THE REVIEW OF SYMBOLIC LOGIC
Volume 13, Number 2, June 2020

ON MORITA EQUIVALENCE AND INTERPRETABILITY

PAUL ANH MCELDOWNEY
Department of Philosophy, University of Notre Dame

Abstract. In a recent article, Barrett & Halvorson (2016) define a notion of equivalence for first-order theories, which they call "Morita equivalence." To argue that Morita equivalence is a reasonable measure of "theoretical equivalence," they make use of the claim that Morita extensions "say no more" than the theories they are extending. The goal of this article is to challenge this central claim by raising objections to their argument for it and by showing why there is good reason to think that the claim itself is false. In light of these criticisms, this article develops a natural way for the advocate of Morita equivalence to respond. I prove that this response makes her criterion a special case of bi-interpretability, an already well-established barometer of theoretical equivalence. I conclude by providing reasons why the advocate of Morita equivalence should opt for a notion of theoretical equivalence that is defined in terms of interpretability rather than Morita extensions.

§1. Introduction. How do we know when two theories are the same? Suppose T_1 and T_2 are sets of sentences sharing the same many-sorted first-order signature.[1] To decide whether T_1 and T_2 are the "same theory," we often compare their deductive closures, i.e., their respective sets of theorems. In keeping with widely used conventions, let's call two theories *logically equivalent* if their deductive closures are the same.

Suppose now that T_1 and T_2 consist of sentences in distinct signatures. Since logically equivalent theories must share the same signature, the notion of logical equivalence will be largely uninformative in this context. As a response to the restrictive nature of logical equivalence, many logicians and philosophers have proposed various formal criteria of equivalence which aim to clarify the idea that theories need not share the same signature in order for them to be equivalent.[2]

One way to approach this issue is to introduce the notion of a "theory extension." If T is a first-order theory, one would like, for a variety of reasons, to extend T to a theory T^+ such that $T \subseteq T^+$ and the signature of T is a subset of the signature of T^+. Let's call such a T^+ a *theory extension* (or more concisely, an *extension*) of T and say that T^+ *extends* T if T^+ is an extension of T.

A strong case can be made that some theory extensions do not "add more theory," i.e., do not add any substantive claims to the theories they are extending. For example, an extension that extends a theory to itself does not "add more theory."

Received: March 08, 2018.
2010 *Mathematics Subject Classification*: 03A10, 03F25, 00A30, 03A05.
Key words and phrases: Morita equivalence, interpretability, many-sorted logic, theoretical equivalence.

[1] A first-order signature gives rise to what is commonly called a "first-order language."
[2] See Visser (2006) and Button & Walsh (2018) for surveys of such attempts.

doi:10.1017/S1755020319000303

For a less trivial example, if T^+ is a definitional extension of T, then it is natural to think that T^+ does not add new substantive claims to T.[3] Let's call theory extensions that do not "add more theory" *admissible*.

In a recent article, Barrett & Halvorson (2016) define a notion of equivalence for first-order theories called "Morita equivalence" which, they define in terms of theory extensions called "Morita extensions." According to Barrett and Halvorson, Morita equivalence provides a plausible signature-independent criterion for theoretical equivalence especially suited for comparing theories with different sorts. As a result, Barrett and Halvorson apply their notion of equivalence to a variety of philosophical problems.[4]

To argue that Morita equivalence is a reasonable measure of "theoretical equivalence," Barrett and Halvorson make use of the claim that Morita extensions "say no more" than the theories they are extending. In this article, I attempt to accomplish two things. First, I challenge this central claim by raising objections to Barrett and Halvorson's argument for it and by showing that there are good reasons to think that the claim itself is false. Second, I demonstrate that a natural way for the advocate of Morita equivalence to address these criticisms will make Morita equivalence a special case of bi-interpretability, an already well-established notion of theoretical equivalence.

In light of this result, I conclude that the reasonableness of Morita equivalence as a notion of theoretical equivalence depends on how much it strengthens bi-interpretability.[5] Moreover, as part of my concluding remarks, I provide reasons why the advocate of Morita equivalence should opt for a notion of theoretical equivalence that is defined in terms of interpretability rather than Morita extensions.[6]

§2. Preliminaries. I begin with technical preliminaries. The setting of this article is classical finitary many-sorted first-order logic.[7] Nothing in this article requires special background in logic: all of the techniques and concepts used in the following proofs are elementary.[8]

2.1. Many-sorted logic. A (many-sorted first-order) signature, $\mathcal{L}$, is given by a quadruple $(\mathcal{S}, \mathcal{R}, \mathcal{F}, \mathcal{C})$, where $\mathcal{S}$ is a set of sort symbols, $\mathcal{R}$ is a set of sorted relation symbols, $\mathcal{F}$ is a set of sorted function symbols, and $\mathcal{C}$ is a set of sorted constant symbols. This means that each relation symbol $R \in \mathcal{R}$ comes equipped with a tuple of sorts symbols $s_1 \times \cdots \times s_n$ from $\mathcal{S}$ which is called the *arity* of R. Analogously, each function symbol $f \in \mathcal{F}$ comes equipped with an arity $s_1 \times \cdots \times s_n \to s_{n+1}$, and each constant symbol $c \in \mathcal{C}$ comes equipped with an arity s_c.

[3] The notion of a definitional extension will be discussed at length in §2.2.

[4] See, for example, Barrett & Halvorson (2017a), where the notion of Morita equivalence is used to address the issue of conventionality in the philosophy of science.

[5] Let $E_1(x, y)$ and $E_2(x, y)$ be notions of equivalence for many-sorted first-order theories. Then E_1 is said to *strengthen* E_2 if whenever A and B are many-sorted first-order theories, $E_1(A, B)$ implies $E_2(A, B)$.

[6] In particular, I will discuss how these reasons motivated Tarski's student Szczerba (1975) to define a notion of theoretical equivalence in terms of interpretability rather than theory extensions.

[7] There are a variety of mathematical contexts where working with a many-sorted logic is more natural and fruitful than a single-sorted one. For example, the model theory of compact complex spaces admits a convenient presentation in many-sorted logic; see Moosa (2005). Moreover, first-order categorical logic finds a natural treatment in the many-sorted context; see Makkai & Reyes (1967).

[8] The following presentation of first-order logic will not be exhaustive. The interested reader is encouraged to refer to Marker (2002) for a proper introduction to first-order logic.

As usual, the set of $\mathcal{L}$-formulas is defined inductively in the usual way using the symbols in $\mathcal{L}$, propositional connectives, sorted first-order quantifiers, sorted variable symbols, and sorted identity symbols (one for every $\mathcal{L}$-sort s). A *many-sorted first-order language* is simply the set of $\mathcal{L}$-formulas for some signature $\mathcal{L}$. For convenience, we shall denote the many-sorted first-order language generated by the signature $\mathcal{L}$ also by $\mathcal{L}$. Intuitively, each identity symbol $=_s$ denotes the identity relation between objects of sort s; we do not allow (as logical constants) identity symbols relating objects of different sorts. When there is no ambiguity, the subscript of an identity symbol $=_s$, which signifies its *arity*, will be omitted. As mentioned, variables and quantifiers are assumed to be sorted as well. This means that a variable or a quantifier binding a variable ranges over precisely one sort. Thus, an existential quantifier will be denoted by $\exists_s$ and a universal quantifier will be denoted by $\forall_s$, where s is a sort symbol of $\mathcal{L}$. The sort of a variable symbol will be called its *arity*. An $\mathcal{L}$-sentence is an $\mathcal{L}$-formula with no free variables. The *arity* of an $\mathcal{L}$-formula $\phi(x_1, \dots, x_n)$ with free variables $x_1, \dots, x_n$ of sort $s_1, \dots, s_n$ is defined to be $s_1 \times \cdots \times s_n$.

Intuitively, an $\mathcal{L}$-structure $\mathcal{M}$ assigns values to the symbols in $\mathcal{L}$. More specifically, an *$\mathcal{L}$-structure* $\mathcal{M}$ is given by the following:

1. For each sort symbol $s \in \mathcal{S}$, a non-empty set M_s, which we call a *sort* of $\mathcal{M}$.
2. For each relation $R \in \mathcal{R}$ of arity $s_1 \times \cdots \times s_n$, $R^{\mathcal{M}} \subseteq M_{s_1} \times \cdots \times M_{s_n}$.
3. For each function $f \in \mathcal{F}$ of arity $s_1 \times \cdots \times s_n \to s_{n+1}$, $f^{\mathcal{M}}$ is a total function from $M_{s_1} \times \cdots \times M_{s_n}$ to $M_{s_{n+1}}$.
4. For each constant $c \in \mathcal{C}$ of arity s_c, $c^{\mathcal{M}}$ is an element of M_{s_c}.

If $\mathcal{M}$ is an $\mathcal{L}$-structure and s is a sort symbol of $\mathcal{L}$, then the identity symbol $=_s$ is always interpreted in $\mathcal{M}$ as equality among objects in M_s. Moreover, finite tuples from $\mathcal{M}$ are denoted by $\bar{a}$, $\bar{b}$, etc. Finite tuples of variables are denoted by $\bar{x}$, $\bar{y}$, etc.

Central to model theory is the notion of being "true in a model." Let $\mathcal{M}$ be an $\mathcal{L}$-structure and $\phi(\bar{x})$ be an $\mathcal{L}$-formula, where $\bar{x}$ is the tuple $(x_1, \dots, x_n)$ and each x_i is a variable of sort s_i of $\mathcal{L}$. If $\bar{a}$ is $(a_1, \dots, a_n)$, where $a_i \in M_{s_i}$, then $\mathcal{M} \models \phi(\bar{a})$ is defined inductively and classically in the usual way (so, in particular, either $\mathcal{M} \models \phi(a_1, \dots a_n)$ or $\mathcal{M} \models \neg\phi(a_1, \dots, a_n)$). When $\mathcal{M} \models \phi(\bar{a})$, one says that $\mathcal{M}$ *satisfies* $\phi(\bar{a})$, or that $\mathcal{M}$ is a *model* of $\phi(\bar{a})$.

Let $\mathcal{M}$ be an $\mathcal{L}$-structure and let $A \subseteq M_{s_1} \times \cdots \times M_{s_n}$, where each M_{s_i} is a sort of $\mathcal{M}$. Then, A is *definable* in $\mathcal{M}$ if there is an $\mathcal{L}$-formula $\phi(\bar{x})$ such that $A = \{\bar{a} \in M_{s_1} \times \cdots \times M_{s_n} : \mathcal{M} \models \phi(\bar{a})\}$. We say that a set A is a *definable set* of $\mathcal{M}$, or $\mathcal{M}$ *defines* A, if A is definable in $\mathcal{M}$.[9] If $\phi(\bar{x})$ is an $\mathcal{L}$-formula of arity $s_1 \times \cdots \times s_n$, then $\phi(\mathcal{M})$ denotes the set of realizations of $\phi(\bar{x})$ in $\mathcal{M}$, i.e., the set $\{\bar{a} \in M_{s_1} \times \cdots \times M_{s_n} : \mathcal{M} \models \phi(\bar{a})\}$.

An *$\mathcal{L}$-theory* T is a set of $\mathcal{L}$-sentences. If T is an $\mathcal{L}$-theory, then T is assumed to be satisfiable, i.e., there is an $\mathcal{L}$-structure $\mathcal{M}$ such that whenever $\phi \in T$, $\mathcal{M} \models \phi$. If T is an $\mathcal{L}$-theory and ϕ is an $\mathcal{L}$-sentence, then we say that T (*logically*) *implies* ϕ (denoted by $T \models \phi$) if every model $\mathcal{M}$ of T is a model of ϕ. If T and T' have the same set of theorems, then we say that T and T' are *logically equivalent* (denoted by $T \equiv T'$). We shall assume that if T is an $\mathcal{L}$-theory, then T is closed under logical implication. Thus, in this article, $T = T'$ if and only if $T \equiv T'$.

2.2. Definitional equivalence. To argue that Morita equivalence is a plausible criterion for theoretical equivalence, Barrett and Halvorson draw analogies between it and

[9] Note that in this article, a definable set is one that is definable *without parameters*.

the notion of definitional equivalence (Barrett & Halvorson, 2017b, p. 13). Thus, before turning to Morita equivalence, I will define and briefly discuss the latter notion.[10] In particular, I will focus on why definitionally equivalent theories might reasonably count as being "theoretically equivalent." In this section, I work only with signatures that share the same sort symbols. So, if $\mathcal{L}$ and $\mathcal{L}'$ are signatures, then $\mathcal{S} = \mathcal{S}'$.

In order to define definitional equivalence, I will need to introduce a few preliminary notions first.

DEFINITION 2.1. *Let $\mathcal{L} \subset \mathcal{L}^+$. If $\mathcal{M}^+$ is an $\mathcal{L}^+$-structure, then ignoring the interpretations of the symbols in $\mathcal{L}^+ - \mathcal{L}$ gives us an $\mathcal{L}$-structure $\mathcal{M}$. The $\mathcal{L}$-structure $\mathcal{M}$ is called a* reduct *of $\mathcal{M}^+$ (more specifically, the* $\mathcal{L}$-reduct of $\mathcal{M}^+$*) and $\mathcal{M}^+$ is called an* expansion *of $\mathcal{M}$ (more specifically, an* $\mathcal{L}^+$-expansion *of $\mathcal{M}$).*

If $\mathcal{L} \subset \mathcal{L}^+$, and if $\mathcal{M}^+$ is an $\mathcal{L}^+$-structure, then we denote the $\mathcal{L}$-reduct of $\mathcal{M}^+$ by $\mathcal{M}$ or $\mathcal{M}^+|_{\mathcal{L}}$.

If $\mathcal{M}^+$ is an expansion of $\mathcal{M}$, then $\mathcal{M}^+$ is called a definitional expansion *of $\mathcal{M}$ if every $\mathcal{L}^+$-definable set of $\mathcal{M}^+$ is $\mathcal{L}$-definable.*

The idea behind a definitional expansion is simple. A definitional expansion $\mathcal{M}^+$ names a definable set A of a given structure $\mathcal{M}$ with some nonlogical symbol not yet in the signature of $\mathcal{M}$. The following syntactic definition allows one to definitionally expand a whole class of structures simultaneously.

DEFINITION 2.2. *Let $\mathcal{L} \subset \mathcal{L}^+$, and let $R \in \mathcal{L}^+$. Then, an* explicit definition of R in terms of $\mathcal{L}$ *is an $\mathcal{L}^+$-sentence of the form*

$$\forall \bar{x}(R(\bar{x}) \leftrightarrow \phi(\bar{x})), \tag{1}$$

where ϕ is an $\mathcal{L}$-formula. Similarly, if f is a function symbol in $\mathcal{L}^+$, then an explicit definition of f in terms of $\mathcal{L}$ *is an $\mathcal{L}^+$-sentence of the form*

$$\forall \bar{x}, y(f(\bar{x}) = y \leftrightarrow \phi(\bar{x}, y)), \tag{2}$$

where $\phi(\bar{x}, y)$ is an $\mathcal{L}$-formula. Lastly, if c is a constant symbol in $\mathcal{L}^+$, then an explicit definition of c in terms of $\mathcal{L}$ *is an $\mathcal{L}^+$-sentence of the form*

$$\forall x(c = x \leftrightarrow \phi(x)), \tag{3}$$

where $\phi(x)$ is an $\mathcal{L}$-formula.

Note that (2) and (3) respectively, imply

$$\forall \bar{x} \exists_{=1} y \phi(\bar{x}, y) \tag{4}$$

and

$$\exists_{=1} x(\phi(x)). \tag{5}$$

These sentences are called the *admissibility conditions* for (2) and (3), respectively.

We are now in a position to define the notion of a definitional extension.

DEFINITION 2.3. *Let $\mathcal{L} \subset \mathcal{L}^+$. Then, a* definitional extension T^+ *of T is an $\mathcal{L}^+$-theory of the form $T \cup \{\delta_\sigma : \sigma \in \mathcal{L}^+\}$ satisfying the following conditions:*

1. *If σ is a nonsortal symbol in $\mathcal{L}^+ - \mathcal{L}$, then δ_σ is an explicit definition of σ in terms of $\mathcal{L}$; and*

[10] My presentation, as does Barrett and Halvorson's, follows Hodges (1993, sec. 2.6).

https://doi.org/10.1017/S1755020319000303 Published online by Cambridge University Press

2. *if σ is a constant or function symbol in $\mathcal{L}^+ - \mathcal{L}$ and χ is an admissibility condition for δ_σ, then $T \models \chi$.*

Essentially, definitional extensions of a theory simply add explicit definitions to theories that already imply their associated admissibility conditions. The following elementary example will help clarify the notion.

EXAMPLE 2.4. *Let $\mathcal{L}$ be the language of arithmetic $\{+, \times, s, 1\}$, let T be* PA*, i.e., first-order Peano Arithmetic, and let $\phi(x)$ be the $\mathcal{L}$-formula expressing that x is a prime number. An example of a definitional extension of T is the $\mathcal{L}^+$-theory T^+, which adds to T an $\mathcal{L}^+$-sentence δ_P of the form* (1) *with $\phi(x)$ in the relevant position. Intuitively, δ_P defines a new one-place predicate symbol P that picks out the elements satisfying $P(x)$, i.e., all of the prime numbers.*

One might ask whether definitional extensions add any substantial claims to the theories they are extending, i.e., whether they are admissible extensions. As I will discuss, since definitional extensions satisfy the following conditions, the view that definitional extensions are admissible is well founded.

DEFINITION 2.5. *Let T^+ be an extension of T. Then T^+ is* conservative *over T if whenever ϕ is an $\mathcal{L}$-sentence, $T^+ \models \phi$ if and only if $T \models \phi$.*

Similarly, T^+ is eliminable *over T if whenever $\phi(\bar{x})$ is an $\mathcal{L}^+$-formula, there is an $\mathcal{L}$-formula $\psi(\bar{x})$ such that*

$$T^+ \models \forall \bar{x}(\phi(\bar{x}) \leftrightarrow \psi(\bar{x})) \tag{6}$$

Given this, the following proposition substantiates the claim that definitional extensions do not "add more theory" to the theories they are extending:

PROPOSITION 2.6. *If T^+ is a definitional extension of T, then the following are true.*[11]

1. *Every model $\mathcal{M}$ of T has a unique definitional expansion $\mathcal{M}^+$ that is a model of T^+.*
2. *Every model of T^+ is a definitional expansion of some model of T.*
3. *T^+ is conservative over T.*
4. *T^+ is eliminable over T.*

As a consequence of Proposition 2.6, there is good reason to think that T^+ does not add substantive truths to T. In particular, this proposition shows that explicit definitions are in line with a tradition of thinking about the logic of definitions on which definitions for new terms (*i*) can only define expressions that are substitutable with expressions in the original language; and (*ii*) cannot prove new substantive truths concerning expressions in the original language.[12]

By eliminability, explicit definitions can only define expressions that can be substituted for expressions in the original signature, thereby satisfying condition (*i*). By conservativity, explicit definitions (in the sense of Definition 2.2) do not prove anything about terms in the original signature that is not already provable in the original theory, thereby satisfying condition (*ii*). According to the traditional view of definitions then, explicit definitions do not "add more theory." Thus, definitional extensions do not either.

[11] For a proof of this result, see Hodges (1993, pp. 58–62).

[12] See, in particular, Hodges (2008, p. 104) and Suppes (1957, p. 153) for more on the traditional view regarding definitions.

For those who prefer to think semantically, there is another way of convincing oneself that definitional extensions do not "add more theory." Given any model $\mathcal{M}^+$ of T^+, note that any definable set in $\mathcal{M}^+$ is already definable in $\mathcal{M}^+$ by an $\mathcal{L}$-formula. This follows from eliminability. Since every $\mathcal{M}^+$ is a definitional expansion of an $\mathcal{L}$-structure $\mathcal{M}$ satisfying T, it follows that A is already definable in $\mathcal{M}$. Thus, if the "semantic content" of a theory is given by the definable sets in its models, then definitional extensions do not increase semantic content. And if "theoretical content" is understood in terms of semantic content, then it follows that definitional extensions do not add "further theory" to the theories they are extending.

The notion of a definitional extension induces a natural notion of equivalence for first-order theories:

DEFINITION 2.7. *Let T and T' be first-order theories. Then, T and T' are* definitionally equivalent *if there exists a common definitional extension T^+ of T and T', i.e., if there exists a theory T^+ which is a definitional extension of both T and T'.*

A strong case can be made that definitionally equivalent theories are theoretically equivalent. Suppose T_1 and T_2 are definitionally equivalent theories. Then, we can definitionally extend them to get identical theories. But since we have good reason to think that definitional extensions do not "add more theory," it follows that we have good reason to think that the common definitional extension of T_1 and T_2 does not *add more theory* to either of them. Thus, if T_1 and T_2 are definitionally equivalent, then we are justified in regarding them as the same theory.[13]

§3. Morita.

Given a first-order $\mathcal{L}$-theory T, a *Morita extension* T^+ generalizes the notion of a definitional extension of T by allowing T^+ to add new *sort* symbols to $\mathcal{L}$ not already in $\mathcal{L}$.[14] As we have seen, definitional extensions add explicit definitions only for new relation, function and constant symbols. They do not extend the original signature by adding new sort symbols. This implies that two theories with different sort symbols cannot be definitionally equivalent. Motivating Barrett and Halvorson's notion of Morita equivalence is the plausible idea that two first-order theories may be equivalent even if they do not share the same sort symbols.

3.1. Morita extension. Let $\mathcal{L}$ be a first-order many-sorted signature. Barrett and Halvorson consider four different ways of adding new sort symbols to an $\mathcal{L}$-theory T.

3.1.1. Product. Let $\mathcal{L} \subset \mathcal{L}^+$ such that s is a sort symbol in $\mathcal{L}^+ - \mathcal{L}$, s_1 and s_2 are sort symbols in $\mathcal{L}$, and π_1 and π_2 are function symbols in $\mathcal{L}^+ - \mathcal{L}$ of arity $s \to s_1$ and $s \to s_2$, respectively. Then, an explicit definition of s, π_1 and π_2 as a *product sort* in terms of $\mathcal{L}$ is an $\mathcal{L}^+$-sentence of the form

$$\forall_{s_1} x \forall_{s_2} y \exists_{s=1} z[\pi_1(z) = x \wedge \pi_2(z) = y] \tag{7}$$

3.1.2. Coproduct. Let $\mathcal{L} \subset \mathcal{L}^+$ such that s is a sort in $\mathcal{L}^+ - \mathcal{L}$, s_1 and s_2 are sort symbols in $\mathcal{L}$, and ρ_1 and ρ_2 are function symbols in $\mathcal{L}^+ - \mathcal{L}$ of arities $s_1 \to s$ and $s_2 \to s$,

[13] For more on the philosophical significance of definitional equivalence, see, for example, Glymour (1971), which proposes definitional equivalence as a measure of theoretical equivalence in the context of scientific theories.

[14] As Barrett and Halvorson note, the name "Morita equivalence" originally comes from ring theory. Two rings R and S are *Morita equivalent* if there is an equivalence of categories between the category $\mathsf{R}\text{-}\mathsf{Mod}$ of (left) modules over R and the category $\mathsf{S}\text{-}\mathsf{Mod}$ of (left) modules over S.

respectively. Then, an explicit definition of s, ρ_1, and ρ_2 as a *coproduct* (or *disjoint union sort*) in terms of $\mathcal{L}$ is an $\mathcal{L}^+$-sentence of the form

$$\forall_s z[\exists_{s_1=1} x(\rho_1(x) = z) \vee \exists_{s_2=1} y(\rho_2(y) = z)] \wedge \forall_{s_1} x \forall_{s_2} y[\rho_1(x) \neq \rho_2(y)]. \tag{8}$$

3.1.3. Subsort. Let $\mathcal{L} \subset \mathcal{L}^+$ such that s is a sort symbol in $\mathcal{L}^+ - \mathcal{L}$, s_0 is a sort symbol in $\mathcal{L}$, i is a function symbol in $\mathcal{L}^+ - \mathcal{L}$ of arity $s \to s_0$. Then, an explicit definition of s and i as a *subsort* in terms of $\mathcal{L}$ is an $\mathcal{L}^+$-sentence of the form

$$\forall_{s_0} x[\phi(x) \leftrightarrow \exists_s z(i(z) = x)] \wedge \forall_s z_1 \forall_s z_2[i(z_1) = i(z_2) \to z_1 = z_2], \tag{9}$$

where $\phi(x)$ is an $\mathcal{L}$-formula. Note that (9) implies the $\mathcal{L}$-sentence $\exists_{s_0} x[\phi(x)]$, which is called the *admissibility condition* for (9).

3.1.4. Quotient. Let $\mathcal{L} \subset \mathcal{L}^+$ such that s is a sort symbol in $\mathcal{L}^+ - \mathcal{L}$, s_1 is a sort symbol in $\mathcal{L}$, and μ is a function in $\mathcal{L}^+$ of arity $s_1 \to s$. Then, an explicit definition of s and μ as a *quotient sort* in terms of $\mathcal{L}$ is an $\mathcal{L}^+$-sentence of the form

$$\forall_{s_1} x_1 \forall_{s_1} x_2[\mu(x_1) = \mu(x_2) \leftrightarrow \phi(x_1, x_2)] \wedge \forall_s z \exists_{s_1} x[\mu(x) = z], \tag{10}$$

where $\phi(x_1, x_2)$ is an $\mathcal{L}$-formula. Note that (10) implies the following $\mathcal{L}$-sentences:

$$\forall_{s_1} x(\phi(x, x))$$
$$\forall_{s_1} x_1 \forall_{s_1} x_2[\phi(x_1, x_2) \to \phi(x_2, x_1)]$$
$$\forall_{s_1} x_1 \forall_{s_1} x_2 \forall_{s_1} x_3[(\phi(x_1, x_2) \wedge \phi(x_2, x_3)) \to \phi(x_1, x_3)]$$

These $\mathcal{L}$-sentences are called the *admissibility conditions* for (10).

Using the above, the definition of a Morita extension is given as follows.

DEFINITION 3.1. *Let T be an $\mathcal{L}$-theory. Then, a* Morita extension *T^+ of T to $\mathcal{L}^+$ is a an $\mathcal{L}^+$-theory of the form $T \cup \{\delta_\sigma : \sigma \in \mathcal{L}^+ - \mathcal{L}\}$ satisfying the following conditions*:

1. *For each symbol $\sigma \in \mathcal{L}^+ - \mathcal{L}$, the sentence δ_σ is an explicit definition of σ in terms of $\mathcal{L}$.*[15]
2. *If α_σ is an admissibility condition for a definition δ_σ, then $T \models \alpha_\sigma$.*
3. *For each sort symbol $s \in \mathcal{L}^+ - \mathcal{L}$ and function symbol $f \in \mathcal{L}^+ - \mathcal{L}$, if f appears in the sort definition $\delta_s \in T^+$, then $\delta_f = \delta_s$.*

Morita extensions generalize the notion of a definitional extension by allowing one to add new sort symbols to a given theory. The following example will help clarify the notion of a Morita extension.

EXAMPLE 3.2. *Let $\mathcal{L}$ be the language of groups $\{\cdot, e\}$ and let T be the theory of groups. Moreover, let $Z(x)$ be the $\mathcal{L}$-formula $\forall y(x \cdot y = y \cdot x)$. The formula $Z(x)$ picks out what's called the "center" of a group. Let $E(x, y)$ be the formula $Z(x \cdot y^{-1})$. One can verify that for any group G, $E(x, y)$ defines an equivalence relation on G.*

An example of a Morita extension of T is the $\mathcal{L}^+$-theory T^+, which adds to T a sentence δ_s of the form (10) with $E(x, y)$ in the relevant position. Intuitively, δ_s adds a new quotient sort s to models of T. More specifically, when interpreted by a group G, s will denote the

15 Note that σ may be a relation, constant, or function symbol. Thus, δ_σ may be an explicit definition for σ in the sense of Definition 2.2 (i.e., in the sense of a definitional extension).

underlying set of the quotient group $G/Z(G)$, which is isomorphic to the group of inner automorphisms of G.[16]

3.2. Morita equivalence. Using the notion of a Morita extension, Morita equivalence is defined accordingly.

DEFINITION 3.3. *Let T and T' be many-sorted first-order theories. The theories T and T' are* Morita equivalent *if there are finite sequences of Morita extensions $T, T_1, \ldots, T_n$ and $T', T'_1, \ldots, T'_m$ such that $T_n \equiv T'_m$.*[17]

According to Barrett and Halvorson, what makes Morita equivalence a particularly attractive notion of equivalence is that it can relate theories equipped with different sorts, which one is unable to do with the notion of definitional equivalence.[18] Moreover, they argue that, like the notion of definitional equivalence, Morita equivalence captures the intuitive idea that two theories are the same if "they each can, in compatible ways, define all the vocabulary that the other uses" (Barrett & Halvorson, 2017b, p. 3568). However, whether Morita equivalence captures this intuition will depend on whether Morita extensions add substantial claims to the theories they extend, i.e., whether they in fact, just "add 'abbreviations' of old statements" (Barrett & Halvorson, 2017b, p. 3568).

In the next section, I will turn to Barrett and Halvorson's argument that Morita extensions are admissible over the theories they extend. I will then discuss why their argument is unsatisfactory and show that there is good reason to think that Morita extensions sometimes do add substantial claims.

§4. Criticisms.

Consider the following view.

> M-EQ: Morita equivalent theories are theoretically equivalent.

As I have briefly discussed, to argue for this view, Barrett and Halvorson make use of the following claim.

> M-EXT: If T is an $\mathcal{L}$-theory and T^+ is a Morita extension of T, then T^+ "says no more" than T.

In other words, M-EXT asserts that T^+ does not add any substantial claims to T.

Barrett and Halvorson attempt to substantiate M-EXT by proving three theorems that purport to provide a precise sense in which M-EXT is true. First, they prove that if T is an $\mathcal{L}$-theory and T^+ is a Morita extension of T, then every model of T extends uniquely to a model of T^+. Second, they show that T^+ is a conservative extension of T in the same sense

[16] An *inner automorphism of* G is an automorphism $\sigma_g : G \to G$ of the form $\sigma_g(a) = g^{-1} \cdot a \cdot g$ where $g \in G$.

[17] One might wonder why Morita equivalence is defined in terms of *finite sequences* of Morita extensions. Why not just say that two theories T and T' are Morita equivalent if there are logically equivalent theories T^+ and T'^+ such that T^+ is a Morita extension of T and T'^+ is a Morita extension of T'? One reason is that this proposed definition is less general. To see this, note that sequences of Morita extensions are not transitive, i.e., if T_3 is a Morita extension of T_2, and if T_2 is a Morita extension of T_1, then T_3 is not necessarily a Morita extension of T_1. Thus, sequences of Morita extensions of arbitrary finite length are not reducible to sequences of length 1. I will discuss this failure of transitivity in greater detail later in §6.1.

[18] See Example 4.8 of Barrett & Halvorson (2016) for an example of a pair of many-sorted theories which are Morita equivalent, but not definitionally equivalent. This example is discussed further in §5.1.

found in Definition 2.5. Third, they show that if ϕ is an $\mathcal{L}^+$-formula, then for any "code" ξ, which is an $\mathcal{L}^+$-formula relating sorts in $\mathcal{L}^+ - \mathcal{L}$ with sorts in $\mathcal{L}$, there is an $\mathcal{L}$-formula ϕ^* such that ϕ is equivalent to ϕ^* modulo T^+ and ξ.[19] Let's call this last theorem the *translation theorem*.

Since I will be referring to the translation theorem throughout my evaluation of Barrett and Halvorson's argument, it will be useful to state it here. As a note, it won't be important to know the precise definition of Halvorson and Barrett's definition of a "code" in order to understand the basic gist of the theorem.[20]

THEOREM 4.1 (Translation theorem). *Let T^+ be a Morita extension of T, and let $\phi(x_1, \ldots, x_n; y_1, \ldots, y_m)$ be any $\mathcal{L}^+$-formula. Then given any code $\xi(x_1, \ldots, y_n^2)$ for variables $x_1, \ldots, x_n$, there exists an $\mathcal{L}$-formula $\phi^*(y_1, \ldots, y_m; y_1^1, \ldots, y_n^2)$ such that*

$$T^+ \models \forall_{s_1} x_1 \cdots \forall_{s_n} x_n \forall_{t_1} y_1 \cdots \forall_{t_m} y_m \forall_{s_1^1} y_1^1 \cdots \forall_{s_n^2} y_n^2 [\xi(x_1, \ldots, y_n^2) \\ \rightarrow (\phi(x_1, \ldots, x_n; y_1, \ldots, y_m) \leftrightarrow \phi^*(y_1, \ldots, y_m; y_1^1, y_1^2, \ldots, y_n^1, y_n^2))]. \quad (11)$$

Let's call a theory extension T^+ of T satisfying this kind of translation theorem *eliminable** over T. To better understand the content of Theorem 4.11, it will be helpful to consider the special case where $n = 1$ and $m = 1$.

COROLLARY 4.2 (Special case of Theorem 4.11). *Let T^+ be a Morita extension of T, and let $\phi(x; y)$ be an $\mathcal{L}^+$-formula. Then for every code $\xi(x, z_1, z_2)$ for x, there is an $\mathcal{L}$-formula $\phi^*(y; z_1, z_2)$ such that*

$$T^+ \models \forall_s x \forall_t y \forall_{s_1} z_1 \forall_{s_2} z_2 [\xi(x, z_1, z_2) \rightarrow (\phi(x; y) \leftrightarrow \phi^*(y; z_1, z_2))]. \quad (12)$$

According to Corollary 4.2, given any code $\xi(x, z_1, z_2)$, which relates the $\mathcal{L}^+$-variable x with the $\mathcal{L}$-variables z_1 and z_2, it follows that any $\mathcal{L}^+$-formula $\phi(x; y)$ is equivalent to some $\mathcal{L}$-formula $\phi^*(y, z_1, z_2)$.

4.1. The argument. Why should we think that by satisfying conservativity and eliminability*, Morita extensions "say no more" than the theories they are extending? (Barrett & Halvorson, 2016, p. 565) A charitable way of interpreting Barrett and Halvorson's defense of M-EXT can be given as follows:

P1. Morita extensions are conservative and eliminable* over the theories they extend.

P2. If an extension T^+ is conservative and eliminable* over T, then T^+ merely adds in abbreviations to T.

P3. If an extension T^+ merely adds in abbreviations to T, then T^+ does not "add theory" to T, i.e., T^+ does not add substantial truths to T.

C4. Morita extensions do not "add theory" to the theories they are extending, i.e., M-EXT.

Together, the statements P1–C4 form a valid argument. Barrett and Halvorson provide a mathematical proof for the first premise P1. The third premise P3 follows from a traditional view regarding the logic of definitions (see discussion on p. 392).

This leaves the second premise P2, which will be the focus of my criticisms. According to Barrett and Halvorson, the plausibility of P2 rests on an analogy between

[19] These theorems are Theorems 4.2, 4.4, and 4.6 of Barrett & Halvorson (2016).

[20] For a fuller discussion, the interested reader is encouraged to refer to Appendix 7.1.

eliminability* and eliminability. In my evaluation of P2, not only will I argue that this analogy is tenuous but also I will provide reasons to think that P2 (and C4 as well) is actually false.

4.2. Against eliminability*. One of the main issues regarding P2 is that it is insufficiently explained and argued for. In fact, Barrett and Halvorson only briefly allude to a defense of P2. According to them, if conservative and eliminable theory extensions merely add abbreviations to the theories they are extending, and if eliminability* is similar (enough) to eliminability, then the same is true of extensions satisfying conservativity and eliminablility*. Whether this inference goes through, however, depends on whether eliminability* is, in fact, sufficiently similar to eliminability. Clearly, a further argument is needed since the correspondence described in the translation theorem is importantly different in logical form from the correspondence described by the usual notion of eliminability for definitional extensions. In particular, the translation theorem features a correspondence that is conditionalized by an $\mathcal{L}^+$-formula, i.e., what Barrett and Halvorson call a "code." Given the conditional nature of the translation theorem, it is not immediately clear how $\mathcal{L}^+$-expressions can be genuinely replaced with $\mathcal{L}$-expressions. Thus, it is unclear why one should take eliminability* to share any of the philosophical consequences that eliminability enjoys.

In addition to being insufficiently argued for, there is good reason to think that P2 and in fact C4 (i.e., M-EXT) are simply false. To see why Morita extensions do not merely add abbreviations, it will be useful to recall the ways in which eliminable and conservative extensions only add abbreviations to the theories they are extending. Recall that if T^+ is conservative and eliminable over T, then T^+ does not add nondefinable sets to models of T. In other words, any expression in $\mathcal{L}^+ - \mathcal{L}$ can be replaced by an expression in $\mathcal{L}$ that is, according to T^+, equivalent to expressions in $\mathcal{L}^+ - \mathcal{L}$. In contrast, if T^+ *were* to add non-definable sets to the models of T, then T^+ would not just add abbreviations to T, but would seem to add genuinely "new theory." In particular, if each model $\mathcal{M}^+$ of T^+ definably asserts the existence of a set X (say, by using the $\mathcal{L}^+$-formula $\phi(x)$) that cannot be defined in any model of T by a first-order formula in $\mathcal{L}$, then it would seem that $\phi(x)$ is an $\mathcal{L}^+$-formula that does *not* abbreviate anything in T.

An example will help to illustrate my point. Let $\mathcal{L}$ denote the signature of rings, and let T be the complete $\mathcal{L}$-theory of the complex field $\mathbb{C}$.[21] Moreover, let $\mathcal{L}^+$ denote the signature $\mathcal{L} \cup \{i\}$, where i is a constant symbol, and let T^+ denote the theory $T \cup \{i^2 = -1\}$. However, recall that i and $-i$ (in any model of T^+) are indiscernible in $\mathcal{L}$; they satisfy the same $\mathcal{L}$-formulas. Since no model $\mathcal{M}$ of T can define $i^{\mathcal{M}^+}$, we are inclined to say that T^+ adds more theory to T.

In light of these considerations, we may formulate the following definability constraint, which explains why examples like this one seem to involve some addition of "new theory:"

DC: If σ is a symbol in $\mathcal{L}^+$, then for any model $\mathcal{M}^+$ of T^+, there is an $\mathcal{L}^+$-structure $\mathcal{N}^+$ isomorphic to $\mathcal{M}^+$ and a model $\mathcal{M}'$ of T such that $\sigma^{\mathcal{N}^+}$ is definable in $\mathcal{M}'$, i.e., there is an $\mathcal{L}$-formula $\phi(\bar{x})$ such that, whenever $\bar{a} \in \mathcal{N}^+$, $\mathcal{M}' \models \phi(\bar{a})$ if and only if $\bar{a} \in \sigma^{\mathcal{N}^+}$.

[21] T is also known as the theory ACF_0, the theory of algebraically closed fields with characteristic 0.

https://doi.org/10.1017/S1755020319000303 Published online by Cambridge University Press

One might wonder what role $\mathcal{N}^+$ is playing in DC, and whether this role can be performed by $\mathcal{M}^+$. The following example will help illustrate why $\mathcal{N}^+$ is necessary in formulating a useful definability condition in the many-sorted context. Let T be the empty theory with one sort s. Suppose T^+ is the Morita extension of T adding a product sort s^+ for s. Note that the set of natural numbers $\mathbb{N}$ (with no further structure) is a model of T. Let N_1^+ be a model of T^+ which extends $\mathbb{N}$ and which interprets s^+ as $\mathbb{N} \times \mathbb{N}$. Similarly, let $\mathcal{N}_2^+$ be the extension of $\mathbb{N}$ which interprets s^+ as $\mathbb{Z}$. Observe that N_1^+ and N_2^+ are isomorphic (as $\mathcal{L}^+$-structures) and that both are models of T^+. However, while $\mathbb{N} \times \mathbb{N}$ is definable in $\mathbb{N}$, there is no sense in which $\mathbb{Z}$ is definable in $\mathbb{N}$ since $\mathbb{Z}$ is not a subset of $\mathbb{N}^n$ for any n. This example shows how theory extensions adding new sorts can only determine their model extensions *up to isomorphism* at best, and thus, if we are interested in definability in theory extensions that add new sorts, then we'll have to make reference to a model extension $\mathcal{N}^+$ whose new sorts are actually subsets of the original sorts.

In short, DC merely states that the interpretation of any new symbol added while extending T to T^+ is already definable in some model of T. What's problematic about extending the theory T of the complex field to the theory $T \cup \{i^2 = -1\}$ is that it specifies a way of extending theories that violate DC.

In practice, it is common for logicians to understand definability in a more general sense, where a structure's interpretation of a symbol is definable in some quotient object of some model of the original theory.[22] Thus, one may state the following modest generalization of DC:

DC*: If σ is a symbol in $\mathcal{L}^+ - \mathcal{L}$, then for any model $\mathcal{M}^+$ of T^+, there is a model $\mathcal{N}^+$ isomorphic to $\mathcal{M}^+$ and a model $\mathcal{M}'$ of T such that $\sigma^{\mathcal{N}^+}$ is definable in some definable quotient structure of $\mathcal{M}'$.

Intuitively, DC* generalizes DC by allowing the interpretation of new symbols to be definable in definable quotient structures of models of T. Satisfying DC* ensures that T^+ does not add any new structure to the models of T that isn't, in some way or another, already definable in some model of T.

I contend that DC* is a necessary condition for theory extensions to count as "not adding theory." More concisely, I hold the following position:

> If a theory extension T^+ of T does not add substantial claims to T, then T^+ satisfies DC*.

This position strikes me as being difficult to deny for reasons that have already been discussed. If T^+ is an extension that adds non-definable sets to models of T, then T^+ seems to be adding genuinely new theory to T. If adding nondefinable sets to models does not count as an example of "adding more theory," then I am not sure what would count.

If this is right, then the advocate of M-EQ is in trouble. As I will show, there are examples of Morita extensions that fail to satisfy DC* (and thus, DC as well). In particular, if T is an $\mathcal{L}$-theory and T^+ is a Morita extension which adds a coproduct sort to T, then T^+ does not, in general, satisfy DC*.

To see this, consider the following example. Let $\mathcal{L}$ have one sort s, and no relation, function, or constant symbols. Let T be the $\mathcal{L}$-theory consisting of the single axiom $\exists_s x \forall_s y(y = x)$, i.e., T states there is exactly one object. Let T^+ denote the Morita extension

[22] The notion of a definable quotient structure will be discussed further in §5.1.

which adds a new sort, s', for the disjoint union (i.e., coproduct) of s with itself. Then, T^+ implies that there are *two* s'-objects, i.e., $T^+ \models \exists_{s'} x \exists_{s'} y(x \neq y)$.

To see that T^+ does not satisfy DC*, note that the disjoint union of s with itself is not definable in a first-order way; it is not even definable in a definable quotient structure of some model of T. This is because any definable set of a one element structure is either empty or has cardinality one. However, every model $\mathcal{M}^+$ of T^+ defines a two element set, namely, $s'^{\mathcal{M}^+}$. Thus, $s'^{\mathcal{M}^+}$ is not definable in any definable quotient structure of the $\mathcal{L}$-reduct $\mathcal{M}$ of $\mathcal{M}^+$. In other words, T^+ does not satisfy DC*.

This gives us a precise sense in which Morita extensions do not merely "add in 'abbreviations' of old statements into the theory T" (Barrett & Halvorson, 2017b, p. 3568) as Barrett and Halvorson claim. And if the only alternative to adding abbreviations is for Morita extensions to add substantial truths to the theories they are extending, then we have strong reason to think that C4, i.e., M-EXT, is, in fact, false.[23]

4.3. Fixing Morita. In light of the issues facing M-EQ, there are ways for its advocate to respond. One way is to argue against the view that theory extensions have to satisfy DC* in order to count as not "adding more theory." In going this route, the advocate of M-EQ would have to provide an account of what makes P2 plausible and, in particular, what distinguishes Morita extensions from various other ways of extending theories that involve adding arbitrary or otherwise problematic non-definable sets as new sorts.

One tempting way to proceed along these lines is to propose eliminability* as a dividing line that separates Morita extensions from various "bad company" alternatives. Unfortunately, as the next example shows, there are theory extensions that satisfy eliminability*, yet seem to add a significant amount of theory.

Let T be an $\mathcal{L}$-theory where $\mathcal{L}$ is single-sorted but has no relation, function, or constant symbols. Let $\mathcal{L}^+$ be the signature $\mathcal{L} \cup \{s_\omega\} \cup \{\rho_i : i < \omega\}$, where s_ω is a sort symbol not appearing in $\mathcal{L}$, and each ρ_i is a function symbol of arity $s \to s_\omega$ for the same fixed sort $s \in \mathcal{L}$. Moreover, let $\phi_{i,j}$ be the $\mathcal{L}^+$-sentence

$$\forall_s x \forall_s y[\rho_i(x) \neq \rho_j(y)] \tag{13}$$

and let ψ_i be the $\mathcal{L}^+$-sentence stating that ρ_i is injective. Define T^ω to be the $\mathcal{L}^+$-theory $T \cup \{\phi_{i,j} : i, j < \omega \text{ and } i \neq j\} \cup \{\psi_i : i < \omega\}$.

Intuitively, T^ω extends T by defining a new sort s_ω that contains the disjoint union of s with itself ω-many times.[24] Clearly, the interpretation of s_ω in models of T^ω cannot be defined in models of T unless one allows for infinitary formulas. Since no interpretation of s_ω is definable in a model of T or even in a definable quotient structure of a model of T it follows that T^ω violates DC*. In addition to violating DC*, there are other reasons to think that adding the $\mathcal{L}^+$-sentences $\phi_{i,j}$ to T constitutes a substantial increase in "theory." For example, if T is axiomatized by the sentence stating that there is one thing, its extension T^ω implies that there is a sort with infinitely many things. Intuitively, this seems like a great increase in "theory."

One can straightforwardly show that T^ω is eliminable* and conservative over T by formulating a notion of a "code" and proving a translation-type theorem analogous to the

[23] In other words, if Morita extensions do not merely add in abbreviations, and moreover, if the *converse* of P3 is also true, then the negation of C4 validly follows. However, one might ask: why think that the converse of P3 is true? The answer is that both P3 and its converse follow from the traditional view regarding definitions (see p. 392).

[24] Intuitively, the union of the images of the ρ_i is the disjoint union of s with itself ω-many times.

https://doi.org/10.1017/S1755020319000303 Published online by Cambridge University Press

one given in Theorem 4.11 (see Appendix 7.2 for a full detailed proof). Examples like T^{ϖ} provide insight into how badly eliminability* can fail to control the amount of "new theory" that gets added to a given first-order theory. Since the construction of T^{ϖ} generalizes to any arbitrary infinite cardinal, it is tempting to conclude that this failure might be limitless.[25]

It seems, then, that the advocate of M-EQ who wants to push back against DC* cannot appeal to eliminability* as a dividing line between Morita extensions and their bad company extensions. Neither can she appeal to the fact that T^{ϖ} adds infinitely many new statements to T, since Morita extensions themselves often result from adding infinitely many new statements to an original theory.

Rather than pushing back against DC*, I suggest that a more natural way of responding is to modify the definition of a Morita extension so that it satisfies DC* in full generality. This can be done (*i*) by restricting our attention to theories that imply the existence of (at least) two objects within the same sort (see Corollary 5.8); or (*ii*) by giving up on disjoint union (i.e., coproduct) as a kind of Morita extension (see Corollary 5.9). However, with either modification, as I will show in the next section, Morita equivalence will turn out to be a special case of the model-theoretic notion of "bi-interpretability," a widely used measure of theoretical equivalence.

§5. Morita and bi-interpretability. In the previous section, I enumerated two ways that an advocate of M-EQ can modify the definition of a Morita extension so that it satisfies DC*. The main goal of this section is to prove that either modification makes Morita equivalence a strengthening of bi-interpretability, i.e., Morita equivalence implies bi-interpretability. In light of this, I conclude that the tenability of Morita equivalence as a measure of theoretical equivalence depends on how much it strengthens the notion of bi-interpretability.

Before sketching out the proofs of these results, it is important to define the notion of bi-interpretability.

5.1. Interpretability. Constructing quotient objects is ubiquitous in mathematics: the complex numbers are often represented as pairs of reals; the real projective plane is often represented as a quotient structure of $\mathbb{R}^3$; a field extension $F[\alpha]$ of F can be represented as quotient structure of the polynomial ring $F[x]$. *Interpretations* are ways of generalizing these kinds of constructions while maintaining a concern for first-order definability. The following presentation of interpretations is inspired by Button & Walsh (2018) and Hodges (1993).

DEFINITION 5.1. *Let $\mathcal{M}$ be an $\mathcal{L}$-structure and $\mathcal{N}$ be an $\mathcal{L}'$-structure. We say that $\mathcal{M}$ is* a (definable) quotient structure *of $\mathcal{N}$ if the following conditions hold*:

1. *Each sort M_s of $\mathcal{M}$ has the form $X_s/E_s = \{\bar{a}/E_s : \bar{a} \in X_s\}$ where X_s is a definable set in $\mathcal{N}$ and E_s is a definable (in $\mathcal{N}$) equivalence relation on X_s.*[26]
2. *For every constant symbol c of arity s, relation symbol R of arity $s_1 \times \cdots \times s_n$, and function symbol f of arity $s_1 \times \cdots \times s_n \to s_{n+1}$ of $\mathcal{L}$, the following sets are definable in $\mathcal{N}$:*

[25] Observe that the construction of T^{ϖ} generalizes to any many-sorted first-order language $\mathcal{L}$. Moreover, the appropriate analogue of Theorem 4.11 should hold for any such generalization of T^{ϖ}.

[26] Recall that, in this article, a definable set is a set that is definable *without parameters*.

https://doi.org/10.1017/S1755020319000303 Published online by Cambridge University Press

$$(c^{\mathcal{M}})^{-1} \equiv_{\text{def}} \{\bar{a} \in X_s : \mathcal{M} \models c = \bar{a}/E_s\}$$
$$(R^{\mathcal{M}})^{-1} \equiv_{\text{def}} \{(\bar{a}_1, \dots, \bar{a}_n) \in X_{s_1} \times \cdots \times X_{s_n} : \mathcal{M} \models R(\bar{a}_1/E_{s_1}, \dots, \bar{a}_n/E_{s_n})\}$$
$$(f^{\mathcal{M}})^{-1} \equiv_{\text{def}} \{(\bar{a}_1, \dots, \bar{a}_{n+1}) \in X_{s_1} \times \cdots \times X_{s_{n+1}} : \mathcal{M} \models f(\bar{a}_1/E_{s_1}, \dots, \bar{a}_n/E_{s_n}) = \bar{a}_{n+1}/E_{s_{n+1}}\}.$$

We say that $\mathcal{M}$ is interpretable *in $\mathcal{N}$ if $\mathcal{M}$ is isomorphic to a quotient structure of $\mathcal{N}$. We say that $\mathcal{M}$ and $\mathcal{N}$ are* mutually interpretable *if each is interpretable in the other.*

In short, $\mathcal{N}$ interprets $\mathcal{M}$ if the sorts, relations, functions, and constants of $\mathcal{M}$ are all definable in some quotient structure of $\mathcal{N}$.

Given the duality between definable sets and their defining formulas, one can show that $\mathcal{M}$ is interpretable in $\mathcal{N}$ if and only if there is an *interpretation* Γ of $\mathcal{M}$ in $\mathcal{N}$ in the following sense:[27]

DEFINITION 5.2. *Let $\mathcal{L}$ and $\mathcal{L}'$ be many-sorted first-order languages. An* interpretation Γ *of $\mathcal{L}$ in $\mathcal{L}'$ is given by the following two items*:

1. Domain formulas: *For every sort symbol s of $\mathcal{L}$, an $\mathcal{L}'$-formula $\delta_s(\bar{x})$ called a* domain formula for s.
2. Translation formulas: *For each unnested atomic $\mathcal{L}$-formula $\phi(y_1, \dots, y_m)$, where y_i is a variable of sort s_i, an $\mathcal{L}'$-formula $\phi^\Gamma(\bar{x_1}, \dots, \bar{x_m})$, where the arity of $\bar{x_i}$ is given by the arity of the domain formula $\delta_{s_i}(\bar{x})$ for s_i.*[28]

Let $\mathcal{M}$ be an $\mathcal{L}$-structure, and $\mathcal{N}$ be an $\mathcal{L}'$-structure. Then, an interpretation Γ of $\mathcal{M}$ in $\mathcal{N}$ *is given an interpretation Γ of $\mathcal{L}$ in $\mathcal{L}'$ with the following additional property*: *for each sort symbol $s \in \mathcal{L}$, there is a surjective map* (*called a* coordinate map for s)

$$f_s : \delta_s(\mathcal{N}) \to M_s$$

such that for any unnested atomic $\mathcal{L}$-formula $\phi(\bar{x})$ of arity $s_1 \times \cdots \times s_n$ and any finite tuple $(\bar{a_1}, \dots, \bar{a_n}) \in \delta_{s_1}(\mathcal{N}) \times \cdots \times \delta_{s_n}(\mathcal{N})$

$$\mathcal{M} \models \phi(f_{s_1}(\bar{a_1}), \dots, f_{s_n}(\bar{a_n})) \Leftrightarrow \mathcal{N} \models \phi^\Gamma(\bar{a_1}, \dots, \bar{a_n}).$$

Let Γ be an interpretation of $\mathcal{M}$ in $\mathcal{N}$, where $\mathcal{M}$ is an $\mathcal{L}$-structure and $\mathcal{N}$ is an $\mathcal{L}'$-structure. In light of the equivalence between interpretability and the existence of an interpretation map, we may denote $\Gamma\mathcal{N}$ as the quotient structure of $\mathcal{N}$ defined by Γ which is isomorphic to $\mathcal{M}$. Note that $\Gamma\mathcal{N}$ is an $\mathcal{L}$-structure and each sort ΓN_s of $\Gamma\mathcal{N}$ is of the form $X_s/E_s = \{\bar{a}/E_s : \bar{a} \in X_s\}$, where $X_s = \delta_s(\mathcal{N})$ and E_s is given by the realizations of $(x =_s y)^\Gamma$ in $\mathcal{N}$. Moreover, the formulas in Γ determine the interpretations of the rest of the non-logical symbols of $\mathcal{L}$ in $\Gamma\mathcal{N}$ in an analogous way.

[27] A proof of this claim can be adapted from the proof of Theorem 5.3.1 of Hodges (1993).

[28] An *unnested atomic* $\mathcal{L}$-formula is an atomic formula of one of the following forms (cf. Hodges (1993, sec. 2.6)):

1. $x = y$, where x and y are variables of the same sort.
2. $x = c$, where c is a constant symbol of $\mathcal{L}$ (where x and c have the same sort).
3. $f(\bar{x}) = y$, where f is a function symbol of $\mathcal{L}$ (where the arity of f determines the arity of $\bar{x}$ and y).
4. $R(\bar{x})$, where R is a relation symbol of $\mathcal{L}$ (where the arity of R determines the arity of $\bar{x}$).

https://doi.org/10.1017/S1755020319000303 Published online by Cambridge University Press

If Γ is an interpretation of $\mathcal{L}$ in $\mathcal{L}'$, then the domain and translation formulas of Γ induce a *translation map* $\phi \mapsto \phi^\Gamma$ from the *entire* set of $\mathcal{L}$-formulas to $\mathcal{L}'$-formulas as follows. If ϕ is an atomic $\mathcal{L}$-formula, then ϕ is mapped to the translation formula ϕ^Γ given by Γ. To lift this mapping on atomic formulas to the entire set of $\mathcal{L}$-formulas, we let the map commute with propositional connectives and quantifiers in the natural way, where the translated quantifiers are relativized to the appropriate domain formulas. So, for instance, the $\mathcal{L}$-formula $\forall_s y(\phi(y))$ is mapped to the $\mathcal{L}'$-formula $\forall_{s_1'} x_1 \cdots \forall_{s_n'} x_n(\delta_s(x_1, \dots, x_n) \to \phi^\Gamma(x_1, \dots, x_n))$ where δ_s is the domain formula (of arity $s_1' \times \cdots \times s_n'$) for s given by Γ. When there is no ambiguity, interpretations will be identified with their associated translation maps.[29]

Using the notion of interpretability for structures, we may define a notion of interpretability for theories as follows:

DEFINITION 5.3. *Let T be an $\mathcal{L}$-theory and T' be an $\mathcal{L}'$-theory. We say that T is* interpretable *in T' if every model $\mathcal{M}'$ of T' uniformly interprets a model of T, i.e., each model $\mathcal{M}'$ of T' interprets a model $\mathcal{M}$ of T using the same set of formulas.*

Equivalently, T is interpretable *in T' if and only if there is an interpretation Γ of $\mathcal{L}$ in $\mathcal{L}'$ such that for any $\mathcal{L}$-sentence ϕ, if $T \models \phi$, then $T' \models \phi^\Gamma$.*[30]

An interpretation Γ of $\mathcal{L}$ in $\mathcal{L}'$ witnessing the interpretability of T in T' is called an interpretation *of T in T'.*

The theories T and T' are mutually interpretable *if T and T' interpret each other.*

In sum, there are two equivalent ways of defining the notion of interpretability for first-order theories. One way is semantic: interpretability is characterized in terms of uniform definability of models. The other is syntactic: interpretability is characterized in terms of the existence of an interpretation map which preserves logical form and provability. Given this equivalence, it follows that if Γ is an interpretation of $\mathcal{L}$ in $\mathcal{L}'$ witnessing the interpretability of T in T', then for any model $\mathcal{M}'$ of T', there is some model $\mathcal{M}$ of T such that Γ witnesses the interpretability of $\mathcal{M}$ in $\mathcal{M}'$. In this case, we may denote $\Gamma\mathcal{M}'$ as the quotient structure of $\mathcal{M}'$ defined by Γ which is isomorphic to $\mathcal{M}$. Note that $\Gamma\mathcal{M}'$ is a model of T and each sort $\Gamma M_s'$ of $\Gamma\mathcal{M}'$ is of the form $X_s/E_s = \{\bar{a}/E_s : \bar{a} \in X_s\}$, where $X_s = \delta_s(\mathcal{M}')$ and E_s is the set of realizations of $(x =_s y)^\Gamma$ in $\mathcal{M}'$. As before, the formulas in Γ determine the interpretations of the rest of the non-logical symbols of $\mathcal{L}$ in $\Gamma\mathcal{M}'$.

There is a widespread belief that interpretability offers insight into issues regarding theory reduction and theoretical equivalence.[31] For our current purposes, it will be important to focus on how interpretability relates to the notion of admissibility. Recall that if T^+ is an admissible extension of T, then the concepts and theorems of T^+ must be translatable to those of T. Given the equivalence between interpretability (for theories) and the existence of an interpretation map which preserves logical form and provability, interpretability offers a promising signature-independent way of making the notion of translatability precise. Thus, it is natural to think that if T^+ is an admissible extension of T, then T^+ must be interpretable in T.

29 See Visser (2006, sec. 2.2) for a rigorous definition of a translation map.

30 Recall that the translation map $\phi \mapsto \phi^\Gamma$ induced by Γ will often be identified with Γ itself. For more on this "syntactic" approach to interpretability, the interested reader is encouraged to refer to Button & Walsh (2018, sec. 5.5) and Visser (2006).

31 See, for example, Niebergall (2013), Visser (2006), and van Fraassen (2014).

If this is right, then we have additional reasons for thinking that M-EQ and M-EXT are false: if T^+ is a Morita extension of T, then T^+ is not necessarily interpretable in T. The example which shows that Morita extensions do not always satisfy DC* demonstrates this (see p. 398). More generally, let T have a model $\mathcal{M}$ such that every sort M_s of $\mathcal{M}$ has precisely one element. Let T^+ be a Morita extension adding a new coproduct sort to T. Then every model $\mathcal{M}^+$ of T^+ has a sort with more than one element. However, any definable set of $\mathcal{M}$ has at most one element. Thus, T^+ is not interpretable in T. As a result, those who believe that interpretability is a necessary condition on admissibility have reason to reject M-EQ and M-EXT.

As a corollary, if T and T' are Morita equivalent, then they are not necessarily mutually interpretable. In fact, Example 4.8 of Barrett & Halvorson (2016) features a pair of theories T_1 and T_2 which are Morita equivalent but not mutually interpretable. Let $\mathcal{L}_1 = \{s_1, P, Q\}$ and $\mathcal{L}_2 = \{s_2, s_3\}$, where each s_i is a sort symbol, and P and Q are (unary) predicate symbols. Let T_1 be the $\mathcal{L}_1$-theory which states that P and Q are non-empty, mutually exclusive and exhaustive. Let T_2 be the empty theory in $\mathcal{L}_2$. Note that T_2 has a model $\mathcal{M}$ with a pair of one-element sorts M_2 and M_3. The structure $\mathcal{M}$ cannot interpret any model of T_1 since each model of T_1 has at least two elements. Thus, T_2 cannot interpret T_1, and so T_1 and T_2 cannot be mutually interpretable. This example illustrates how Morita extensions allow us to "define" the predicates P and Q of T_1 by using the sorts of T_2 even though the sort of T_1 (on which P and Q are defined) *cannot* be uniformly defined in models of T_2. As a consequence, this example demonstrates another sense in which Morita equivalent theories may fail to define "all the vocabulary that the other uses" (Barrett & Halvorson, 2017b, p. 3586).

As I have discussed, it is natural to think that interpretability is a necessary condition on admissibility. However, it is arguably not sufficient. If T is interpretable in T', then every model of T' uniformly defines a model of T. However, there is no guarantee that *every* model of T can be defined by a model of T'.[32] Thus, interpretable theory extensions do not satisfy DC* in full generality. To address this, one may strengthen the notion of interpretability as follows.

DEFINITION 5.4. *Let T be an $\mathcal{L}$-theory and T' be an $\mathcal{L}'$-theory. We say that T is* (*semantically*) *faithfully interpretable in T' if T is interpretable in T', and moreover, by the same interpretation, every model of T is uniformly interpreted in a model of T'.*[33]

Equivalently, T is (*semantically*) *faithfully interpretable in T' if there is an interpretation Γ of T in T' such that for any model $\mathcal{M}$ of T, there is a model $\mathcal{M}'$ of T' such that $\mathcal{M} \cong \Gamma\mathcal{M}'$.*

The theories T and T' are mutually faithfully interpretable *if T and T' faithfully interpret each other.*

If T is semantically faithfully interpretable in T', then one can define *any* model of T in some model of T'. Thus, faithfully interpretable theories *do* satisfy DC*.

The reason for calling such interpretations "semantically faithful" is because they imply the syntactic way of defining faithful interpretability: that is, T' *syntactically faithfully interprets* T if there is a translation map Γ of $\mathcal{L}$ in $\mathcal{L}'$ such that for any $\mathcal{L}$-sentence ϕ,

32 See Example 2 of Example 5.6.

33 I should note that Hodges (1993, sec. 5.5) calls semantically faithful interpretations "total interpretations." Moreover, it is important to note that faithful interpretability corresponds to the relation $\leq_2$ defined by Szczerba (1975, p. 136).

$T \models \phi$ if and only if $T' \models \phi^\Gamma$.[34] However, it is important to note that the converse does not hold.[35] Nevertheless, we will restrict our attention to faithful interpretability as defined in Definition 5.4 because it satisfies DC*. Thus, for the remainder of this article, by "faithful interpretability," I will mean semantic faithful interpretability.

Lastly, we define the notion of bi-interpretability, which is a special case of mutual (faithful) interpretability.[36]

DEFINITION 5.5. *Let T be an $\mathcal{L}$-theory and T' be an $\mathcal{L}'$-theory. Then T and T' are* bi-interpretable *if the two classes $V = \{\mathcal{M} : \mathcal{M} \models T\}$ and $W = \{\mathcal{N} : \mathcal{N} \models T'\}$ satisfy the following*:

1. *Every structure $\mathcal{M}$ from V uniformly defines a structure $\mathcal{M}^\tau$ from W which is a quotient structure of $\mathcal{M}$.*[37]
2. *Every structure $\mathcal{N}$ from W uniformly defines a structure $\mathcal{N}^\gamma$ from V which is a quotient structure of $\mathcal{N}$.*
3. *For every structure $\mathcal{M}$ from V, there are uniformly definable (in $\mathcal{M}$) bijections $g_s : (M^\tau)^\gamma_s \to M_s$ (for every sort symbol $s \in \mathcal{L}$) that induce an isomorphism between $(\mathcal{M}^\tau)^\gamma$ and $\mathcal{M}$.*
4. *For every structure $\mathcal{N}$ from W, there are uniformly definable (in $\mathcal{N}$) bijections $h_{s'} : (N^\gamma)^\tau_{s'} \to N_{s'}$ (for every sort symbol $s' \in \mathcal{L}'$) that induce an isomorphism between $(\mathcal{N}^\gamma)^\tau$ and $\mathcal{N}$.*

In many cases, $\mathcal{M}^\tau$ will be a *proper* quotient structure of $\mathcal{M}$, i.e., there is a sort $s' \in \mathcal{L}'$ such that the equivalence relation $E^\tau_{s'}(\bar{x}, \bar{y})$ used to define τ's interpretation of s' is not simply the identity (with respect to some sort s in $\mathcal{L}$). Similarly, $\mathcal{N}^\gamma$ will often be a *proper* quotient structure of $\mathcal{N}$. In this situation, a bijection $g_s : (M^\tau)^\gamma)_s \to M_s$ is definable in $\mathcal{M}$ if there is an $\mathcal{L}$-formula which defines $(g_s^{\mathcal{M}})^{-1}$ in the sense of Definition 5.1, where the sort M_s is viewed as being the realizations of the formula $x_s = x_s$ in $\mathcal{M}$ (where x_s is a variable of sort s). The analogous statement holds, by definition, for each $h_{s'} : (N^\gamma)^\tau_{s'} \to N_{s'}$.

Intuitively, bi-interpretability strengthens mutual (faithful) interpretability by requiring that T and T' define each other's models in a *definably* invertible way.

---

[34] There are other semantic analogues of syntactic faithful interpretability. For example, one can show that Γ is a syntactic faithful interpretation of T_1 in T_2 if and only if for every model $\mathcal{M}_1$ of T_1, there is a model $\mathcal{M}_2$ of T_2 such that $\Gamma\mathcal{M}_2 \equiv \mathcal{M}_1$, i.e., $\Gamma\mathcal{M}_2$ and $\mathcal{M}_1$ are elementarily equivalent.

[35] I thank Sean Walsh for the following counter-example. Let $\mathcal{L}_1$ be the language of Peano Arithmetic and $\mathcal{L}_2$ be $\mathcal{L}_1 \cup \{c\}$, where c is a constant symbol not in $\mathcal{L}_1$. Let T_1 be PA and T_2 be PA^*, where PA^* is $\mathsf{PA} \cup \{c > s^n(1) : n \in \mathbb{N}\}$. An interpretation Γ of T_1 in T_2 is given by $(\phi)^\Gamma = \phi$, i.e., the inclusion map from $\mathcal{L}_1$ to $\mathcal{L}_2$. One checks that Γ is a syntactically faithful. Note that if $\mathcal{M}_2$ is a model of T_2, then $\Gamma\mathcal{M}_2$ is just the $\mathcal{L}_1$-reduct of $\mathcal{M}_2$, which is a non-standard model of PA. Thus, there is no model $\mathcal{M}_2$ of T_2 such that $\Gamma\mathcal{M}_2 \cong \mathbb{N}$.

[36] The definition of bi-interpretability given here (which can be found on p. 90 of Walsh (2014)) is relatively semantic in nature (since it makes explicit reference to the models of a theory). It is important to emphasize that, given the equivalence between syntactic and semantic forms of interpretability, bi-interpretability may be (equivalently) defined in a syntactic fashion. For more on the syntactic formulation of bi-interpretability, see sec. A.7.2. of Visser (2015).

[37] Here, "uniformly" means that the same formulas are used for each structure in V, and similarly for structures in W.

https://doi.org/10.1017/S1755020319000303 Published online by Cambridge University Press

EXAMPLE 5.6. *The following are examples of interpretations.*

1. *If $\mathcal{M}^+$ is a definitional expansion of $\mathcal{M}$, then $\mathcal{M}^+$ is interpretable in $\mathcal{M}$. If T^+ is a definitional extension of T, then T^+ is interpretable in T. If T and T' are definitionally equivalent, then they are bi-interpretable.*
2. *The theory* $\mathsf{PA}+\neg\mathsf{Con}(\mathsf{PA})$ *is mutually interpretable with* PA *but there is no faithful interpretation of* PA *in* $\mathsf{PA}+\neg\mathsf{Con}(\mathsf{PA})$.[38]
3. *First-order Peano arithmetic* PA *is bi-interpretable with the theory* $\mathsf{ZFfin}+\mathsf{TC}$*, where* ZFfin *is* $(\mathsf{ZF}-\mathsf{Infinity})+\neg\mathsf{Infinity}$ *and* TC *states that every set has a transitive closure.*[39]

5.2. Morita implies bi-interpretability. Let's recall the two ways of modifying the definition of a Morita extension that I suggested earlier. The first is to (*i*) restrict the definition to theories that assert the existence of (at least) two objects (within the same sort). The second is to (*ii*) drop disjoint union (i.e., coproduct) as a way of adding new sorts.

I begin by showing that (*i*) entails that Morita equivalent theories are bi-interpretable. I then use the proof of this to establish the analogous claim with regards to (*ii*). As a corollary, I note that, under either (*i*) or (*ii*), Morita extensions satisfy DC*. I begin by proving a lemma that reduces the notion of a Morita extension to that of bi-interpretability.

LEMMA 5.7. *Let T be an $\mathcal{L}$-theory such that $T \models \exists_s x \exists_s y(x \neq y)$ for some sort symbol $s \in \mathcal{L}$. Let T^+ be a Morita extension of T, then T^+ and T are bi-interpretable.*

Proof Sketch. Due to the length of the proof, it has been placed in Appendix 7.3. The basic idea behind the proof is worth discussing since the proof can be used to show that Morita extensions satisfy DC*.

The proof begins with the observation that there are four possible ways for a Morita extension T^+ of a given theory T to define a new sort symbol. There is the case where T^+ defines a new product sort, the case where it defines a new coproduct sort, the case where it defines a new quotient sort and the case where it defines a new subsort. I then show that, in each case, every model $\mathcal{M}^+$ of T^+ is interpretable in its $\mathcal{L}$-reduct $\mathcal{M}$ using the same set of formulas. Thus, in each case, the new sort symbol will be uniformly definable by definable quotient structures of T. It follows by definition that T^+ is (faithfully) interpretable in T. Let's τ denote this uniform construction of models of T^+ from models of T.

Conversely, T^+ (faithfully) interprets T via the interpretation γ which maps each model $\mathcal{M}^+$ of T^+ to its $\mathcal{L}$-reduct $\mathcal{M}^+ \restriction_{\mathcal{L}}$.

Lastly, one defines a bijection $g_s : (M^\tau)^\gamma_s \to M_s$ for every $s \in \mathcal{L}$ such that the collection $\{g_s : s \in \mathcal{L}\}$ induces an isomorphism between $(\mathcal{M}^\tau)^\gamma$ and $\mathcal{M}$. One defines an analogous collection of bijections $\{h_s : s \in \mathcal{L}^+\}$ which induce an isomorphism between $((\mathcal{M}^+)^\gamma)^\tau$ and $\mathcal{M}^+$. □

According to the proof of Lemma 5.7, a Morita extension T^+ of T can only add new sorts that are definable in some definable quotient structure of a model of T. It follows that Morita extensions satisfy DC* (with respect to the theories they are extending). To put things more concisely, the following is an immediate corollary of the proof.

---

38 See Proposition 5.10 of Button & Walsh (2018).

39 This result is proven in Kaye & Wong (2007).

COROLLARY 5.8. *Let T be an $\mathcal{L}$-theory such that $T \models \exists_s x \exists_s y(x \neq y)$ for some sort symbol $s \in \mathcal{L}$. If T^+ is a Morita extension of T, then T^+ satisfies* DC* (*with respect to T*).

Moreover, in the proof of Lemma 5.7, the assumption that $T \models \exists_s x \exists_s y(x \neq y)$ is only used to handle coproduct sorts. Thus, the proof of Lemma 5.7 also shows that if one drops coproduct sorts from the definition of Morita extensions, then *all* Morita extensions satisfy DC*.

COROLLARY 5.9. *If T^+ be a Morita extension of T which does not add any new coproduct sorts, then T^+ satisfies* DC* (*with respect to T*).

For the same reason, the proof of Lemma 5.7 can be used to show the following.

LEMMA 5.10. *If T^+ is a Morita extension of T which does not add any new coproduct sorts to T, then T and T^+ are bi-interpretable.*

Proof. As remarked, the proof is the same as in Lemma 5.7, except that we do not include the case covering coproduct sorts. □

It is a straightforward consequence of Lemma 5.7 that, if we restrict ourselves to theories that satisfy $T \models \exists_s x \exists_s y(x \neq y)$, Morita equivalent theories are bi-interpretable.

PROPOSITION 5.11. *Let T and T' imply $\exists_s x \exists_s y(x \neq y)$ for some sort s, respectively. Then if T and T' are Morita equivalent, then they are bi-interpretable.*

Proof. Let T and T' be Morita equivalent. Then there are finite sequences of Morita extensions $T, T_1, \ldots, T_n$ and $T', T'_1, \ldots, T'_m$ such that $T_n \equiv T'_m$. By Lemma 5.7, each T_{i+1} is bi-interpretable in T_i for $i = 1, \ldots, n$. Similarly, each T_{j+1} is bi-interpretable in T_j for $j = 1, \ldots, m$. Since bi-interpretability is transitive, T_n is bi-interpretable with T and T'_m is bi-interpretable with T'. Moreover, since logical equivalence is a special case of bi-interpretable, T_n is bi-interpretable with T'_m. Again, by transitivity, T and T' are bi-interpretable. □

The above proposition shows that if we restrict ourselves to theories that imply $\exists_s x \exists_s y(x \neq y)$ for some sort s, Morita equivalence is a special case of bi-interpretability. Similarly, we can use Lemma 5.10 to prove the following analogue of Proposition 5.11.

PROPOSITION 5.12. *Let T and T' be Morita equivalent by way of finite sequences of Morita extensions which define no new coproduct sorts. Then T and T' are bi-interpretable.*

Proof. The proof is the same as the proof of Proposition 5.11, except that we use Lemma 5.10 and omit the coproduct case. □

Thus, if the advocate of M-EQ chooses to drop coproduct sorts as a way of defining new sorts via Morita extensions, Morita equivalent theories will end up being bi-interpretable.

In general, the converse does not hold: if T and T' are bi-interpretable, then T and T' may not be Morita equivalent. A counter-example can be given as follows. Let T be the empty theory with one sort s and let $T' = \bigcup_{i=0}^{\omega} T_i$ where each T_i defines a product sort for the i-fold product of s with itself. Then T and T' are bi-interpretable.[40] However, T and T' are not Morita equivalent for the simple reason that there is no finite sequence of Morita extensions which gives one T' from T. Thus, Morita equivalence *properly* strengthens

[40] Given any model $\mathcal{M}$ of T and extension $\mathcal{M}'$ of T' such that $s^{\mathcal{M}} = s^{\mathcal{M}'}$, $\mathcal{M}^n$ is definable in $\mathcal{M}$ by the $\mathcal{L}$-formula $x_1 = x_1 \wedge x_2 = x_2 \wedge \cdots \wedge x_n = x_n$. Thus, T is bi-interpretable in T'.

bi-interpretability. It is left as an open question whether the converse holds for theories with finite signatures.[41]

§6. Conclusion. In this article, I have argued that the central claim that Barrett and Halvorson use to argue for the reasonableness of Morita equivalence as a measure of theoretical equivalence is problematic. I have shown that the claim is insufficiently argued for, and moreover, that we have good reason to think that the claim is actually false. In light of these problems, I have shown that one natural way of responding will make Morita equivalence a strengthening of bi-interpretability, i.e., Morita equivalence implies bi-interpretability.

There is a fairly widespread belief among logicians that bi-interpretability deserves a privileged status as a "signature-independent" notion of equivalence for many-sorted first-order theories.[42] A thorough assessment of whether this sentiment is ultimately justified is outside the scope of this article.[43] However, it should be noted that bi-interpretability preserves a variety of desirable meta-theoretic properties such as finite axiomatizability, decidability, and κ-categoricity.[44] Moreover, bi-interpretability captures a precise sense in which two theories define each other's models in a definably invertible way. Syntactically, this implies that bi-interpretable theories (syntactically) faithfully interpret one another (see p. 403). Thus, whether we think of many-sorted first-order theories "semantically" or "syntactically," these considerations count in favor of bi-interpretability's privileged status as a notion of theoretical equivalence.[45]

As a result, I conclude that the tenability of Morita equivalence as a measure of theoretical equivalence depends on how much it strengthens the notion of bi-interpretability. If it strengthens bi-interpretability too much, Morita equivalence runs the risk of systematically and arbitrarily deeming classes of theories inequivalent that should actually count as equivalent. In this case, Morita equivalence would lose practical value as measure

41 An earlier version of this article contained a purported proof of a partial converse. In particular, this purported proof attempted to establish that if we restrict ourselves to theories with finite signatures, then mutual faithful interpretability implies Morita equivalence. In private correspondence, Professor Halvorson notified me of a gap in this proof. For this, I am deeply grateful to Professor Halvorson. In fact, it is important to note that mutual faithful interpretability does *not* imply Morita equivalence (even assuming finite signatures) since Morita equivalent theories have the isomorphic automorphism groups while mutually faithfully interpretable theories need not.

42 See, for example, Visser (2015) and Slaman (2008).

43 For a critical discussion on bi-interpretability as a notion of theoretical equivalence, see Button & Walsh (2018).

44 For more on the properties preserved by bi-interpretability, see Visser (2015).

45 It should be mentioned that from a category-theoretic point of view, bi-interpretability seems like the most natural weakening of definitional equivalence (cf. Friedman & Visser (2014) and Visser (2006)). Let INT_0 be the category of interpretations and maps between interpretations. In this category, two interpretations $\tau, \gamma : S \to T$ are defined to be *equal* if whenever $\mathcal{M}$ is a model of the interpreting theory T, $\mathcal{M}^\tau = \mathcal{M}^\gamma$. As a result, isomorphism in INT_0 is definitional equivalence. The category INT_1 is the same as INT_0 except it defines equality as the existence of a "definable isomorphism" rather than strict identity. Two interpretations $\tau, \gamma : S \to T$ are defined to be *equal* in INT_1 if, whenever $\mathcal{M}$ is a model of T, there is a formula F in the language of T such that $F^{\mathcal{M}}$ induces an isomorphism between $\mathcal{M}^\tau$ and $\mathcal{M}^\gamma$. Isomorphism in INT_1 is bi-interpretability. Unless there is some natural notion of equality sitting between definable isomorphism and strict identity, then it follows that from this category-theoretic point of view, bi-interpretability seems like the most natural weakening of definitional equivalence.

https://doi.org/10.1017/S1755020319000303 Published online by Cambridge University Press

of theoretical equivalence. If not strong enough, Morita equivalence would not seem to possess any theoretical advantages over bi-interpretability.

In the remainder of these concluding remarks, I discuss whether the advocate of Morita equivalence might be better off opting for a notion of theoretical equivalence defined in terms of interpretability rather than the notion of a Morita extension. To this end, I enumerate a number of conceptual and practical worries concerning the notion of a Morita extension, and then discuss how the notion of interpretability is not subject to the same sorts of worries. Thus, regardless of how much Morita equivalence strengthens bi-interpretability, there are compelling reasons to adopt bi-interpretability as a measure of theoretical equivalence over Morita equivalence.

6.1. Interpretability versus morita. A general worry concerning the notion of Morita equivalence is that its definition appears *ad hoc*. There is little to motivate the claim that the only kinds of sort symbols that we can admissibly add to a theory are the ones that Barrett and Halvorson happen to discuss. It is not even clear whether a more motivated list can be enumerated.

Another general worry stems from the failure of Morita extensions to be *transitive*. That is, if T_2 is a Morita extension of T_1, and T_3 is a Morita extension of T_2, then it is not necessarily the case that T_3 is a Morita extension of T_1. In fact, this lack of transitivity is pervasive. For example, if T_1 is the empty theory with one sort s, T_2 defines a product sort $s \times s$ of s with itself, and T_3 defines a product sort $(s \times s) \times s$ of $s \times s$ with s, then T_3 is not a Morita extension of T_1.

Failing to satisfy transitivity would have been seen as problematic from the point of view of someone like Tarski, who influenced his student Szczerba to *avoid* defining a formal notion of equivalence in terms of theory extensions along the lines used by Barrett and Halvorson. According to Tarski, if we extend theories based on rules of definition, then our theory extensions will depend on the order in which rules of definition are used (or, in our current case, the order in which Morita extensions are introduced). To return to the example just given, note that the rule we used to extend T_2 to T_3 can't be used to extend T_1 to T_3 since T_1 has no sort for $s \times s$. Thus, there is no way to introduce a sort $(s \times s) \times s$ to get to T_3 from T_1. This dependency on the order of extension was viewed by Tarski as an "unpleasant property" (Szczerba, 1975, p. 129) because it reveals a deep disanalogy between rules of definition and rules of inference.

For instance, one often considers the closure of a first-order theory under the standard rules of deduction. Regardless of the order by which one implements these rules, one will (ultimately) arrive at the same theory. So, it makes perfect sense to speak of *the* closure of a first-order theory. However, since extending theories by rules of definition depends on the order in which one implements those rules, we cannot similarly talk about "the" definitional closure of a theory without first specifying an order.

Given all this, a major theoretical advantage of the notion of interpretability over the notion of a Morita extension is that the former is general and well motivated: interpretability captures the idea that a theory T_1 is a "part" of a theory T_2 if the the primitive symbols of T_2 can uniformly define the primitive symbols of T_1.

Moreover, since interpretability is transitive, it escapes the "dependency of order" that Tarski considered to be problematic. It was this consideration that motivated Szczerba himself to define a notion of equivalence in terms of interpretability rather than in terms of extensions. Therefore, there are compelling reasons to think that methodological and conceptual gains will result from opting for the notion of interpretability over the notion of a Morita extension.

It is no surprise that interpretability has enjoyed wide currency in mathematical logic over the past half-century. At the end of the day, I hope to have shown how, in view of the various issues facing Morita equivalence, we may now appreciate the conceptual leaps and technical innovations of Tarski and his students in a brand new light.[46]

§7. Appendix.

7.1. The translation theorem. In this appendix, I discuss Theorem 4.11 in greater detail. Before the theorem can be precisely stated, I'll need to define the notion of a "code."

Intuitively, a code is an $\mathcal{L}^+$-formula which relates variables from sorts in $\mathcal{L}^+$ with variables from sorts in $\mathcal{L}$. More formally, given variables $x_1, \dots, x_n$ of sorts $s_1, \dots, s_n \in \mathcal{L}^+ - \mathcal{L}$, a *code* is an $\mathcal{L}^+$ formula of the form

$$\xi(x_1, y_1^1, y_1^2; \dots; x_n, y_n^1, y_n^2) \equiv_{\text{def}} (\xi_1(x_1, y_1^1, y_1^2) \wedge \cdots \wedge \xi_n(x_n, y_n^1, y_n^2)) \tag{14}$$

where y_i^1 and y_i^2 are variables from sorts in $\mathcal{L}$, and where each conjunct ξ_i is defined according to the kind of sort x_i belongs to.

If $x_i \in s_i$ and T^+ defines s_i as a product sort with projection maps π_1 and π_2 of arities $s_i \to s_i^1$ and $s_i \to s_i^2$, respectively, then the conjunct $\xi_i(x_i, y_i^1, y_i^2)$ is just the $\mathcal{L}^+$ formula $\pi_1(x_i) = y_i^1 \wedge \pi_2(x_i) = y_i^2$. If T^+ defines s_i as a coproduct sort of T with projections ρ_1 and ρ_2 of arities $s_1 \to s$ and $s_2 \to s$, then ξ_i is either the $\mathcal{L}^+$ formula $\rho_1(y_i^1) = x_i$ or the $\mathcal{L}^+$ formula $\rho_2(y_i^2) = x_i$. If T^+ defines s_i as a subsort with inclusion map i of arity $s_i \to s^1$, then ξ_i is the $\mathcal{L}^+$ formula $i(x_i) = y_i^1$. Lastly, if T^+ defines s_i as a quotient sort with projection map ϵ of arity $s_i^1 \to s_i$, then ξ is the $\mathcal{L}^+$ formula $\epsilon(y_i^1) = x_i$.

In what follows, variables in sorts s_1, s_2, etc. from $\mathcal{L}^+ - \mathcal{L}$ will be denoted by x_1, x_2, etc. and variables in sorts t_1, t_2, etc. from $\mathcal{L}$ will be denoted by y_1, y_2, etc. Theorem 4.11 can now be precisely stated.

THEOREM 4.11 (Translation theorem). *Let T^+ be a a Morita extension of T. Moreover, suppose that $\phi(x_1, \dots, x_n; y_1, \dots, y_m)$ is an $\mathcal{L}^+$-formula. Then, for any code $\xi(x_1, y_1^1, y_1^2; \dots; x_n, y_n^1, y_n^2)$ for variables $x_1, \dots, x_n$, there exists an $\mathcal{L}$-formula $\phi^*(y_1, \dots, y_m; y_1^1, \dots, y_n^2)$ such that*

$$\begin{aligned} T^+ \models \forall_{s_1} x_1, \dots \forall_{s_n} x_n \forall_{t_1} y_1 \dots \forall_{t_m} y_m \forall_{s_1^1} y_1^1 \dots \forall_{s_n^2} y_n^2 [\xi(x_1, \dots, y_n^2) \\ \to (\phi(x_1, \dots, x_n; y_1, \dots, y_m) \leftrightarrow \phi^*(y_1, \dots, y_m; y_1^1, y_1^2, \dots, y_n^1, y_n^2))]. \end{aligned} \tag{15}$$

7.2. Bad company. In this appendix, I discuss an example given earlier in the article (see p. 399), which I used to argue that Morita extensions had "bad company," i.e., there are instances of theory extensions that share many features with Morita extensions, yet seem to add substantial truths to the theories they extend. First, let's briefly recall the example.

Let T be an $\mathcal{L}$-theory where $\mathcal{L}$ is empty and single-sorted. Let $\mathcal{L}^+$ be the signature $\mathcal{L} \cup \{s_\omega\} \cup \{\rho_i : i < \omega\}$, where s_ω is a sort symbol not appearing in $\mathcal{L}$, and each ρ_i is a function symbol of arity $s \to s_\omega$ for the sort $s \in \mathcal{L}$.

We define T^ω to be the $\mathcal{L}^+$-theory

$$T \cup \{\phi_{i,j} : i, j < \omega \text{ and } i \neq j\} \cup \{\psi_i : i < \omega\}, \tag{16}$$

[46] These points provide support for van Fraassen's suggestion (cf. van Fraassen (2014)) that the conceptual framework given by the notion of interpretability can help usher in a promising new formal approach in the philosophy of science.

https://doi.org/10.1017/S1755020319000303 Published online by Cambridge University Press

where each $\phi_{i,j}$ is the $\mathcal{L}^+$-sentence

$$\forall_s x \forall_s y[\rho_i(x) \neq \rho_j(y)], \tag{17}$$

and each ψ_i is an $\mathcal{L}^+$-sentence stating that ρ_i is injective. Intuitively, T^ω extends T by defining a new sort s_ω that contains the disjoint union of s with itself ω-many times.

Next, I will show how a notion of a "code" can be defined for T^+ which allows us to state a translation theorem in the style of Theorem 4.11. Like before, a *code* for variables $x_1, \ldots, x_n$ of sort s_ω is an $\mathcal{L}^+$-formula of the form

$$\xi(x_1, y_1; \ldots; x_n, y_n) \equiv_{\text{def}} (\xi_1(x_1, y_1) \wedge \cdots \wedge \xi_n(x_n, y_n)), \tag{18}$$

where each y_i is a variable of sort $s \in \mathcal{L}$, and where each conjunct $\xi_i(x_i, y_i)$ is defined to be any $\mathcal{L}^+$-formula of the form $\rho_k(y_i) = x_i$ for some $k < \omega$ (where x_i is a variable of sort s_ω). With this slight modification of the notion of a code, we can formulate the following analogue of Theorem 4.11:

PROPOSITION 7.1. *Let T^ω be defined as above. Let $\phi(x_1, \ldots, x_n; \bar{z})$ be an $\mathcal{L}^+$-formula where $x_1, \ldots, x_n$ are variables of sort $s_\omega \in \mathcal{L}^+ - \mathcal{L}$ and $\bar{z}$ is a finite tuple of variables of sort $s \in \mathcal{L}$. Then given any code $\xi(x_1, y_1; \ldots; x_n, y_n)$ for the variables $x_1, \ldots, x_n$, there exists an $\mathcal{L}$-formula $\phi^*(z_1, \ldots, z_m; y_1, \ldots, y_n)$ such that*

$$\begin{aligned} T^+ \models \forall_{s_\omega} x_1, \ldots, x_n \forall_s \bar{z}, y_1, \ldots, y_n[\xi(x_1, y_1; \ldots; x_n, y_n) \\ \to (\phi(x_1, \ldots, x_n; \bar{z}) \leftrightarrow \phi^*(\bar{z}; y_1, \ldots, y_n))]. \end{aligned} \tag{19}$$

To prove this, I begin by proving an analogue of Lemma A.1 from Barrett & Halvorson (2016).

LEMMA 7.2. *Let $t(x_1, \ldots, x_n, \bar{z})$ be an $\mathcal{L}^+$-term of sort $s' \in \mathcal{L}^+$ and x is a variable of sort s'. Let $\xi(x, y; x_1, y_1; \ldots; x_n, y_n)$ be a code for the variables $x, x_1, \ldots, x_n$. Then there is an $\mathcal{L}$-formula $\phi^*(x, \bar{z}, y_1, \ldots, y_n)$ such that*

$$\begin{aligned} T^+ \models \forall_{s'} x \forall_{s_\omega} x_1, \ldots, x_n \forall_s \bar{z}, y_1, \ldots, y_n[\xi(x, y; \ldots; x_n, y_n) \\ \to (t(x_1, \ldots, x_n, \bar{z}) = x \leftrightarrow \phi^*(x, \bar{z}, y_1, \ldots, y_n)]. \end{aligned} \tag{20}$$

If $s' \in \mathcal{L}$, then x will not appear in ξ. If $s' \in \mathcal{L}^+ - \mathcal{L}$, then x will not appear in ϕ^*.

Proof. The proof is the same as the proof of Lemma A.1 in Barrett & Halvorson (2016), and proceeds by induction on the complexity of t. All but two cases are covered in the original proof.

The first case is when x is a variable of sort s_ω, and t is a variable x_i of sort s_ω for some $i \leq n$. A code for x and x_i is an $\mathcal{L}^+$-formula of the form $\rho_k(y_i) = x_i \wedge \rho_l(y) = x$ for $k, l < \omega$. We define the $\mathcal{L}$-formula ϕ^* to be $y_i = y$ in the case where $k = l$, and any contradiction in the case where $k \neq l$.

The second case is when t is a term of the form

$$f(t_1(x_1, \ldots, x_n, \bar{z}), \ldots, t_k(x_1, \ldots, x_n, \bar{z})), \tag{21}$$

where f is a function symbol of the form ρ_k for some $k < \omega$ and ρ_k is of arity $s \to s_\omega$. Then, t is of the form $\rho_k(y_i)$ where y_i is of sort s.

Let $\xi(x, y) \equiv_{\text{def}} \rho_l(y) = x$ be a code for x, where y is of sort s. If $l \neq k$, then let ϕ^* be any contradiction, for example $y \neq y$. If $l = k$, then let ϕ^* be the formula $y_i = y$. Either way, ϕ^* satisfies condition (20). □

We now return to the proof of Proposition 7.1.

Proof of Proposition 7.1. The proof follows the proof of Theorem 4.11 in Barrett & Halvorson (2016) and proceeds by induction on the complexity of ϕ. Barrett and Halvorson's proof of Theorem 4.11. Note that since $\mathcal{L}$ is empty, the only nontrivial case to consider is when ϕ is an $\mathcal{L}^+$-formula of the form

$$t(x_1, \ldots, x_n, \bar{z}) = u(x_1, \ldots, x_n, \bar{z}), \tag{22}$$

for s_ω-terms t and u. If t or u is a variable symbol of sort s_ω, then we simply apply Lemma 7.2. Otherwise, t is of the form $\rho_k(y_1)$ for some $k < \omega$ and u is of the form $\rho_l(y_2)$ for some $l < \omega$. If $k \neq l$, then let ϕ^* be any contradiction in the signature $\mathcal{L}$, for example, $y_1 \neq y_1$. If $k = l$, then we let ϕ^* be the $\mathcal{L}$-formula $y_1 = y_2$. One verifies that, in either case, ϕ^* satisfies the condition (19).

This covers the base cases. For the inductive step, we consider the cases of $\neg$, $\wedge$, and $\forall$. Suppose the result holds for $\mathcal{L}^+$-formulas ϕ_1 and ϕ_2. Then the result holds for $\neg\phi_1$ by letting $(\neg\phi)^*$ be $\neg(\phi^*)$. Similarly, the result holds for $\phi_1 \wedge \phi_2$ by letting $(\phi_1 \wedge \phi_2)^*$ be $\phi_1^* \wedge \phi_2^*$. Lastly, to check $\forall$, consider the $\mathcal{L}^+$ formula $\forall_{s_\omega} x\phi(x, x_1, \ldots, x_n)$ such that the result holds for $\phi(x, x_1, \ldots, x_n)$. Let $\xi(x, y, x_1, y_1, \ldots, x_n, y_n)$ be a code for the s_ω-variables $x, x_1, \ldots, x_n$. By the inductive hypothesis, there is an $\mathcal{L}$-formula $\phi^*(y, y_1, \ldots, y_n)$. We define $(\forall_{s_\omega} x\phi(x))^*$ to be the $\mathcal{L}$-formula $\forall_s y\phi^*(y, y_1, \ldots, y_n)$. One verifies that this formula satisfies the condition (19). □

7.3. Proof of Lemma 5.7. In this appendix, I provide a rigorous proof of Lemma 5.7. The proof is broken up into two parts. First, we prove that T and T^+ are mutually (faithfully) interpretable by defining a pair of (faithful) interpretations. Then, we show that these interpretations witness the bi-interpretability of T and T^+, which establishes Lemma 5.7. The following Lemma establishes the first part.

LEMMA 7.3. *Let T be an $\mathcal{L}$-theory such that $T \models \exists_{s'} x, y(x \neq y)$ for some sort symbol $s' \in \mathcal{L}$. Then T and T^+ are mutually faithfully interpretable.*

Proof. Let T^+ be a Morita extension of T, where T satisfies the hypothesis of the Lemma. We define a pair of interpretations γ and τ such that, whenever $\mathcal{M}$ is a model of T, $\mathcal{M}^\tau$ is a quotient structure of $\mathcal{M}$ that is a model of T; and similarly, whenever $\mathcal{M}^+$ is a model of T^+, $(\mathcal{M}^+)^\gamma$ is a quotient structure of $\mathcal{M}^+$ that is a model of T^+ (see Definition 5.3). Faithfulness of these interpretations follow from the fact that every model $\mathcal{M}$ of T uniquely extends to a model $\mathcal{M}^+$ of T^+.[47]

We begin by defining an interpretation γ which defines models of $\mathcal{M}$ from $\mathcal{M}^+$. Given a model $\mathcal{M}^+$ of T^+, let $(\mathcal{M}^+)^\gamma$ be the $\mathcal{L}$-reduct $\mathcal{M}^+|_\mathcal{L}$ of $\mathcal{M}^+$, which is a model of T. Note that the $\mathcal{L}$-reduct $\mathcal{M}^+|_\mathcal{L}$ of $\mathcal{M}^+$ is trivially definable in $\mathcal{M}^+$.

Next, we define an interpretation τ, which defines a quotient structure $\mathcal{M}^\tau$ from any model $\mathcal{M}$ of T such that $\mathcal{M}^\tau$ is a model of T^+. First, if σ is a symbol in $\mathcal{L}$, then let $\sigma^{\mathcal{M}^\tau}$ be $\sigma^{\mathcal{M}}$, which is trivially definable in $\mathcal{M}$. Suppose then that σ is a symbol in $\mathcal{L}^+ - \mathcal{L}$. When σ is a sort symbol s of $\mathcal{L}^+ - \mathcal{L}$, there are four cases, depending on the type of sort s is:

Case 1. s is a subsort sort. Then, the explicit definition δ_s of s is of the form

$$\forall_{s_0} x[\phi(x) \leftrightarrow \exists_s z(i(z) = x)] \wedge \forall_s z_1, z_2[i(z_1) = i(z_2) \rightarrow z_1 = z_2], \tag{23}$$

[47] This is Theorem 4.2 of Barrett & Halvorson (2016).

https://doi.org/10.1017/S1755020319000303 Published online by Cambridge University Press

where $\phi(x)$ is an $\mathcal{L}$-formula, x is a variable of $\mathcal{L}$-sort s_0, and i is a function of arity $s \to s_0$.

We define M_s^τ and $i^{\mathcal{M}^\tau}$ as follows. Let M_s^τ be $\phi(\mathcal{M})$, which is definable in $\mathcal{M}$. Moreover, let the graph of $i^{\mathcal{M}^\tau}$ be the set of realizations of $\phi(x_0) \wedge x_0 = x$ in $\mathcal{M}$. One verifies that (23) and the associated admissibility condition are satisfied by the quotient structure $\mathcal{M}^\tau$.

Case 2. s is a product sort of $\mathcal{L}$-sorts s_1 and s_2. Then, the explicit definition δ_s for s is of the form

$$\forall_{s_1} x \forall_{s_2} y \exists_{s=1} z[\pi_1(z) = x \wedge \pi_2(z) = y] \tag{24}$$

where π_1 is a function of arity $s \to s_1$ and π_2 is a function of arity $s \to s_2$.

We define M_s^τ, $\pi_1^{\mathcal{M}^\tau}$, and $\pi_2^{\mathcal{M}^\tau}$ as follows. Let M_s^τ be the set of realizations of the following $\mathcal{L}$-formula in $\mathcal{M}$:

$$\phi(x_1, x_2) \equiv_{\text{def}} (x_1 = x_1 \wedge x_2 = x_2)$$

where x_1 is a variable of sort s_1 and x_2 is a variable of sort s_2. Lastly, we define the maps $\pi_1^{\mathcal{M}^\tau}$ and $\pi_2^{\mathcal{M}^\tau}$, respectively, to be the sets of realizations of the $\mathcal{L}$-formulas $\phi(x_1, x_2) \wedge x_1 = x$ and $\phi(x_1, x_2) \wedge x_2 = x$ in $\mathcal{M}$. One verifies that $\mathcal{M}^+$ satisfies (24).

Case 3. s is a quotient sort. Then, the explicit definition δ_s for s is of the form

$$\forall_{s_0} x_1 \forall_{s_0} x_2[\epsilon(x_1) = \epsilon(x_2) \leftrightarrow \phi(x_1, x_2)] \wedge \forall_s z \exists_{s_1} x[\epsilon(x) = z] \tag{25}$$

where s_0 is an $\mathcal{L}$-sort, ϵ is a function of arity $s_0 \to s$, and $\phi(x, y)$ is an $\mathcal{L}$-formula defining an equivalence relation on every model of T.

We define M_s^τ and $\epsilon^{\mathcal{M}^\tau}$ as follows. Let M_s^τ be $M_{s_1}/\phi(\mathcal{M})$ (i.e., the quotient of $\mathcal{M}_{s_1}$ under the equivalence relation given by $\phi(\bar{x}, \bar{y})$ on $M_{s_1} \times M_{s_1}$ in $\mathcal{M}$). Moreover, we define $[(\epsilon)^{-1}]^{\mathcal{M}^\tau}$ as the set of realizations of $\phi(x_1, x_2)$ in $\mathcal{M}$. One verifies that $\mathcal{M}^+$ satisfies (25) and the associated admissibility conditions for s.

Case 4. s is a coproduct of $\mathcal{L}$-sorts s_1 and s_2. Then, the explicit definition δ_s for s is of the form

$$\forall_s z[\exists_{s_1=1} x(\rho_1(x) = z) \vee \exists_{s_2=1} y(\rho(y) = z)] \wedge \forall_{s_1} x \forall_{s_2} y[\rho_1(x) \neq \rho_2(y)] \tag{26}$$

where ρ_1 and ρ_2 are functions of arity $s_1 \to s$ and $s_2 \to s$, respectively.

We define M_s^τ, $\rho_1^{\mathcal{M}^\tau}$, and $\rho_2^{\mathcal{M}^\tau}$ as follows. Since $T \models \exists_{s'} x \exists_{s'} y(x \neq y)$ for some $\mathcal{L}$-sort s', we may define M_s^τ in $\mathcal{M}$ in the following way. Let M' be the product $M_{s_1} \times M_{s_2} \times M_{s'} \times M_{s'}$, and let $(x, y, z_1, z_2) \sim (x', y', z_1', z_2')$ be a relation on M' defined by the $\mathcal{L}$-formula

$$(x = x' \wedge z_1 = z_2 \wedge z_1' = z_2') \vee (y = y' \wedge z_1 \neq z_2 \wedge z_1' \neq z_2'). \tag{27}$$

Note that $\sim$ is, by construction, a definable relation on M'. Moreover, it is not hard to check that $\sim$ is an equivalence relation on M' such that $M'/\sim$ is isomorphic to the usual way of defining the disjoint union of M_{s_1} and M_{s_2}. Thus, we define M_s^τ as $M'/\sim$.

To complete the proof of the coproduct case, we have to define the the interpretation of the coproduct maps in M_s^τ. To this end, we define $\rho_1^{\mathcal{M}^\tau} : M_{s_1} \to M_s^\tau$ as $\rho_1(a) = ([(a, b, c, c)])$, where $b \in M_{s_2}$ and $c \in M_{s'}$. Similarly, define $\rho_2^{\mathcal{M}^\tau} : M_{s_2} \to M_s^\tau$ as $\rho_2(b) = ([a, b, c, d)])$, where $a \in M_{s_1}$ and c, d are distinct elements in $M_{s'}$. Note that $[(\rho_1)^{-1}]^{\mathcal{M}^\tau}$ and $[(\rho_2)^{-1}]^{\mathcal{M}^\tau}$ are definable in $\mathcal{M}$: the map $[(\rho_1)^{-1}]^{\mathcal{M}^\tau}$ is definable by the $\mathcal{L}$-formula $x_1 = x \wedge z_1 = z$. Similarly, the map $[(\rho_2)^{-1}]^{\mathcal{M}^\tau}$ is definable in $\mathcal{M}$ by the $\mathcal{L}$-formula $x_2 = y \wedge z_1 \neq z$. Thus, these maps are definable in $\mathcal{M}^\tau$ as a quotient structure of $\mathcal{M}$. One verifies that, given these definitions, $\mathcal{M}^\tau$ satisfies (26).

Cases 1–4 cover all possible ways T^+ can define new sorts to T. What remains to be checked is the case where the symbol σ is a relation, function, or constant symbol in $\mathcal{L}^+ - \mathcal{L}$. In this case, s is definable in terms of some $\mathcal{L}$-formula $\phi(\bar{x})$ and the function symbols used to define new sort symbols, which we have already defined in $\mathcal{M}^\tau$ as a quotient structure of $\mathcal{M}$.

Since the $\mathcal{L}$-formulas used to define $\mathcal{M}^\tau$ from $\mathcal{M}$ can be used on *any* model of T to construct a model of T^+, we have provided a uniform way of defining models of T^+ from T. Thus, τ defines an interpretation (in fact, a faithful interpretation) of T^+ in T. □

Next, we show τ and γ as defined in Lemma 7.3 witness the bi-interpretability of T and T^+.

LEMMA 5.7. *Let T be an $\mathcal{L}$-theory such that $T \models \exists_s x \exists_s y(x \neq y)$ for some sort symbol $s \in \mathcal{L}$. Let T^+ be a Morita extension of T, then T^+ and T are bi-interpretable.*

Proof. Let T be an $\mathcal{L}$-theory such that $T \models \exists_s x \exists_s y(x \neq y)$ for some sort symbol s of $\mathcal{L}$. Let T^+ be a Morita extension of T. In what follows, a variable of sort s in $\mathcal{L}$ will be denoted as x, y, x', y', etc. Variables of sort s^+ in $\mathcal{L}^+ - \mathcal{L}$ will be denoted x_+, y_+, etc.

Recall that the proof of Lemma 7.3 establishes that there is a uniform definable faithful inner model construction τ of models of T^+ from models of T. That is, whenever $\mathcal{M}$ is a model of T, $\mathcal{M}^\tau$ is a model of T^+. Conversely, there is a definable faithful uniform inner model construction γ such that for any model $\mathcal{M}$ of T, there is some model $\mathcal{M}^+$ of T^+ such that $(\mathcal{M}^+)^\gamma \cong \mathcal{M}$, which is just given by the $\mathcal{L}$-reduct of $\mathcal{M}^+$, which is definable in $\mathcal{M}^+$. Thus, to show that T and T^+ are bi-interpretable, it suffices to show that there is a definable map (in $\mathcal{M}$) that induces an isomorphism between $\mathcal{M}$ and $((\mathcal{M})^\tau)^\gamma$ and likewise, there is a definable map (in $\mathcal{M}^+$) inducing an isomorphism between $\mathcal{M}^+$ and $((\mathcal{M}^+)^\gamma)^\tau$. Note that $((\mathcal{M})^\tau)^\gamma$ is identical to $\mathcal{M}$, and so we will denote $((\mathcal{M})^\tau)^\gamma$ as $\mathcal{M}$. Similarly, $(\mathcal{M}^+)^\gamma)^\tau$ is identical to $\mathcal{M}^\tau$, and so we will denote $(\mathcal{M}^+)^\gamma)^\tau$ as $\mathcal{M}^\tau$.

Since $((\mathcal{M})^\tau)^\gamma$ is identical to $\mathcal{M}$, the $\mathcal{L}$-formulas of the form $x_s = y_s$, where s is a sort symbol in $\mathcal{L}$, define (in $\mathcal{M}$) bijections which induce an isomorphism between $\mathcal{M}$ and $((\mathcal{M})^\tau)^\gamma$.

Now we show that, for each sort $s \in \mathcal{L}^+$, there is a definable (in $\mathcal{M}^+$) bijection h_s between M_s^τ and M_s^+ such that the collection $(h_s : s \in \mathcal{L}^+)$ induces an isomorphism h between $\mathcal{M}^\tau$ and $\mathcal{M}^+$.

First, if s is a sort symbol in $\mathcal{L}$, then $x_s = y_s$ defines a bijection from M_s^τ to M_s^+. That is, h_s will be defined to be the identity map when s is an $\mathcal{L}$-sort. If s is a sort symbol s^+ in $\mathcal{L}^+ - \mathcal{L}$, then there are four cases depending on how T^+ defines s^+. For the sake of readability, I will often repeat the construction of τ:

Case 1. T^+ defines s^+ as a subsort corresponding to the $\mathcal{L}$-formula $\phi(x)$ with one variable x of sort s_1. Then, $\mathcal{M}^\tau$ is the $\mathcal{L}^+$-structure whose interpretation of s^+ in $\mathcal{M}^\tau$ is $\phi(\mathcal{M})$, i.e., the realizations of $\phi(x)$ in $\mathcal{M}$ (i.e., the $\mathcal{L}$-reduct of $\mathcal{M}^+$). Moreover, the interpretation of the inclusion map i is $\phi(x) \wedge x = y$.

Note that the $\mathcal{L}^+$-formula

$$\phi(x_1') \wedge (i(x_1') = y_+)$$

defines (in $\mathcal{M}^+$) a bijection $h_{s^+} : M_{s_1}^\tau \to M_{s^+}^+$. The bijections $(h_s : s \in \mathcal{L}^+)$ induce an isomorphism h from $\mathcal{M}^\tau$ to $\mathcal{M}^+$. Intuitively, h is the identity on $\mathcal{M}^\tau \restriction_{\mathcal{L}}$, and h defines a bijection on the realizations of ϕ in $\mathcal{M}^\tau$ to the domain of $M_{s^+}^+$.

https://doi.org/10.1017/S1755020319000303 Published online by Cambridge University Press

Case 2. T^+ defines s^+ as a product sort of s_1 and s_2, where s_1 and s_2 are sort symbols in $\mathcal{L}$. Then for any model $\mathcal{M}$ of T, $\mathcal{M}^\tau$ is the $\mathcal{L}^+$-structure whose interpretation of s^+ in $\mathcal{M}^\tau$ is the set of realizations (in $\mathcal{M}$) of the $\mathcal{L}$-formula $\phi(x_1, x_2) \equiv_{\text{def}} (x_1 = x_1 \wedge x_2 = x_2)$, and the interpretation (in $\mathcal{M}^\tau$) of π_1 and π_2 are given by the $\mathcal{L}$-formulas $\phi(x_1, x_2) \wedge x_1 = x$ and $\phi(x_1, x_2) \wedge x_2 = x$, respectively.

Note that the $\mathcal{L}^+$-formula

$$\exists y[(\pi_1(y) = x_1') \wedge (\pi_2(y) = x_2') \wedge (y = y^+)]$$

defines (in $\mathcal{M}^+$) a bijection of $h_{s^+} : M^\tau_{s_1} \times M^\tau_{s_2} \to M^+_s$. The maps which induces an isomorphism h between $\mathcal{M}^\tau$ and $\mathcal{M}^+$. Intuitively, h is the identity on $\mathcal{M}^\tau \restriction_{\mathcal{L}}$, and h maps ordered pairs (a, b) from $\mathcal{M}$ to the unique element c^+ in M_{s^+} such that $\pi_1(c^+) = a$ and $\pi_2(c^+) = b$.

Case 3. T^+ adds a quotient sort s^+ from an $\mathcal{L}$-formula $\phi(x_1, x_2)$, where x_1 and x_2 are variables of sort s_1 in $\mathcal{L}$. Then, $\mathcal{M}^\tau$ interprets s^+ as $M_1/\phi(\mathcal{M})$ (i.e., the quotient of M_1 under the equivalence relation given by $\phi(x, y)$ on M_1). Similarly, the interpretation of ϵ by $\mathcal{M}^\tau$ is given by the set of realizations of $\phi(x_1, x_2)$ in $\mathcal{M}$.
Note that the $\mathcal{L}^+$-formula

$$h^{-1}_{s^+}(x, y^+) \equiv_{\text{def}} \mu(x) = y^+$$

defines (in $\mathcal{M}^+$) a bijection from $M^\tau_{s^+}$ to $M^+_{s^+}$. Together, the bijections $\{h_s : s \in \mathcal{L}^+\}$ induce an isomorphism from $\mathcal{M}^\tau$ to $\mathcal{M}^+$. Intuitively, h is the identity on $\mathcal{M}^\tau \restriction_{\mathcal{L}}$ and h maps each $\phi(x, y)$-class $[a]$, where a is an element of M_1, to $\mu(a)$ which is an element of M_{s^+}. One easily checks that h is well-defined.

Case 4. T^+ defines s^+ as a coproduct sort of the $\mathcal{L}$-sorts s_1 and s_2. Let s_0 be the sort symbol in $\mathcal{L}$ such that $T \models \exists_{s_0} x \exists_{s_0} y(x \neq y)$. Then, $\mathcal{M}^\tau$ interprets s^+ as the product $M_{s_1} \times M_{s_2} \times M_{s_0} \times M_{s_0}$ modulo the equivalence relation given by the $\mathcal{L}$-formula:

$$(x_1 = x_1' \wedge z_0 = w_0 \wedge z_0' = w_0') \vee (y_2 = y_2' \wedge z_0 \neq w_0 \wedge z_0' \neq w_0').$$

Similarly, the projection maps ρ_1 of arity $s_1 \to s^+$ and ρ_2 of arity $s_2 \to s^+$ are definable by the $\mathcal{L}$-formulas $x_1 = x \wedge z_0 = w_0$ and $y_2 = y \wedge z_0 \neq w_0$, respectively.

Note that the $\mathcal{L}^+$-formula

$$\begin{aligned} h^{-1}_{s^+}(x_1', y_2', z_0, z_0', y^+) \equiv_{\text{def}} & [(x_1' = x_1' \wedge y_2' = y_2' \wedge z_0 = z_0') \to \rho(x_1') = y^+] \\ & \wedge [(x_1' = x_1' \wedge y_1' = y_1' \wedge z_0 \neq z_0') \to \rho(y_1') = y^+]. \end{aligned}$$

defines (in $\mathcal{M}^+$) a function of arity $(s_1 \times s_2 \times s_0 \times s_0) \to s^+$, which induces a bijection h_{s^+} from $\mathcal{M}^\tau_{s^+}$ to $\mathcal{M}^+_{s^+}$. The bijections $(h_s : s \in \mathcal{L}^+)$ induce an isomorphism $h : \mathcal{M}^\tau \to \mathcal{M}^+$. Intuitively, h is the identity on $\mathcal{M}^\tau \restriction_{\mathcal{L}}$ and h maps the $\sim$-class $[(a, b, c, d)]$, where (a, b, c, d) is a tuple of arity $M_1 \times M_2 \times M_0 \times M_0$, to $\rho_1(a)$ if $c = d$ and to $\rho_2(b)$ if $c \neq d$. $\square$

§8. Acknowledgments. I am grateful to Matteo Bianchetti, Patricia Blanchette, Michael Detlefsen, Sean Walsh, and two anonymous referees for their helpful comments and suggestions. I am particularly grateful to Hans Halvorson for his comments and for discovering a technical error in an earlier version of this article. Lastly, I especially thank Timothy Bays and Curtis Franks for their guidance and encouragement.

BIBLIOGRAPHY

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

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

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

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

Friedman, H. M. & Visser, A. (2014). When bi-interpretability implies synonymy. *Logic Group Preprint Series*, **320**, 1–19.

Glymour, C. (1971). Theoretical realism and theoretical equivalence. In Buck, R. C., and Cohen, R. S., editors. *PSA*: *Proceedings of the Biennial Meeting Of The Philosophy Of Science Association, Vol. 1970*. Dordrecht: Springer, pp. 275–288.

Hodges, W. (1993). *Model Theory*. Cambridge: Cambridge University Press.

Hodges, W. (2008). Tarski's theory of definition. In Patterson, D., editor. *New Essays On Tarski*, Chapter 5. Oxford: Oxford University Press, pp. 94–132.

Kaye, R. & Wong, T. L. (2007). On interpretations of arithmetic and set theory. *Notre Dame Journal of Formal Logic*, **48**(4), 497–510.

Makkai, M. & Reyes, G. E. (1967). *First Order Categorical Logic*. Berlin: Springer-Verlag.

Marker, D. (2002). *Model Theory*: *An Introduction*. New York: Springer Verlag.

Moosa, R. (2005). The model theory of compact complex spaces. In Matthias, B., Friedman, S.-D., and Krajícek, J., editors. *Logic Colloquium '01*. Lecture Notes in Logic, Vol. 20. Wellesley: A K Peters, Ltd./CRC Press, pp. 317–349.

Niebergall, K.-G. (2013). On the logic of reducibility: Axioms and examples. *Erkenntis*, **53**(1), 27–61.

Slaman, T. A. (2008). Global properties of the Turing degrees and the Turing jump. *Computational Prospects of Infinity. Part I. Tutorials*, **14**, 83–101.

Suppes, P. (1957). *Introduction to Logic*. New York: Van Nostrand Heinhold Company.

Szczerba, L. W. (1975). Interpretability of elementary theories. In Butts, R. E. and Hintikka, J., editors. *Logic, Foundations of Mathematics, and Computability Theory*. Dordrecht: D. Reidel Publishing Company, pp. 129–145.

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 Enayat, A., Kalantari, I., and Moniri, M., editors. *Logic in Tehran: Proceedings of the Workshop and Conference on Logic, Algebra, and Arithmetic, held October 18–22, 2003*, Lecture Notes in Logic, Volume 26. Cambridge: Cambridge University Press, pp. 284–341.

Visser, A. (2015). Extension & interpretability. Logic Group preprint series, volume 329. Available at https://dspace.library.uu.nl/handle/1874/327588.

Walsh, S. (2014). Logicism, interpretability, and knowledge of arithmetic. *The Review of Symbolic Logic*, **7**(1), 84–119.

DEPARTMENT OF PHILOSOPHY
UNIVERSITY OF NOTRE DAME
100 MALLOY HALL
NOTRE DAME, INDIANA 46556, USA
E-mail: pmceldow@nd.edu or pmceldow@gmail.com

https://doi.org/10.1017/S1755020319000303 Published online by Cambridge University Press