Notre Dame Journal of Formal Logic
Volume 59, Number 1, 2018

The Logical Strength of Compositional Principles

Richard G. Heck Jr.

Abstract This paper investigates a set of issues connected with the so-called *conservativeness argument against deflationism*. Although I do not defend that argument, I think the discussion of it has raised some interesting questions about whether what I call "compositional principles," such as "a conjunction is true iff its conjuncts are true," have substantial content or are in some sense logically trivial. The paper presents a series of results that purport to show that the compositional principles for a first-order language, taken together, have substantial logical strength, amounting to a kind of abstract consistency statement.

In 1998, Shapiro [29], and later, in 1999, Ketland [22], independently developed what is now known as the "conservativeness argument" against deflationary views of truth. Attempting to understand in what sense a deflationary truth-predicate is "insubstantial," they proposed that the principles concerning truth that a deflationist accepts should conservatively extend whatever nonsemantic theories the deflationist also accepts: no "insubstantial" theory of truth ought to allow us to prove things about nonsemantic matters that we cannot prove without it. In particular, the thought was, adding principles about truth to Peano arithmetic (PA) should yield a conservative extension of PA. And, indeed, if we add only what Horwich [21] called a "minimal" theory of truth, consisting simply of the T-sentences[1] for the language of arithmetic, then the result is indeed a conservative extension of PA.

By the time Shapiro and Ketland were writing, however, Gupta [12] had made it clear that the minimal theory of truth is too weak to do the work that even a deflationist needs truth to do. Generalizations about truth, such as:

(1) A conjunction is true if and only if both of its conjuncts are true,

are going to be required, as well. But if we add all of the various principles of that sort to PA—that is, if we add a theory of truth of the kind Tarski [31] showed us

Received September 4, 2014; accepted February 28, 2015
First published online July 5, 2017
2010 Mathematics Subject Classification: Primary 03A05
Keywords: truth, compositionality, Tarski, deflationism
 10.1215/00294527-2017-0011

how to formulate—then the result is not a conservative extension of PA, since the resulting theory proves that PA is consistent, via the following sort of argument. Every axiom of PA is true; the rules of inference preserve truth; so every theorem of PA is true; but there is at least one sentence that is not true, for example, $0 = 1$,[2] which is therefore not a theorem of PA; so PA is consistent.

Shapiro's version of the conservativeness argument attracted a direct response from Field, who had by then emerged as one of the leaders of the deflationist uprising. Field [9] notes that, if we *only* add "compositional principles" like (1) to PA, then the result is again a conservative extension (see [26, pp. 5–7]). It is only if we also extend the induction scheme to permit semantic vocabulary that we get a nonconservative extension (see also [16]). And so Field writes:

> Since truth can be added in ways that produce a conservative extension. . . , there is no need to disagree with Shapiro when he says that "conservativeness is essential to deflationism". . . . Shapiro's position, however, is that a deflationist must hold that adding 'true' to number theory in the full-blooded way that involves [extending the induction axioms also] produces a conservative extension. ([9, p. 536])

Field then goes on to argue that a deflationist need hold no such thing. At most, the deflationist should hold that the principles about truth that "flow from its disquotational nature" are conservative over number theory; she need not hold that *all* principles about truth are conservative. But, Field claims, the induction principles flow not from the nature of truth, but from the nature of the natural numbers. They are not semantical but arithmetical in character, so whether adding them yields a conservative extension is irrelevant to the issue at hand.

In what seems to me to be the crucial passage, Field quotes Shapiro [29, p. 499] as asking: "How thin can the notion of arithmetical truth be if, by invoking it, we can learn more about the natural numbers?" Field then replies:

> . . . [T]he way in which we "learn more about the natural numbers by invoking truth" is that in having that notion we can rigorously formulate a more powerful arithmetical theory than we could rigorously formulate before. There is nothing very special about truth here: using any other notion not expressible in the original language we can get new instances of induction, and in many cases these lead to nonconservative extensions. ([9, p. 536])

This is right, so far as it goes, but it is also extremely misleading.

What does Field mean by "using [a] notion not expressible in the original language"? The natural way to read him would be as talking about definability: about what happens if we add a new predicate whose extension is not definable in the original language. In that case, Field would be saying something like this:

> If we add a new predicate whose extension is not definable in the original language, then we will get new instances of induction, which may lead to new theorems in the original language.

That is of course right. We *will* get new instances of induction that *may* lead to new theorems. But the case of the truth-predicate is precisely not one of those cases. Tarski's theorem tells us that the set of truths of the language of PA is not definable in the language of arithmetic. But if we add a truth-predicate $\mathsf{T}(x)$ to the language of PA and extend PA by adding the T-sentences, then that is enough to guarantee that $\mathsf{T}(x)$ defines a set not definable in the original language, namely: the set of true sentences of the language of arithmetic. But the result is still a conservative extension of PA even if we extend induction. It follows that the nonconservativity result is *not*

due just to the presence of "new instances of induction" formulated using a "notion not expressible in the original language." It is also necessary that we have a fully compositional truth-theory, and not just the T-sentences.

So, again: It is only if we *both* add a compositional truth-theory to PA *and* extend the induction axioms to permit semantic vocabulary that we get a nonconservative extension. Neither is sufficient by itself. That makes the dialectical situation complicated. Field wants to allow that compositional principles about truth "flow from truth's disquotational nature" and so should be conservative;[3] he can do so by blaming the nonconservativity result on the extension of induction. Shapiro, by contrast, blames the nonconservativity result on the compositional principles, because he thinks we are independently committed to induction for any well-defined predicate we can understand. Are we, then, at a standoff?

Not necessarily. As we shall see, the situation is not entirely symmetrical. Adding a compositional truth-theory does yield a conservative extension if we do not extend the induction axioms, but the resulting theory is, in most cases, still logically stronger than the original theory. On the other hand, if we just add the T-sentences, then, in almost all cases, the resulting theory is, in a well-defined sense, *not* logically stronger than the original theory even if we *do* extend the induction axioms. And that already seems to me to be out of the spirit of deflationism. Deflationists routinely deride compositional principles like (1) as trivial (see [10, p. 24]) and of "no interest in their own right" (see [8, p. 269]). In fact, however, such principles, taken together, have significant logical strength, independent of the extension of the induction axioms, and for reasons that are closely connected with the sort of consistency proof on which the conservativeness argument was originally based.

In the remainder of this article, the results to which I have just alluded will be stated precisely and, in some cases, proved.[4] In Section 1, I shall present the background material from logic that is necessary for the rest of the discussion. In Section 2, I shall present a first form of the results. In Section 3, we shall encounter two natural worries about the significance of those results. In Section 4, I will present the results in a different form, one that should assuage such concerns. There is a different worry about that form of the results, however, which we shall discuss in Section 5. I shall close, in Section 6, by returning to the philosophical issues we have just been discussing and explaining how I think the technical results presented bear upon them.

1 Preliminaries

In an effort to make what follows as widely accessible as possible, I will first present a brief overview of the machinery from logic that we will be using.

The *theories* we will be discussing will all be recursively axiomatized. And since we will be discussing consistency statements, we need, for reasons famously made clear by Feferman [7], to think of theories *intensionally*: not as sets of theorems, nor even as sets of axioms, but as particular presentations of sets of axioms. Officially, we identify a theory with a formula that is true of (the Gödel numbers of) its axioms. Where we are dealing with finitely axiomatized theories, we shall assume that their axioms are presented in the simplest possible way: as a list or, if you prefer, a disjunction.

A theory is "stated in" a *language*. The languages in which we will be interested here are first-order languages, constructed from terms, function-symbols, and

predicate-letters in the usual way. These languages are assumed to be finite, in the sense that they have only finitely many atomic expressions. It is convenient to identify a language with the set of its atomic expressions, together with some indication of their logical type, that is, with what is sometimes called the *signature* of the language.

1.1 Interpretatability There are a number of ways of comparing the logical strength of theories. If the theories are stated in the same language, then the obvious question is whether one proves all the results the other proves. Comparison is more difficult when the theories are stated in different languages. In that case, the theories will trivially prove different theorems: if A is in the language of the one but not the other, then $\ulcorner A \vee \neg A \urcorner$ will be a theorem of the one but not of the other.

If the language of one theory contains that of the other, then one way to compare them is to ask if the first is a "conservative extension" of the second, that is, whether the theory in the extended language proves any new theorems that can be stated in the original language.[5] But even this fails if the theories are not so related. In that case, the established method of comparison uses the notion of interpretation, which was first explored in a systematic way by Tarski [30], although the basic idea is much older.

Let theories $\mathcal{B}$ (for "base") and $\mathcal{T}$ (for "target") be given, stated in languages $\mathcal{L}_\mathcal{B}$ and $\mathcal{L}_\mathcal{T}$, respectively. A *relative interpretation*[6] of $\mathcal{T}$ in $\mathcal{B}$ consists of two parts: a translation of $\mathcal{L}_\mathcal{T}$ into $\mathcal{L}_\mathcal{B}$, and proofs in $\mathcal{B}$ of the translations of the axioms of $\mathcal{T}$. The translation is compositional, in the sense that the only thing we actually need to do is define the (nonlogical) atomic expressions of $\mathcal{L}_\mathcal{T}$ in terms of those of $\mathcal{L}_\mathcal{B}$ and to specify a "domain" for the interpretation in terms of a formula $\delta(x)$ of $\mathcal{L}_\mathcal{B}$. This can then be extended to a complete translation of $\mathcal{L}_\mathcal{T}$ into $\mathcal{L}_\mathcal{B}$ in the obvious way, where quantifiers are "relativized" to $\delta(x)$: $\forall x(\phi(x))$ is translated as: $\forall x(\delta(x) \to \phi^*(x))$, where $\phi^*(x)$ translates $\phi(x)$; $\exists x(\phi(x))$, as: $\exists x(\delta(x) \wedge \phi^*(x))$.[7]

Note that interpretability is transitive and reflexive.

If $\mathcal{T}$ is relatively interpretable in $\mathcal{B}$, then it follows that, if $\mathcal{B}$ is consistent, so is $\mathcal{T}$. If a contradiction could be derived from the axioms of $\mathcal{T}$, that proof could be mimicked in $\mathcal{B}$: Just prove the translations of the axioms of $\mathcal{T}$ used in the proof of the contradiction, then append a modified version of the proof given in $\mathcal{T}$. Indeed, quite generally, if $\vdash_\mathcal{T} A$, then $\vdash_\mathcal{B} A^*$, where, again, the asterisk means: translation of. Moreover, if $\mathcal{B}$ and $\mathcal{T}$ are not *too* terribly weak,[8] then all of this will be provable in $\mathcal{B}$ and $\mathcal{T}$ themselves. So, in particular, $\mathcal{T}$ will prove $\mathsf{Con}(\mathcal{B}) \to \mathsf{Con}(\mathcal{T})$ and so, by the second incompleteness theorem, cannot prove $\mathsf{Con}(\mathcal{B})$.[9] By contrast, $\mathcal{B}$ perfectly well could prove $\mathsf{Con}(\mathcal{T})$.

One way to give content to the idea that $\mathcal{B}$ is at least as strong as $\mathcal{T}$ is therefore to take it to mean: $\mathcal{T}$ is relatively interpretable in $\mathcal{B}$. That this is a useful way to make the intuitive idea of relative strength rigorous emerged only after a good deal of hard work, beginning with Tarski, Mostowski, and Robinson [32] and continuing through work by Feferman [7] to the present day. And, while the notion of interpretation is particularly useful when we are dealing with theories stated in different languages, we can still ask whether $\mathcal{T}$ can be interpreted in $\mathcal{B}$ even when $\mathcal{L}_\mathcal{T}$ and $\mathcal{L}_\mathcal{B}$ are the same: the interpretation of the atomic vocabulary does not have to be the identity function. So the notion is very general, and it links up nicely with what we know about the strength of theories from the second incompleteness theorem.

Now, a couple definitions that apply (sensibly) only to nonfinitely axiomatized theories.

Definition $\mathcal{T}$ is *locally interpretable* in $\mathcal{B}$ if every finite subset of $\mathcal{T}$ is interpretable in $\mathcal{B}$.

Local interpretability obviously follows from "global" interpretability, but not conversely. Local interpretability is also transitive and reflexive, and it relates to relative consistency just as global interpretability does: if $\mathcal{T}$ is locally interpretable in $\mathcal{B}$, then $\mathcal{T}$ is consistent if $\mathcal{B}$ is. The reason is that any proof of a contradiction in $\mathcal{T}$ will use only finitely many of $\mathcal{T}$'s axioms.

Definition $\mathcal{T}$ is *reflexive* if $\mathcal{T}$ proves the consistency of each of its finite subtheories. That is, for each finite $\mathcal{U} \subseteq \mathcal{T}$, $\mathcal{T} \vdash \mathsf{Con}(\mathcal{U})$.

A theory's being reflexive can cause all sorts of unexpected phenomena as regards the interpretability of other theories in it (see [7]). What will matter most to us here is the fact that reflexive theories collapse the distinction between local and global interpretability.[10]

Theorem (Orey's compactness theorem) *Suppose that $\mathcal{T}$ is locally interpretable in $\mathcal{B}$ and that $\mathcal{B}$ is reflexive. Then $\mathcal{T}$ is globally interpretable in $\mathcal{B}$.*

Since PA is reflexive (see Mostowski [24]), then we can expect PA to be something of a special case. Which, indeed, we shall see that it is.

1.2 Fragments of arithmetic We shall mostly be concerned here with PA and certain of its subtheories.

Robinson arithmetic, or Q, is the theory whose axioms are the universal closures of the following eight formulae:

Q1: $x \neq 0$,
Q2: $\mathsf{S}x = \mathsf{S}y \rightarrow x = y$,
Q3: $x + 0 = x$,
Q4: $x + \mathsf{S}y = \mathsf{S}(x + y)$,
Q5: $x \times 0 = 0$,
Q6: $x \times \mathsf{S}y = (x \times y) + x$,
Q7: $x \neq 0 \rightarrow \exists y(x = \mathsf{S}y)$,
Q8: $x < y \equiv \exists z(y = \mathsf{S}z + x)$.

The last is often considered a definition of $<$; it is convenient in the present context to regard $<$ as just part of the language. The language of Q, $\{0, \mathsf{S}, +, \times, <\}$, is what we call the "language of arithmetic" and denote by $\mathcal{A}$.

A formula is said to be Δ_0 if all quantifiers contained in it are "bounded," that is, if all of its quantified subformulae are of the form $\forall x(x < t \rightarrow \cdots)$ or $\exists x(x < t \wedge \cdots)$, where t is a term.[11] A formula is Σ_1 (resp., Π_1) if it is of the form $\exists x_1 \ldots \exists x_n(\phi)$ (resp., $\forall x_1 \ldots \forall x_n(\phi)$), where ϕ is Δ_0. A formula is Σ_n (resp., Π_n) if it is $\exists x_1 \ldots \exists x_n(\phi)$ (resp., $\forall x_1 \ldots \forall x_n(\phi)$), where ϕ is Π_{n-1} (resp., Σ_{n-1}). A formula A is Σ_n in a theory $\mathcal{T}$ if $\mathcal{T}$ proves $A \equiv \phi$, for some Σ_n formula ϕ, and similarly for other notions.

An important class of subtheories of PA is characterized in terms of the induction axioms these theories permit. PA itself is Q plus the full induction scheme:

$$A(0) \wedge \forall x\big(A(x) \rightarrow A(\mathsf{S}x)\big) \rightarrow \forall x\big(A(x)\big),$$

where $A(x)$ is any formula at all.[12] The theory $\mathsf{I}\Theta$ is Q plus induction for formulae in the set Θ: so $A(x)$ has to be in Θ. Thus, $\mathsf{I}\Delta_0$ is Q plus induction for Δ_0 formulae; $\mathsf{I}\Sigma_1$ is Q plus induction for Σ_1 formulae; $\mathsf{I}\Sigma_n$ is Q plus induction for Σ_n formulae. Note that $\mathsf{I}\Sigma_n$, though not finitely axiomati*zed*, as I have described it, is finitely axiomati*zable* (see Hájek and Pudlák [13, pp. 77ff]). We assume a finite axiomatization. It is not known whether $\mathsf{I}\Delta_0$ is finitely axiomatizable.

$\mathsf{I}\Delta_0$ is in one sense clearly stronger than Q: it proves lots of important generalizations about the natural numbers that Q does not, such as $x \neq \mathsf{S}x$. But in another sense $\mathsf{I}\Delta_0$ is still very weak: It is interpretable in Q.[13] Another respect in which $\mathsf{I}\Delta_0$ is weak is that, although one can define the relation $y = 2^x$ by means of a Δ_0-formula $\mathsf{exp}(x, y)$, we cannot prove in $\mathsf{I}\Delta_0$ that exponentiation is total: $\forall x \exists y(\mathsf{exp}(x, y))$. The obvious proof uses induction on $\exists y(\mathsf{exp}(x, y))$, which is Σ_1. But for that very reason, the totality of exponentiation is provable in $\mathsf{I}\Sigma_1$, as is the totality of every other primitive recursive function. So $\mathsf{I}\Sigma_1$ is much stronger than $\mathsf{I}\Delta_0$: Indeed, $\mathsf{I}\Sigma_1$ proves $\mathsf{Con}(\mathsf{I}\Delta_0)$.

2 The Logical Strength of Compositional Principles

In this section, I present and discuss the technical results to which I alluded at the end of the introduction. First, we need to talk about exactly what a theory of truth is.

2.1 Theories of truth Since the semantic axioms for the quantifiers, as Tarski formulated them, make use of sequences of elements from the domain, we shall need a nice theory of sequences if we are to formalize theories of truth. Technically, we will need our base theory to be *sequential*,[14] which essentially means that it can code (and decode) finite sequences of its elements. Q is not sequential, but there are lots of sequential theories that are interpretable in Q. For example, $\mathsf{I}\Delta_0$ is sequential, and it is interpretable in Q. We can also just add a simple theory of sequences to Q to get a new theory, which we might call Q_{seq}, which is also interpretable in Q and is sequential by construction. We will assume something like that done.

A *compositional theory of truth* consists of Tarski-style axioms for the logical and nonlogical vocabulary. The axioms for the logical part of the language will always be the same:

(v) $\mathsf{Den}_\sigma(v_i, x) \equiv \mathsf{val}(\sigma, i) = x$, where v_i is the ith variable,
($=$) $\mathsf{Sat}_\sigma(\ulcorner t = u \urcorner) \equiv \exists x \exists y[\mathsf{Den}_\sigma(t, x) \wedge \mathsf{Den}_\sigma(u, y) \wedge x = y]$,
($\neg$) $\mathsf{Sat}_\sigma(\ulcorner \neg A \urcorner) \equiv \neg \mathsf{Sat}_\sigma(A)$,
($\wedge$) $\mathsf{Sat}_\sigma(\ulcorner A \wedge B \urcorner) \equiv \mathsf{Sat}_\sigma(A) \wedge \mathsf{Sat}_\sigma(B)$,
($\forall$) $\mathsf{Sat}_\sigma(\ulcorner \forall v_i (A(v_i)) \urcorner) \equiv \forall \tau[\tau \overset{i}{\sim} \sigma \rightarrow \mathsf{Sat}_\sigma(\ulcorner A(v_i) \urcorner)]$.

There are similar clauses for the other logical constants. Here, $\mathsf{val}(\sigma, i)$ means the value that σ assigns to the ith variable;[15] $\mathsf{Den}_\sigma(t, x)$ means that t denotes x with respect to the sequence σ; $\mathsf{Sat}_\sigma(A)$ means that σ satisfies A; and $\tau \overset{i}{\sim} \sigma$ means that τ and σ agree on what they assign to each variable, with the possible exception of v_i; that is,

$$\forall k < \mathsf{lh}(\sigma)\big[k \neq i \rightarrow \mathsf{val}(\sigma, k) = \mathsf{val}(\tau, k)\big],$$

where $\mathsf{lh}(\sigma)$ is the length of the sequence σ.

In the case of the language of arithmetic, we will also have these axioms for the nonlogical constants:

(0) $\mathsf{Den}_\sigma(\ulcorner 0 \urcorner, x) \equiv x = 0$,

(S) $\mathsf{Den}_\sigma(\ulcorner \mathsf{S}t \urcorner, x) \equiv \exists y(\mathsf{Den}_\sigma(t, y) \wedge y = \mathsf{S}x)$,
$(+)$ $\mathsf{Den}_\sigma(\ulcorner t + u \urcorner, x) \equiv \exists y \exists z[\mathsf{Den}_\sigma(t, y) \wedge \mathsf{Den}_\sigma(u, z) \wedge x = y + z]$,
$(\times)$ $\mathsf{Den}_\sigma(\ulcorner t \times u \urcorner, x) \equiv \exists y \exists z[\mathsf{Den}_\sigma(t, y) \wedge \mathsf{Den}_\sigma(u, z) \wedge x = y \times z]$,
$(<)$ $\mathsf{Sat}_\sigma(\ulcorner t < u \urcorner) \equiv \exists y \exists z[\mathsf{Den}_\sigma(t, y) \wedge \mathsf{Den}_\sigma(u, z) \wedge y < z]$.

The pattern should be clear.[16]

Finally, then, we need to define the notion of truth itself:

(T) $\mathsf{T}(A) \equiv A$ is a sentence $\wedge\ \forall \sigma(\mathsf{Sat}_\sigma(A))$.

That is Tarski's definition.

So that is what a theory of truth is. Now for some notation.

Definition Let $\mathcal{T}$ be sequential. Then

(1) $\mathsf{DT}^-[\mathcal{T}]$ is $\mathcal{T}$ plus all T-sentences for the language of $\mathcal{T}$;
(2) $\mathsf{DS}^-[\mathcal{T}]$ is the result of adding not just the T-sentences for the language of $\mathcal{T}$ but also the Sat-sentences, such as:

$$\mathsf{Sat}_\sigma(v_0 = v_1) \equiv \mathsf{val}(\sigma, 0) = \mathsf{val}(\sigma, 1);$$

(3) $\mathsf{CT}^-[\mathcal{T}]$ is the theory that extends $\mathcal{T}$ by adding the truth-theoretic axioms just described for the logical and nonlogical vocabulary of the language of $\mathcal{T}$.

Here, DT stands for *disquotational truth*, DS for *disquotational satisfaction*, and CT for *compositional truth*.

Note that none of these theories extends any induction scheme that might be present in $\mathcal{T}$. There is no real chance, then, that even $\mathsf{CT}^-[\mathcal{T}]$ is going to prove the consistency of $\mathcal{T}$. One might therefore suspect that $\mathsf{CT}^-[\mathcal{T}]$ would logically be no stronger than $\mathcal{T}$. If so, then, as we shall see, one would suspect wrongly, at least in general.

We shall also be interested in theories that do extend the induction scheme.

Definition Suppose that $\mathcal{T}$ is among $\mathsf{I}\Delta_0$, $\mathsf{I}\Sigma_n$, and so forth. Then

(1) $\mathsf{DT}[\mathcal{T}]$ is like $\mathsf{DT}^-[\mathcal{T}]$ except that it extends the induction scheme to permit semantic vocabulary;
(2) $\mathsf{DS}[\mathcal{T}]$ is like $\mathsf{DS}^-[\mathcal{T}]$ except that it extends the induction scheme;
(3) $\mathsf{CT}[\mathcal{T}]$ is the result of adding a fully compositional truth-theory and extending the induction scheme.

To be frank, it is not at all obvious, in general, what it means to "extend the induction scheme." But in the cases in which we shall be interested, it is obvious enough: One simply treats the semantic vocabulary as being among the primitives of the language.[17] So, for example, $\exists x(\mathsf{Den}_\sigma(t, x))$ counts as Σ_1, and $\forall x \exists \sigma(\mathsf{Den}_\sigma(t, x))$ counts as Π_2.

2.2 Induction versus the compositional principles We are now ready to state—and in some cases prove—the results I have been promising. We will begin by exploring the various disquotational theories.

As noted earlier, $\mathsf{DT}[\mathsf{PA}]$ is a conservative extension of PA. Here are some similar results, but stated in terms of interpretability.[18]

Theorem 2.1 $\mathsf{DT}^-[\mathcal{T}]$ *is locally interpretable in* $\mathcal{T}$.

Proof Let S be a finite subset of the axioms of $\mathsf{DT}^-[\mathcal{T}]$. Then S will contain at most finitely many T-sentences, say, for $A_1, \ldots, A_n$. We interpret $\mathsf{T}(x)$ in terms of a "listlike" theory of truth, that is, as

$$(x = \ulcorner A_1 \urcorner \wedge A_1) \vee \cdots \vee (x = \ulcorner A_n \urcorner \wedge A_n).$$

With $\mathsf{T}(x)$ so defined, $\mathcal{T}$ will prove the T-sentences for $A_1, \ldots, A_n$. For example, the translation of the T-sentence for A_1 is

$$(\ulcorner A_1 \urcorner = \ulcorner A_1 \urcorner \wedge A_1) \vee \cdots \vee (\ulcorner A_1 \urcorner = \ulcorner A_n \urcorner \wedge A_n) \equiv A_1.$$

But the first conjunct of the first disjunct is provable, and the first conjunct of the other disjuncts is refutable. So this is provably equivalent to $A_1 \equiv A_1$, and so is itself provable. □

This result extends smoothly to the case of satisfaction.

Theorem 2.2 $\mathsf{DS}^-[\mathcal{T}]$ *is locally interpretable in* $\mathcal{T}$.

Proof Essentially the same proof works as in the case of Theorem 2.1. To interpret the Sat-sentences for $A_1(\mathrm{v}_1, \mathrm{v}_2)$ and $A_2(\mathrm{v}_2, \mathrm{v}_3)$, say, simply define $\mathsf{Sat}_\sigma(x)$ as

$$\big[x = \ulcorner A_1(\mathrm{v}_1, \mathrm{v}_2) \urcorner \wedge A_1(\mathsf{val}(\sigma, 1), \mathsf{val}(\sigma, 2))\big] \vee$$
$$\big[x = \ulcorner A_2(\mathrm{v}_1, \mathrm{v}_2) \urcorner \wedge A_2(\mathsf{val}(\sigma, 2), \mathsf{val}(\sigma, 3))\big].$$

Then the translation of the Sat-sentence for $A_1(\mathrm{v}_1, \mathrm{v}_2)$ is

$$\big[\ulcorner A_1(\mathrm{v}_1, \mathrm{v}_2) \urcorner = \ulcorner A_1(\mathrm{v}_1, \mathrm{v}_2) \urcorner \wedge A_1(\mathsf{val}(\sigma, 1), \mathsf{val}(\sigma, 2))\big] \vee$$
$$\big[\ulcorner A_1(\mathrm{v}_1, \mathrm{v}_2) \urcorner = \ulcorner A_2(\mathrm{v}_1, \mathrm{v}_2) \urcorner \wedge A_2(\mathsf{val}(\sigma, 2), \mathsf{val}(\sigma, 3)\big] \equiv$$
$$A_1(\mathsf{val}(\sigma, 1), \mathsf{val}(\sigma, 2))],$$

which is again provable. □

Theorem 2.3 $\mathsf{DT}[\mathsf{PA}]$ *is interpretable in* PA*, and so is* $\mathsf{DS}[\mathsf{PA}]$.

Proof We will prove the more inclusive case. Let S be a finite subset of the axioms of $\mathsf{DS}[\mathsf{PA}]$. Then S will contain at most finitely many Sat-sentences. We interpret $\mathsf{Sat}_\sigma(x)$ as in the previous proof. So those Sat-sentences are all provable. But S may also contain some extended induction axioms. However, under our definition of $\mathsf{Sat}_\sigma(x)$, those induction axioms simply become induction axioms of PA.

So $\mathsf{DS}[\mathsf{PA}]$ is locally interpretable in PA. Orey's compactness theorem then implies that $\mathsf{DS}[\mathsf{PA}]$ is *globally* interpretable in PA. □

The proof of Theorem 2.3 does not extend to subsystems of PA such as $\mathsf{I}\Sigma_1$. The reason is that the A_i's may be of any complexity, and so, if we have an induction axiom for some Σ_1-formula $A(x)$ containing semantic vocabulary, the result of replacing $\mathsf{T}(x)$ or $\mathsf{Sat}_\sigma(x)$ by its "listlike" definition in $A(x)$ may yield a formula that is not itself Σ_1. But there is a slightly more complicated proof that does work in the case of truth.

Theorem 2.4 $\mathsf{DT}[\mathsf{I}\Sigma_n]$ *is locally interpretable in* $\mathsf{I}\Sigma_n$.

Proof Let S be a finite subset of the axioms of $\mathsf{DT}[\mathsf{I}\Sigma_n]$. Then S contains only finitely many T-sentences. For illustration, say these are for A and B. As before, we interpret $\mathsf{T}(x)$ as: $(x = \ulcorner A \urcorner \wedge A) \vee (x = \ulcorner B \urcorner \wedge B)$. We can then easily prove the T-sentences for A and B. But, of course, S may also contain some extended

induction axioms from $\mathsf{DT}[\mathsf{I}\Sigma_n]$. We need to see that these are also going to be provable.

Suppose that one of these induction axioms is the axiom for the formula $\phi(x) \vee \mathsf{T}(\mathsf{sb}(\ulcorner\psi(x)\urcorner, x))$, where $\phi(x)$ is Σ_n but $\psi(x)$ need not be. Here, $\mathsf{sb}(y, x)$ means: the result of substituting the numeral for x for the sole free variable in y. I choose this example because the threat is that the ability to substitute in this way will allow us to get the induction axiom for $\phi(x) \vee \psi(x)$, which need not be Σ_n. But, in fact, the threat is idle, because the induction axiom for the mentioned formula

$$\begin{aligned}\big[\phi(0) \vee \mathsf{T}\big(\mathsf{sb}\big(\ulcorner\psi(x)\urcorner, 0\big)\big)\big] \wedge \\ \forall x\big[\phi(x) \vee \mathsf{T}\big(\mathsf{sb}\big(\ulcorner\psi(x)\urcorner, x\big)\big) \to \phi(\mathsf{S}x) \vee \mathsf{T}\big(\mathsf{sb}\big(\ulcorner\psi(x)\urcorner, \mathsf{S}x\big)\big)\big] \to \\ \forall x\big(\phi(x) \vee \mathsf{T}\big(\mathsf{sb}\big(\ulcorner\psi(x)\urcorner, x\big)\big)\big)\end{aligned}$$

can be proven under our interpretation of $\mathsf{T}(x)$. I am claiming, that is, that we can prove

$$\begin{aligned}\big[\phi(0) \vee \big(\mathsf{sb}\big(\ulcorner\psi(x)\urcorner, 0\big) = \ulcorner A\urcorner \wedge A\big) \vee \big(\mathsf{sb}\big(\ulcorner\psi(x)\urcorner, 0\big) = \ulcorner B\urcorner \wedge B\big)\big] \wedge \\ \forall x\big[\phi(x) \vee \big(\mathsf{sb}\big(\ulcorner\psi(x)\urcorner, x\big) = \ulcorner A\urcorner \wedge A\big) \vee \big(\mathsf{sb}\big(\ulcorner\psi(x)\urcorner, x\big) = \ulcorner B\urcorner \wedge B\big) \to \\ \phi(\mathsf{S}x) \vee \big(\mathsf{sb}\big(\ulcorner\psi(x)\urcorner, \mathsf{S}x\big) = \ulcorner A\urcorner \wedge A\big) \vee \big(\mathsf{sb}\big(\ulcorner\psi(x)\urcorner, \mathsf{S}x\big) = \ulcorner B\urcorner \wedge B\big)\big] \to \\ \forall x\big[\phi(x) \vee \big(\mathsf{sb}\big(\ulcorner\psi(x)\urcorner, x\big) = \ulcorner A\urcorner \wedge A\big) \vee \big(\mathsf{sb}\big(\ulcorner\psi(x)\urcorner, x\big) = \ulcorner B\urcorner \wedge B\big)\big]. \quad (2.1)\end{aligned}$$

(Sorry about that.) The crucial point is that A and B are *sentences*, so the quantifier $\forall x$ cannot bind any variables in A or B. Hence, they can be "pulled out" in the following way.

Abbreviate (2.1) as $\Phi(A, B)$. Then it is logically equivalent to

$$\begin{aligned}\big[A \wedge B \to \Phi(\top, \top)\big] \wedge \big[A \wedge \neg B \to \Phi(\top, \bot)\big] \wedge \\ \big[\neg A \wedge B \to \Phi(\bot, \top)\big] \wedge \big[\neg A \wedge \neg B \to \Phi(\bot, \bot)\big], \quad (2.2)\end{aligned}$$

where $\top$ is $0 = 0$ and $\bot$ is $0 \neq 0$. By completeness, this equivalence is provable. Now $\Phi(\top, \top)$ is

$$\begin{aligned}\big[\phi(0) \vee \big(\mathsf{sb}\big(\ulcorner\psi(x)\urcorner, 0\big) = \ulcorner A\urcorner \wedge \top\big) \vee \phi(0) \vee \big(\mathsf{sb}\big(\ulcorner\psi(x)\urcorner, 0\big) = \ulcorner B\urcorner \wedge \top\big)\big] \wedge \\ \forall x\big[\phi(x) \vee \big(\mathsf{sb}\big(\ulcorner\psi(x)\urcorner, x\big) = \ulcorner A\urcorner \wedge \top\big) \vee \phi(x) \vee \big(\mathsf{sb}\big(\ulcorner\psi(x)\urcorner, x\big) = \ulcorner B\urcorner \wedge \top\big) \to \\ \phi(\mathsf{S}x) \vee \big(\mathsf{sb}\big(\ulcorner\psi(x)\urcorner, \mathsf{S}x\big) = \ulcorner A\urcorner \wedge \top\big) \vee \phi(\mathsf{S}x) \vee \big(\mathsf{sb}\big(\ulcorner\psi(x)\urcorner, \mathsf{S}x\big) = \ulcorner B\urcorner \wedge \top\big)\big] \to \\ \forall x\big[\phi(x) \vee \big(\mathsf{sb}\big(\ulcorner\psi(x)\urcorner, x\big) = \ulcorner A\urcorner \wedge \top\big) \vee \phi(x) \vee \big(\mathsf{sb}\big(\ulcorner\psi(x)\urcorner, x\big) = \ulcorner B\urcorner \wedge \top\big)\big]\end{aligned}$$

and that is the induction axiom for the formula

$$\phi(x) \vee \big(\mathsf{sb}\big(\ulcorner\psi(x)\urcorner, x\big) = \ulcorner A\urcorner \wedge \top\big) \vee \phi(x) \vee \big(\mathsf{sb}\big(\ulcorner\psi(x)\urcorner, x\big) = \ulcorner B\urcorner \wedge \top\big),$$

which is Σ_n. So $\Phi(\top, \top)$ is provable, and hence so is $A \wedge B \to \Phi(\top, \top)$. The same goes for the other cases. So (2.2) is provable; so (2.1) is provable.

Nothing hinges on the details of this particular example. □

We thus see that, for theories $\mathcal{T}$ in the usual arithmetical hierarchy—Q, $\mathsf{I}\Delta_0$, $\mathsf{I}\Sigma_n$, PA—the deflationary theory $\mathsf{DT}[\mathcal{T}]$ is always locally interpretable in $\mathcal{T}$.

The situation with compositional truth-theories is different.

Theorem 2.5 *Suppose that $\mathcal{T} \supseteq \mathsf{Q}$ is sequential and finitely axiomatized. Then $\mathsf{CT}^-[\mathcal{T}]$ interprets $\mathsf{Q} + \mathsf{Con}(\mathcal{T})$.*

It follows that, if $\mathcal{T}$ is finitely axiomatized, then $\mathsf{CT}^-[\mathcal{T}]$ is logically stronger than $\mathcal{T}$. This is a consequence of a beautiful form of the second incompleteness theorem due to Pudlák.

Theorem 2.6 ([27, Corollary 3.5]) *Suppose that $\mathcal{T}$ is finitely axiomatized, sequential, and consistent. Then $\mathcal{T}$ does not interpret $\mathsf{Q} + \mathsf{Con}(\mathcal{T})$.*

The original form of the second incompleteness theorem says merely that $\mathcal{T}$ does not *prove* $\mathsf{Con}(\mathcal{T})$. Building on earlier work by Feferman [7, Theorem 6.5], however, Pudlák improves on Gödel by showing that $\mathcal{T}$ cannot even *interpret* $\mathcal{T} + \mathsf{Con}(\mathcal{T})$, or even $\mathsf{Q} + \mathsf{Con}(\mathcal{T})$. So, we have the following.

Corollary 2.7 *Suppose that $\mathcal{T} \supseteq \mathsf{Q}$ is sequential and finitely axiomatized. Then $\mathsf{CT}^-[\mathcal{T}]$ is not interpretable in $\mathcal{T}$. In particular, for no $n \geq 1$ does $\mathsf{I}\Sigma_n$ interpret $\mathsf{CT}^-[\mathsf{I}\Sigma_n]$.*

Proof If $\mathsf{I}\Sigma_n$ interpreted $\mathsf{CT}^-[\mathsf{I}\Sigma_n]$, then, since $\mathsf{CT}^-[\mathsf{I}\Sigma_n]$ interprets $\mathsf{Q}+\mathsf{Con}(\mathsf{I}\Sigma_n)$, so would $\mathsf{I}\Sigma_n$, contradicting Theorem 2.6—on the assumption, of course, that $\mathsf{I}\Sigma_n$ is consistent. □

The key result here is obviously Theorem 2.5. The proof of it turns out to be quite messy and so is presented elsewhere (see [18, Section 3.2]).[19] But certain features of the proof will be important below, and the basic idea behind it is easy enough to explain. As said above, there is no real hope that $\mathsf{CT}^-[\mathcal{T}]$ will prove $\mathsf{Con}(\mathcal{T})$, since whatever induction axioms might be present in $\mathcal{T}$ have not been extended. But it turns out that we can get very close.

One argument for $\mathsf{Con}(\mathcal{T})$ would proceed as follows.

> Call a proof *good* if all of its lines are true. Proofs with 0 lines are trivially good. So suppose that n line proofs are good, and consider some $n + 1$ line proof. If the last line is an axiom, then it is true, since all $\mathcal{T}$'s axioms are true and all the logical axioms are true, too. If it is not an axiom, then it must follow by one of the rules of inference from some of the earlier lines. But those lines are true, by the induction hypothesis, and the rules of inference preserve truth, so the last line is again true. Hence, *by induction*, n line proofs are good, for all n; hence all proofs are good; hence, all proofs have true conclusions. Since there is a sentence that is not true, namely $0 = 1$, it cannot be a theorem of $\mathcal{T}$, and so $\mathcal{T}$ is consistent.

The emphasized use of induction is unavailable in $\mathsf{CT}^-[\mathcal{T}]$, but the rest of the proof turns out to be perfectly fine. Proving that it is fine is what gets messy, for reasons connected with such logical principles as universal instantiation. But, if $\mathcal{T}$ is finitely axiomatized, then we can indeed show in $\mathsf{CT}^-[\mathcal{T}]$ that

(i) 0 line $\mathcal{T}$-proofs are good;
(ii) if n line $\mathcal{T}$-proofs are good, then $n + 1$ line $\mathcal{T}$-proofs are good.

The formula expressing "n line $\mathcal{T}$-proofs are good" is thus what Russell called "inductive." Quite general techniques, known as "shortening of cuts,"[20] can then be used to show that $\mathsf{CT}^-[\mathcal{T}]$ interprets Q plus the statement "$\forall n(n$ line $\mathcal{T}$-proofs are good)." And from that it follows that $\mathsf{CT}^-[\mathcal{T}]$ interprets Q plus $\mathsf{Con}(\mathcal{T})$—if, again, $\mathcal{T}$ is finitely axiomatized.

One thing about this argument that it is important to appreciate is that it is absolutely essential that we be able to prove that *all* of $\mathcal{T}$'s axioms are true. If we do not know that *all* of $\mathcal{T}$'s axioms are true, but only know, of each of them, that it is true,

then we cannot even prove that *all* one-line proofs are good. The best way to think of this is to see Theorem 2.5 as a consequence of the following.[21]

Theorem 2.8 *Let $\mathcal{U}$ be any theory in the language of arithmetic. Then $\mathsf{CT}^-[\mathcal{T}] + \mathsf{T}(\mathcal{U})$ interprets $\mathsf{Q} + \mathsf{Con}(\mathcal{U})$.*

Here, $\mathsf{T}(\mathcal{U})$ is the formalization of: all axioms of $\mathcal{U}$ are true.

Proposition 2.9 *If $\mathcal{T}$ is finitely axiomatized, then $\mathsf{CT}^-[\mathcal{T}]$ proves $\mathsf{T}(\mathcal{T})$.*[22]

The action, unsurprisingly, is in the proof of Theorem 2.8 (see [18, Section 3.2]). Proposition 2.9 is fairly trivial. Indeed, it is easy to see that $\mathsf{DT}^-[\mathcal{T}]$ already proves that all axioms of $\mathcal{T}$ are true, if $\mathcal{T}$ is finitely axiomatized.

Proposition 2.10 *For each axiom A of $\mathcal{T}$, $\mathsf{DT}^-[\mathcal{T}]$ proves $\mathsf{T}(\ulcorner A \urcorner)$.*

Proof Let A be an axiom of $\mathcal{T}$ and so of $\mathsf{DT}^-[\mathcal{T}]$. Since $\mathsf{T}(\ulcorner A \urcorner) \equiv A$ is also an axiom of $\mathsf{DT}^-[\mathcal{T}]$, it proves $\mathsf{T}(\ulcorner A \urcorner)$. □

Proposition 2.11 *If $\mathcal{T}$ is finitely axiomatized, then $\mathsf{DT}^-[\mathcal{T}]$ proves $\mathsf{T}(\mathcal{T})$.*

Proof Let the axioms of $\mathcal{T}$ be $A_1, \ldots, A_n$. Then by Proposition 2.10, $\mathsf{DT}^-[\mathcal{T}]$ proves $\mathsf{T}(\ulcorner A_1 \urcorner) \wedge \cdots \wedge \mathsf{T}(\ulcorner A_n \urcorner)$. But then $\forall x (x = \ulcorner A_1 \urcorner \vee \cdots \vee x = \ulcorner A_n \urcorner \rightarrow \mathsf{T}(x))$ follows easily. □

Proposition 2.9 now follows from the fact that $\mathsf{CT}^-[\mathcal{T}]$ contains $\mathsf{DT}^-[\mathcal{T}]$.

Lemma 2.12 *For each formula $A(v_1, \ldots, v_n)$, $\mathsf{CT}^-[\mathcal{T}]$ proves the corresponding Sat-sentence:*

$$\mathsf{Sat}_\sigma\big(\ulcorner A(v_1, \ldots, v_n) \urcorner\big) \equiv A\big(\mathsf{val}(\sigma, 1), \ldots, \mathsf{val}(\sigma, n)\big).$$

A fortiori, for each sentence A in the language of $\mathcal{T}$, $\mathsf{CT}^-[\mathcal{T}]$ proves $\mathsf{T}(\ulcorner A \urcorner) \equiv A$.

Proof A rigorous proof would be by induction on the complexity of sentences of $\mathcal{L}$. But this should be fairly obvious.[23] A little experimentation will reveal that proofs of T-sentences need no more than is available in $\mathsf{Q}_{\mathrm{seq}}$: We are not proving any general laws, just a bunch of particular facts, and Q is very good at proving particular facts. □

The crucial thing to note here is the contrast between Proposition 2.9 and Proposition 2.10. If $\mathcal{T}$ is not finitely axiomatizable, then there is no reason whatsoever to suspect that $\mathsf{CT}^-[\mathcal{T}]$ will prove that *all* axioms of $\mathcal{T}$ are true, although it does prove that *each* axiom of $\mathcal{T}$ is true.

To summarize, then: $\mathsf{DT}[\mathsf{I}\Sigma_n]$ is locally interpretable in $\mathsf{I}\Sigma_n$, and so in that sense is no stronger than $\mathsf{I}\Sigma_n$; but $\mathsf{CT}^-[\mathsf{I}\Sigma_n]$ is not interpretable in $\mathsf{I}\Sigma_n$.[24] Thus, $\mathsf{CT}^-[\mathsf{I}\Sigma_n]$ is logically stronger than $\mathsf{DT}[\mathsf{I}\Sigma_n]$. Indeed, $\mathsf{DT}[\mathsf{I}\Sigma_n]$ is at best only very slightly stronger than $\mathsf{I}\Sigma_n$,[25] whereas $\mathsf{CT}^-[\mathsf{I}\Sigma_n]$ is at least as strong as $\mathsf{Q} + \mathsf{Con}(\mathsf{I}\Sigma_n)$, which is the theory that Pudlák's version of the second incompleteness theorem tells us must be stronger than $\mathsf{I}\Sigma_n$. So $\mathsf{CT}^-[\mathsf{I}\Sigma_n]$ is significantly stronger than $\mathsf{DT}[\mathsf{I}\Sigma_n]$.

3 Objections (I)

The moral of the last section is meant to be this: whereas deflationary truth-theories are logically very weak, whether or not induction is extended, compositional truth-theories have significant logical strength, even when induction is not extended. So, as I suggested earlier, there is no symmetry between the compositional axioms and the extension of induction. If we want to "blame" one or the other for the nonconservativity result discussed earlier, then, we should blame the compositional axioms.

There are, however, two sorts of objections that might be made to the interpretation of the mathematical facts that I have just suggested.

3.1 Theories of satisfaction The first objection is that matters look different if we consider satisfaction instead of truth.

We saw earlier that DS[PA] is interpretable in PA. But corresponding results do *not* hold for $\mathsf{DS}[\mathsf{I}\Sigma_n]$, as the following shows.[26]

Theorem 3.1 $\mathsf{DS}[\mathsf{I}\Sigma_1]$[27] *contains* PA.

Proof Let $A(v_0, v_1)$ be a formula in the language of arithmetic. We want to show that we can prove the induction axiom for it. (Extension to the case of extra free variables is straightforward.) Consider the formula

$$\phi(z,\sigma) \stackrel{df}{\equiv} \exists\tau\big[\tau \stackrel{0}{\sim} \sigma \wedge \mathsf{val}(\tau,0) = z \wedge \mathsf{Sat}_\tau\big(\ulcorner A(v_0,v_1)\urcorner\big)\big].$$

Now, $\phi(z,\sigma)$ is Σ_1 in $\mathsf{I}\Sigma_1$,[28] so $\mathsf{DS}[\mathsf{I}\Sigma_1]$ has induction for it. The induction axiom for $\phi(z,\sigma)$ is

$$\begin{aligned}
&\exists\tau\big[\tau \stackrel{0}{\sim} \sigma \wedge \mathsf{val}(\tau,0) = 0 \wedge \mathsf{Sat}_\tau\big(\ulcorner A(v_0,v_1)\urcorner\big)\big] \wedge \\
&\forall v_0\{\exists\tau\big[\tau \stackrel{0}{\sim} \sigma \wedge \mathsf{val}(\tau,0) = v_0 \wedge \mathsf{Sat}_\tau\big(\ulcorner A(v_0,v_1)\urcorner\big)\big] \rightarrow \\
&\exists\tau\big[\tau \stackrel{0}{\sim} \sigma \wedge \mathsf{val}(\tau,0) = \mathsf{S}v_0 \wedge \mathsf{Sat}_\tau\big(\ulcorner A(\mathsf{S}v_0,v_1)\urcorner\big)\big]\} \rightarrow \\
&\forall v_0\exists\tau\big[\tau \stackrel{0}{\sim} \sigma \wedge \mathsf{val}(\tau,0) = v_0 \wedge \mathsf{Sat}_\tau\big(\ulcorner A(v_0,v_1)\urcorner\big)\big]. \qquad (3.1)
\end{aligned}$$

But the Sat-sentence for $A(x,y)$ is[29]

$$\mathsf{Sat}_\tau\big(\ulcorner A(v_0,v_1)\urcorner\big) \equiv A\big(\mathsf{val}(\tau,0),\mathsf{val}(\tau,1)\big).$$

So $\phi(z,\sigma)$ is provably equivalent in $\mathsf{DS}[\mathsf{I}\Sigma_1]$ to[30]

$$A\big(z,\mathsf{val}(\sigma,1)\big).$$

Hence, (3.1) reduces to

$$\begin{aligned}
A\big(0,\mathsf{val}(\sigma,1)\big) \wedge \forall v_0\big[A\big(v_0,\mathsf{val}(\sigma,1)\big) \rightarrow A\big(\mathsf{S}v_0,\mathsf{val}(\sigma,1)\big)\big] \rightarrow \\
\forall v_0\big(A\big(v_0,\mathsf{val}(\sigma,1)\big)\big). \qquad (3.2)
\end{aligned}$$

And this holds for any sequence σ.

But, for any given v_1, there is a sequence σ such that $\mathsf{val}(\sigma,1) = v_1$. Hence, for this σ, we have $A(v_0,v_1) \equiv A(v_0,\mathsf{val}(\sigma,1))$, for any v_0. So (3.2) becomes

$$A(0,v_1) \wedge \forall v_0\big(A(v_0,v_1) \rightarrow A(\mathsf{S}v_0,v_1)\big) \rightarrow \forall v_0\big(A(v_0,v_1)\big).$$

And that is the induction axiom for $A(v_0,v_1)$. □

One might therefore suggest that it is not the compositional nature of $\mathsf{CT}^-[\mathcal{T}]$ that gives it its strength, but its play with the notion of satisfaction.

It goes some way towards answering this objection to note that the extension of induction here is essential. It remains the case that, if we do not extend induction, then, as Theorem 2.2 shows, we get only a marginal increase in logical strength if we add the Sat-sentences,[31] but we get a significant increase in logical strength if we add a compositional truth-theory. So the asymmetry to which I have pointed remains. But it would be nice to have a bit more to say.

3.2 Finite axiomatizability The second worry concerns the fact that Corollary 2.7 does not apply to PA, as the following shows.[32]

Theorem 3.2 ([6, Theorem 5.1]) $\mathsf{CT}^-[\mathsf{PA}]$ *is interpretable in* PA.

The worry here is not just that Corollary 2.7 applies only to finitely axiomatized theories (though one might pursue that point as well). The worry concerns what Theorem 3.2 shows about why "adding a truth-theory" adds logical strength. It is true that $\mathsf{CT}^-[\mathsf{I}\Sigma_n]$ is stronger than $\mathsf{I}\Sigma_n$, but, since $\mathsf{CT}^-[\mathsf{PA}]$ is not stronger than PA, maybe we should conclude that adding a truth-theory adds logical strength only insofar as we pretend not to believe something we ought to believe, namely: full induction. Or, to put it differently: perhaps the reason $\mathsf{CT}^-[\mathsf{I}\Sigma_n]$ is stronger than $\mathsf{I}\Sigma_n$ is because the syntax on which we are building these theories is artificially weak.[33]

3.3 Toward a response Somewhat surprisingly, it turns out that these two objections have a common source: a conflation, common in contemporary work on theories of truth, between the object theory *about which* we propose to reason and the syntactic theory *in which* we propose to reason about that object theory.

As we have seen, $\mathsf{CT}^-[\mathcal{T}]$ is not going to be able to prove the consistency of $\mathcal{T}$: It lacks the necessary induction axioms. If we do want to investigate theories in which the consistency of $\mathcal{T}$ can be proven semantically, then, we need to consider theories in which the induction axioms have been extended. Here, then, are some obvious questions about such theories: How much induction do we need to prove the consistency of Q? How much do we need to prove the consistency of $\mathsf{I}\Sigma_n$? or of PA? Is it always the same amount? Or does it vary with the object theory?

Careful analysis of the structure of semantic consistency proofs shows that the answer is that we always need the same amount of induction: We need it for certain Σ_1-formulae involving semantic vocabulary. More precisely, we have the following.[34]

Theorem 3.3 *Suppose that* $\mathcal{T} \supseteq \mathsf{I}\Sigma_1$. *Then* $\mathsf{CT}[\mathcal{T}] + \mathsf{T}(\mathcal{U})$ *proves* $\mathsf{Con}(\mathcal{U})$.

Corollary 3.4 *Suppose that* $\mathcal{T} \supseteq \mathsf{I}\Sigma_1$ *is finitely axiomatized. Then* $\mathsf{CT}[\mathcal{T}]$ *proves* $\mathsf{Con}(\mathcal{T})$.

The second of these results implies that $\mathsf{CT}[\mathsf{I}\Sigma_1]$ proves $\mathsf{Con}(\mathsf{I}\Sigma_1)$ and that $\mathsf{CT}[\mathsf{I}\Sigma_n]$ proves $\mathsf{Con}(\mathsf{I}\Sigma_n)$. But we do not have any way, if we speak only in the terms in which Corollary 3.4 is formulated, to express the insight that the only induction we need to prove $\mathsf{Con}(\mathsf{I}\Sigma_n)$ is already available in $\mathsf{CT}[\mathsf{I}\Sigma_1]$. The problem is that the "base theory" is playing two roles. On the one hand, through the magic of Gödel numbering, it is our theory of syntax. On the other hand, it is the object theory, and

in that role it is what allows us to prove the basis of the induction: that all the axioms of $\mathcal{T}$ are true.

The formulation in Theorem 3.3, which is ultimately more fundamental, goes some way toward pulling these two roles apart. The "base theory" $\mathcal{T}$ is now providing our theory of syntax, and the formalization of "all axioms of $\mathcal{U}$ are true" is providing the basis for the induction; thus, $\mathcal{U}$ is the object theory. And so, the thought might be, $\mathsf{CT}[\mathsf{I}\Sigma_1]+\mathsf{T}(\mathsf{I}\Sigma_2)$ will prove $\mathsf{Con}(\mathsf{I}\Sigma_2)$, and $\mathsf{CT}[\mathsf{I}\Sigma_1]+\mathsf{T}(\mathsf{I}\Sigma_3)$ will prove $\mathsf{Con}(\mathsf{I}\Sigma_3)$, and so forth; the constant presence of $\mathsf{CT}[\mathsf{I}\Sigma_1]$ now expresses the fact that we only need a limited amount of induction for the argument, no matter what the object theory may be.

In fact, however, the syntax, with its extended induction axioms, can still "infect" the object theory. We have already seen in Theorem 3.1 that $\mathsf{DS}[\mathsf{I}\Sigma_1]$ contains PA since, for each axiom of PA, there is an extended induction axiom that implies it. So $\mathsf{DS}[\mathsf{I}\Sigma_1]$ *by itself* already proves $\mathsf{Con}(\mathsf{I}\Sigma_n)$, for each n, since $\mathsf{DS}[\mathsf{I}\Sigma_1]$ contains PA, and PA is reflexive. And since $\mathsf{CT}[\mathsf{I}\Sigma_1]$ proves all the Sat-sentences, $\mathsf{CT}[\mathsf{I}\Sigma_1]$ contains $\mathsf{DS}[\mathsf{I}\Sigma_1]$ and so also proves $\mathsf{Con}(\mathsf{I}\Sigma_n)$, for each n. So, yes, $\mathsf{CT}[\mathsf{I}\Sigma_1]+\mathsf{T}(\mathsf{I}\Sigma_2)$ proves $\mathsf{Con}(\mathsf{I}\Sigma_2)$, but $\mathsf{CT}[\mathsf{I}\Sigma_1]$ already proves $\mathsf{Con}(\mathsf{I}\Sigma_2)$ by itself.

Indeed, since $\mathsf{CT}[\mathsf{I}\Sigma_1]$ contains PA, it contains $\mathsf{CT}[\mathsf{I}\Sigma_1]+\mathsf{I}\Sigma_n$, for each n. But $\mathsf{CT}[\mathsf{I}\Sigma_1]+\mathsf{I}\Sigma_n$ proves $\mathsf{T}(\mathsf{I}\Sigma_n)$. So $\mathsf{CT}[\mathsf{I}\Sigma_1]$ contains $\mathsf{CT}[\mathsf{I}\Sigma_1]+\mathsf{T}(\mathsf{I}\Sigma_n)$, for each n, which implies that $\mathsf{CT}[\mathsf{I}\Sigma_1]+\mathsf{T}(\mathsf{I}\Sigma_2)$, $\mathsf{CT}[\mathsf{I}\Sigma_1]+\mathsf{T}(\mathsf{I}\Sigma_3)$, and $\mathsf{CT}[\mathsf{I}\Sigma_1]+\mathsf{T}(\mathsf{I}\Sigma_n)$ all have the same theorems as $\mathsf{CT}[\mathsf{I}\Sigma_1]$ itself. It follows that the object theory $\mathcal{U}$ is simply not playing the role it appeared to be.[35]

Things get worse. A modification of the proof of Theorem 3.1 shows that, in $\mathsf{CT}[\mathsf{I}\Sigma_1]$, we can find a *single* extended induction axiom that bundles all the induction axioms of PA together.[36] We thus get the following.

Lemma 3.5 $\mathsf{CT}[\mathsf{I}\Sigma_1]$ *proves that all axioms of* PA *are true.*

So $\mathsf{CT}[\mathsf{I}\Sigma_1]$ actually contains $\mathsf{CT}[\mathsf{I}\Sigma_1]+\mathsf{T}(\mathsf{PA})$. And so Theorem 3.3 implies the following.

Corollary 3.6 $\mathsf{CT}[\mathsf{I}\Sigma_1]$ *proves* $\mathsf{Con}(\mathsf{PA})$.

And now we can see quite clearly that Theorem 3.3 does not really help us to disentangle the different roles being played by the syntactic theory and by the object theory.

If we are going to resolve these problems and get a proper formulation of the insight expressed when we say that only Σ_1 induction is needed for semantic consistency proofs, then what we need to do is explicitly disentangle the syntactic theory from the object theory. That is, we need to find a way to allow ourselves to vary the syntactic theory we use when we talk about the object theory without thereby changing what object theory we are talking about. We will explore how we might do so in the next section. Once we have done so, however, the first objection will have been answered, since, as will then be clear, results like Theorem 3.1 and Corollary 3.6 depend essentially upon the entanglement.

Concerning the second objection, the first thing to note is that it is not the presence of full induction that is driving the proof that $\mathsf{CT}^-[\mathsf{PA}]$ is interpretable in PA. It is the *reflexivity* of PA, as the following more general results show. We first need the following partial converse of Theorem 2.5.[37]

Theorem 3.7 *Suppose that $\mathcal{T}$ is finitely axiomatized. Then $\mathsf{CT}^-[\mathcal{T}]$ is interpretable in $\mathsf{I}\Sigma_1 + \mathsf{Con}(\mathcal{T})$.*

This follows from results proven by Enayat and Visser [5, esp. Theorem 4.5] in their recent explorations of full satisfaction classes.

Corollary 3.8 *$\mathsf{CT}^-[\mathcal{T}]$ is locally interpretable in $\mathsf{I}\Sigma_1 + \bigcup\{\mathsf{Con}(\mathcal{U})$: $\mathcal{U}$ a finite, sequential fragment of $\mathcal{T}\}$.*

Proof Every finite fragment of $\mathsf{CT}^-[\mathcal{T}]$ is contained in $\mathsf{CT}^-[\mathcal{U}]$, for some finite fragment $\mathcal{U} \supseteq \mathcal{T}$. But $\mathsf{CT}^-[\mathcal{U}]$ is interpretable in $\mathsf{I}\Sigma_1 + \mathsf{Con}(\mathcal{U})$, which is of course a subset of $\mathsf{I}\Sigma_1 + \bigcup\{\mathsf{Con}(\mathcal{U})\}$. □

Theorem 3.9 *If $\mathcal{T} \supseteq \mathsf{I}\Sigma_1$ is reflexive, then $\mathsf{CT}^-[\mathcal{T}]$ is interpretable in $\mathcal{T}$.*

Proof Since $\mathcal{T}$ is reflexive, it proves $\mathsf{Con}(\mathcal{U})$, for each finite $\mathcal{U} \supseteq \mathcal{T}$. So $\mathcal{T}$ contains $\mathsf{I}\Sigma_1 + \bigcup\{\mathsf{Con}(\mathcal{U})\}$, so $\mathcal{T}$ locally interprets $\mathsf{CT}^-[\mathcal{T}]$. It then follows from Orey's compactness theorem that $\mathcal{T}$ *globally* interprets $\mathsf{CT}^-[\mathcal{T}]$. □

Theorem 3.2 then follows from Theorem 3.9, since PA is reflexive.

There are plenty of theories that do not have full induction but are nonetheless reflexive.[38] For example, the following sort of construction allows us to build a reflexive theory from any theory with which we wish to begin:

$$\begin{aligned} T_0 &= \mathcal{U}, & C_0 &= \mathsf{Con}(\mathcal{U}), \\ T_1 &= \mathcal{U} + \mathsf{Con}(\mathcal{U}), & C_1 &= \mathsf{Con}\big(\mathcal{U} + \mathsf{Con}(\mathcal{U})\big), \\ T_{n+1} &= T_n + C_n, & C_{n+1} &= \mathsf{Con}(T_{n+1}). \end{aligned}$$

Now let the "reflexive closure" of $\mathcal{U}$, $\mathsf{RCI}(\mathcal{U})$, be $\bigcup T_n$. $\mathsf{RCI}(\mathcal{U})$ is reflexive, since every finite subtheory of $\mathsf{RCI}(\mathcal{U})$ is contained in one of the T_n's, and T_{n+1} proves $\mathsf{Con}(T_n)$ by construction. So, in particular, we have the following from Theorem 3.9.

Corollary 3.10 *$\mathsf{CT}^-[\mathsf{RCI}(\mathsf{I}\Sigma_1)]$ is interpretable in $\mathsf{RCI}(\mathsf{I}\Sigma_1)$.*

That, however, does not look like a result that should call into question the philosophical conclusions drawn from the discussion in Section 2.

Really to answer the second objection, however, we need to understand exactly what role the reflexivity of the base theory is playing in the proof of Theorem 3.9. As we shall see, it is not the reflexivity of the syntactic theory that is responsible for this result, but the reflexivity of the object theory. In particular, the reason we get Theorem 3.2 is because we have taken PA as our *object* theory, not because we have taken PA as our syntactic theory. Indeed, once we have successfully disentangled the syntactic theory from the object theory, we will see that (the relevant analogues of) the results reported in Section 2 are largely insensitive to what syntactic theory we use.

4 Disentangling the Syntactic Theory From the Object Theory

Our goal now is to disentangle the syntactic theory from the object theory. Interestingly enough, we can do so simply by following what Tarski actually did in [31]. Here is his explanation of what a meta-language adequate for his purposes must be like:

> A meta-language... must contain three groups of expressions: (1) expressions of a general logical kind; (2) expressions having the same meaning as all the constants of the language to be discussed...; (3) expressions of the structural-descriptive type which denote single signs and expressions of the language considered, whole classes and sequences of such expressions or, finally, the relations existing between them. [31, pp. 210–11]

The expressions mentioned under (3) belong to syntax; those under (2), to the object language. Tarski does not quite *say* that these two classes are to be disjoint, but it is natural to read him that way, and that is plainly how he conceives the matter in his discussion of the calculus of classes (see [31, Section 3]).

Tarski was of course aware that syntax can be interpreted in arithmetic (at least after reading Gödel); his famous theorem on the indefinability of truth depends upon that fact. But the central purpose of [31] is not "limitative" but positive. Tarski's primary goal in that paper is to show that there is a consistent notion of truth that is adequate for the metamathematical purposes for which truth was then already being deployed. Doing that simply does not require Gödel numbering or any similar technique. The idea of separating syntax from the object theory is thus old, even if the application I propose to make of it is somewhat new.

Let $\mathfrak{L}$ be the object language, that is, the language for which we want to give a truth-theory. Let $\mathfrak{S}$ be a disjoint language in which to formalize syntax. The most natural choice for $\mathfrak{S}$ would be the language of concatenation (see [3], [11], [28]). But so as not to make things too unfamiliar, we may take $\mathfrak{S}$ to be a copy of the language of arithmetic, written in a different font, perhaps. Our theory of syntax can then be taken to be Q, or $\mathsf{I}\Sigma_1$, or whatever we wish.

To formulate a semantics for $\mathfrak{L}$, we of course need to be able to talk about the things $\mathfrak{L}$ talks about. In particular, if we are going to have the usual Tarski-style clauses for the primitive expressions of $\mathfrak{L}$, we need to have the expressive resources of $\mathfrak{L}$ available. So the obvious choice for the language of our semantic theory would be $\mathfrak{S} \cup \mathfrak{L}$ plus whatever semantic machinery we want, and that is what we shall use. Because of complications we need not consider, however, we shall regard the semantic theory as many-sorted. Variables ranging over the domain of $\mathfrak{S}$ will be italic; those ranging over the domain of $\mathfrak{L}$ will be upright.

We also need a theory of sequences or, better, of assignments of objects to variables. There is no hope of coding sequences of objects from the domain of $\mathfrak{L}$ as objects in $\mathfrak{S}$, at least not in general.[39] The details of that theory do not matter here, either. What is important is that assignments live in a third sort. Variables ranging over them will be Greek letters.[40]

A truth-theory for $\mathfrak{L}$ will then be more or less the familiar one, with some adjustments to take account of the present framework. For example, these axioms will be common to all theories, independent of $\mathfrak{L}$:

(v) $\mathsf{var}(v_i) \to \mathsf{Den}_\alpha(v, \mathsf{val}(\alpha, i))$,
($\wedge$) $\mathsf{Sat}_\alpha(\ulcorner A \wedge B \urcorner) \equiv \mathsf{Sat}_\alpha(A) \wedge \mathsf{Sat}_\alpha(B)$,
($\forall$) $\mathsf{Sat}_\alpha(\ulcorner \forall \mathrm{v}_i A(\mathrm{v}_i) \urcorner) \equiv \forall \beta[\beta \overset{i}{\sim} \alpha \to \mathsf{Sat}_\beta(\ulcorner A(\mathrm{v}_i) \urcorner)]$.

The other axioms of the theory will depend upon what $\mathfrak{L}$ is, and it could be anything. If $\mathfrak{L}$ is the language of set theory, then the only other axiom will be

($\in$) $\mathsf{Sat}_\alpha(\ulcorner t \in u \urcorner) \equiv \exists \mathrm{x} \exists \mathrm{y}[\mathsf{Den}_\alpha(t, \mathrm{x}) \wedge \mathsf{Den}_\alpha(u, \mathrm{y}) \wedge \mathrm{x} \in \mathrm{y}]$.

In the case of the language of arithmetic, we will have axioms like

(0) $\mathsf{Den}_\alpha(\text{`0'}, \mathrm{x}) \equiv \mathrm{x} = 0$,
(+) $\mathsf{Den}_\alpha(\ulcorner t + u \urcorner, \mathrm{x}) \equiv \exists \mathrm{y} \exists \mathrm{z}[\mathsf{Den}_\sigma(t, \mathrm{y}) \wedge \mathsf{Den}_\sigma(u, \mathrm{z}) \wedge \mathrm{x} = \mathrm{y} + \mathrm{z}]$.

Note that the *used* expressions $\in$, 0, and $+$ are expressions of $\mathcal{L}$, not of $\mathcal{S}$.

As for notation, we have the following.[41]

Definition Let $\mathcal{T}$ be an arithmetical theory. Then

- $\mathsf{TT}^-_{\mathcal{L}}[\mathcal{T}]$ is the semantics for $\mathcal{L}$ we have just described;
- $\mathsf{TT}_{\mathcal{L}}[\mathcal{T}]$ is $\mathsf{TT}^-_{\mathcal{L}}[\mathcal{T}]$ with the induction axioms in $\mathcal{T}$ extended to permit semantic vocabulary and reference to assignments.[42]

So the induction axioms of $\mathsf{TT}^-_{\mathcal{L}}[\mathcal{T}]$ are "purely syntactic." (TT stands for *Tarskian truth*.)

Our earlier results transfer smoothly to this framework, though often in improved forms, and there are new results available as well. I shall state most of these without proof. Many of the proofs are similar to ones already given; the rest are more complex than it makes sense to present here. Full proofs are presented elsewhere (see [18, Section 4]).

First, we get an analogue of Lemma 2.12.

Lemma 4.1 *For each formula $A(v_0, \ldots, v_n)$ of $\mathcal{L}$, $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{Q}]$ proves the corresponding Sat-sentence*

$$\mathsf{Sat}_\sigma\big(\ulcorner A(v_0, \ldots, v_n)\urcorner\big) \equiv A\big(\mathsf{val}(\sigma, 0), \ldots, \mathsf{val}(\sigma, n)\big).$$

But now the situation is improved: $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{Q}]$ is as weak as it is possible for it to be.[43]

Proposition 4.2 $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{Q}]$ *is interpretable in* Q.

Proof Since no theory stated in $\mathcal{L}$ is so far in evidence, we can give $\mathcal{L}$ the completely trivial interpretation in a one-element domain. A semantic theory for $\mathcal{L}$, so interpreted, is then easily constructed.[44] □

If we develop our truth-theory in the usual way, where syntax and the object theory are intertwined, then the weakest materially adequate truth-theory for the language of arithmetic is $\mathsf{CT}^-[\mathsf{Q}_{\mathrm{seq}}]$, and it follows from Theorem 2.6 that $\mathsf{CT}^-[\mathsf{Q}_{\mathrm{seq}}]$ is *not* interpretable in Q.

As said, no object theory is yet in play here. To add one, we simply add it. Thus, for example, if $\mathcal{T}$ is a theory in $\mathcal{L}$, then $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{Q}] + \mathcal{T}$ is a semantic theory for the language of $\mathcal{L}$, with Q as the syntactic theory, plus the object theory $\mathcal{T}$. Then we get the following analogues of our earlier results.

Proposition 4.3 *For each axiom A of $\mathcal{T}$, $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{Q}] + \mathcal{T}$ proves $\mathsf{T}(\ulcorner A \urcorner)$ (cf. Proposition 2.10).*

Corollary 4.4 *If $\mathcal{T}$ is finitely axiomatized, then $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{Q}] + \mathcal{T}$ proves the obvious, disjunctive formalization of "all axioms of $\mathcal{T}$ are true" (cf. Proposition 2.9).*

Theorem 4.5 $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{Q}]$ *plus "all axioms of $\mathcal{T}$ are true" interprets* $\mathsf{Q} + \mathsf{Con}(\mathcal{T})$ *(cf. Theorem 2.8).*

Corollary 4.6 *Let $\mathcal{T}$ be a finitely axiomatized theory in $\mathcal{L}$. Then $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{Q}] + \mathcal{T}$ interprets $\mathsf{Q} + \mathsf{Con}(\mathcal{T})$ and so is not interpretable in $\mathcal{T}$ (cf. Theorem 2.5 and Corollary 2.7).*

Now, however, we can also prove a converse of Corollary 4.6.

Theorem 4.7 *Let $\mathcal{T}$ be a finitely axiomatized theory in $\mathcal{L}$. Then $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{Q}] + \mathcal{T}$ is interpretable in $\mathsf{Q} + \mathsf{Con}(\mathcal{T})$.*

Thus, we get a precise characterization of just how strong $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{Q}] + \mathcal{T}$ is.

Corollary 4.8 *Let $\mathcal{T}$ be a finitely axiomatized theory in $\mathcal{L}$. Then $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{Q}] + \mathcal{T}$ is mutually interpretable with $\mathsf{Q} + \mathsf{Con}(\mathcal{T})$.*

As said, then, compositional truth-theories have significant logical power, even when the syntax is as weak as possible, and even when we do not extend induction. If we start with a finitely axiomatized theory $\mathcal{T}$ and add an absolutely minimal but still compositional theory of truth for the language of $\mathcal{T}$—and add it in a way that is guaranteed not to "infect" $\mathcal{T}$ itself—then the result is a theory that is logically stronger than $\mathcal{T}$ in the sense that it is not interpretable in $\mathcal{T}$.

Perhaps the nicest way to formulate this point is due to Visser: A compositional theory of truth is like an operator that "up-Gödels" any finitely axiomatized theory you hand it. What is up-Gödeling? It is the operation that maps a finitely axiomatized theory $\mathcal{T}$ to the one that Pudlák's form of the second incompleteness theorem guarantees will always be stronger than it is: $\mathsf{Q} + \mathsf{Con}(\mathcal{T})$. And so, if you think of $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{Q}] + (\cdot)$ as an operator on theories, then what it does, when handed a finitely axiomatized theory $\mathcal{T}$, is precisely to up-Gödel it. It hands you back a theory that is mutually interpretable with $\mathsf{Q} + \mathsf{Con}(\mathcal{T})$. So it is not just that $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{Q}] + \mathcal{T}$ is always stronger than $\mathcal{T}$ (when $\mathcal{T}$ is finitely axiomatized). It is stronger in the very specific, and very important, way that is revealed by the second incompleteness theorem.

Our specific interest here, however, is in the objection raised in Section 3.2: that Corollary 2.7 applies only to finitely axiomatized theories and does not apply to PA. I said in Section 3.3 that, once we had disentangled the syntax from the object theory, it would be possible to see that this is due not to PA's role as syntax, but to its role as object theory. We have done the disentangling now. Let us see what difference it has made.

First, note that we do get the same phenomenon as before when PA is the object theory.

Proposition 4.9 $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{Q}] + \mathsf{PA}$ *is interpretable in* PA *(cf. Theorem 3.2).*

Proof Let $\mathcal{U}$ be a finite fragment of $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{Q}] + \mathsf{PA}$. Then $\mathcal{U}$ is a subtheory of $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{Q}] + \mathsf{I}\Sigma_n$, for some n, and so is interpretable in $\mathsf{Q} + \mathsf{Con}(\mathsf{I}\Sigma_n)$, by Theorem 4.7. But $\mathsf{Q} + \mathsf{Con}(\mathsf{I}\Sigma_n)$ is a subtheory of PA, since PA is reflexive, so $\mathcal{U}$ is interpretable in PA.

That establishes local interpretability, and Orey's compactness theorem does the rest. □

That is essentially the same as the proof of Theorem 3.9, and the proof can easily be extended to the case of reflexive theories generally. So, for example, we also have the following.

Proposition 4.10 $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{Q}] + \mathsf{RCI}(\mathsf{I}\Sigma_1)$ *is interpretable in* $\mathsf{RCI}(\mathsf{I}\Sigma_1)$ *(cf. Corollary 3.10).*

By contrast, Theorem 4.7 extends smoothly to the case of PA as syntactic theory and, indeed, to any theory that contains Q.[45]

Corollary 4.11 *Let $\mathcal{T}$ be a finitely axiomatized theory in $\mathcal{L}$, and suppose that $\mathcal{S} \supseteq \mathsf{Q}$. Then $\mathsf{TT}^-_{\mathcal{L}}[\mathcal{S}] + \mathcal{T}$ is not interpretable in $\mathcal{T}$. Hence, $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{PA}] + \mathcal{T}$ is not interpretable in $\mathcal{T}$.*

Proof We know from Theorem 4.7 that $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{Q}] + \mathcal{T}$ is not interpretable in $\mathcal{T}$. But if $\mathcal{S} \supseteq \mathsf{Q}$, then $\mathsf{TT}^-_{\mathcal{L}}[\mathcal{S}] + \mathcal{T}$ contains $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{Q}] + \mathcal{T}$ and so is not interpretable in $\mathcal{T}$, either. □

Why does it make such a difference whether PA is our syntax or our object theory? The reason, ultimately, is pretty simple. The basic result is Theorem 4.5: $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{Q}]+$ "all axioms of $\mathcal{T}$ are true" interprets $\mathsf{Q} + \mathsf{Con}(\mathcal{T})$. But if we do not know that *all* axioms of $\mathcal{T}$ are true—in particular, if we only know that *each* of them is—then we cannot even prove that all one-line proofs have true conclusions, as noted earlier. So we will be able to prove that $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{Q}] + \mathcal{T}$ interprets $\mathsf{Q} + \mathsf{Con}(\mathcal{T})$ if, *but only if*, we can prove in $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{Q}] + \mathcal{T}$ that all axioms of $\mathcal{T}$ are true. This is trivial if $\mathcal{T}$ is finitely axiomatized. But if it is not even finitely axiomatiz*able*, then there is no evident way for $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{Q}] + \mathcal{T}$ (or even $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{PA}] + \mathcal{T}$) to prove that *all* axioms of $\mathcal{T}$ are true, rather than just that each of them is.

One can see this from the fact that Proposition 4.9 continues to hold not just as the syntactic theory is strengthened...

Proposition 4.12

(i) $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{I}\Sigma_n] + \mathsf{PA}$ *is interpretable in* PA.
(ii) $\mathsf{TT}^-_{\mathcal{L}}[\mathsf{PA}] + \mathsf{PA}$ *is interpretable in* PA.

...but even when we add semantic induction:

Proposition 4.13 ([18, Corollary 4.18]) $\mathsf{TT}_{\mathcal{L}}[\mathsf{PA}] + \mathsf{PA}$ *is interpretable in* PA.

And this is despite the fact that we have the following.

Theorem 4.14 ([18, Theorem 4.11]) $\mathsf{TT}_{\mathcal{L}}[\mathsf{I}\Sigma_1] + \mathsf{T}(\mathcal{T})$ *proves* $\mathsf{Con}(\mathcal{T})$. *In particular, if $\mathcal{T}$ is finitely axiomatized, then* $\mathsf{TT}_{\mathcal{L}}[\mathsf{I}\Sigma_1] + \mathcal{T} \vdash \mathsf{Con}(\mathcal{T})$.

So the reason $\mathsf{TT}_{\mathcal{L}}[\mathsf{PA}] + \mathsf{PA}$ not only does not prove $\mathsf{Con}(\mathsf{PA})$ but cannot even interpret $\mathsf{Q} + \mathsf{Con}(\mathsf{PA})$ is simply that it has no way to prove that *all* of PA's axioms are true, rather than just that each of them is. And the same goes for any other reflexive theory you wish to consider.

There is an odd irony to this situation. Deflationists frequently claim that the truth-predicate is a "device of infinite conjunction." Its function, allegedly, is to allow us to formulate such generalizations as "All axioms of PA are true." But very little effort has been made to tell us precisely what that is supposed to mean. What exactly is the relationship between this generalization and the infinite conjunction of PA's axioms? The only serious attempt known to me to answer this question is due to Halbach [14], who shows that, in certain circumstances, adding such a generalization to a theory is exactly equivalent to adding all of its instances. What we have seen, however, is that, considered as additions to, say, $\mathsf{TT}_{\mathcal{L}}[\mathsf{I}\Sigma_1]$, there is all the difference in the world between the axioms of PA and the generalization stating that all of them are true. The latter is a *lot* stronger than the former.[46]

That is not to say, of course, that there is not some other way of explaining what it means to "use the truth predicate merely as a device of generalization." But I do not know what that would be.

That, then, addresses the objection raised in Section 3.2. The other objection, recall, was based upon the observation that $\mathsf{DS}[\mathsf{I}\Sigma_1]$ contains PA. But that sort of result does not transfer to the present framework, as we shall now see.

Definition Let $\mathcal{U}$ be an arithmetical theory, taken as our theory of syntax. [47]

- $\mathsf{DDT}_{\mathfrak{L}}[\mathcal{U}]$ is the theory of truth for the language of $\mathcal{T}$ that is similar to $\mathsf{TT}_{\mathfrak{L}}[\mathcal{U}]$ but, instead of containing a compositional theory of truth contains just the T-sentences for $\mathfrak{L}$—though it also extends the induction scheme to permit the presence of the truth-predicate.
- $\mathsf{DDS}_{\mathfrak{L}}[\mathcal{U}]$ is the theory of truth for the language of $\mathcal{T}$ that is similar to $\mathsf{DDT}_{\mathfrak{L}}[\mathcal{U}]$ but adds the Sat-sentences for $\mathfrak{L}$ (and extends the induction scheme).

The sorts of results concerning disquotational theories of truth proven earlier transfer to the disentangled setting.

Proposition 4.15

(i) $\mathsf{DDT}_{\mathfrak{L}}[\mathsf{I}\Sigma_n]$ *is interpretable in* $\mathsf{I}\Sigma_n$.
(ii) $\mathsf{DDT}_{\mathfrak{L}}[\mathsf{I}\Sigma_n] + \mathcal{T}$ *is locally interpretable in* $\mathsf{I}\Sigma_n + \mathcal{T}$.
(iii) $\mathsf{DDT}_{\mathfrak{A}}[\mathsf{I}\Sigma_n] + \mathsf{I}\Sigma_m$ *is locally interpretable in* $\mathsf{I}\Sigma_{\max(m,n)}$.

Proof The proof of (i) is similar to that of Proposition 4.2. The proof of (ii) simply mimics that of Theorem 2.4.

For (iii), $\mathsf{DDT}_{\mathfrak{A}}[\mathsf{I}\Sigma_n] + \mathsf{I}\Sigma_m$ is locally interpretable in $\mathsf{I}\Sigma_n + \mathsf{I}\Sigma_m$, where these two theories are formulated in disjoint copies of the language of arithmetic. But $\mathsf{I}\Sigma_n + \mathsf{I}\Sigma_m$ will obviously be interpretable in $\mathsf{I}\Sigma_{\max(m,n)}$. □

So we also get an analogue of Theorem 2.3.

Corollary 4.16 $\mathsf{DDT}_{\mathfrak{A}}[\mathsf{PA}] + \mathsf{PA}$ *is interpretable in* PA.

Proof Any finite fragment of this theory is contained in one or another of the $\mathsf{DDT}_{\mathfrak{A}}[\mathsf{I}\Sigma_n] + \mathsf{I}\Sigma_m$. So each finite fragment is interpretable in $\mathsf{I}\Sigma_n$, for some n, and so is also interpretable in PA. That establishes local interpretability, and now we invoke Orey's compactness theorem. □

Unfortunately, the sorts of techniques used in these proofs do not seem to allow us to prove that $\mathsf{DDS}_{\mathfrak{L}}[\mathsf{I}\Sigma_n] + \mathcal{T}$ is locally interpretable in $\mathsf{I}\Sigma_n + \mathcal{T}$,[48] and I do not know exactly how strong $\mathsf{DDS}_{\mathfrak{L}}[\mathsf{I}\Sigma_n] + \mathcal{T}$ is. But for no n and m does $\mathsf{DDS}_{\mathfrak{A}}[\mathsf{I}\Sigma_n] + \mathsf{I}\Sigma_m$ contain PA. On the contrary, it follows from Theorem 5.2, to be mentioned below, that, $\mathsf{DDS}_{\mathfrak{A}}[\mathsf{I}\Sigma_n] + \mathsf{I}\Sigma_m$ is interpretable in $\mathsf{I}\Sigma_n + \mathsf{Con}(\mathsf{I}\Sigma_m)$. So, in particular, $\mathsf{DDS}_{\mathfrak{A}}[\mathsf{I}\Sigma_1] + \mathsf{I}\Sigma_1$ is no stronger than $\mathsf{I}\Sigma_1 + \mathsf{Con}(\mathsf{I}\Sigma_1)$, which is a proper subtheory of $\mathsf{I}\Sigma_2$ that does not even interpret $\mathsf{I}\Sigma_2$.[49] So there is no danger that $\mathsf{DDS}_{\mathfrak{L}}[\mathsf{I}\Sigma_1] + \mathcal{T}$ is going to be vastly stronger than $\mathcal{T}$, as $\mathsf{DS}[\mathsf{I}\Sigma_1]$ is vastly stronger than $\mathsf{I}\Sigma_1$.

Moreover, the kind of argument that was used to show that $\mathsf{DS}[\mathsf{I}\Sigma_1]$ contains PA—or, more generally, to extract information about the object theory from the theory of truth—is simply unavailable in the disentangled setting. The reason is that the induction that is available in the syntactic theory is over *syntactic* objects: expressions. We can formalize proofs by induction on the complexity of expressions, and object-language expressions may occur in the induction axioms used in those proofs. But the converse is not true: The induction scheme (or other axiom scheme, such as

separation) present in the object theory has *not* been extended, so it is difficult to see how the theory of truth could "infect" the object theory.

In particular, if we look at the instance of induction on which the proof of Theorem 3.1 was based:

$$\exists\tau\big[\tau \overset{0}{\sim} \sigma \wedge \mathsf{val}(\tau,0) = 0 \wedge \mathsf{Sat}_\tau\big(\ulcorner A(v_0,v_1)\urcorner\big)\big] \wedge$$
$$\forall v_0\{\exists\tau\big[\tau \overset{0}{\sim} \sigma \wedge \mathsf{val}(\tau,0) = v_0 \wedge \mathsf{Sat}_\tau\big(\ulcorner A(v_0,v_1)\urcorner\big)\big] \rightarrow$$
$$\exists\tau\big[\tau \overset{0}{\sim} \sigma \wedge \mathsf{val}(\tau,0) = \mathsf{S}v_0 \wedge \mathsf{Sat}_\tau\big(\ulcorner A(\mathsf{S}v_0,v_1)\urcorner\big)\big]\} \rightarrow$$
$$\forall v_0\exists\tau\big[\tau \overset{0}{\sim} \sigma \wedge \mathsf{val}(\tau,0) = v_0 \wedge \mathsf{Sat}_\tau\big(\ulcorner A(v_0,v_1)\urcorner\big)\big],$$

we see that, in the disentangled setting, it is not even well formed. In the second line, for example, the variable v_0 that is bound by the universal quantifier must come from the *syntactic* language: it ranges over expressions. But $\mathsf{val}(\tau,0)$, the value that τ assigns to the first variable, is in the domain of the *object* language. So $\mathsf{val}(\tau,0) = v_0$ makes no sense, and the same is true of the second conjunct on every other line.

Admittedly, then, several issues remain concerning exactly what adding the Sat-sentences to a given theory, even in a disentangled way, gives us, in terms of logical strength, at least in the case when we extend induction. If we do not extend induction, then the same sorts of results as we had earlier are available in the disentangled setting, too. (That is, we have analogues of Theorem 2.1 and Theorem 2.2.) And it seems likely that $\mathsf{DDS}_{\mathcal{A}}[\mathsf{I}\Sigma_1] + \mathsf{I}\Sigma_1$ will prove to be weaker than $\mathsf{TT}_{\mathcal{A}}[\mathsf{I}\Sigma_1] + \mathsf{I}\Sigma_1$, since it is difficult to see how $\mathsf{DDS}_{\mathcal{A}}[\mathsf{I}\Sigma_1] + \mathsf{I}\Sigma_1$ could possibly prove $\mathsf{Con}(\mathsf{I}\Sigma_1)$. At the very least, the sort of proof that is available in $\mathsf{TT}_{\mathcal{A}}[\mathsf{I}\Sigma_1] + \mathsf{I}\Sigma_1$ will not available in $\mathsf{DDS}_{\mathcal{A}}[\mathsf{I}\Sigma_1] + \mathsf{I}\Sigma_1$, since $\mathsf{DDS}_{\mathcal{A}}[\mathsf{I}\Sigma_1] + \mathsf{I}\Sigma_1$ is not even going to be able to prove, say, that modus ponens is valid. To do that, you need to be able to reason about conditionals generally, and $\mathsf{DDS}_{\mathcal{A}}[\mathsf{I}\Sigma_1] + \mathsf{I}\Sigma_1$ has no resources for doing so. We will be able to prove of *each* instance of modus ponens that is it valid, but not that all of them are.

5 Objections (II)

All of that said, there is another worry one might have about the framework we are now using.

As we have seen, if $\mathcal{T}$ is finitely axiomatized, then $\mathsf{TT}_{\mathcal{L}}[\mathsf{I}\Sigma_1] + \mathcal{T}$ proves $\mathsf{Con}(\mathcal{T})$. It is important to understand, however, that the particular sentence $\mathsf{Con}(\mathcal{T})$ that is being proved is a sentence of the *syntactic* language $\mathcal{S}$. If our syntax were stated as a theory of concatenation, then the consistency statement would be formulated using concatenation and other syntactic notions defined in terms of it. Of course, in the sorts of cases in which we are primarily interested, there will also be a sentence of the *object* language $\mathcal{L}$ that expresses the claim that $\mathcal{T}$ is consistent. (There will be many such sentences, in fact.) So we need to distinguish the sentence $\mathsf{Con}_{\mathcal{S}}(\mathcal{T})$ of the syntactic language that I have said can be proven in $\mathsf{TT}_{\mathcal{L}}[\mathsf{I}\Sigma_1] + \mathcal{T}$ from the sentence $\mathsf{Con}_{\mathcal{L}}(\mathcal{T})$ of the object language about which I have so far said nothing.

And, indeed, the object language sentence $\mathsf{Con}_{\mathcal{L}}(\mathcal{T})$ *cannot* be proven in $\mathsf{TT}_{\mathcal{L}}[\mathsf{I}\Sigma_1] + \mathcal{T}$. This follows from a much more general observation, due to

Halbach, that even $\mathsf{TT}_{\mathcal{L}}[\mathsf{PA}] + \mathcal{T}$ is a conservative extension of $\mathcal{T}$ (see [23, Section 3.2]). The thought, then, is that this shows that there is something unnatural about the framework that results from our disentangling the syntactic theory from the object theory. Recall, for example, the following quote from Field:

> . . . [T]he way in which we "learn more about the natural numbers by invoking truth" is that in having that notion we can rigorously formulate a more powerful arithmetical theory than we could rigorously formulate before. There is nothing very special about truth here: using any other notion not expressible in the original language we can get new instances of induction, and in many cases these lead to nonconservative extensions. ([9, p. 536])

Disentangling the syntax from the object theory might have seemed like a good idea, but if we do so, then we *never* get nonconservative extensions! Disentangling thus seems to cost us the ability to use truth to learn more about the natural numbers in the way we thought we could. So maybe we should reconsider.

I understand why one might have such a reaction. To be honest, when Halbach first mentioned his observation to me, I was both surprised and puzzled. On further reflection, however, it has come to seem to me that the situation here is exactly as it should be. There is nothing to *stop* us from using truth to learn more about arithmetic. The interesting questions are (i) what we need to add to $\mathsf{TT}_{\mathcal{L}}[\mathsf{I}\Sigma_1] + \mathcal{T}$ if we are to do so, and (ii) why we should need to add it.

Halbach's original proof of his observation was a straightforward generalization of an earlier model-theoretic proof, due to Craig and Vaught [4, p. 298, Lemma 2.7], that $\mathsf{TT}^{-}_{\mathcal{L}}[\mathsf{Q}]+\mathcal{T}$ is a conservative extension of $\mathcal{T}$. But, if we limit attention to finitely axiomatized theories, then there is an easier proof that is, in the present context, more illuminating.[50]

Proposition 5.1 *If $\mathcal{T}$ is a finitely axiomatized, consistent theory in $\mathcal{L}$, then $\mathsf{TT}_{\mathcal{L}}[\mathsf{PA}] + \mathcal{T}$ is a conservative extension of $\mathcal{T}$.*

Proof Let A be any nontheorem of $\mathcal{T}$. So $\mathcal{T} + \neg A$ is consistent, and it is finitely axiomatized. So, if $\mathsf{TT}_{\mathcal{L}}[\mathsf{PA}] + \mathcal{T}$ proved A, then so would $\mathsf{TT}_{\mathcal{L}}[\mathsf{PA}] + [\mathcal{T} + \neg A]$, which would then be inconsistent. But $\mathsf{TT}_{\mathcal{L}}[\mathsf{PA}] + [\mathcal{T} + \neg A]$ is not inconsistent since, by Theorem 5.2, to be mentioned shortly, it is locally interpretable in $\mathsf{PA} + \mathsf{Con}(\mathcal{T} + \neg A)$, which is not just consistent but true. □

The case of $\mathsf{Con}_{\mathcal{L}}(\mathcal{T})$ is just a special case of this more general result. By the second incompleteness theorem, $\mathcal{T} + \neg\mathsf{Con}_{\mathcal{L}}(\mathcal{T})$ is consistent if $\mathcal{T}$ is, and it is finitely axiomatized if $\mathcal{T}$ is, as well. So it follows from Theorem 4.14 that $\mathsf{TT}_{\mathcal{L}}[\mathsf{PA}] + [\mathcal{T} + \neg\mathsf{Con}_{\mathcal{L}}(\mathcal{T})]$ proves $\mathsf{Con}_{\mathcal{S}}(\mathcal{T} + \neg\mathsf{Con}_{\mathcal{L}}(\mathcal{T}))$. But now it is clear that it had *better* be $\mathsf{Con}_{\mathcal{S}}(\mathcal{T} + \neg\mathsf{Con}_{\mathcal{L}}(\mathcal{T}))$ that we are proving, and not $\mathsf{Con}_{\mathcal{L}}(\mathcal{T} + \neg\mathsf{Con}_{\mathcal{L}}(\mathcal{T}))$. If $\mathsf{TT}_{\mathcal{L}}[\mathsf{PA}] + [\mathcal{T} + \neg\mathsf{Con}_{\mathcal{L}}(\mathcal{T})]$ proved $\mathsf{Con}_{\mathcal{L}}(\mathcal{T} + \neg\mathsf{Con}_{\mathcal{L}}(\mathcal{T}))$, then, since $\mathsf{Con}_{\mathcal{L}}(\mathcal{T})$ trivially follows from $\mathsf{Con}_{\mathcal{L}}(\mathcal{T} + \neg\mathsf{Con}_{\mathcal{L}}(\mathcal{T}))$, $\mathsf{TT}_{\mathcal{L}}[\mathsf{PA}] + [\mathcal{T} + \neg\mathsf{Con}_{\mathcal{L}}(\mathcal{T})]$ would prove $\mathsf{Con}_{\mathcal{L}}(\mathcal{T})$ and so would be inconsistent. Which, again, it is not.

From a model-theoretic point of view, then, what is happening is that the only information we have about the structure of the $\mathcal{L}$-related part of models of $\mathsf{TT}_{\mathcal{L}}[\mathsf{PA}] + \mathcal{T}$ is what is provided by the object theory $\mathcal{T}$. The truth-theory for $\mathcal{L}$—the $\mathsf{TT}_{\mathcal{L}}[\mathsf{PA}]$ part—does not constrain the structure of the part of the model for the object language at all. In particular, there is nothing in the theory of truth that

requires the domain of $\mathcal{L}$ to be in any way "standard" or, to be more precise, to be standard *relative to* the syntax.

A model of $\mathsf{TT}_{\mathcal{L}}[\mathsf{PA}] + \mathcal{T}$ consists, more or less, of a model of PA, considered as our syntax, and a model of $\mathcal{T}$, considered as our object theory, plus some semantic pieces that connect these two parts. It is easy to see that, if $\mathcal{L}$ is the language of arithmetic, then the domain of the syntactic language has to be isomorphic to an initial segment of the domain of the object language. This is because we can prove in $\mathsf{TT}_{\mathcal{L}}[\mathsf{PA}] + \mathcal{T}$ that every numeral denotes a number—a member of the domain of the object language—and because the numbers so denoted will be isomorphic to the numerals that denote them.[51] But the converse need not be true. We have no way to prove that every number is denoted by a numeral. So it is perfectly possible for the domain of the object language *not* to be an initial segment of that of the syntactic language. In particular, the model of the syntactic part of the theory could be standard, and the model of the arithmetical part could be nonstandard. As a result, it is perfectly possible for $\mathsf{Con}_{\mathcal{S}}(\mathcal{T})$ to be true in the model, even though $\mathsf{Con}_{\mathcal{L}}(\mathcal{T})$ is false in the model.

And that, it seems to me, is absolutely as it should be. A theory of truth for the language of arithmetic *should not* tell us anything specific about the domain over which the object-language variables range. It should simply take the domain as given, much as it takes the interpretation of the primitives of the object language as given: "0" denotes 0, whatever that is; $<$ is true of $\langle x, y\rangle$ just in case $x < y$, whatever that means; and the variables range over, well, whatever it is they range over. When we do model theory, we do not take the interpretations of the primitives as given. We may take "0" to refer to $\emptyset$; we may take $<$ to be true of $\langle x, y\rangle$ just in case some complicated condition obtains; and we may take the domain to be whatever we like, sets for the language of arithmetic or numbers for the language of set theory. But we are not doing model theory. We are doing semantics.[52]

Similarly, a theory of truth for the language of arithmetic *should not*, all by itself, allow us to prove that every number is denoted by a numeral, let alone allow us to prove new purely arithmetical theorems. The following is no doubt a plausible argument: 0 is denoted by a numeral; if n is denoted by a numeral, then $n + 1$ is denoted by a numeral; so every number is denoted by a numeral. But to make this argument, we need to use "extended" induction *over the natural numbers*, which is something to which we were not previously committed and to which we cannot be committed simply because we have decided to theorize semantically about the language of arithmetic. To put the point more generally, the mere fact that we have a theory of truth for some language $\mathcal{L}$ cannot, all by itself, force us to accept new principles concerning whatever it is that $\mathcal{L}$ talks about, that is, to add new axioms to whatever theory stated in $\mathcal{L}$ we might antecedently have accepted. In that sense, then, Field is absolutely right. The way we can use truth to learn more about the natural numbers is indeed to use it to "formulate a more powerful *arithmetical* theory" (see [9, p. 536, my emphasis]). And that is something we can do if we wish. If we want, we can extend whatever induction axioms we accept to permit semantic vocabulary. My point, again, is simply that we cannot be committed to doing so simply because we have a theory of truth for the language of arithmetic, even a fully compositional one. That theory is a *semantic* theory, one about expressions and truth. It is not, in its own right, an *arithmetical* theory, one about numbers, and simply having it cannot force us to accept new instances of induction.

Does that mean that Field wins? No, because, once the syntax has been disentangled from the object theory, it becomes clear that the issue should never have concerned conservativity over the *object* theory.[53] Surely one would not expect our theory about the language we use to talk about physical reality, say, to entail new substantive facts about physical reality when added to whatever physical theory we happen to accept. And the same is true for arithmetic and the language we use to talk about it. Everyone, deflationist or otherwise, should therefore agree that a semantics for the language we use to talk about some subject matter should be conservative over our theory of that subject matter.[54] The right question to ask is therefore not whether a semantic theory for a given language $\mathcal{L}$ is conservative over theories stated in $\mathcal{L}$, but whether it is conservative over a purely *syntactic* theory for $\mathcal{L}$. The right question is not what we can learn about *numbers* by using the notion of truth, but what we can learn about *expressions* by using the notion of truth.

With that change, however, the entire dialectic that has surrounded the issue of conservativity transfers smoothly. We can learn a lot about expressions if we have access to semantic notions. If we have a fully compositional theory of truth for a language $\mathcal{L}$, for example, then we can use induction on the complexity of expressions to prove the consistency of any finitely axiomatized theory in $\mathcal{L}$ that we are prepared to accept and, more generally, to prove the consistency of any theory all of whose axioms we regard as true. The statement that a theory is (deductively) consistent is a purely syntactic statement. Semantics is therefore not conservative over syntax.

Of course, someone might respond:

> . . . [T]he way in which we "learn more about [expressions] by invoking truth" is that in having that notion we can rigorously formulate a more powerful [syntactic] theory than we could rigorously formulate before. There is nothing very special about truth here. . . . [9, adapted from p. 536]

But, as before, it is not enough simply to have truth and some extra induction: $\mathsf{DDT}_{\mathcal{A}}[\mathsf{PA}]$ does not allow us to prove $\mathsf{Con}(\mathcal{T})$, even if we add "all axioms of $\mathcal{T}$ are true." The compositional principles are also needed for such a proof. So we may want to know which of these is doing more of the work: Is it the extension of induction that is responsible for the increase in strength? Or is it the compositional principles? The answer to that question emerges from the mathematical facts summarized in Table 1. Adding a compositional theory of truth for a language $\mathcal{L}$ to some finitely axiomatized theory stated in $\mathcal{L}$ adds significant logical strength, whether or

Table 1 The mathematical facts ($\mathcal{U} = \mathsf{I}\Sigma_n, \mathsf{PA}$).

Base: $\mathcal{U} + \mathcal{T}$	No New Induction	Extend Induction
Add the T-sentences	Locally interpretable	$\mathsf{DDT}_{\mathcal{L}}[\mathcal{U}] + \mathcal{T}$ Locally interpretable
Add the Sat-sentences	Locally interpretable	$\mathsf{DDS}_{\mathcal{L}}[\mathcal{U}] + \mathcal{T}$ Unclear
Add a fully compositional truth-theory	$\mathsf{TT}^{-}_{\mathcal{L}}[\mathcal{U}] + \mathcal{T}$ Not interpretable	$\mathsf{TT}_{\mathcal{L}}[\mathcal{U}] + \mathcal{T}$ Not interpretable, and stronger still

not we extend the induction axioms, whereas adding a noncompositional theory adds little if any logical strength, even if we *do* extend the induction axioms.

Now, to be sure, there is a special case in which the object theory is itself a theory of syntax and the object language is the language of that very theory of syntax. In that case, one might think, we have no choice but to entangle syntax with the object theory, so that we collapse back into the more familiar framework used in Section 2, in which case the objections discussed in Sections 3.1 and 3.2 are restored. But, first of all, while this sort of case—in which self-reference is not only possible but natural—is important, it seems to me obvious that it is a special case. And even in this special case, it is still important to distinguish the role played by our theory of syntax *qua* theory of syntax from the role it plays *qua* object theory, for all the reasons given in Section 3.3. That is what disentangling allows us to do.

For some, the disentangled framework may still feel unnatural somehow. If so, then consider the fact that, if we do disentangle the syntactic theory from the object theory, we not only get improved results like the ones discussed in Section 4 but results like the following.

Theorem 5.2 ([18, Section 4.5]) *Suppose that $\mathcal{T}$ is finitely axiomatized. Then*

(i) *for all $n \geq 1$, $\mathsf{TT}_{\mathscr{L}}[\mathsf{I}\Sigma_n] + \mathcal{T}$ is mutually interpretable with $\mathsf{I}\Sigma_n + \mathsf{Con}(\mathcal{T})$;*
(ii) *$\mathsf{TT}_{\mathscr{L}}[\mathsf{PA}] + \mathcal{T}$ is mutually locally interpretable with $\mathsf{PA} + \mathsf{Con}(\mathcal{T})$.*

I for one trust mathematical elegance much more than I trust intuitive judgments about what seems natural, especially when those judgments have been shaped by decades of doing things one particular way. And the elegance of the results just mentioned, in my opinion, makes a very strong case for any framework that permits them to be formulated and proved.

Even in the special case in which our syntactic theory and our object theory are formulated in the same language, then, my response to the objections discussed in Sections 3.1 and 3.2 is the same: what is responsible for the phenomena on which they rest is the interaction of our syntactic theory with our object theory. We can allow these theories to interact if we like, but the familiarity of the usual setting that does not even distinguish them should not blind us to what we are doing. Even in this case, we can still distinguish between the two roles a single theory might play and investigate them formally, using the framework developed in Section 4. And we *ought* to distinguish those roles, too, since it is only if we do so that certain insights can be properly stated. Such facts do not lapse simply because we choose, for different reasons, to work in a setting in which the questions to which the relevant results provide answers cannot even be formulated.

6 Closing

I began this article by recalling the history of the debate over the conservativeness argument against deflationary theories of truth, and we have just had reason to recall that history again. Nonetheless, my purpose here has not been to revive that debate. My purpose, rather, has been to argue that compositional theories of truth are non-trivial, in the sense that they have significant logical strength. Although such theories can only be used to *prove* consistency when we extend the induction axioms, they allow us to *interpret* consistency statements even when we do not, and that is known to be logically significant.

For what it is worth, I do not myself take this result to show that Shapiro and Ketland were right and that Field was wrong. Shapiro and Ketland are to be applauded for trying to find some concrete content in the gnomic pronouncements some deflationists have made about truth's "insubstantiality," but I tend to agree with Halbach [16, p. 188] that "... it is hard to see why the deflationist should be committed to conservativeness at all." Since it also seems hard to see why a deflationist should be committed to the logical vacuity of whatever truth-theoretic principles she might accept,[55] I do not take the results proven here to "refute" deflationism. But they certainly do show that the compositional principles are not the trivialities they are often taken to be.[56]

It thus becomes an important question what right deflationists have to such compositional principles and how they should understand the role the notion of truth plays in them. Those questions, however, are ones I will have to discuss elsewhere (see [20]).[57]

Notes

1. By a T-sentence, I of course mean one of the form: $\ulcorner A \urcorner$ is true if and only if A.

2. I shall omit quotes where they are not absolutely necessary, so as not to clutter the exposition.

3. Field [10] has argued that the compositional principles follow from the T-scheme, if it is understood in the right way. I criticize this claim elsewhere (see [20]).

4. Results not proven here are proven in a companion paper (see [18]).

5. Strictly, $\mathcal{T}'$ conservatively extends $\mathcal{T}$ if (i) whenever $\mathcal{T} \vdash A$, then $\mathcal{T}' \vdash A$, and (ii) whenever $\mathcal{T}' \vdash A$ and A is in the language of $\mathcal{T}$, then $\mathcal{T} \vdash A$.

6. In fact, there are several different notions of interpretation. We shall only need this one.

7. As well as proofs of the translations of the axioms, we also need proofs of $\delta(t^*)$, for each atomic term t of $\mathcal{L}_\mathcal{T}$, and of the closure condition

$$\forall x_1 \cdots x_n \big(\delta(x_1) \wedge \cdots \wedge \delta(x_n) \rightarrow \delta\big(f^*(x_1, \ldots, x_n)\big)\big)$$

for each primitive function-symbol f, of however many places. We also need (if this is not already covered) a proof that the domain is nonempty: $\exists x \delta(x)$. It is also convenient to allow terms and function-symbols to be translated using descriptions, which can then be eliminated as Russell taught. In that case, we need $\mathcal{B}$ to prove that the descriptions are proper.

8. Facts concerning interpretability can generally be verified in the theory known as $\mathsf{I}\Delta_0 + \Omega_1$, which is itself interpretable in Q. Note that any theory this strong will be subject to the second incompleteness theorem (see [35]).

9. As Feferman [7, Theorem 5.9] famously showed, if $\mathcal{B}$ is reflexive, in the sense to be mentioned shortly, then there are ways $\mathcal{B}'$ of specifying the same set of axioms so that $\mathcal{B}'$ will prove $\mathsf{Con}(\mathcal{B}')$. We will ignore this complication here, however, and assume that all our theories are specified in "nice" ways.

10. The proof of this result was first published by Feferman [7, Theorem 6.9].

11. These are customarily abbreviated: $\forall x < t(\cdots)$ and $\exists x < t(\cdots)$.

12. Q7 is then redundant and is typically omitted.

13. That $\mathsf{I}\Delta_0$ is *locally* interpretable in Q was first proven by Nelson [25]. Wilkie proved that it is globally interpretable in Q. The proof is discussed by both Hájek and Pudlák [13, pp. 366–70] and by Burgess [2, Section 2.2].

14. Visser [33] gives lots of information about sequentiality.

15. What we would normally have, in the language of arithmetic, is a formula $\mathsf{val}(x, y, z)$ meaning "z is the value x assigns to the yth variable," rather than a functional expression as in the text. But these complications affect nothing that follows and clutter the exposition, so I shall ignore them.

16. It appears to have been Wang [34] who first worked out the details of this sort of construction.

17. More formally, the theories in which we are interested can be characterized in terms of the relativized arithmetical hierarchy (see [13, pp. 81ff]).

18. There are some interesting questions still open here. Enayat has asked, for example, whether $\mathsf{DT}^-[\mathcal{T}]$ can ever be *globally* interpretable in $\mathcal{T}$, if $\mathcal{T}$ is finitely axiomatized, and the same sort of question arises for $\mathsf{DT}[\mathcal{T}]$, as well. If not, then that would show that even just adding the T-sentences to a finitely axiomatized theory always increased the theory's logical strength (though, as we shall see, not nearly as much as adding a fully compositional theory). See endnote 31 for some related remarks.

19. I learned Theorem 2.5 from Visser, who tells me that he regards it as "folklore."

20. This technique is originally due to Solovay. Burgess [2, Section 2.2] gives an accessible treatment. A more complete treatment is presented in Hájek and Pudlák [13, pp. 366ff]. The basic idea is that, if $A(x)$ is inductive in some theory $\mathcal{U}$—that is, if $\mathcal{U} \vdash A(0)$ and $\mathcal{U} \vdash A(x) \rightarrow A(Sx)$—then, from $\mathcal{U}$'s point of view, the natural numbers might as well just *be* the numbers that satisfy $A(x)$. Unsurprisingly, that is not quite right, but the details can be made to work.

21. Note that, if $\mathcal{T}$ is infinitely axiomatized, we will have to choose some specification of its axioms, *both* in order to formalize "all axioms of $\mathcal{T}$ are true" *and* to formalize $\mathsf{Con}(\mathcal{T})$. In Theorem 2.8, then, we are using the same specification both times.

22. This will also be true in many cases when $\mathcal{T}$ is just finitely axiomatiz*able*. It is an interesting exercise to prove this. But we will officially stick to the case of theories that are actually finitely axiomatized.

23. Leigh and Nicolai [23, Section 3.1] give a detailed proof.

24. Since $\mathsf{CT}^-[\mathsf{I}\Sigma_n]$ is itself finitely axiomatized, it is not *locally* interpretable in $\mathsf{I}\Sigma_n$.

25. It will be slightly stronger if, as mentioned in endnote 18, $\mathsf{DT}[\mathsf{I}\Sigma_n]$ turns out not to be globally interpretable in $\mathsf{I}\Sigma_n$.

26. This result is relevant to Halbach's claim that the "uniform disquotation scheme"—our $\mathsf{DS}[\cdot]$—is plausibly analytic, since $\mathsf{DS}[\mathsf{PA}]$ is a conservative extension of PA (see [15, Section 2]). What we are about to see is that this result depends crucially upon the choice of PA as base theory. Whether one takes conservativity to be required for analyticity or regards it as merely indicative of it, the uniform disquotation scheme appears to be logically quite strong, transforming $\mathsf{I}\Sigma_1$ into a theory containing PA. It is only in very special cases that it gives us nothing we did not already have.

27. This result can surely be improved by bounding the quantifier $\exists\tau$ in the first displayed formula in the proof. It is not clear to me just how good the bound can be made to be, however—that will depend upon exactly how we code sequences—so I am not sure whether we can show that $\mathsf{DS}[\mathsf{I}\Delta_0]$ contains PA. But it seems almost certain that $\mathsf{DS}[\mathsf{I}\Delta_0 + \Omega_1]$ contains PA.

28. The crucial point here is that, except for Sat, everything here is primitive recursive and so is Δ_1 in $\mathsf{I}\Sigma_1$.

29. This is where an attempt to use truth and substitution to prove a similar result would break down. Let $\mathsf{num}(x)$ be the numeral for x. Then we can certainly consider the formula
$$\mathsf{T}(\ulcorner A\big(\mathsf{num}(z), \mathsf{num}(v_1)\big)\urcorner)$$
and the induction axiom for it will be available in $\mathsf{DT}[\mathsf{I}\Sigma_1]$. But there is no T-sentence for $\mathsf{T}(\ulcorner A(\mathsf{num}(z), \mathsf{num}(v_1))\urcorner)$, with variable z and v_1. There are only T-sentences for the various instances of this formula, and we can use only finitely many of them at a time.

30. By the following calculation:
$$\begin{gathered}\exists\tau\big[\tau \overset{0}{\sim} \sigma \wedge \mathsf{val}(\tau, 0) = z \wedge \mathsf{Sat}_\tau\big(\ulcorner A(v_0, v_1)\urcorner\big)\big] \\ \exists\tau\big[\tau \overset{0}{\sim} \sigma \wedge \mathsf{val}(\tau, 0) = z \wedge A\big(\mathsf{val}(\tau, 0), \mathsf{val}(\tau, 1)\big)\big] \\ \exists\tau\big[\tau \overset{0}{\sim} \sigma \wedge A\big(z, \mathsf{val}(\tau, 1)\big)\big] \\ A\big(z, \mathsf{val}(\sigma, 1)\big)\end{gathered}$$
The last step uses the fact that, since $\tau \overset{0}{\sim} \sigma$, $\mathsf{val}(\sigma, 1) = \mathsf{val}(\tau, 1)$.

31. Visser has observed that $\mathsf{DS}^-[\mathcal{T}]$ is never *globally* interpretable in $\mathcal{T}$, if $\mathcal{T}$ is finitely axiomatized. Even the Sat-sentences by themselves, then, are not logically trivial.

32. Note that it follows from Theorem 3.2 that $\mathsf{CT}^-[\mathsf{PA}]$ does not prove that all axioms of PA are true.

33. An objection along these lines was communicated to me by Burgess, whose student Noel Swanson had made it in response to an earlier manuscript of mine that covered much of the material we are discussing (see [19]). Thanks to both of them for the objection.

34. Something like this result has presumably been known for some time, though I do not know of any previous statement of the result in this sort of form: as applied to axiomatic

theories of truth. I present a detailed proof elsewhere (see [18, Section 3.4]). It may be that this result is not the best possible and that it can be strengthened to $\mathcal{T} \supseteq \mathsf{I}\Delta_0$. It is unclear, however, whether that is true, for reasons I discuss in the paper just cited. Note that the remarks made in endnote 21 apply here, too.

35. Of course, this is true only so long as $\mathcal{U}$ is a subtheory of PA. In that respect, then, Theorem 3.3 can be used to "pull apart" the roles played by the object theory and the base theory, so long as $\mathcal{T}$ is an arithmetical theory that PA does not already prove consistent. But that seriously limits the scope of the result, and Corollary 3.4 then has limited application, since there are no finitely axiomatizable extensions of PA in the same language.

36. Wang [34, p. 260] credits Rosser with the observation that we do need to ask how CT[PA] manages to prove that *all* of PA's axioms are true, and not just that each of them is, and he gives the first detailed proof of that fact. I give a more modern presentation of this result elsewhere (see [18, Section 3.4]).

37. It is, at present, unclear whether this result can be strengthened to give us a proper converse of Theorem 2.5. It would be really nice if it could, since we could then conclude that $\mathsf{CT}^{-}[\mathcal{T}]$ was mutually interpretable with $\mathsf{Q} + \mathsf{Con}(\mathcal{T})$. As we shall see below, though, we do get this result in the disentangled setting.

38. One well known such theory is primitive recursive arithmetic, but its language is not finite, and it is not clear how the results proven here apply to such theories.

39. Simply because $\mathcal{L}$ might be the language of set theory, and there are way too many sets to code even finite sequences of them as numbers.

40. The missing details are provided elsewhere (see [18, Section 4.1]).

41. There are again questions about what exactly it means to extend the induction scheme, in general. But we will limit our attention to cases where it is clear what it means.

42. Because our theory is many-sorted, quantifiers of any of the three types could now appear in the induction axioms. That leads to the question what exactly we mean by a Σ_n-formula in the present setting. It turns out that we can ignore the differences between types of quantifiers for our purposes. Thus, for example, $\exists \mathrm{x}(\mathsf{Den}_\sigma(t, \mathrm{x}))$ counts as Σ_1 for our purposes, and $\forall\sigma\exists t\exists \mathrm{x}(\mathsf{Den}_\sigma(t, \mathrm{x}))$ counts as Π_2.

43. This means, among other things, that there is a materially adequate, fully compositional theory of truth for the language of ZFC that is interpretable in Q.

44. If $\mathcal{L}$ contains no terms other than variables, then we may not be able to specify a one-element domain via a formula $\delta(x)$ with just x free. In that case, we will have to use a parameter, which means that $\mathsf{TT}^{-}_{\mathcal{L}}[\mathsf{Q}]$ will only be parametrically interpretable in Q.

45. If $\mathcal{T}$ is some finitely axiomatizable subtheory of PA, then this is of course a boring result. There is no reason whatsoever to expect, say, $\mathsf{TT}^{-}_{\mathcal{L}}[\mathsf{PA}] + \mathsf{I}\Sigma_2$ to be interpretable in $\mathsf{I}\Sigma_2$, and in fact PA is not interpretable in any of its finitely axiomatizable subtheories (see [7, Theorem 6.8]). But of course $\mathcal{T}$ need not be a subtheory of PA. It could, for example, be $\mathsf{Q} + \mathsf{Con}(\mathsf{PA})$, and then Corollary 4.11 tells us that $\mathsf{TT}^{-}_{\mathcal{L}}[\mathsf{PA}] + (\mathsf{Q} + \mathsf{Con}(\mathsf{PA}))$ is not interpretable in $\mathsf{Q} + \mathsf{Con}(\mathsf{PA})$. It is one of the advantages of the present way of

proceeding, however, that $\mathcal{T}$ need not even be formulated in the language of arithmetic. We know that $\mathsf{ZF} \vdash \mathsf{Con}(\mathsf{PA})$. So let $\mathcal{C}$ be the finite set of axioms of ZF that are used in that proof. Since $\mathcal{C} \vdash \mathsf{Con}(\mathsf{PA})$, it follows from results of Feferman's that $\mathcal{C}$ is not interpretable in PA. Yet it still follows from Corollary 4.11 that $\mathsf{TT}^{-}_{\mathcal{L}}[\mathsf{Q}] + \mathcal{C}$, and so $\mathsf{TT}^{-}_{\mathcal{L}}[\mathsf{PA}] + \mathcal{C}$, is not interpretable in $\mathcal{C}$. Clearly, it is the theory of truth that is responsible for the extra strength.

46. I have made a similar complaint elsewhere (see [17, Section 3]).

47. DDT stands for *disentangled disquotational truth*, DDS for *disentangled disquotational satisfaction*.

48. It is true that $\mathsf{DDS}_{\mathcal{L}}[\mathsf{I}\Sigma_n]$ is interpretable in $\mathsf{I}\Sigma_n$: we can give $\mathcal{L}$ a trivial interpretation again. But we can do that for $\mathsf{TT}_{\mathcal{L}}[\mathsf{I}\Sigma_n]$, as well.

49. This is because $\mathsf{I}\Sigma_2$ is the same theory as $\mathsf{I}\Sigma_1$ plus reflection for Σ_3 formulas (see [1, p. 231, Theorem 7]). So $\mathsf{I}\Sigma_2$ proves $\mathsf{Con}(\mathsf{I}\Sigma_1 + \mathsf{Con}(\mathsf{I}\Sigma_1))$. Thanks to Volodya Shavrukov for confirming my suspicion and for the reference.

50. The same proof works for subtheories of PA such as $\mathsf{I}\Sigma_1$, taken as the syntactic theory.

51. Note that this will be true even if there are nonstandard numerals.

52. I will not pursue the issue here, but if one wanted to formalize model-theoretic reasoning in the sort of framework in which we are working, then what one would need to do is add a third sort of language, that in which the model is to be described, and a "theory of models" that allows to reason about their structure. The truth-theory would then no longer be homophonic, but would interpret the object language by using the language of the theory of models. What was the object theory then becomes part of our theory about the structure of the model. Its role is simply to ensure that certain statements of the object language come out true in that model.

53. A similar point is made by Leigh and Nicolai [23, Section 4.1].

54. Yes, there is a special case. We will get to it.

55. Field [9, p. 534] emphasizes that no deflationist has ever held that truth is "expressively" insubstantial.

56. In my view (see [17, Section 4]), the T-sentences themselves are not trivialities, either, but for quite different reasons.

57. This paper is one of many to emerge from an earlier manuscript, "The Strength of Truth Theories" [19], that ultimately became unmanageable. Thanks to Volker Halbach and Jeff Ketland for conversations early in the history of my work on this topic, and to Josh Schechter for conversations later on, that helped greatly. Comments on the earlier manuscript from Cezary Cieśliński and Ali Enayat were also very helpful. Thanks also to two anonymous referees for their remarks. Talks incorporating some of these ideas were given at a conference on philosophical logic, organized by Delia Graff Fara and held at Princeton University in April 2009; at the New England Logic and Language Colloquium and at the Philosophy of Mathematics Seminar at Oxford University, both in May 2011;

and at a meeting of the Logic Group at the University of Connecticut, in April 2012. Thanks to everyone present for their questions and comments, especially J. C. Beall, John P. Burgess, Hartry Field, Daniel Isaacson, Graham Leigh, Carlo Nicolai, Charles Parsons, Agustín Rayo, and Lionel Shapiro, as well as Volker and Josh, again. Special thanks to my commentator at Princeton, Josh Dever, whose comments were insightful and lucid, as well as helpful. I owe the greatest debt, however, to Albert Visser. Just as my ideas were starting to come together, discussions with Albert transformed the direction of this project. It was from him that I learned of Theorem 2.8 and its attendant corollaries, which led, of course, to the idea that we should focus on interpretability, not on conservativity, which is pretty much the central idea of this paper. Albert has also read many drafts along the way and provided extensive feedback. Obviously, he bears no responsibility for what I have done with the idea, and I am not sure he would agree with how I have developed it. Nonetheless, this paper would never have been written without Albert's assistance, for which I am extremely grateful.

References

[1] Beklemishev, L. D., "Reflection schemes and provability algebras in formal arithmetic," *Russian Mathematical Surveys*, vol. 60 (2005), pp. 197–268. Zbl 1097.03054. MR 2152943. DOI 10.1070/RM2005v060n02ABEH000823. 30

[2] Burgess, J. P., *Fixing Frege*, Princeton University Press, Princeton, 2005. Zbl 1089.03001. MR 2157847. 27

[3] Corcoran, J., W. Frank, and M. Maloney, "String theory," *Journal of Symbolic Logic*, vol. 39 (1974), pp. 625–37. Zbl 0298.02011. MR 0398771. DOI 10.2307/2272846. 16

[4] Craig, W., and R. L. Vaught, "Finite axiomatizability using additional predicates," *Journal of Symbolic Logic*, vol. 23 (1958), pp. 289–308. Zbl 0085.24601. MR 0106175. 22

[5] Enayat, A., and A. Visser, "Full satisfaction classes in a general setting (Part I)," preprint, https://pdfs.semanticscholar.org/730d/be402772e16926179b92bfa1416f636ce340.pdf (accessed 21 May 2017). 15

[6] Enayat, A., and A. Visser, "New constructions of satisfaction classes," pp. 321–35 in *Unifying the Philosophy of Truth*, edited by T. Achourioti, H. Galinon, J. M. Fernández, and K. Fujimoto, Springer, New York, 2015. DOI 10.1007/978-94-017-9673-6. 13

[7] Feferman, S., "Arithmetization of metamathematics in a general setting," *Fundamenta Mathematicae*, vol. 49 (1960/1961), pp. 35–92. Zbl 0095.24301. MR 0147397. 3, 4, 5, 10, 26, 27, 29

[8] Field, H., "Deflationist views of meaning and content," *Mind*, vol. 103 (1994), pp. 249–85. MR 1297713. DOI 10.1093/mind/103.411.249. 3

[9] Field, H., "Deflating the conservativeness requirement," *Journal of Philosophy*, vol. 96 (1999), pp. 533–40. MR 1718770. DOI 10.2307/2564613. 2, 22, 23, 24, 30

[10] Field, H., "Compositional principles vs. schematic reasoning," *The Monist*, vol. 89 (2006), pp. 9–27. 3, 26

[11] Grzegorczyk, A., "Undecidability without arithmetization," *Studia Logica*, vol. 79 (2005), pp. 163–230. Zbl 1080.03004. MR 2135033. DOI 10.1007/s11225-005-2976-1. 16

[12] Gupta, A., "A critique of deflationism," *Philosophical Topics*, vol. 21 (1993), pp. 57–81. 1

[13] Hájek, P., and P. Pudlák, *Metamathematics of First-order Arithmetic*, Springer, New York, 1993. Zbl 0781.03047. MR 1219738. DOI 10.1007/978-3-662-22156-3. 6, 27

[14] Halbach, V., "Disquotationalism and infinite conjunction," *Mind*, vol. 108 (1999), pp. 1–22. MR 1682631. DOI 10.1093/mind/108.429.1. 19

[15] Halbach, V., "Disquotational truth and analyticity," *Journal of Symbolic Logic*, vol. 66 (2001), pp. 1959–73. Zbl 1002.03007. MR 1877034. DOI 10.2307/2694987. 28

[16] Halbach, V., "How innocent is deflationism?," *Synthese*, vol. 126 (2001), pp. 167–94. Zbl 0980.03004. MR 1813400. DOI 10.1023/A:1005275222332. 2, 26

[17] Heck, R. G., Jr., "Truth and disquotation," *Synthese*, vol. 142 (2004), pp. 317–52. Zbl 1072.03008. MR 2118062. DOI 10.1007/s11229-005-3719-6. 30

[18] Heck, R. G., Jr., "Consistency and the theory of truth," *Review of Symbolic Logic*, vol. 8 (2015), pp. 424–66. Zbl 06496481. MR 3388729. DOI 10.1017/S1755020314000549. 10, 11, 17, 19, 25, 26, 29

[19] Heck, R. G., Jr., "The strength of truth-theories," preprint, http://rgheck.frege.org/pdf/unpublished/StrengthOfTruthTheories.pdf (accessed 21 May 2017). 28, 30

[20] Heck, R. G., "Disquotationalism and the compositional principles," preprint, http://rgheck.frege.org/pdf/unpublished/CompositionalPrinciples.pdf, 2013 (accessed 21 May 2017). 26

[21] Horwich, P., *Truth*, Blackwell, Oxford, 1990. 1

[22] Ketland, J., "Deflationism and Tarski's paradise," *Mind*, vol. 108 (1999), pp. 69–94. MR 1682633. DOI 10.1093/mind/108.429.69. 1

[23] Leigh, G. E., and C. Nicolai, "Axiomatic truth, syntax and metatheoretic reasoning," *Review of Symbolic Logic*, vol. 6 (2013), pp. 613–36. Zbl 1350.03010. MR 3150688. DOI 10.1017/S1755020313000233. 22, 27, 30

[24] Mostowski, A., "On models of axiomatic systems," *Fundamenta Mathematicae*, vol. 39 (1952), pp. 133–58. Zbl 0053.20102. MR 0054547. 5

[25] Nelson, E., *Predicative Arithmetic*, vol. 32 of *Mathematical Notes*, Princeton University Press, Princeton, 1986. Zbl 0617.03002. MR 0869999. DOI 10.1515/9781400858927. 27

[26] Parsons, C., "Sets and classes," *Noûs*, vol. 8 (1974), pp. 1–12. MR 0472526. DOI 10.2307/2214641. 2

[27] Pudlák, P., "Cuts, consistency statements and interpretations," *Journal of Symbolic Logic*, vol. 50 (1985), pp. 423–41. Zbl 0569.03024. MR 0793123. DOI 10.2307/2274231. 10

[28] Quine, W. V. O., "Concatenation as a basis for arithmetic," *Journal of Symbolic Logic*, vol. 11 (1946), pp. 105–14. Zbl 0063.06362. MR 0018618. DOI 10.2307/2268308. 16

[29] Shapiro, S., "Proof and truth: Through thick and thin," *Journal of Philosophy*, vol. 95 (1998), pp. 493–521. MR 1651807. DOI 10.2307/2564719. 1, 2

[30] Tarski, A., "A general method in proofs of undecidability," pp. 1–35 in *Undecidable Theories*, edited by A. Tarski, A. Mostowski, and A. Robinson, vol. 13 of *Studies in Logic and the Foundations of Mathematics*, North-Holland, Amsterdam, 1953. 4

[31] Tarski, A., "The concept of truth in formalized languages," pp. 152–278 in *Logic, Semantics, and Metamathematics*, edited by J. Corcoran, Hackett, Indianapolis, 1958. MR 0736686. 1, 15, 16

[32] Tarski, A., A. Mostowski, and A. Robinson, *Undecidable Theories*, vol. 13 of *Studies in Logic and the Foundations of Mathematics*, North-Holland, Amsterdam, 1953. Zbl 0053.00401. MR 0058532. 4

[33] Visser, A., "Growing commas: A study of sequentiality and concatenation," *Notre Dame Journal of Formal Logic*, vol. 50 (2009), pp. 61–85. Zbl 1190.03052. MR 2536701. DOI 10.1215/00294527-2008-028. 27

[34] Wang, H., "Truth definitions and consistency proofs," *Transactions of the American Mathematical Society*, vol. 73 (1952), pp. 243–75. Zbl 0047.01302. MR 0049136. DOI 10.2307/1990668. 27, 29

[35] Wilkie, A. J., and J. B. Paris, "On the scheme of induction for bounded arithmetic formulas," *Annals of Pure and Applied Logic*, vol. 35 (1987), pp. 261–302. Zbl 0647.03046. MR 0904326. DOI 10.1016/0168-0072(87)90066-2. 26

Department of Philosophy
Brown University
Providence, Rhode Island
USA
rgheck@brown.edu