x_n (p(x_1, \dots x_n) \leftrightarrow \phi(x_1, \dots, x_n))$$

where $\phi(x_1, \dots, x_n)$ is a Σ-formula. A **definitional extension** of a Σ-theory T to the signature Σ^+ is a Σ^+-theory

$$T^+ = T \cup \{\delta_s : s \in \Sigma^+ - \Sigma\},$$

such that for each predicate symbol $s \in \Sigma^+ - \Sigma$, the sentence δ_s is an explicit definition of s in terms of Σ. One can also define new function and constant symbols, but for our purposes this will not be important. There are two familiar facts about definitional extensions that we need to mention.[6] First, if T^+ is a definitional extension of T, then every model of T has a unique expansion that is a model of T^+. When $\Sigma \subset \Sigma^+$, a Σ^+-structure M^+ is said to be an **expansion** of a Σ-structure M if M is obtained from M^+ by 'forgetting about' the interpretations of the symbols in $\Sigma^+ - \Sigma$. And second, a definitional extension T^+ of T is also a conservative extension of T. Recall that T^+ is an **extension** of T if $T \vDash \phi$ entails that $T^+ \vDash \phi$, and that it is a **conservative extension** if for any sentence ϕ in the signature of T, $T^+ \vDash \phi$ if and only if $T \vDash \phi$.

We say that two theories are **definitionally equivalent** if they have a 'common definitional extension'. More precisely, if T is a Σ-theory and T' is a Σ'-theory, T and T' are definitionally equivalent if there is a definitional extension T^+ of T to the signature $\Sigma \cup \Sigma'$ and a definitional extension T'^+ of T' to the signature $\Sigma \cup \Sigma'$ such that T^+ and T'^+ are logically equivalent (i.e. they have the same class of models).

We now have the following simple theorem.

Theorem 1. *T_1 and T_2 are not definitionally equivalent.*

[6] See Hodges (2008) for proof of these facts.

Proof. Suppose for contradiction that T is a common definitional extension of T_1 and T_2. Since T defines each of the predicate symbols of T_2, there is a Σ_1-sentence ϕ such that $T \vDash \forall y q_0(y) \leftrightarrow \phi$. Recall that models of each of these three theories have one element.

We begin by showing that the sentence ϕ has the following property.

($\star$) If ψ is a Σ_1-sentence and $T_1 \vDash \psi \rightarrow \phi$, then either (i) $T_1 \vDash \neg\psi$ or (ii) $T_1 \vDash \phi \rightarrow \psi$.

So let ψ be a Σ_1-sentence such that $T_1 \vDash \psi \rightarrow \phi$ and suppose that $T_1 \nvDash \neg\psi$. This means that there is a model M of T_1 such that $M \vDash \psi$. By assumption this means that $M \vDash \phi$ too. Now consider the model M^+ of T. Since $M^+ \vDash \phi$, we know that $M^+ \vDash \forall y q_0(y)$. So it must be the case that $M^+ \vDash \forall y q_i(y)$ for each i because T is an extension of T_2. Now we want to show that $T_1 \vDash \phi \rightarrow \psi$. So suppose N is a model of T_1 such that $N \vDash \phi$. Consider the model N^+ of T. We know that $N^+ \vDash \phi$, so $N^+ \vDash \forall y q_0(y)$, which — since T is an extension of T_2 — implies that $N^+ \vDash \forall y q_i(y)$ for each i. This means that $N^+|_{\Sigma_2}$ and $M^+|\Sigma_2$ are isomorphic. Every model of T_2 has a *unique* expansion that is a model of T, so this implies that N^+ and M^+ must be isomorphic as $\Sigma_1 \cup \Sigma_2$-structures. That immediately implies that $N^+ \vDash \psi$, so $N \vDash \psi$ too. This means that the sentence ϕ does indeed have property ($\star$). But ϕ cannot have this property. Consider the Σ_1-sentence

$$\phi \wedge \forall x p_i(x)$$

where p_i is a predicate symbol that does not occur in ϕ. We trivially see that $T_1 \vDash (\phi \wedge \forall x p_i(x)) \rightarrow \phi$, but one can verify that neither (i) nor (ii) hold of $\phi \wedge \forall x p_i(x)$. This is a contradiction, so T_1 and T_2 are not definitionally equivalent. □

It has recently been suggested that definitional equivalence is too strict a standard of equivalence between theories, in the sense that it judges theories to be inequivalent that we have good reason to consider equivalent. For example, Euclidean geometry can be formulated with only a sort of points (Tarski, 1959), with only a sort of lines (Schwabhäuser and Szczerba, 1975), or with both a sort of points and a sort of lines (Hilbert, 1930).[7] Since these formulations use different sort symbols, and we have so far provided no way of defining new sort symbols, definitional equivalence does not capture any sense in which they are equivalent. In order to address this shortcoming of definitional equivalence, a more liberal standard of equivalence has been proposed. It has been called "Morita equivalence" (by Barrett and Halvorson (2016b, 2017a,b)) and "generalized definitional equivalence" (by Andréka et al. (2008)).[8]

[7] See Szczerba (1977) and Schwabhäuser et al. (1983, Proposition 4.59, Proposition 4.89).

[8] See also Hudetz (2017a,b) and Tsementzis (2015). Note that Morita equivalence does not collapse into definitional equivalence in the single-sorted setting. There are single-sorted theories that are Morita equivalent but not definitionally equivalent, like geometries with points and geometries with lines (Barrett and Halvorson, 2017a). For more on the relationship between Morita equivalence and single-sorted theories, see Barrett and Halvorson (2017b).

The precise details of Morita equivalence are not important for our purposes here, but the basic idea is simple. Morita equivalence allows one to define new *sort* symbols — in addition to new predicate, function, and constant symbols — using some basic construction rules. Two theories are then said to be Morita equivalent if they have a 'common Morita extension', which is just like a common definitional extension except that it might define new sorts. One can show that geometry formulated in terms of points is Morita equivalent to geometry formulated in terms of lines; one uses the sort of lines to build the sort of points, and vice versa (Barrett and Halvorson, 2017a).

One might wonder whether our theories T_1 and T_2 are Morita equivalent. The following simple result has already been demonstrated. The proof proceeds in essentially the same manner as the proof of Theorem 1. Barrett and Halvorson (2016b, Theorem 5.2) give the precise details.[9]

Theorem 2. *T_1 and T_2 are not Morita equivalent.*

4 Translation

Our next three claims are about the kinds of 'translations' that exist between these theories. We need some basic preliminaries.

Let Σ and Σ' be signatures. A **reconstrual** F of Σ into Σ' is a map from the predicates in the signature Σ to Σ'-formulas that takes an n-ary predicate symbol $p \in \Sigma$ to a Σ'-formula $Fp(x_1, \ldots, x_n)$ with n free variables.[10] One can think of the Σ'-formula $Fp(x_1, \ldots, x_n)$ as the 'translation' of the Σ-formula $p(x_1, \ldots, x_n)$ into the signature Σ'. We will use the notation $F : \Sigma \to \Sigma'$ to denote a reconstrual F of Σ into Σ'.

A reconstrual $F : \Sigma \to \Sigma'$ extends to a map from arbitrary Σ-formulas to Σ'-formulas in the natural recursive manner. In the case where one is only considering signatures with predicate symbols (as we are here), this map is particularly easy to describe. Let $\phi(x_1, \ldots, x_n)$ be a Σ-formula. We define the Σ'-formula $F\phi(x_1, \ldots, x_n)$ recursively as follows.

- If $\phi(x_1, \ldots, x_n)$ is $x_i = x_j$, then $F\phi(x_1, \ldots, x_n)$ is the Σ'-formula $x_i = x_j$.
- If $\phi(x_1, \ldots, x_n)$ is $p(x_1, \ldots, x_n)$, where $p \in \Sigma_1$ is an n-ary predicate symbol, then $F\phi(x_1, \ldots, x_n)$ is the Σ'-formula $Fp(x_1, \ldots, x_n)$.
- If $F\phi$ and $F\psi$ have already been defined for Σ-formulas ϕ and ψ, then we define the Σ'-formula $F(\neg\phi)$ to be $\neg F\phi$, $F(\phi \wedge \psi)$ to be $F\phi \wedge F\psi$, $F(\forall x\phi)$ to be $\forall x F\phi$, etc.

Suppose that T and T' are theories in the signatures Σ and Σ', respectively. We say that a reconstrual $F : \Sigma \to \Sigma'$ is a **translation** $F : T \to T'$ if $T \vDash \phi$

[9] The result goes through even if one formulates T_1 and T_2 using two different sort symbols, instead of the same one sort symbol.

[10] This notion naturally extends to signatures that contain function and constant symbols, but that will be unimportant for our purposes. See Hodges (2008), Button and Walsh (2018), and Barrett and Halvorson (2016a) for details.

implies that $T' \vDash F\phi$ for every Σ-sentence ϕ. A translation F gives rise to a map $F^* : \mathrm{Mod}(T') \to \mathrm{Mod}(T)$, which takes models of the theory T' to models of the theory T. For every model A of T' we first define a Σ-structure $F^*(A)$ as follows.

- $\mathrm{dom}(F^*(A)) = \mathrm{dom}(A)$.
- $(a_1, \ldots, a_n) \in p^{F^*(A)}$ if and only if $A \vDash Fp[a_1, \ldots, a_n]$.

A straightforward argument demonstrates that $F^*(A)$ is indeed a model of T (Barrett and Halvorson, 2016a, §4).

A translation $F : T \to T'$ is **conservative** if $T' \vDash F\phi$ implies that $T \vDash \phi$ for any Σ-sentence ϕ. One can easily verify that if a translation $F : T \to T'$ is such that $F^* : \mathrm{Mod}(T') \to \mathrm{Mod}(T)$ is surjective, then F is conservative. In the following section we will argue that conservative translations can be thought of as 'embeddings' or 'injections' between theories.

We now have our first simple result concerning the existence of translations between our theories T_1 and T_2.

Theorem 3. *There are conservative translations $F : T_1 \to T_2$ and $G : T_2 \to T_1$.*

Proof. Consider the reconstruals $F : \Sigma_1 \to \Sigma_2$ and $G : \Sigma_2 \to \Sigma_1$ defined by

$$F : p_i \longmapsto q_{i+1} \qquad\qquad G : q_i \longmapsto p_0 \vee p_i$$

It is trivial that $F : T_1 \to T_2$ is a translation. Since G maps the Σ_2-sentence $\forall x(q_0(x) \to q_i(x))$ to $\forall x((p_0(x) \vee p_0(x)) \to (p_0(x) \vee p_i(x)))$ and T_1 entails this latter sentence, it follows that G is a translation too.

It remains to show that F and G are conservative. One does this by showing that F^* and G^* are surjective. Suppose that M is a model of T_1. Then M is completely determined by which of the p_i hold of the one thing. We let N be the model of T_2 defined as follows: N has the same domain as M, q_0 does not hold of the one thing in N, and q_{i+1} holds of the one thing in N if and only if p_i holds of the only thing in M. One trivially sees that $F^*(N) = M$, so F^* is surjective. A similar argument shows that G^* is surjective. □

Theorem 3 shows us that T_1 and T_2 are 'mutually conservatively translatable'. As we will argue in the following section, since they are conservative, the translations F and G are naturally thought of as injections or embeddings between the theories T_1 and T_2. But there is a natural sense in which neither is 'surjective'. For example, F does not map any Σ_1-formula to a formula which is equivalent modulo T_2 to the Σ_2-formula $q_0(x)$. The following property makes this thought precise. We say that a translation $F : T \to T'$ is **essentially surjective** if for every Σ'-formula ψ there is a Σ-formula ϕ such that $T' \vDash \forall x_1 \ldots \forall x_n(\psi(x_1, \ldots, x_n) \leftrightarrow F\phi(x_1, \ldots, x_n))$. One can easily verify that neither F nor G from Theorem 3 are essentially surjective. But we do have the following simple result.

Theorem 4. *There are essentially surjective translations $H : T_1 \to T_2$ and $K : T_2 \to T_1$.*

Proof. Consider the reconstruals $H : \Sigma_1 \to \Sigma_2$ and $K : \Sigma_2 \to \Sigma_1$ defined by

$$H : p_i \longmapsto q_i \qquad\qquad K(q_i) = \begin{cases} p_0 \wedge \neg p_0 & \text{if } i = 0 \\ p_{i-1} & \text{otherwise} \end{cases}$$

It is trivial that H is an essentially surjective translation. Since K maps the Σ_2-sentence $\forall x(q_0(x) \to q_i(x))$ to $\forall x((p_0(x) \wedge \neg p_0(x)) \to p_{i-1}(x))$ and T_1 entails this latter sentence, it follows that K is a translation. It is easy to see that K is essentially surjective. □

Theorem 4 shows us that T_1 and T_2 are 'mutually surjectively translatable'. But neither H nor K is a conservative translation. One can verify that the theory T_1 does not entail the sentence $\forall x(p_0(x) \to p_1(x))$, but T_2 does entail $H(\forall x(p_0(x) \to p_1(x)))$, i.e. $\forall x(q_0(x) \to q_1(x))$. Similarly, the theory T_2 does not entail $\neg\forall x q_0(x)$, but T_1 does entail $K(\neg\forall x q_0(x))$, i.e. $\neg\forall x(p_0(x) \wedge \neg p_0(x))$.

Given that these two theories are mutually conservatively translatable and mutually surjectively translatable, one wonders whether there is any translation between T_1 and T_2 that is essentially surjective and conservative. It would be natural to think of such a translation as an 'isomorphism' between the two theories. Theorem 5 settles this issue. The most straightforward proof proceeds via the following lemma.

Lemma 1. *Let Σ and Σ' be signatures, with T and T' theories in those signatures, respectively. If T and T' are definitionally equivalent, then there is a conservative and essentially surjective translation $F : T \to T'$. Conversely, if Σ and Σ' are disjoint and there is a conservative and essentially surjective translation $F : T \to T'$, then T and T' are definitionally equivalent.*

Proof. This follows from Propositions 4.5.26, 4.5.27, 4.6.17, and 6.6.21 of Halvorson (2019). □

Theorem 5. *There is no essentially surjective and conservative translation between T_1 and T_2 (in either direction).*

Proof. This follows immediately from the second half of Lemma 1. □

5 The Cantor-Bernstein and co-Cantor-Bernstein properties of theories

We now draw out the philosophical payoffs that these technical claims yield. The first two payoffs that we will discuss concern two peculiar features that the collection of first-order theories exhibits. First, it can be that one theory is embeddable in another and the other is embeddable in the one, but the two theories are not equivalent. And second, it can be that one theory 'posits all of the structure' of another and the other 'posits all of the structure' of the one, but the two theories do not posit the same structure. This means that both of the principles mentioned above are false. We restate them here for convenience.

Principle 1. If T posits all of the structure of T' and T' posits all of the structure of T, then T and T' are equivalent.

Principle 2. If T can be embedded in T' and T' can be embedded in T, then T and T' are equivalent.

For reasons that we will discuss below, it is natural to call Principle 2 the 'Cantor-Bernstein property' of theories and Principle 1 the 'co-Cantor-Bernstein' property of theories. These results show that theories lack both of these properties. It will take a moment to explain these two claims. The falsity of Principle 1 is particularly relevant to recent discussions of structure, so we will start with Principle 2 as a 'warm up'.

The Cantor-Bernstein property

We begin with the following claim.

Claim 1. The existence of a conservative translation $F : T \to T'$ captures a sense in which T can be embedded into T'.

One can think of a conservative translation as a kind of 'injection' or 'embedding' on sentences of the two theories. The most natural argument for Claim 1 relies on the following proposition.

Proposition 1. *A translation $F : T \to T'$ is conservative if and only if the following condition holds: for any Σ-sentences ϕ_1 and ϕ_2, if $T' \vDash F\phi_1 \leftrightarrow F\phi_2$ then $T \vDash \phi_1 \leftrightarrow \phi_2$.*

Proof. It is immediate that the condition holds whenever F is conservative. Suppose that the condition holds and that $T' \vDash F\phi$. This means that $T' \vDash F\phi \leftrightarrow \exists x(x = x)$. The condition implies that $T \vDash \phi \leftrightarrow \exists x(x = x)$, so $T \vDash \phi$. □

The condition that appears in Proposition 1 provides another characterization of F's conservativity. And it is clearly capturing a sense in which F is an injection or embedding on sentences: If F maps two Σ-sentences to equivalent Σ'-sentences, then they must have been equivalent to begin with.

As further evidence for Claim 1, we show how the existence of a conservative translation $F : T \to T'$ demonstrates that T can be 'viewed as a part of' T', or in other words, that T can be thought of as a 'sub-theory' of T'.[11] The following proposition substantiates this idea.

[11] A clarification about terminology will be useful. Note that by saying that T is a 'sub-theory' or 'a part of' T' we do not mean to say that the axioms of T are a subset of the axioms of T'. Rather, the idea is that if we think of a theory as a kind of 'inferential structure' over the collection of formulas in the theory's signature, then a conservative translation from T to T' shows that the inferential structure of T' is not more 'coarse-grained' than that of T. This is what Proposition 2 shows. It is also in this sense that if T' is a conservative extension of T, then T is a 'part' of T'; the set of all formulas in the signature of T' will be strictly greater than the set of all formulae in the signature of T, but because the extension is conservative, the inferential relations among these formulas of T will be the same in both cases.

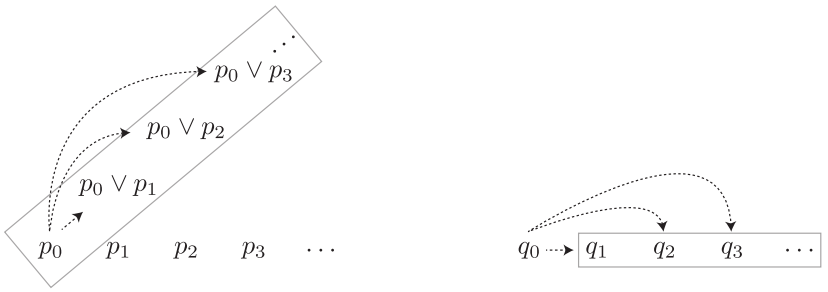
Proposition 2. *Let Σ and Σ' be disjoint signatures with T and T' theories in those signatures, respectively. If $F : T \to T'$ is a conservative translation, then there is a conservative extension of T that is definitionally equivalent to T'.*

Proof. We show that there is a definitional extension of T'^{+} of T' that is a conservative extension of T. Consider the $\Sigma \cup \Sigma'$-theory $T'^{+} = T' \cup \{\forall x(Fp(x) \leftrightarrow p(x)) : p \in \Sigma\}$. It is clear that T'^{+} is a definitional extension of T' to the signature $\Sigma \cup \Sigma'$. Let ϕ be a Σ-sentence and suppose that $T \vDash \phi$. Then $T' \vDash F\phi$, since F is a translation. By the way that we have defined T'^{+} it follows that $T'^{+} \vDash \phi$, so T'^{+} is an extension of T. Suppose now that $T'^{+} \vDash \phi$. Then once again by the definition of T'^{+} we see that $T'^{+} \vDash F\phi$. Since T'^{+} is a definitional extension of T', it is also a conservative extension, so it must be that $T' \vDash F\phi$. Since F is conservative, $T \vDash \phi$. □

One can understand this proposition in the following way. The existence of a conservative translation from T to T' tells us that, up to definitional equivalence, T' is a conservative extension of T. This captures a strong sense in which T is a sub-theory of T'. If T' is a conservative extension of T, then T can be thought of as exactly the part of T' that is formulated in the language Σ, since T entails a Σ-sentence if and only if T' does.

Note that Proposition 2 implies that T_1 is (up to definitional equivalence) a conservative extension of T_2 and vice versa. The $\Sigma_1 \cup \Sigma_2$-theory T_2^{+} that results from adding the axioms $\forall x(p_i(x) \leftrightarrow q_{i+1}(x))$ for each i to T_2 is definitionally equivalent to T_2 and a conservative extension of T_1. And the $\Sigma_1 \cup \Sigma_2$-theory T_1^{+} that results from adding the axioms $\forall x(q_i(x) \leftrightarrow (p_0(x) \vee p_i(x)))$ to T_1 is definitionally equivalent to T_1 and a conservative extension of T_2. This captures a sense in which we can obtain T_1 from T_2 (and vice versa) simply by adding vocabulary and axioms 'on this new vocabulary', in the sense that the new axioms do not result in any new entailments in purely the old vocabulary. So simply by 'adding structure', and saying how this new structure relates to the old structure, we can move from T_1 to T_2. We can move from T_2 to T_1 in precisely the same fashion: simply by adding structure. This is a surprising result; one can add structure to *either* of the two theories in order to arrive at the other. We will return to this fact in the following section, when we discuss whether or not structure can be 'counted'.

With Claim 1 now in hand, Theorem 3 captures a sense in which our theory T_1 can be embedded in T_2 and our theory T_2 can be embedded in T_1. There are conservative translations — which, as we just argued, can be thought of as injections or embeddings — in both directions between our two theories. The following figure illustrates the sense in which T_1 is a sub-theory of T_2 and vice versa. The dotted arrows indicate theory-relative entailments between the formulas in each of our theories.



theory T_1 theory T_2

The existence of the conservative translation $F : T_1 \to T_2$ shows that T_1 can be viewed as the part of T_2 that is constructed from the predicate symbols $q_1, q_2, \ldots$, and so on — the sub-theory of T_2 that is surrounded by the box in the above figure. And similarly the existence of the conservative translation $G : T_2 \to T_1$ shows us that one can view T_2 as a part of T_1; it is that part of T_1 that is constructed from the Σ_1-formulas $p_0, p_0 \vee p_1, p_0 \vee p_2, \ldots$, and so on — the part of the theory T_1 that is surrounded by the box in the above figure.

Despite the fact that these theories are mutually embeddable, however, they are certainly not equivalent. Theorems 1, 2, and 5 make this precise. Our theories T_1 and T_2 are neither definitionally nor Morita equivalent, and there does not exist an essentially surjective and conservative translation between them. So we have shown that, on particularly natural understandings of the relations 'is embeddable in' and 'is equivalent to', Principle 2 is false.

We should take a moment to unravel why the falsity of Principle 2 is surprising. Perhaps the most natural way to state this result is by saying that *theories lack the Cantor-Bernstein property.* The Cantor-Bernstein theorem famously says that if there are injections $f : X \to Y$ and $g : Y \to X$ between sets X and Y, then there is a bijection between X and Y. Because of this theorem we say that sets have the Cantor-Bernstein property. And indeed, it makes sense to talk about the Cantor-Bernstein property for any category. One says that a category C has the **Cantor-Bernstein property** if for any objects c and d of C whenever there is a monomorphism (i.e. a generalization of the concept of an injection or embedding) from $c \to d$ and a monomorphism $d \to c$, the objects c and d are isomorphic. The property captures a basic intuition one might have about the notion of 'being a part of': If X is a part of Y and Y is a part of X, then X and Y are the same. Our example here demonstrates that the category of theories does not have the Cantor-Bernstein property. In other words, the analogue of the Cantor-Bernstein theorem does not hold of theories: T_1 and T_2 can be embedded into one another, but nonetheless they are not the same. So this basic intuition about 'being a part of' does not hold of theories: T_1 can be viewed as a part of T_2 and T_2 can be viewed as a part of T_1, but they are not the same theory.[12]

[12]The fact that theories lack the Cantor-Bernstein property is already known by certain

The Co-Cantor-Bernstein property

Theories also lack what we will call the *co-Cantor-Bernstein property*. This is particularly relevant to recent discussions of structure in philosophy of science. We show first that Principle 1 is false. In other words, it can be that one theory 'posits all of the structure' of another and the other 'posits all of the structure' of the one, but the two theories are inequivalent and do not posit the same structure.[13]

We begin with the following claim.

Claim 2. The existence of an essentially surjective translation $F : T \to T'$ captures a sense in which T posits all of the structure of T'.

The idea here is simple. When a translation F is essentially surjective, any formula ψ in the language of T' is expressible using the language of T; that is precisely what the essential surjectivity of F guarantees. There is some formula ϕ in the language of T that F translates to a formula that is equivalent modulo T' to ψ. Intuitively, this means that the existence of an essentially surjective F shows that the theory T can define or 'build' all of the structure that T' has. T can express all of the 'concepts' that T' employs.

One can grasp the basic idea here by considering the following two simple examples. Suppose that T is a Σ-theory and consider the signature $\Sigma^+ = \Sigma \cup \{p\}$, where p is a new unary predicate symbol not contained in Σ.

Example 1. Let T^+ be the Σ^+-theory that has precisely the same axioms as T. There is a natural sense in which T^+ has all of the structure of T, but not vice versa. In particular, T^+ has the new piece of structure p that T lacks. One can capture this basic idea by looking to facts about essentially surjective translations between these two theories. First, there is an essentially surjective translation $T^+ \to T$. The translation simply maps each symbol in Σ to itself and maps p to any Σ-formula with one free variable. It is trivial to verify that this is indeed an essentially surjective translation. This makes precise our basic intuition that T^+ has all of the structure of T. Second, since there is no Σ-formula that is logically equivalent to p, the translation $T \to T^+$ that maps

logicians. See, for example, the work of Visser (2006, p. 313) and Andréka et al. (2005). It is not, however, well known by the broader philosophical community. Another example of the failure of the Cantor-Bernstein property is the relation between classical and intuitionistic logic. Intuitionistic logic can obviously be embedded in classical logic. The famous Gödel translation shows that intuitionistic logic can be embedded in classical logic. But it is nonetheless unnatural to think of the two logics as equivalent, as was recently pointed out, for example, by Dewar (2018).

[13] Another way to put this is as follows: It can be that one theory 'is as ideologically rich as' another and the other 'is as ideologically rich as' the one, but the two theories are not ideologically equivalent. The ideology of a theory is usually thought of as the range of concepts that are expressible in the language in which the theory is formulated (Quine, 1951). Quine (1951, p. 15) himself remarks that one can investigate the ideology of a theory by, as we do here, examining the kinds of translations that exist between theories: "Much that belongs to ideology can be handled in terms merely of the translatability of notations from one language into another; witness the mathematical work on definability by Tarski and others."

every element of Σ to itself is not essentially surjective. This makes precise our intuition that T does *not* have all of the structure of T^+. ⌟

Example 2. Now suppose instead that T^+ is a definitional extension of T to the signature Σ^+. In this case there is a strong sense in which T and T^+ have precisely the same structure; p is not a piece of structure that is new to T^+. Rather, it is explicitly definable in terms of those structures that T posits, so there is a strong sense in which T itself already posits the structure p. This basic intuition can be made precise by noticing that there are essentially surjective translations $T \to T^+$ and $T^+ \to T$. It follows from our discussion in Example 1 that there is an essentially surjective translation $T^+ \to T$. And since T^+ is a definitional extension of T, the translation $T \to T^+$ that maps every element in Σ to itself is in this case essentially surjective since the Σ-formula ϕ that defines p is logically equivalent to p. ⌟

Examples 1 and 2 provide is our first argument for Claim 2: It is a direct generalization of the kinds of simple examples that formed our intuitions about amounts of structure in the first place. We can, however, do better than an argument by example. We will call a Σ-theory T^+ that is an extension of a Σ-theory T a **specification** of T. Note that since both T^+ and T are formulated in the same signature, a specification of T results from merely 'adding axioms' (and no new vocabulary) to T. We have the following proposition.

Proposition 3. *Let Σ and Σ' be disjoint signatures and T and T' theories in those signatures, respectively. If $F : T \to T'$ is an essentially surjective translation, then there is a specification of T that is definitionally equivalent to T'.*

Proof. Suppose that we have such an essentially surjective translation $F : T \to T'$ from the Σ-theory T to the Σ'-theory T'. This allows us to define a specification of T as follows:

$$T^F = \{\phi : T' \vDash F\phi\}$$

In other words, T^F is the Σ-theory that has as axioms those Σ-sentences that, once translated via F into the signature Σ', are entailed by T'. It is easy to verify that T^F is indeed a specification of T. If $T \vDash \phi$, then $T' \vDash F\phi$ since F is a translation, so ϕ is among the axioms of T^F and we therefore trivially see that $T^F \vDash \phi$. Furthermore, it is also easy to see that $F : T^F \to T'$ is both essentially surjective and conservative, simply by the way we defined T^F. So T^F and T' are equivalent, in the sense that there is an essentially surjective and conservative translation between them. Since Σ and Σ' are disjoint, Lemma 1 implies that the two theories are definitionally equivalent. □

One can understand this proposition in the following way. The existence of an essentially surjective translation from T to T' tells us that, up to definitional equivalence, T' is a specification of T. Note that Proposition 3 implies that T_1 is (up to definitional equivalence) a specification of T_2 and that T_2 is (up to definitional equivalence) a specification of T_1. By adding the axioms $\forall x(p_0(x) \to$

$p_i(x)$) for each i to T_1, and no new vocabulary, one arrives at a theory that is definitionally equivalent to T_2. And by adding the axiom $\forall x(\neg q_0(x))$ to T_2, and no new vocabulary, one arrives at a theory that is definitionally equivalent to T_1. Each of the theories from our main example can therefore be obtained from the other simply by adding axioms. This in a sense the complement to the fact, implied earlier by Proposition 2, that each of the theories can be obtained from one another simply by adding vocabulary and some axioms on that new vocabulary.

We can now unravel why Proposition 3 provides evidence for Claim 2. Since equivalent theories must posit the same structure, the specification T^F of T posits the same structure as T'. And since it is a specification, T^F is obtained from T by merely adding axioms and no new vocabulary or structure. Since no new vocabulary or structure has been added, a specification of T cannot posit *more* structure than T posited itself. Putting this all together, we have that T posits all of the structure of T^F, which in turn posits the same structure as T', so T posits all of the structure of T'. This argument yields our claim. We supposed that there exists an essentially surjective translation $F : T \to T'$ and showed that captures a strong sense in which T posits all of the structure of T'.

It is worth taking a moment to further explain why a specification of a theory does not posit more structure than the original theory did. In particular, one might worry that sometimes specifications *do* posit more structure. For example, one might argue that the theory of Abelian groups — which is just like the theory of groups but also says that multiplication is commutative — posits more structure than the mere theory of groups does. The former is in some sense more 'restrictive'. More generally, the thought is that adding axioms can add structure to a theory too; it is not just new vocabulary that adds structure.

The issue here is that there are two distinct ways in which we can 'add' something to a theory. On the one hand, we can add new vocabulary; on the other hand, we merely can add new axioms (and no new vocabulary). The question is then whether the latter way of adding something should count as adding *structure*. If so, then specifications would add structure. There are, however, a few reasons to think that adding axioms is not a case of adding structure. In brief, this idea is simply not in line with the concept of structure that is usually appealed to in the philosophy of science literature.

We provide some evidence that adding axioms is not a paradigm case of adding structure. The first piece of evidence for this has to do with the relationship that is often emphasized between structure and symmetries. The amount of structure that a theory posits is standardly thought to be inversely related to the 'number of symmetries' that the theory admits. Earman (1989, p. 36), for example, says that "[a]s the space-time structure becomes richer, the symmetries become narrower." And North (2009, p. 87) writes that "stronger structure [...] admits a smaller group of symmetries." The same basic idea is suggested by Friedman (1983), and has been appealed to in much of the recent literature on how to 'compare' amounts of structure between theories. Indeed, it has become standard to think of the move from Newtonian to Galilean spacetime as a paradigm case of removing structure from a theory. In that case no

new axioms are added; rather, the new theory admits more symmetries than the old one, since the concept of 'absolute rest' has been excised. Note that moving from a theory to a specification of it is *not* like the move from Newtonian to Galilean spacetime. It does not reduce the collection of symmetries in the relevant sense. One can trivially see that for every model of the specification, there is a model of the original theory that has precisely the same automorphism group. Indeed, every model of the specification is itself a model of the original theory. That is certainly not the case for the Newtonian and Galilean spacetime theories. Moreover, the move from Galilean to Newtonian spacetime is clearly an addition of *vocabulary* (and some new axioms involving that vocabulary) rather than an addition of merely axioms; we have intuitively added in a new 'is at rest' predicate. This is evidence that, on the standard understanding of structure, adding axioms is not a paradigm case of adding structure.

North (2009, p. 65) presents a collection of examples that are similar in this regard to the case of Newtonian and Galilean spacetime:

> Note that we can also compare different degrees, or amounts, of structure. Compare a Euclidean plane with a similar plane that has a preferred spatial direction. The Euclidean plane without a preferred direction has less structure than the one with a preferred spatial direction. Picking out a preferred direction requires additional structure (an orientation). In building up a mathematical space, some objects will presuppose others, in that some of the mathematical objects cannot be defined without assuming others. Starting from a structureless set of points, we can add on different 'levels' of structure. A bare set of points has less structure than a topological space, a set of points together with a topology (specifying the open subsets). A topological space has less structure than a metric space: in order to define a metric, the space must already have a topology.

All of these examples — the Euclidean plane and the Euclidean plane with a preferred direction, a set and a topological space, a topological space and a metric space — involve an addition of vocabulary. (These mirror the examples presented by Barrett (2021).) None of them involve merely an addition of axioms. So once again we have evidence that adding axioms is not a paradigm case of adding structure. If it were, then some of the simple examples that philosophers give to illustrate the concept would no doubt involve a move from a theory to a specification of it.

It is worth expanding further on this point. The idea that symmetries can be used as a guide to structure yields two more conceptual arguments for Claim 2 and for the idea that philosophers who discuss amounts of structure likely do not mean to imply that merely adding more axioms (and no new vocabulary) adds structure. In brief, one is led to Claim 2 by taking seriously a collection of tools that have recently been used to compare the structure of theories. As we mentioned above, it has been suggested — by Earman (1989), Friedman (1983) and more recently by North (2009), Swanson and Halvorson (2012), and Barrett (2015a,b) — that the size of an object's automorphism group can be used as

a guide to the amount of structure that the object has. An automorphism of an object is a structure-preserving bijection from the object to itself. If an object has more automorphisms, therefore, that suggests that the object has less structure that these automorphisms are required to preserve. The important fact for our purposes here is the following: If there is an essentially surjective translation from T to T', this captures a sense in which the models of these two theories have automorphism groups of precisely the same size.

It only takes a moment to make this claim precise. An **automorphism** of a Σ-structure M is a bijection $f : M \to M$ that satisfies $M \vDash p[a_1, \ldots, a_n]$ if and only if $M \vDash p[f(a_1), \ldots f(a_n)]$ for any predicate symbol $p \in \Sigma$ and elements $a_1, \ldots, a_n \in M$. One can easily prove the following.

Proposition 4. *Let $F : T \to T'$ be an essentially surjective translation with N a model of T'. Then $F^*(N)$ and N have the same automorphism group.*

Proof. If $f : N \to N$ is an automorphism of N, one can easily verify using the definition of F^* that f is also an automorphism of $F^*(N)$. Suppose that $f : F^*(N) \to F^*(N)$ is an automorphism and let $q \in \Sigma_2$ be a predicate symbol in the signature of T'. We immediately see that for any $a \in N$:

$$N \vDash q[a] \iff F^*(N) \vDash \phi[a] \iff F^*(N) \vDash \phi[f(a)] \iff N \vDash q[f(a)]$$

Here ϕ is a Σ_1-formula such that $H\phi$ is logically equivalent to q, whose existence is guaranteed by the essential surjectivity of F. The first and third biconditionals follow from this choice of ϕ and the definition of F^*, while the second follows from the fact that f is an automorphism of $F^*(N)$. This means that f is an automorphism of N. □

Suppose that there is an essentially surjective translation $F : T \to T'$. Then Proposition 4 shows that each model N of T' can be paired up with a corresponding model $F^*(N)$ of T in such a way that these two models have precisely the same amount of structure, according to one of our best methods of comparing amounts of structure.

Moreover, the essential surjectivity of F is actually related to 'symmetries' between objects of the two theories even more closely than Proposition 4 suggests. It has recently been shown that a translation $F : T \to T'$ is essentially surjective if and only if the map F^* is a full functor from the category of models $\mathrm{Mod}(T')$ to $\mathrm{Mod}(T)$ (Barrett, 2021). We take a moment to unravel this result. The category of models $\mathrm{Mod}(T)$ for a theory T has as objects the models of T, and the arrows between the objects are elementary embeddings between these models.[14] A functor $F : C \to D$ is **full** if for all objects c_1, c_2 in C and arrows $g : Fc_1 \to Fc_2$ in D there exists an arrow $f : c_1 \to c_2$ in C with $Ff = g$. The existence of a full functor from C to D captures a sense in which there are not 'more' arrows between the objects of D than there are between the objects of C. This has led Baez et al. (2006) to classify functors that are *not* full as those that

[14] For further preliminaries on the relevant category theory see Halvorson (2019).

'forget structure', capturing a sense in which the objects of D have less structure (since they admit more symmetries) than the objects of C. This method of comparing amounts of structure has been employed widely in philosophy of physics in recent years.[15]

Since essentially surjective translations $F : T \to T'$ correspond to underlying functors F^* that are full, this captures an even more robust sense in which the essential surjectivity of F shows that T posits all the structure of T'. When F is essentially surjective, $F^* : \mathrm{Mod}(T') \to \mathrm{Mod}(T)$ is full. So according to the Baez method of comparing structure, F^* has not 'forgotten' structure and hence T has all of the structure of T'. This is in fact intuitive, since the fullness of F^* is telling us that the theory T' has just as many symmetries as the theory T. So if more symmetries is what indicates less structure, it cannot be that T has less structure than T'.

We have been arguing that the existence of an essentially surjective translation $F : T \to T'$ captures a sense in which T' has just as many, if not more, symmetries as T. And therefore Claim 2 respects the basic idea that many philosophers have expressed about structure: that more symmetries should indicate less structure. One might object to this by arguing that there is also a sense in which a specification has 'fewer' symmetries than the original theory. Let $\mathrm{Aut}(T)$ be the class of all automorphisms of models of T. When we move from T to a specification T' it will — insofar as the two theories are not logically equivalent — be the case that $\mathrm{Aut}(T')$ is properly contained in $\mathrm{Aut}(T)$. This is because $\mathrm{Mod}(T')$ is contained in $\mathrm{Mod}(T)$, so any automorphisms of the models that have been 'lost' in the move from T to T' will no longer be in $\mathrm{Aut}(T)$. This is indeed a sense in which a specification has 'fewer' symmetries than the original theory did. We would like to suggest, however, that this sense does not yield a satisfactory understanding of amounts of structure. In particular, if we look to classes of automorphisms of models of theories to determine amounts of structure, we are led to many unintuitive verdicts about which theories posit more structure than which others.

The following examples provide further evidence.

- Let T be an inconsistent theory. Then since $\mathrm{Aut}(T)$ is empty, it is properly contained in $\mathrm{Aut}(T')$ for every consistent theory T'. So T posits more structure than every consistent theory, no matter how rich the vocabulary is that the other theory employs. (Note that there is no translation, let alone an essentially surjective one, from T to any consistent theory.)

- Let T be a single-sorted theory that says there is exactly one thing, and let T' be a two-sorted theory that says that there is exactly one thing of the first sort, and has all the axioms of ZF set theory in the second sort. $\mathrm{Aut}(T)$ has a single member, while $\mathrm{Aut}(T')$ is quite rich. This is because automorphisms of models of T' consist of pairs of maps (one on elements of the first sort, one on elements of the second sort) that preserve the

[15] It is also well known in the category theory community. See Baez and Shulman (2010). For further details on it see Barrett (2021) and the references therein.

vocabulary of T', and there are some non-trivial automorphisms of models of ZF. And so this would seem to imply that the trivial theory T posits more structure than ZF set theory. (There is, indeed, a compelling sense in which T' is equivalent to ZF set theory; the two theories are Morita equivalent.)

- Let T' be the theory of groups, and let T be the theory of groups plus the additional axiom saying that there is exactly one thing. Note that T is definitionally equivalent to the theory in the empty signature with the one axiom saying there is exactly one thing. It is clear that the class of automorphisms of models of T is a subclass of the class of automorphisms of T' since the class of models of T is a subclass of the class of models of T'. But it is unappealing to say that the trivial theory T posits more structure than all of group theory.

Both of these cases yield unintuitive verdicts. This suggests that this way of comparing numbers of symmetries between theories is not what philosophers who emphasize the connection between symmetry and structure have in mind. In particular, the intuition about amount of structure being inverse to symmetry group size comes from comparing individual models — as we did above when motivating Claim 2 — and not from comparing all symmetries of all models by looking at $\mathrm{Aut}(T)$ as this counterargument suggests.

We can now return properly to the question of specifications. Suppose that T' is a specification of T. There is clearly a full functor from $\mathrm{Mod}(T')$ to $\mathrm{Mod}(T)$. The functor that 'includes' the category $\mathrm{Mod}(T')$ into $\mathrm{Mod}(T)$ is full — or in other words, the functor i^*, where i is the identity translation $i : T \to T'$ — since every model of T' is a model of T. According to all of the proposals for how to compare amounts of structure that are on the table, therefore, a specification does not have more structure than the theory we began with. The point here is a simple one. There is a conceptual difference between, on the one hand, the relationship that holds between topological spaces and sets, and on the other hand, the relationship that holds between Abelian groups and groups. In the former case, the difference lies in the 'number of symmetries' the objects admit. In the latter case, the difference lies in the 'number of models' that the theories have.

One can also understand this conceptual difference as we described it above. In the first case the difference is that the one theory has more vocabulary (and some new axioms involving that new vocabulary) and therefore fewer symmetries since they must preserve all of this new vocabulary. In the second case the one theory has no new vocabulary, only some new axioms in the old vocabulary. Philosophers of science have traditionally described the former kind of case as exhibiting a difference in 'structure'; one can see this from the fact that they emphasize the close relationship between structure and symmetry. And clearly a specification and its original theory do not differ in that sense. Rather, they only differ in the latter sense. A specification will most often have 'fewer' models than the original theory, in the sense that some models of the original theory

will not be models of the specification, but there is no substantive difference in the symmetries that these models admit.

This leads us to another piece of evidence that adding only axioms does not add structure in the sense that the literature has focused on. It is standard to say that adding axioms increases 'logical strength', not structure. The fact that there is a different term for what happens when we add axioms — we add logical strength — suggests that this is not what philosophers have in mind when they talk about structure. The concept of logical strength goes back at least to van Fraassen (1980), who says that "logical strength is determined by the class of models (inversely: the fewer the models the (logically) stronger the theory!)." Adding axioms to a theory is what results in fewer models, so we have good reason to think that logical strength in van Fraassen's sense is what is increased when axioms are added. This lines up with our idea from above. Adding axioms to a theory results in fewer models, while adding *structure* to a theory in the form of new vocabulary results in fewer symmetries admitted by these models. The same concept of logical strength has been employed in more recent literature. Williamson (2017, p. 336) writes that "in the context of logic [...] a theory T is stronger than a theory T^* if and only if T entails T^* but T^* does not entail T: every theorem of T^* is a theorem of T, but not every theorem of T is a theorem of T^*." Williamson's concept of logical strength lines up perfectly with the notion of a specification. We should therefore think that *logical strength*, and not the amount of structure posited, is what is increased when we move to a specification.

There is a potential objection that one might raise at this point. One can provide an example of a specification T' of a theory T that is definitionally equivalent to a theory that appears to add vocabulary to T. This would mean that specifications (up to definitional equivalence) add vocabulary, and so can be said to posit more structure than the original theory. We have the following example. Let Σ be a signature containing a unary predicate symbol p, and consider a Σ-theory T. Let T' be a specification of T that contains some new axiom involving p. Let T'^+ be the $\Sigma \cup \{p'\}$-theory where p' is a new unary predicate symbol, and T'^+ has all of the axioms of T' but also the axiom $\forall x(p(x) \leftrightarrow p'(x))$. We immediately see that T'^+ and T' are definitionally equivalent; indeed, the former is a definitional extension of the latter. But T'^+ is obtained from T by adding both vocabulary and axioms. So if definitionally equivalent theories posit the same amount of structure, and every addition of vocabulary to a theory added structure, then specifications add structure.

This example helps us to clarify exactly what it is to 'add vocabulary' to a theory. A mere addition of vocabulary to a theory does add structure, but adding axioms can then sometimes take away that structure. That is what happens in the above example: the new structure was 'collapsed' or 'reduced' back into the old by the addition of the axiom defining the new structure p' as being equivalent to the old piece of structure p. Just as not every axiom added is a 'new axiom' — if, for example, the axiom was already entailed by the original theory — not every piece of vocabulary added is new structure. When there is no essentially surjective translation from T to T', that captures

a robust sense in which T' resulted from a *genuine* addition of vocabulary to T. If there is no essentially surjective translation from T to T', that means that the former cannot 'construct' or 'define' all of the vocabulary that the latter employs. This guarantees that a specification T' of T is not definitionally equivalent to any theory that genuinely adds vocabulary to T.[16] This does mean that a specification might posit *less* structure than the original theory, if the specification collapses some of the structures of the old theory. Consider, for example, a theory with two unary predicate symbols p and q and no axioms. The specification of this theory that says $\forall x(p(x) \leftrightarrow q(x))$ has taken structure away from the original theory; it has in effect reduced the genuine vocabulary that the theory employs.

With this clarification in hand, we have a collection of compelling reasons to think that specifications do not posit more structure than the original theory. First, the operant notion of structure in the literature meshes better with the understanding of 'more structure' as 'more vocabulary' rather than 'more axioms'. And second, the concept of 'more axioms' has been explicitly discussed by philosophers in the guise of 'logical strength' and not structure. That said, one may still think that philosophers *should* think that more axioms means more structure, despite our claim that they do not as a descriptive matter of fact. We will return to this issue in the following section. We will show that even if one thinks that specifications posit more structure than the original theory, amounts of structure *still* cannot be 'counted'.[17]

We have now argued for Claim 2, and we can return to consider our pair of theories T_1 and T_2. As we showed in Theorem 4, there are essentially surjective translations $H : T_1 \to T_2$ and $K : T_2 \to T_1$. Claim 2 implies that T_1 posits all of the structure of T_2 and T_2 posits all of the structure of T_1.[18] But Theorems 1, 2, and 5 again imply that T_1 and T_2 are inequivalent and do not posit the *same structure*. This shows that Principle 1 is false and theories lack what one might call the 'co-Cantor-Bernstein property'. It is easy to verify that if there

[16] If there were a theory U that was definitionally equivalent to T' and genuinely added vocabulary to T, then there would be an essentially surjective translation from T to T' (the identity translation) and an essentially surjective translation from T' to U (by Lemma 1). This contradicts the fact that (since U genuinely adds vocabulary to T) there is no essentially surjective translation from T to U.

[17] And moreover, even if one thinks that specifications posit more structure than the original theory, the conclusion of this present section still follows: there can be theories such that the one posits all of the structure of the other and vice versa, but they do not posit the same structure. It only takes a moment to see this. Suppose that adding axioms to a theory adds structure. The existence of the conservative translations between T_1 and T_2, in combination with Proposition 2, shows that we can obtain the theory T_1 from T_2 (add vice versa) just by adding new predicates to T_1 and some new axioms that involve those predicates. This is what Proposition 2 illustrated. This captures a sense in which T_2 posits all of the structure of T_1; it can be obtained from T_1 by adding axioms and vocabulary. And the same holds in the other direction. T_1 can be obtained from T_2 by adding axioms and vocabulary. So T_1 posits all of the structure of T_2 and vice versa, but they do not posit the same structure, and the Co-Cantor-Bernstein property still does not hold.

[18] Intuitively, H shows that T_1 defines q_0 in the following sense: H 'translates' T_1 to a theory to which merely adding axioms turns it into T_2. This is implied by the proposition above and our discussion of specifications.

are surjections $f : X \to Y$ and $g : Y \to X$ between sets X and Y, then there is a bijection between X and Y. One might call this the 'co-Cantor-Bernstein theorem', and therefore say that sets have the co-Cantor-Bernstein property. Once again, it makes sense to talk about the co-Cantor-Bernstein property for any category. We can say that a category C has the **co-Cantor-Bernstein property** if for any objects c and d of C whenever there is an epimorphism (i.e. a generalization of the concept of an surjection) from $c \to d$ and an epimorphism $d \to c$, the objects c and d are isomorphic. Our example here demonstrates that the category of theories does not have this property. Our two theories T_1 and T_2 can be 'surjected onto' one another, but nonetheless, as Theorems 1, 2, and 5 tell us, they are not equivalent. The structural commitments that theories make therefore behave in a much more subtle manner than has so far been recognized.

6 Structure and equivalence

We now turn our attention to two payoffs that our results yield for recent discussions of structure and equivalence. First, they allow us to clarify the overall geography of standards of equivalence that have recently been proposed. And second, they allow us to raise a concern about recent discussion of the 'amount of structure' that different theories posit.

Morita equivalence and translation

The first payoff concerns the recent debate about when two theories should be considered equivalent, and in particular, the relationship that Morita equivalence bears to 'translation' criteria for equivalence — those standards of equivalence that require the existence of suitable translations between the theories in question. The simple example that we have considered here allows us to further clarify the overall geography of standards of equivalence between theories.[19]

Two theories T and T' are **mutually faithfully interpretable** if there are conservative translations $F : T \to T'$ and $G : T' \to T$.[20] Mutual faithful interpretability is one standard of equivalence that requires a kind of 'mutual translatability' between the two theories in question. But one can easily imagine others. For example, we say that two theories are **mutually surjectively interpretable** if there are essentially surjective translations $F : T \to T'$ and $G : T' \to T$. Now it is already well known that definitional equivalence is a stricter standard of equivalence than mutual faithful interpretability. More

[19] For discussion of these relationships, see for example Barrett and Halvorson (2016a,b) and Button and Walsh (2018, p. 118). In addition to the recent articles on equivalence already cited, see Barrett (2017), Coffey (2014), Curiel (2014), Halvorson (2013), Glymour (2013), Hudetz (2015, 2017a), Knox (2011, 2014), North (2009), Rosenstock et al. (2015), Rosenstock and Weatherall (2016), Teh and Tsementzis (2017), Van Fraassen (2014), and Weatherall (2016, 2017), and the references therein. See Weatherall (2019a) for a survey of recent work.

[20] See Button and Walsh (2018, p. 117). Our presentation here is less general. Button and Walsh work with a general notion of an interpretation, which allows one theory to define the domain of another theory through definable equivalence relations.

precisely, any pair of theories that are definitionally equivalent are mutually faithfully interpretable, but there are mutually faithful interpretable theories that are not definitionally equivalent (Andréka et al., 2005; Button and Walsh, 2018). One wonders exactly where these standards of mutual translatability fall in the broader geography of standards of equivalence.[21]

In particular, Morita equivalence is a more liberal notion of equivalence than definitional equivalence. So it is natural to ask the following questions: *If two theories are mutually faithfully interpretable, then are they also Morita equivalent? If two theories are mutually surjectively interpretable, then are they also Morita equivalent? If two theories are mutually faithfully interpretable and mutually surjectively interpretable, then are they also Morita equivalent?* Beyond simply being important for better understanding the relations between different standards of equivalence, these questions help us to evaluate these standards of equivalence on their own merits, especially since standards of equivalence that merely require mutual translatability between theories tend to be implausibly liberal. For example, if the answer to the first question were yes, then that would be a mark against Morita equivalence as a plausible standard of equivalence between theories. It is widely accepted that mutual faithful interpretability is too liberal a standard of equivalence; it considers theories to be equivalent that we have good reason to consider inequivalent (Szczerba, 1977). So if mutual faithful interpretability entailed Morita equivalence, that would immediately imply that Morita equivalence is too liberal a standard as well.[22]

The answer to all of the above questions, however, is no. It is not the case that Morita equivalence is entailed by mutual faithful interpretability, nor by mutual surjective interpretability, nor by their conjunction. Our discussion here has demonstrated precisely this. It follows from Theorem 3 that our theories T_1 and T_2 are mutually faithfully interpretable and from Theorem 4 that they are mutually surjectively interpretable, but as was stated in Theorem 2, they are not Morita equivalent. While this is a mark in favor of Morita equivalence — or rather, the dismissal of a potential charge *against* it — these facts do raise some important questions. Morita equivalence is more liberal than definitional equivalence, which itself is the same as a particular kind of translation criterion. But on the other hand, it is less liberal than a handful of other translation criteria for equivalence — mutual faithful interpretability and mutual surjective interpretability. One therefore wonders *how exactly Morita equivalence relates to translatability between theories.* When two theories are Morita equivalent, what kind of translation does that imply exists between them? And conversely, what kind of translation between two theories will imply that they are Morita equivalent? Is Morita equivalence the same as some translation criterion? This

[21] We will discuss how they relate to Morita equivalence below, but one also wonders how they relate to categorical equivalence, a strictly weaker standard than Morita equivalence. In particular, one wonders whether there is an example of theories that are mutually faithfully or surjectively translatable, but not categorically equivalent. The theories T_1 and T_2 that we consider above are categorically equivalent (Barrett and Halvorson, 2016b), so the question remains open.

[22] See McEldowney (2020), for example, for worries about Morita equivalence along these lines.

question is of obvious philosophical importance. Morita equivalence conceptually resembles definitional equivalence since it requires the existence of a particular kind of extension of the two theories in question. But it can nonetheless be difficult to see how Morita equivalent theories 'say the same thing'. If one can make precise the kind of 'intertranslatability' that Morita equivalence corresponds to, this would take a significant step towards capturing the sense in which Morita equivalent theories 'say the same thing'.

It is worth taking a moment to mention some work that has recently been done on these questions, and to raise a question of our own.[23] Recall that we have the following result about definitional equivalence. If Σ and Σ' are disjoint signatures and T is a Σ-theory and T' is a Σ'-theory, then the following conditions are equivalent:

- T and T' are definitionally equivalent.
- There is an essentially surjective and conservative translation $F : T \to T'$.
- There are translations $F : T \to T'$ and $G : T' \to T$ such that i) $T \vDash \phi \leftrightarrow GF\phi$ for every Σ-formula ϕ and ii) $T' \vDash \psi \leftrightarrow FG\psi$ for every Σ'-formula ψ.

The equivalence of the first two conditions follows from Lemma 1, while Theorems 4.6.17 and 6.6.21 of Halvorson (2019) imply that the third is equivalent to the first two. It is worth taking a moment to unravel this result. The third condition is telling us that definitional equivalence is capturing a particular kind of 'isomorphism' between theories; there exist translations between T and T' that are 'almost inverse' to one another (in the precise sense given by the third condition) if and only if the two theories are definitionally equivalent. When the third condition holds we say that the theories T and T' are **intertranslatable**, and we say that F and G are each **one half** of an intertranslation. The second condition provides us with a necessary and sufficient condition that we can use to tell, simply by examining a translation $F : T \to T'$, whether it is an 'isomorphism' in this sense, i.e. whether it is one half of an intertranslation between T and T'.

One would like to have an analogous result about Morita equivalence. Such a result would clarify the sense in which Morita equivalent theories are saying the same thing. One immediately notices, however, that the variety of translation that we have been working with so far is too restrictive to be closely related to Morita equivalence; two theories that employ different sort symbols can be Morita equivalent, but our variety of translation does not allow one to translate sorts to other sorts. Fortunately, the concept of a 'generalized reconstrual' exists. A generalized reconstrual $F : \Sigma \to \Sigma'$ provides one with the flexibility to translate between many-sorted signatures. Corresponding to this more liberal notion of translation, one obtains a weakening of intertranslatability, which

[23] In addition to the work cited below, see the following work on Morita equivalence and translation: (Barrett and Halvorson, 2016b, Theorem 4.6), McEldowney (2020), Washington (2018), and D'Arienzo et al. (2020).

has naturally come to be called 'weak intertranslatability'; it is a weakening of the third condition in the above result about definitional equivalence. It has recently been shown by Washington (2018) that if two theories are weakly intertranslatable, then they are Morita equivalent. And conversely, for 'proper' theories, Morita equivalence entails weak intertranslatability. (See Halvorson (2019) for discussion and full details on generalized reconstruals.) This provides us with a result that is analogous to the equivalence of the first and third conditions in the above fact about definitional equivalence.

One would still like, however, a condition that corresponds to the second condition. In particular, one would like to be able to tell, simply by examining a generalized translation $F : T \to T'$, whether it is an 'isomorphism' in the sense that its existence implies that T and T' are Morita equivalent. The following example illustrates that it is not the case that the existence of an essentially surjective and conservative generalized translation implies that the two theories are Morita equivalent. We first note that a generalized translation can map each sort symbol in a signature Σ to another single sort symbol in Σ'; such a map can be thought of as 'translating' quantification over the former sorts to quantification over the latter. Generalized translations are much more flexible than this, but this simple kind will serve our purpose in what follows.[24]

Example 3. Let Σ be a signature that contains just the two sort symbols σ and σ'. Let Σ' contain just σ'. We define the Σ-theory T to have two axioms; one says that there are two things of sort σ, the other says that there are two things of sort σ'. The Σ'-theory T' says that there are two things of sort σ'.

We now describe a generalized translation $F : T \to T'$ that is conservative and essentially surjective, but does not witness the Morita equivalence of T and T'. F maps both of the sorts σ and σ' in Σ to the one sort σ' in Σ'. This map extends to a map on all Σ-formulas in a natural manner: it simply changes all quantification over σ into quantification over σ'. We note that this is indeed a translation. Both of the axioms of T translate to the single axiom of T'. The substitution theorem for generalized translations (Halvorson, 2019, 5.4.7) immediately implies that if $T \vDash \phi$, then $T' \vDash F\phi$. Furthermore, F is both conservative and essentially surjective. It is conservative because T is complete. Suppose that $T' \vDash F\phi$. Then either $T \vDash \phi$ or $T \vDash \neg\phi$; since F is a translation, the latter would imply that T' is inconsistent, which it is not. So $T \vDash \phi$, and hence F is conservative. It is trivial that F is essentially surjective, since there are no predicate symbols in Σ'.

Finally, we note that T and T' are not Morita equivalent. One can easily verify that these two theories do not have equivalent categories of models. The model of T has four automorphisms, while the model of T' has only two. Since Morita equivalence entails categorical equivalence (Barrett and Halvorson, 2016b), this implies that the two theories are not Morita equivalent, and therefore F should not be considered an 'isomorphism'; it is not one 'half of a weak intertranslation' between them. ⌟

[24] For preliminaries on many-sorted logic, the reader is encouraged to consult Halvorson (2019).

This example shows that in order for a generalized translation $F : T \to T'$ to witness the Morita equivalence of T and T', it must satisfy some third condition in addition to essential surjectivity and conservativity. One would like to be able to clearly state this third condition. It is not surprising that there should be three distinct requirements on F for it to indicate that two theories are Morita equivalent. Morita equivalence is closely related to categorical equivalence (Hudetz, 2017a). And for a functor — a map that one can roughly think of as a 'translation' between categories — to realize the categorical equivalence of two categories of models, it must be full, faithful, and essentially surjective. We have already said above what it is for a functor to be full. A functor F is **faithful** if $Ff = Fg$ implies that $f = g$ for all arrows $f : c_1 \to c_2$ and $g : c_1 \to c_2$ in C. F is **essentially surjective** if for every object d in D there exists an object c in C such that $Fc \cong d$. A functor $F : C \to D$ that is full, faithful, and essentially surjective is called an **equivalence of categories**. As we remarked above, it has recently been shown that a translation F is essentially surjective if and only if its underlying functor F^* is full. The conservativity of F corresponds roughly with the essential surjectivity of the underlying functor; both conditions are capturing a sense in which the two theories have the 'same number of models'. And therefore, this leaves a third condition on F that should correspond to the underlying functor being faithful. This condition should capture a sense in which T' can define the sorts of T. Note that this is precisely what is failing in the above example; T' does not have the resources to define both of the sort symbols of T. If one could isolate precisely what this condition is, we would take a step towards better understanding exactly how Morita equivalence relates to translation.

Can we 'count' structure?

We conclude with a remark about the extent to which one can 'count' the amount of structure that a theory posits. In recent years, many metaphysicians, philosophers of physics, and physicists have adopted a version of the following methodological principle.

Structural parsimony. All other things equal, we should prefer theories that posit less structure.

North (2009, p. 64), for example, puts this idea as follows:

> This is a principle informed by Ockham's razor; though it is not just that, other things being equal, it is best to go with the ontologically minimal theory. It is not that, other things being equal, we should go with the fewest entities, but that we should go with the least structure.

There are some fairly straightforward cases where one theory seems to posit 'more' structure than another, and we have good reason to prefer the theory that posits less. The most famous example, as we mentioned above, is the case

of Newtonian and Galilean spacetime. It is standard to claim that the Galilean spacetime theory posits less structure than its Newtonian counterpart, and it is also standard to prefer the former theory.

The structural parsimony principle, however, is only useful if it can be applied in many (or most) cases where we want to adjudicate between rival theories. For example, in a case of 'incomparable' structure — where two theories posit different structure, but neither posits less structure than the other — the principle would not help us.[25] The way that we often talk about the structure that theories posit, however, suggests that the structural parsimony principle *will* be applicable in most cases where we want to choose between theories. Indeed, our use of phrases like 'more structure', 'less structure', and 'amount of structure' suggests that the amount of structure that a theory posits can be represented by a real number.[26] Other quantities that we talk about using words like 'amount', 'less', and 'more' tend to behave this way. We can speak, for example, about the 'amount of money' that an individual has and often say that some individual has 'more' or 'less' money than another. And in that case we *can* attach a real number quantity to the amount of money that an individual has; amounts of money are something that we can genuinely count.

Since we talk about amounts of structure that theories posit in much the same terms, one might suspect that we can attach a real number to a theory that represents the 'amount of structure' that the theory posits and genuinely count amounts of structure. If so, then the structural parsimony principle would be applicable quite generally. If two theories posit different structure, the principle would always weigh in favor of one or the other of them. And the kind of situation mentioned above — in which two theories posit 'incomparable' amounts of structure — would be impossible. There is already some compelling evidence that such cases *are* possible. For example, there is a sense in which Galilean and Minkowski spacetimes posit incomparable amounts of structure and in which symplectic and metric spaces posit incomparable amounts of structure (Barrett, 2015a,b).

Our main example provides another case of this. Recall that a binary relation $\leq$ on a set X is a **partial order** if it is reflexive (i.e. $x \leq x$ for all x), antisymmetric (i.e. if $x \leq y$ and $y \leq x$, then $x = y$ for all x and y), and transitive. It is a **total order** if, in addition, any two elements are comparable (i.e. $x \leq y$ or $y \leq x$ for all x and y). In what follows we will employ the following setup. We will let X be a set equipped with a binary relation $\leq$. We will think of the elements of X as the possible 'amounts of structure' that theories might posit and the relation $\leq$ as capturing the 'is no more structure than' relation. We will then let s be a map that assigns to each theory some element of X, which we think of as the amount of structure that the theory posits. We can then ask how these amounts of structure behave. For example, if X is $\mathbb{R}$ and $\leq$ is the standard total order on $\mathbb{R}$, we would be attaching a real number to each theory that quantifies its amount of structure. The following proposition,

[25] See (Barrett, 2015a,b) for discussion.

[26] For other discussion of 'amounts of structure' see Barrett (2015a,b, 2018), Weatherall (2017), Bradley (2020), Bradley and Weatherall (2020), and the references therein.

however, shows that $\leq$ cannot behave like the standard ordering on the real numbers.

Proposition 5. *Let X be a set equipped with a binary relation $\leq$ and suppose that s is a map from first-order theories to X. Then the following three conditions cannot be jointly maintained:*

1. *$\leq$ is a total order.*
2. *If T^+ is a conservative extension of T and $s(T^+) \leq s(T)$, then T is definitionally equivalent to T^+.*
3. *If T and T' are definitionally equivalent, then $s(T) = s(T^+)$.*

Proof. We know by Proposition 2 and Theorem 3 that T_1 is definitionally equivalent to a conservative extension T_1^+ of T_2 and T_2 is definitionally equivalent to a conservative extension T_2^+ of T_1. By condition 3, $s(T_1^+) = s(T_1)$ and $s(T_2^+) = s(T_2)$. Suppose that $s(T_1) \leq s(T_2)$. This implies that $s(T_1^+) \leq s(T_2)$. But T_1^+ is a conservative extension of T_2, so condition 2 implies that T_1^+ is definitionally equivalent to T_2. That cannot be the case, since if it were, T_2 and T_1 would be definitionally equivalent, contradicting Theorem 1. This implies that it is not the case that $s(T_1) \leq s(T_2)$. One shows in precisely the same manner that it is not the case that $s(T_2) \leq s(T_1)$. So $\leq$ is not a total order, contradicting condition 1. □

The first condition captures the idea that we can genuinely 'count' the structure that a theory posits. Every theory T has an associated 'amount of structure' that is represented by the element $s(T)$ in the totally ordered set X. The totally ordered set X could, for example, be the real numbers $\mathbb{R}$ with their standard ordering, in which case we would be 'counting' amounts of structure in the most natural way possible. Since the three conditions cannot all be the case, if there are compelling arguments for the second and third conditions, we should reject the first. And indeed, this is the case.

The argument for condition 2 is straightforward. Suppose that T^+ is a conservative extension of T. That means that T^+ and T entail precisely the same sentences in the language of T. If T^+ and T are going to be inequivalent, therefore, the difference between them must concern something that T^+ says using the conceptual resources that are available to it, but not to T. The requirement that $s(T^+) \leq s(T)$, however, rules out this possibility. T^+ posits less structure than or the same amount of structure as T does, so T^+ does not have access to more conceptual resources than those available to T. This means that the two theories must actually be saying the same thing, and so they should be definitionally equivalent.[27] And condition 3 is intuitive. If two theories are

[27] It is important to mention that this same kind of argument would go through even if one moved to a more general standard of equivalence like Morita equivalence. In that case, the equivalence relation in the consequent of condition 3 would simply be changed to Morita equivalence, rather than definitional equivalence. The same argument given in our proof would then imply that T_1 and T_2 are Morita equivalent, and that would contradict Theorem 2.

definitionally equivalent, then each can define or 'build' the structures of the other. So in a strong sense they actually posit *same structure*, not merely the same *amount* of structure. We therefore have compelling arguments for the second and third conditions, so we should reject the first: structure cannot be counted.

The basic idea behind this result is the following. The theories T_1 and T_2 posit incomparable amounts of structure, in the sense that either theory can be obtained from the other by 'adding' some structure, but the two are not equivalent. This is what was shown in the discussion surrounding Proposition 2. T_1 is (up to definitional equivalence) a conservative, but not definitional, extension of T_2, and T_2 is (up to definitional equivalence) a conservative, but not definitional, extension of T_1. This implies that we can obtain T_1 from T_2 by simply 'adding' structure to T_1, along with some new axioms dictating how that new structure behaves. There are theories with incomparable amounts of structure, so the relation $\leq$ is not total.

In addition to not being totally ordered by 'amount of structure', our example demonstrates that theories cannot even be *partially* ordered by the amount of structure they posit.

Proposition 6. *Let X be a set equipped with a binary relation $\leq$ and suppose that s is a map from first-order theories to X. Then the following three conditions cannot be jointly maintained:*

1. *$\leq$ is a partial order.*
2. *If T^+ is a conservative extension of T and $s(T^+) \leq s(T)$, then T and T^+ are definitionally equivalent.*
3. *If there is an essentially surjective translation from a theory T to a theory T', then $s(T') \leq s(T)$.*

Proof. Suppose for contradiction that conditions 1, 2, and 3 are all true. Consider our two theories T_1 and T_2. Condition 1 implies that $\leq$ is antisymmetric, so condition 3 and Theorem 4 entail that $s(T_1) = s(T_2)$. Theorem 3 and Proposition 2 imply that there is a theory T_2^+ that is a definitional extension of T_2 and a conservative extension of T_1. Since T_2^+ is trivially definitionally equivalent to T_2, condition 3 and Lemma 1 imply that $s(T_2^+) = s(T_2)$. So $s(T_2^+) = s(T_1)$. Condition 2 then implies that T_2^+ and T_1 are definitionally equivalent. Since T_2^+ is a definitional extension of T_2, it must be that T_1 and T_2 are definitionally equivalent, which contradicts Theorem 1. □

We take a moment here to unravel what Proposition 6 is telling us. The result is closely related to Proposition 5. Condition 1 again captures the idea that we can 'count' the structure that a theory posits, but in a weaker sense than in the previous proposition. The partially ordered set X could still be the real numbers $\mathbb{R}$ with their standard ordering, but condition 1 is more general than that, however, since $\mathbb{R}$ is *totally* ordered. Condition 1 does not require

that. Conditions 2 and 3 are then natural requirements about how this map s should behave, and they contradict the fact that $\leq$ is antisymmetric.

Condition 2 is the same as in Proposition 5, and condition 3 is slightly stronger than its counterpart in Proposition 5. But arguments for condition 3 have already been given above in the course of arguing for Claim 2. As that claim states, an essentially surjective translation $F : T \rightarrow T'$ captures a sense in which T has all of the structure of T'. This means that T has at least as much structure as T', and so, insofar as there is a map s that encodes amounts of structure in terms of real numbers, $s(T') \leq s(T)$. Claim 2 therefore implies condition 3 in Proposition 6. And moreover, even without Claim 2, the discussion surrounding Proposition 4 showed that all of the tools that have recently been used to compare amounts of structure between theories imply that when there is an essentially surjective translation $F : T \rightarrow T'$, $s(T') \leq s(T)$.

Now one might try to resist condition 3 in Proposition 6 — and, for that matter, our earlier Claim 2 — by claiming once again that a specification posits more structure than the original theory did; adding axioms does add structure. Recall that this was a crucial premise in our argument for Claim 2. One might at this point be inclined to revisit this idea in order to resist the conclusion that we are drawing from Proposition 6 — that structure cannot be partially ordered, much less 'counted'. We argued above that this thought is not in line with the standard understanding of structure that is present in the literature. But we can here go one step further. If specifications posit more structure than the original theory did, then there is a pair of theories that each posit more structure than the other. The basic idea behind this point is simple. There exist essentially surjective translations back and forth between our theories T_1 and T_2. Proposition 3 implies then that T_1 is definitionally equivalent to a specification of T_2 and that T_2 is definitionally equivalent to a specification of T_1. Insofar as definitionally equivalent theories posit the same amount of structure, then if specifications posit *more* structure than the original theory, this would imply that T_1 posits more structure than T_2 and T_2 posits more structure than T_1.

We make this idea precise with the following result. Recall that we say that $x < y$ just in case $x \leq y$ and $x \neq y$.

Proposition 7. *Let X be a set equipped with a binary relation $\leq$ and suppose that s is a map from first-order theories to X. Then the following three conditions cannot be jointly maintained:*

1. *$\leq$ is a partial order.*
2. *If T' is a specification of T that is not logically equivalent to T, then $s(T) < s(T')$.*
3. *If two theories T and T' are definitionally equivalent, then $s(T) = s(T')$.*

Proof. Consider our pair of theories T_1 and T_2. Theorem 4 implies that there is an essentially surjective translation $H : T_1 \rightarrow T_2$. Proposition 3 then implies that there is a Σ_1-theory T_1^H that is a specification of T_1 and definitionally

equivalent to T_2. It cannot be logically equivalent to T_1, because that would imply that T_1 and T_2 are definitionally equivalent, and we already know that they are not. So conditions 2 and 3 imply that $s(T_1) < s(T_1^H)$ and $s(T_1^H) = s(T_2)$. This means that $s(T_1) < s(T_2)$. Similarly, by Theorem 4 there is a Σ_2-theory T_2^K that is a specification of T_2 and definitionally equivalent to T_1. Conditions 2 and 3 again imply that $s(T_2) < s(T_2^K)$ and $s(T_2^K) = s(T_1)$, so $s(T_2) < s(T_1)$. Now $s(T_1)$ is both less than and greater than $s(T_2)$, which contradicts condition 1. □

The basic idea behind this result is the following. Conditions 1 and 3 are the same as in Proposition 5. Condition 2 is the idea that a specification — that is not logically equivalent to the original theory, perhaps obtained by adding an axiom that was not a theorem of the original theory — posits more structure than the original theory did. Proposition 7 demonstrates that these three conditions cannot all be the case. If conditions 2 and 3 hold, then there are theories T_1 and T_2 that each posit more structure than the other. This provides a compelling reason to think that adding only axioms and no new vocabulary to a theory does not add structure. And more to the point, it shows that one cannot resist the conclusion that we are drawing from Proposition 6 — namely, that structure cannot be 'counted' or even partially ordered — by claiming that adding axioms adds structure. If adding axioms adds structure, then insofar as definitionally equivalent theories posit the same amount of structure, we see that once again structure cannot be counted.

Given our discussion surrounding Claim 2, we of course find it more natural to think of the relation $\leq$ in Proposition 7 as ordering theories by their *logical strength*, rather than by the amount of structure that they posit. In recent years logical strength has been put forward by philosophers of logic variously as both a vice and a virtue of theories. Williamson (2017, p. 336), for example, thinks that logical strength is a virtue among logical theories, just as it is among scientific theories, since stronger theories are "more specific or informative". In the case of scientific theories, Williamson writes that we consider strength a virtue because it "provide[s] a minimal threshold of informativeness below which theories do not even come up for serious abductive evaluation. We want scientific theories to inform us about their subject matters; weak theories do too little of that to give us what we want. Furthermore, strength contributes to explanatory power [...], the capacity to bring our miscellaneous information under generalizations that unify it in illuminating ways" (Williamson, 2017, p. 336).[28] Hjortland (2017) argues that strength is a vice. Russell (2019) argues that it is neither a virtue nor a vice. Proposition 7 here yields a cautionary remark about using logical strength to decide between theories: vice or virtue, it cannot be counted in the way that we might have thought. The relation 'has no more logical strength than' does not partially order theories. This means that it is somewhat

[28] Williamson does not consider logical theories in different languages, and the proof of our Proposition 7 does require this. But the way we have generalized the proposal to different languages is by appeal to condition 3, which is a perfectly intuitive desiderata on any notion of logical strength.

misleading to talk about theories as having 'more' or 'less' logical strength than others.

The previous three propositions show that theories are not totally or even partially ordered by the 'amount of structure' that they posit. Amounts of structure are therefore *not* something that we can genuinely count, or even order in the most natural way. This isolates an important difference between the structural parsimony principle and other 'ontological' parsimony principles that license us to prefer the theory that commits to fewer *entities*. It is natural to suppose that we can genuinely count — and therefore attach a real number quantity to — the number of entities (or the number of kinds of entities) that a theory is committed to. We cannot do so with structure. It is important to emphasize that we still can, however, *compare* structure between theories, insofar as what mean by the phrase 'X has less structure than Y' is just that Y has all of the structures that X has, and possibly others in addition. And indeed, this understanding of the relation 'has less structure than' captures the kind of cases that motivated us to compare amounts of structure in the first place. Newtonian spacetime, for example, has all of the structure that Galilean spacetime has, but it also has a preferred standard of rest.[29] We can look to translations between theories to assess whether this relation holds between them. But as Propositions 5, 6, and 7 demonstrate, this understanding of the relation 'posits less structure than' does not end up ordering theories by the amount of structure that they posit. Much care must therefore be taken when comparing structure between theories.

7 Conclusion

One might wonder how these purely logical results come to bear on the current debates about structure and equivalence in philosophy of science. Real scientific theories are, of course, much more complicated than theories in first-order logic. We therefore should be careful not to assume that the 'category of all scientific theories' has all the features that the category of first-order theories has. Nonetheless, we should take 'no go' results like the ones we have presented here seriously. If the category of first-order theories fails to have some simple feature — like the Cantor-Bernstein or co-Cantor-Bernstein property — then that gives us some reason to doubt that the category of all scientific theories will have that feature.

There is, however, an important question about how prevalent counterexamples to the Cantor-Bernstein and co-Cantor-Bernstein properties of theories are. One wonders whether all such counterexamples are pathological in some sense that would prevent them from occurring in the case of real scientific theories. On the one hand, if counterexamples are rare, then our results here are more surprising. If they are not rare, then one has reason to suspect they will arise for scientific theories as well, and therefore these results will have to be

[29] This is the concept that Barrett (2021) investigates, and arguably the one that bears the closest relationship to symmetry Barrett (2018).

taken seriously when reasoning about structure and equivalence in physics. We conclude by contributing two further results. Our aim is to begin a discussion of how common counterexamples like the one presented above are.

In particular, we show that if a pair of theories is a counterexample to either the Cantor-Bernstein or co-Cantor-Bernstein property, then both theories must, roughly, have infinitely many models. For convenience, we will slightly abuse our earlier terminology and say that a pair of theories (T_1, T_2) **has the Cantor-Bernstein property** if the existence of conservative translations $T_1 \to T_2$ and $T_2 \to T_1$ implies that T_1 and T_2 are intertranslatable. We will say that (T_1, T_2) **has the co-Cantor-Bernstein property** if the existence of essentially surjective translations $T_1 \to T_2$ and $T_2 \to T_1$ implies that T_1 and T_2 are intertranslatable.

We have the following two results, whose proofs have been placed in the appendix. We say that a category C is **object-finite** just in case it has only finitely many objects up to isomorphism.

Theorem 6. *If T_1 and T_2 are theories such that Mod(T_1) and Mod(T_2) are object-finite, then (T_1, T_2) has the Cantor-Bernstein property.*

Theorem 7. *If (T_1, T_2) does not have the co-Cantor-Bernstein property, then there is a cardinal number κ such that T_1 and T_2 have infinitely many non-isomorphic models of size κ.*

These results do not suggest that counterexamples like ours are all pathological, in the sense that they can only occur in strange cases that one would not expect to arise for real scientific theories. In particular, most fundamental theories in physics do have infinitely many models. But on the other hand, the results do begin to explain why it is difficult to come up with counterexamples to the Cantor-Bernstein and co-Cantor-Bernstein properties of theories. If theories are complicated enough to have infinitely many non-isomorphic models, then it will most likely be difficult to prove that such theories are mutually translatable in the relevant sense. For example, Andréka et al. (2005), who give a simple example of two theories that violate the Cantor-Bernstein property, construct the relevant conservative translations in a way that is quite complex. There is certainly room for more work on these questions, but at present it seems that counterexamples like the one considered in this paper must be taken seriously when theorizing about structure and equivalence.

References

Andréka, H., Madarasz, J., and Németi, I. (2008). Defining new universes in many-sorted logic. *Mathematical Institute of the Hungarian Academy of Sciences, Budapest*, 93.

Andréka, H., Madarász, J. X., and Németi, I. (2005). Mutual definability does not imply definitional equivalence, a simple example. *Mathematical Logic Quarterly*, 51(6):591–597.

Baez, J., Bartels, T., Dolan, J., and Corfield, D. (2006). Property, structure and stuff. Available at http://math.ucr.edu/home/baez/qg-spring2004/discussion.html.

Baez, J. and Shulman, M. (2010). Lectures on n-categories and cohomology. In Baez, J. and May, P., editors, *Towards Higher Categories*. Springer-Verlag New York.

Barrett, T. (2019). Structure and equivalence. *Forthcoming in Philosophy of Science*.

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

Barrett, T. W. (2015b). Spacetime structure. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 51:37–43.

Barrett, T. W. (2017). Equivalent and inequivalent formulations of classical mechanics. *Forthcoming in the British Journal for the Philosophy of Science*.

Barrett, T. W. (2018). What do symmetries tell us about structure? *Philosophy of Science*, 85:617–639.

Barrett, T. W. (2021). How to count structure. *Forthcoming in Noûs*.

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

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

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

Barrett, T. W. and Halvorson, H. (2017b). Quine's conjecture on many-sorted logic. *Synthese*, 194(9):3563–3582.

Bradley, C. (2020). The non-equivalence of Einstein and Lorentz. *Forthcoming in the British Journal for the Philosophy of Science*.

Bradley, C. and Weatherall, J. O. (2020). On representational redundancy, surplus structure, and the hole argument. *Forthcoming in Foundations of Physics*.

Button, T. and Walsh, S. (2018). *Philosophy and Model Theory*. Oxford University Press.

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

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

D'Arienzo, A., Pagano, V., and Johnson, I. (2020). The 2-categorical structure of predicate theories. *Manuscript: arXiv:2011.14056*.

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

Dewar, N. (2019). Ramsey equivalence. *Erkenntnis*, 84:77–99.

Dewar, N. (2021). *Structure and Equivalence*. Manuscript.

Earman, J. (1989). *World Enough and Spacetime: Absolute versus Relational Theories of Space and Time*. MIT.

Friedman, M. (1983). *Foundations of space-time theories: Relativistic physics and philosophy of science*. Princeton University Press.

Geroch, R. (1978). *General Relativity from A to B*. Chicago University Press.

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.

Halvorson, H. (2019). *The Logic in Philosophy of Science*. Cambridge University Press.

Hilbert, D. (1930). *Grundlagen der Geometrie*. Teubner.

Hjortland, O. T. (2017). Anti-exceptionalism about logic. *Philosophical Studies*, 174(3):631–658.

Hodges, W. (2008). *Model Theory*. Cambridge University Press.

Hudetz, L. (2015). Linear structures, causal sets and topology. *Studies in History and Philosophy of Modern Physics*, pages 294–308.

Hudetz, L. (2017a). Definable categorical equivalence: Towards an adequate criterion of theoretical intertranslatability. *Forthcoming in Philosophy of Science*.

Hudetz, L. (2017b). The semantic view of theories and higher-order languages. *Forthcoming in Synthese*.

Knox, E. (2011). Newton-Cartan theory and teleparallel gravity: The force of a formulation. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 42(4):264–275.

Knox, E. (2014). Newtonian spacetime structure in light of the equivalence principle. *The British Journal for the Philosophy of Science*, 65(4):863–880.

Makkai, M. and Reyes, G. E. (2006). *First order categorical logic: model-theoretical methods in the theory of topoi and related categories*, volume 611. Springer.

Maudlin, T. (2012). *Philosophy of Physics: Space and Time*. Princeton University Press.

McEldowney, P. A. (2020). On Morita equivalence and interpretability. *The Review of Symbolic Logic*, 13(2):388–415.

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

Quine, W. V. O. (1951). Ontology and ideology. *Philosophical Studies*, 2(1):11–15.

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

Rosenstock, S. and Weatherall, J. O. (2016). A categorical equivalence between generalized holonomy maps on a connected manifold and principal connections on bundles over that manifold. *Journal of Mathematical Physics*, 57(10). arXiv:1504.02401 [math-ph].

Russell, G. (2019). Deviance and vice: Strength as a theoretical virtue in the epistemology of logic. *Philosophy and Phenomenological Research*, 99(3):548–563.

Schwabhäuser, W. and Szczerba, L. (1975). Relations on lines as primitive notions for Euclidean geometry. *Fundamenta Mathematicae*.

Schwabhäuser, W., Szmielew, W., and Tarski, A. (1983). *Metamathematische Methoden in der Geometrie*. Springer.

Swanson, N. and Halvorson, H. (2012). On North's 'The structure of physics'. *Manuscript*.

Szczerba, L. (1977). Interpretability of elementary theories. In *Logic, Foundations of Mathematics, and Computability Theory*. Springer.

Tarski, A. (1959). What is elementary geometry? In *The Axiomatic Method With Special Reference to Geometry and Physics*. North-Holland.

Teh, N. and Tsementzis, D. (2017). Theoretical equivalence in classical mechanics and its relationship to duality. *Forthcoming in Studies in History and Philosophy of Modern Physics*.

Tsementzis, D. (2015). A syntactic characterization of Morita equivalence. *Manuscript*.

van Fraassen, B. C. (1980). *The Scientific Image*. Oxford.

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.

Visser, A. (2006). Categories of theories and interpretations. In *Logic in Tehran. Proceedings of the workshop and conference on Logic, Algebra and Arithmetic, held October 18–22, 2003*. ASL.

Washington, E. (2018). *On the Equivalence of Logical Theories*. Princeton University Bachelor's Thesis.

Weatherall, J. O. (2016). Are Newtonian gravitation and geometrized Newtonian gravitation theoretically equivalent? *Erkenntnis*, 81(5):1073–1091.

Weatherall, J. O. (2017). Category theory and the foundations of classical field theories. In Landry, E., editor, *Forthcoming in Categories for the Working Philosopher*. Oxford University Press.

Weatherall, J. O. (2019a). Classical spacetimes. In Knox, E. and Wilson, A., editors, *The Routledge Companion to Philosophy of Physics*. Routledge.

Weatherall, J. O. (2019b). Theoretical equivalence in physics. *Forthcoming in Philosophy Compass*.

Williamson, T. (2017). Semantic paradoxes and abductive methodology. In Armour-Garb, B., editor, *Reflections on the liar*. Oxford University Press.

Appendix

The purpose of this appendix is to provide proofs of Theorems 6 and 7. We will begin by the Cantor-Bernstein property. First, we show that if T_1 or T_2 has only finitely many non-isomorphic models, then (T_1, T_2) has the Cantor-Bernstein property. We begin with a result that is conceptually close to what Makkai and Reyes (2006) call "conceptual completeness".[30] They use sophisticated category theory to prove it, however, and the following result is more streamlined, in particular because it relies on a recent corollary to Beth's Theorem (Barrett, 2021, Prop. 2).

Proposition 8. *Suppose that $F : T_1 \to T_2$ is a translation and $F^* : Mod(T_2) \to Mod(T_1)$ is one half of an equivalence of categories. Then F is one half of an intertranslation.*

[30] What they show is deeply wrapped up in category-theorical language: if P is a pretopos and $I : P \to S$ is a logical functor such that I^* is an equivalence of categories, then so is I. Strictly speaking, their result does not imply the following proposition, because our theory T does not have to correspond to a pretopos. But the two are conceptually close to one another.

Proof. If F^* is one half of an equivalence of categories, then F^* is full. By Proposition 2 of Barrett (2021), F is essentially surjective. F^* is also essentially surjective. By Proposition 6.6.17 of Halvorson (2019) F is conservative. Since it is essentially surjective and conservative, Proposition 4.5.27 of Halvorson (2019), implies that F is one half of an intertranslation. □

We say that a category C is **object-finite** just in case it has only finitely many objects up to isomorphism. We say that C is **totally-finite** if it has finitely many objects up to isomorphism, and also finitely many arrows.

Lemma 2. *If Mod(T) is object-finite, then for each model M of T there is a sentence ϕ_M such that $M \vDash \phi_M$, but $N \vDash \neg\phi_M$ for all models N of T that are not isomorphic to M.*

Proof. Let T be a theory. If $\mathrm{Mod}(T)$ is object-finite then the Löwenheim-Skolem theorem implies that every model of T has finite cardinality. From this it follows that there are only finitely many arrows between any pair of objects in $\mathrm{Mod}(T)$. Let M be a model of T. Choose one model N_i from each of the finitely many isomorphism classes of model that are disjoint from the isomorphism class of M. Since for finite structures, elementary equivalence implies isomorphism, it follows that for each N_i, there is a sentence ϕ_i such that $M \vDash \phi_i$ and $N_i \vDash \neg\phi_i$. Now let ϕ_M be the conjunction $\phi_1 \wedge \ldots \wedge \phi_n$. It follows immediately that $M \vDash \phi_M$ but $N_i \vDash \neg\phi_M$ for each i. □

Proposition 9. *Let T_1 be a theory such that Mod(T_1) is object-finite and let $F : T_1 \to T_2$ be a translation. If F is conservative then $F^* : Mod(T_2) \to Mod(T_1)$ is essentially surjective.*

Proof. We prove the contrapositive. If $F^* : \mathrm{Mod}(T_2) \to \mathrm{Mod}(T_1)$ is not essentially surjective, then there is a model M of T_1 that is not isomorphic to any model of the form $F^*(N)$ where N is a model of T_2. By Lemma 2, there is a sentence ϕ_M such that $M \vDash \phi_M$ but $F^*(N) \vDash \neg\phi_M$ for all models N of T_2. Hence $N \vDash F(\neg\phi_M)$ for all models N of T_2, so $T_2 \vDash F(\neg\phi_M)$. But clearly it is not the case that $T_1 \vDash \neg\phi_M$. Therefore F is not conservative. □

Let C and D be categories with respective object sets C_0 and D_0. Let $[C_0]$ and $[D_0]$ be the corresponding sets of equivalence classes of isomorphic objects. Each functor $F : C \to D$ induces a function $F_0 : [C_0] \to [D_0]$, and F_0 is surjective iff F is essentially surjective. If $F : C \to D$ and $G : D \to C$ are both essentially surjective, then the Cantor-Bernstein theorem for finite sets implies that F_0 is a bijection.

Lemma 3. *Let F be a finite set, let $f : F \to \mathbb{N}$ be a function, and let $\phi : F \to F$ be a bijection such that $f(x) \leq f(\phi(x))$ for all $x \in F$. Then $f(x) = f(\phi(x))$ for all $x \in F$.*

Proof. We sketch the proof. The function f corresponds to a fibration of F over $\mathbb{N}$. Since ϕ is a bijection, the size of the fibers remain constant, i.e., $|f^{-1}(n)| = |(f \circ \phi)^{-1}(n)|$. Since ϕ is monotonic, it cannot move an element to a lower fiber.

Thus no element can be moved out of the highest fiber, nor the next highest fiber, etc. □

Proposition 10. *Let C and D be totally-finite categories. If there are faithful, essentially surjective functors $F : C \to D$ and $G : D \to C$, then C and D are equivalent categories. In fact, F itself is one half of an equivalence.*

Proof. By the remark preceding Lemma 3, $F_0 : [C_0] \to [D_0]$ is a bijection. Since F is automatically faithful, it will suffice to show that F is full. For simplicity, we may henceforth replace C and D with the corresponding skeletal categories.

Consider the (finite) set $C_0 \times C_0$ and the function $f : C_0 \times C_0 \to \mathbb{N}$ that assigns the cardinality of the corresponding hom set. That is, $f(a, b) = |\hom(a, b)|$. Let $g : D_0 \times D_0 \to \mathbb{N}$ be the corresponding function for D. Since F is faithful, it induces a bijection $\eta : C_0 \to D_0$ such that

$$f(a, b) \leq g(\eta(a), \eta(b)).$$

And since G is faithful, it induces a bijection $\theta : D_0 \to C_0$ such that

$$g(a, b) \leq f(\theta(a), \theta(b)).$$

If we let $\phi = \theta \circ \eta$ then

$$f(a, b) \leq f(\phi(a), \phi(b)),$$

for all $a, b \in C_0$. By the above lemma, $f(a, b) = f(\phi(a), \phi(b))$, and it follows that F is full. □

We now have the following result. In brief, it says that in order for a pair of theories to be a counterexample to the Cantor-Bernstein property, the theories must have (up to isomorphism) infinitely many models.

Theorem 6. *Let T_1 and T_2 be theories such that Mod(T_1) and Mod(T_2) are object-finite. Then (T_1, T_2) has the Cantor-Bernstein property.*

Proof. If Mod(T) is object-finite then the Löwenheim-Skolem theorem implies that every model of T has finite cardinality. From this it follows that there are only finitely many arrows between any pair of objects in Mod(T). So T is totally-finite. So we know that both of these categories of models for T_1 and T_2 are totally-finite. Let $F : T_1 \to T_2$ and $G : T_2 \to T_1$ be conservative translations. By Proposition 9, F^* and G^* are essentially surjective. They are trivially faithful. Proposition 10 implies that the two categories of models are equivalent, and that F^* and G^* are equivalences. Proposition 8 then implies that F is one half of an intertranslation. □

We now turn to the co-Cantor-Bernstein property. We will show that if either T_1 or T_2 has only finitely many models of some fixed cardinality, then (T_1, T_2) has the co-Cantor-Bernstein property. As a warm-up result, we note that complete theories always have the co-Cantor-Bernstein property.

Proposition 11. *If T_1 or T_2 is complete, then (T_1, T_2) has the co-Cantor-Bernstein property.*

Proof. As we noted above in Example 3, if T_1 is complete, then every translation $F : T_1 \to T_2$ is conservative. So if $F : T_1 \to T_2$ is essentially surjective, then F is one half of an intertranslation. □

For a theory T and a cardinal number κ, let $I(T, \kappa)$ be the number of non-isomorphic models of T of cardinality κ.

Proposition 12. *If $F : T_1 \to T_2$ is essentially surjective, then $I(T_2, \kappa) \leq I(T_1, \kappa)$ then for any cardinal number κ.*

Proof. Recall that the dual functor F^* is always faithful for single-sorted theories. If $F : T_1 \to T_2$ is essentially surjective, then F^* is also full (Halvorson, 2019, Prop 6.6.13). In particular, for any models M, N of T_2, if $F^*(M)$ is isomorphic to $F^*(N)$, then M is isomorphic to N. Now fix a cardinal number κ, and let $[\mathrm{Mod}(T_i)]_\kappa$ be the set of isomorphism classes of models of T_i of cardinality κ. Then F^* induces a one-to-one mapping from $[\mathrm{Mod}(T_2)]_\kappa$ into $[\mathrm{Mod}(T_1)]_\kappa$. This immediately implies that $I(T_2, \kappa) \leq I(T_1, \kappa)$. □

Proposition 13. *Suppose that T_2 has finitely many non-isomorphic models of each cardinality. If $F : T_1 \to T_2$ and $G : T_2 \to T_1$ are essentially surjective, then F^* and G^* are both equivalences of categories.*

Proof. It will suffice to show that F^* is essentially surjective, since F^* is guaranteed to be full because F is essentially surjective (Halvorson, 2019, Prop 6.6.13). By the previous proof, G^* induces an injection of $[\mathrm{Mod}(T_1)]_\kappa$ into $[\mathrm{Mod}(T_2)]_\kappa$. Since the latter is finite, so is the former. Since F^* is an injection of one finite set into a not-larger finite set, it follows that F^* is bijection. Therefore F^* is essentially surjective: every model of T_1 is isomorphic to a model of the form $F^*(N)$ where N is a model of T_2. □

We now have our final result. In brief, it says that in order for a pair of theories to be a counterexample to the co-Cantor-Bernstein property, the theories must have (up to isomorphism) infinitely many models of some fixed cardinality.

Theorem 7. *If (T_1, T_2) does not have the co-Cantor-Bernstein property, then there is a cardinal number κ such that T_1 and T_2 have infinitely many non-isomorphic models of size κ.*

Proof. We prove the contrapositive. Suppose that for every cardinal κ, either T_1 or T_2 has only finitely many non-isomorphic models of size κ. And suppose for contradiction that (T_1, T_2) violates the co-Cantor-Bernstein property. This means that there must be essentially surjective translations $F : T_1 \to T_2$ and $G : T_2 \to T_1$. Note that Proposition 12 now implies that if either T_1 or T_2 has only finitely many models of size κ, then both of them do. So it must be that for every cardinal κ both T_1 and T_2 have only finitely many non-isomorphic

models of size κ. By Proposition 13, F^* and G^* are equivalence of categories. In particular, F^* is essentially surjective, and by Proposition 6.6.17 of (Halvorson, 2019), F is conservative. Therefore, by Proposition 4.5.27 of (Halvorson, 2019), F is one half of an intertranslation. □