Stable and Unstable Theories of Truth and Syntax

Beau Madison Mount and Daniel Waxman

Draft version: please don't cite

Abstract

Recent work on formal theories of truth has revived an approach, due originally to Tarski, on which syntax and truth theories are sharply distinguished—'disentangled'—from mathematical base theories. In this paper, we defend a novel philosophical constraint on disentangled theories. We argue that these theories must be epistemically stable: they must possess an intrinsic motivation justifying no strictly stronger theory. In a disentangled setting, even if the base and the syntax theory are individually stable, they may be jointly unstable. We contend that this flaw afflicts many proposals discussed in the literature; we defend a new, stable disentangled theory, *double second-order arithmetic*.

Introduction

According to deflationist accounts of truth, truth is fundamentally *lightweight*—it serves as a device of generalization, but contributes nothing deeper to our theorizing about the world (Horwich 1980; Field 1994, 1999; Cieśliński 2017). Deflationism has a great deal of appeal: it trades on the idea that an assertion that "'snow is white" is true" has little more content than an assertion that 'Snow is white', and that grasping a generalized version of this connection—and possibly a few other, similarly straightforward principles—is all that is required to acquire the concept of truth.

But there is a strong argument against deflationism, developed independently by Leon Horsten, Stewart Shapiro, and Jeffrey Ketland in the late 1990s (Horsten 1995; Shapiro 1998; Ketland 1999): the *conservativeness argument*. In rough outline, the argument proceeds as follows. If truth is a lightweight property, then a conservativeness condition is plausible: no truth-free sentence is proved by the extended theory unless it is already proved by the truth-free base theory.[1] But theories that entail reasonable

[1]Whether deflationism is committed to any such constraint is a vexed question; see for instance Field (1999) and Halbach (2001). Waxman (2017) argues that deflationists are better served by accepting a conservativeness constraint formulated in terms of semantic, not proof-theoretic, consequence. For the purposes of this paper, we put these issues to one side, and address only those deflationists who are committed to proof-theoretic conservativeness.

compositional claims about truth—such as the claim that a conjunction is true if and only if both conjuncts are true—are ordinarily non-conservative, for they entail the arithmetical consistency statement for the base theory, which (by the Second Incompleteness Theorem) the base theory itself does not. If these compositional claims are required by an adequate truth theory, then deflationism fails.

One of the key background assumptions in the conservativeness argument is that truth theories should be evaluated over an arithmetical base theory containing extended induction—induction for formulas involving the truth predicate, which applies to natural numbers (understood as Gödel codes of sentences). This opens up room for a response on the part of the deflationist: she can distinguish between instances of induction which are *genuinely mathematical* and instances whose motivation comes from the theory of truth and syntax and argue that non-conservativeness derives from the former, not the latter. If this response can be made out, then the 'guilt' for non-conservativeness can be laid at the door of arithmetic; truth will emerge innocent (Field 1999).

Here, however, a difficulty arises: the standard contemporary way of formulating truth theories—using an arithmetical base theory whose objects play a double role, functioning both as mathematical entities and (via Gödel coding) as surrogates for syntactic entities—obstructs the attempt to draw this distinction precisely.[2] The *entanglement* of syntax and arithmetic induced by coding limits the ability of the deflationist and her opponent to clarify their commitments.

As a result, Richard Kimberly Heck, Carlo Nicolai, and others have developed the *disentanglement programme* (Heck MS, 2015, 2018; Leigh and Nicolai 2013; Nicolai 2015; Fujimoto MS §5): an approach to formal theories of truth on which syntax theory and base theory are fully distinct. In so doing, they revive an idea found in Tarski's 'The Concept of Truth in Formalized Language' (Tarski 1935), which inaugurated the systematic logical and mathematical study of truth. In Tarski's system, as in the contemporary disentanglement programme, truth is predicated of a domain of objects wholly disjoint from the objects of the base theory; these *syntactic objects*, which represent sentences in the language of the base theory, are handled by a separate theory of syntax, sharply distinguished from the base theory.

In this paper, we discuss a novel set of philosophical issues arising within the disentanglement programme. Drawing on an idea due to Walter Dean (2015), we advocate an *epistemic stability* constraint. Roughly, a framework is epistemically stable when it is well-motivated on a basis that motivates no strictly stronger framework. Epistemic stability is especially interesting in the disentangled setting, since a collection of indi-

[2]See McGee (1990), Horsten (2011), Halbach (2014), and Cieśliński (2017) for extensive discussion of truth theories of the standard type.

vidually stable theories can prove to be jointly unstable. On our view, this problem afflicts some influential proposals in the literature. In particular, we focus on two systems discussed by Leigh and Nicolai (2013: 628).

The first system, which we call $\mathsf{PA}_b + \mathsf{PA}_s^{CT}$, appears friendly to deflationists, for it conservatively extends its arithmetical base. While it proves the *syntactic* consistency statement $\mathrm{Con}_s(\mathsf{PA}_b)$, it does not prove the coded arithmetical consistency statement $\mathrm{Con}_b(\mathsf{PA}_b)$, which would lead to nonconservativeness. As Leigh and Nicolai point out, however, this system "cannot capture our common metatheoretic practice" (2013: 635), in which we move seamlessly from coded to non-coded claims about syntax. To rectify this, they propose a theory $\mathsf{PA}_s + \mathsf{PA}_b^{CT} + \mathsf{CodAx}$, which restores harmony by adding "coding axioms" that bridge the base and syntax theories—but this theory is nonconservative over the arithmetical base and thus inhospitable to the deflationist.

Although Leigh and Nicolai ultimately advocate $\mathsf{PA}_s + \mathsf{PA}_b^{CT} + \mathsf{CodAx}$, we do not think that the dialectic should end there. We agree with them that $\mathsf{PA}_s + \mathsf{PA}_b^{CT}$ is unsatisfactory, but offer our own explanation: it exhibits a failure of joint epistemic stability. However, we go on to argue that a similar failure afflicts $\mathsf{PA}_s + \mathsf{PA}_b^{CT} + \mathsf{CodAx}$ as well. Taking the epistemic stability constraint seriously motivates a much stronger theory—a second-order system we call DZ^{2CT}. We show that DZ^{2CT} is not only adequate to our informal mathematical practice, but places the required bridge principles on the strongest explanatory footing: they are *derivable* as theorems. However, the deflationist should take no comfort from the failure of $\mathsf{PA}_s + \mathsf{PA}_b^{CT} + \mathsf{CodAx}$, for DZ^{2CT} is itself a non-conservative extension over its arithmetical base theory. We conclude that deflationism faces a serious dilemma: the most promising conservative truth theories are unstable, but the best stable theory on offer in the vicinity is non-conservative.

I. Disentangling Truth and Syntax

Before introducing the notion of a disentangled truth theory, we sketch the usual, entangled approach as a point of comparison. Our base theory is first-order Peano arithmetic (PA). The signature of the language of PA, $\mathscr{L}_{\mathsf{PA}}$, contains a constant 0, a singulary function S, and binary function symbols $+$ and $\times$, with the obvious intended interpretations. The axioms of the theory are:

(PA1) $\quad \forall x \forall y (\mathsf{S}x = \mathsf{S}y \rightarrow x = y)$,

(PA2) $\quad \neg \exists x\, \mathsf{S}x = 0$,

(PA3) $\quad \forall x\, x + 0 = x$,

(PA4) $\quad \forall x \forall y\, x + \mathsf{S}y = \mathsf{S}(x + y)$,

(PA5) $\forall x\, x \times 0 = 0$,

(PA6) $\forall x \forall y\, x \times \mathsf{S}y = (x \times y) + x$,

(PA7) $\Phi(0) \wedge \forall x(\Phi(x) \rightarrow \phi(\mathsf{S}x)) \rightarrow \forall x\, \Phi(x)$ where $\Phi(x)$ is a formula of $\mathscr{L}_{\mathsf{PA}}$.

Here and throughout, we use capital Greek letters as schematic letters in our metalanguage for formulas of the language under consideration; lower-case Greek letters will be reserved for variables ranging over *codes* of formulas, in accordance with conventions introduced below.

To add a theory of truth to PA, we extend $\mathscr{L}_{\mathsf{PA}}$ to $\mathscr{L}^T_{\mathsf{PA}}$ by adding a one-place predicate T. We fix a Gödel coding $\ulcorner\urcorner$ on strings of $\mathscr{L}^T_{\mathsf{PA}}$; T is intended to apply to a number if it codes a true sentence of $\mathscr{L}_{\mathsf{PA}}$. The details of the coding do not matter, provided it is recursive and reasonably natural.

One simple, standard truth theory in $\mathscr{L}^T_{\mathsf{PA}}$ is the *compositional theory of truth for* $\mathscr{L}_{\mathsf{PA}}$, which comes in two forms: $\mathsf{PA}^{CT\restriction}$ (with restricted induction) and PA^{CT} (with full induction).

Before stating the theory, we introduce some notational conventions: $\forall \phi \cdots$ abbreviates $\forall x\, (\mathsf{Sent}_{\mathsf{PA}} x \rightarrow \cdots)$, where $\mathsf{Sent}_{\mathsf{PA}}$ applies to a number just in case it is the code of a sentence of $\mathscr{L}_{\mathsf{PA}}$; $\forall \mathsf{t} \cdots$ and $\forall \mathsf{v} \cdots$ are defined similarly for terms and variables. We also use the Feferman dot convention: for example, $\dot{\neg}$ expresses the function which yields, when applied to the Gödel number of a sentence, the Gödel number of its negation.

The functor $^\circ$ abbreviates an expression taking the code of a closed term to its value; for each number n, $\bar{n}$ denotes the numeral representing n; and $\phi^{\mathsf{t}_1}/_{\mathsf{t}_2}$ denotes the code of the result of performing capture-free substitution of t_1 for t_2 in ϕ. All of these syntactic operations are primitively recursively definable in PA using standard techniques (see, e.g., Smith 2013); the displayed expressions should be viewed as metalinguistic abbreviations.

The compositional axioms for truth are:

(CT1) $\forall \mathsf{t}_1 \forall \mathsf{t}_2 (T(\mathsf{t}_1 \doteq \mathsf{t}_2) \leftrightarrow \mathsf{t}_1^\circ = \mathsf{t}_2^\circ)$,

(CT2) $\forall \phi (T \dot{\neg} \phi \leftrightarrow \neg T\phi)$,

(CT3) $\forall \phi \forall \psi (T(\phi \dot{\vee} \psi) \leftrightarrow T\phi \vee T\psi)$,

(CT4) $\forall \phi \forall \psi (T(\phi \dot{\wedge} \psi) \leftrightarrow T\phi \wedge T\psi)$,

(CT5) $\forall \phi \forall \mathsf{v} (T(\dot{\forall} \mathsf{v} \phi) \leftrightarrow \forall \mathsf{t} T\phi^{\mathsf{t}}/_{\mathsf{v}})$.

Notice that (CT1)—(CT5) govern the application of T only to sentences in $\mathscr{L}_{\mathsf{PA}}$—i.e., sentences that do not themselves contain T; in the standard terminology, they form the basis for a *typed* truth theory. Adding (CT1)-–(CT5) to (PA1)-–(PA7) yields the theory of compositional truth over PA (with restricted induction), $\mathsf{PA}^{CT\restriction}$, a conservative extension of PA.

But $\mathsf{PA}^{CT\restriction}$ is plausibly an unnatural theory: it seems inherent in our understanding of the natural numbers that induction holds for every property, not just those we can express in our current language.[3] If we regard our commitment to mathematical induction as open-ended in this sense and replace (PA7) by

$$\text{(PA7}') \qquad \Phi(0) \wedge \forall x(\Phi(x) \rightarrow \Phi(Sx)) \rightarrow \forall x\, \Phi(x) \text{ for } \Phi(x) \in \mathsf{Fmla}(\mathscr{L}^{T}_{\mathsf{PA}}),$$

we obtain the theory of compositional truth over PA with full induction, PA^{CT}. PA^{CT} is *not* a conservative extension of PA; in fact, PA^{CT} proves $\mathrm{Con}(\mathsf{PA})$, the canonical arithmetized consistency statement for PA. The proof is a formalization of the 'semantic argument': all proper axioms of PA and all logical axioms are true, and all inference rules preserve truth, so all theorems of PA are true. Thus if Φ is provable in PA, then Φ is true. Contraposing: $0 \neq 1$; so '$0 = 1$' is not true; so PA does not prove '$0 = 1$'.[4]

We turn now to the disentangled setting. One philosophically natural framework for disentanglement is that of Tarski (1935) and Grzegorczyk (2005), where the language contains primitive names for each of the symbols in the alphabet of the base language and a binary function symbol $^\frown$ for concatenation. Substitution and the other usual syntactic operations can then be defined.[5]

This approach effects disentanglement in two different ways. First, it has a feature we call *separation*: the syntax theory and the base theory range over disjoint domains. Second, it has a feature we call *de-arithmetization*: the domain of the syntax theory consists of genuinely linguistic or syntactic objects— linguistic symbols or symbol-types and concatenations thereof—not mathematical objects of some distinct ontological kind. In a de-arithmetized setting, no coding apparatus is needed, since syntactic claims can be taken at face value without any detour through arithmetic.

In what follows, we shall consider theories that are separated but *not* de-arithmetized. Such theories make use of two disjoint domains of mathematical objects—in the simplest case, two 'copies' of the natural numbers—one for the properly mathematical entities and the other used to provide coded surrogates for syntactic entities. In such a setting, as Heck puts it, one can think of the language of the syntax theory as "the language of arithmetic written in boldface" (2015: 451).

[3] See McGee (1997) and Feferman (1991).

[4] For extensive discussion of PA^{CT} and $\mathsf{PA}^{CT\restriction}$, see Halbach (2014: 63–102).

[5] See also Quine (1946).

Proceeding in this way has substantial practical advantages. Theories of arithmetic, unlike theories involving a primitive concatenation operation, have been extensively studied. Virtually all previous work on disentanglement (Leigh and Nicolai 2013, Nicolai 2015, Heck 2015) has focussed on such theories; we follow the literature in this regard. It might be objected that logical and ontological hygiene requires both separation and de-arithmetization. But no generality is lost, since results can normally be carried over. In particular, as shown by Corcoran et al. (1974), the second-order theory of syntax over an alphabet with any finite number of letters is *synonymous* (in the sense of de Bouvère 1965a, 1965b) with full second-order arithmetic, which we shall later use as our separated syntax theory. As Friedman and Visser (2014: 1) note, synonymy "appears to be the strictest notion of sameness of theories except strict identity of signature and set of theorems". It would be a purely mechanical exercise to reconstruct the results below in a de-arithmetized setting.[6]

The most developed disentangled system is due to Graham Leigh and Carlo Nicolai (2013). We shall call it $\mathsf{PA}_b + \mathsf{PA}_s^{CT}$.[7] It is a three-sorted theory; (i) entities of sort *O* are 'genuine' natural numbers; (ii) entities of sort *S* are the 'duplicate' natural numbers coding syntactic objects; (iii) entities of sort *Seq* are 'mixed'—sequences of sort-*O* objects serving as assignments of values to variables (where variables are syntactic, i.e. sort-*S*, entities).

The theory of sort-*O* objects is PA, which we call PA_b (for 'base') in this context; its primitives are $0_b, \mathsf{S}_b, +_b$, and $\times_b$. The theory of sort-*S* objects is again PA—we call this PA_s (for 'syntax'), with primitives $0_s, \mathsf{S}_s, +_s$, and $\times_s$. Sort-*Seq* objects are handled by introducing three additional primitive notions: $\alpha[i]$ returns the value of the ith variable, v_i, on the assignment α; $\mathrm{Den}_\alpha \mathsf{t}$ returns the denotation of the term t on the assignment α; and $\mathrm{Sat}_\alpha \phi$ holds just in case the assignment α satisfies the formula coded by ϕ.[8]

The axioms of $\mathsf{PA}_b + \mathsf{PA}_s^{CT}$ are as follows (Leigh and Nicolai 2013: 620). We temporarily use the convention that the variables $x, y, \ldots$ range over sort-*O*; $i, j, \ldots$ range over sort-*S*; and $\alpha, \beta, \ldots$ range over sort-*Seq*:

(I) All axioms of PA (for objects of sort *O*);

(II) All axioms of PA (for objects of sort *S*);

(III) Axiom for sequences:

[6] Proponents of weaker theories can generally appeal to analogous bi-interpretability results (Visser 2009, Švejdar 2009, Friedman and Visser 2014); bi-interpretability is, however, sometimes weaker than synonymy.

[7] Leigh and Nicolai call the theory $CTD[PA] + TAX$.

[8] The functions denoted by these symbols are total, but their values on objects other than those specified are irrelevant.

(SQ) $\forall\alpha\forall x\forall j\exists\beta(\forall i(i \neq j) \rightarrow \alpha[i] = \beta[i] \wedge \beta[j] = x)$;

(IV) Axioms for denotation and satisfaction:

(CTD$_v$) $\forall\alpha\forall i\ \mathrm{Den}_\alpha \mathsf{v}_i = \alpha[i]$,

(CTD$_c$) $\forall\alpha\forall i\ \mathrm{Den}_\alpha \mathsf{c}_i = c_i$,

(CTD$_f$) $\forall\alpha\forall \mathsf{t}_1 \cdots \forall \mathsf{t}_n\ (\mathrm{Den}_\alpha \dot{f}(\mathsf{t}_1, \cdots, \mathsf{t}_n) = f(\mathrm{Den}_\alpha \mathsf{t}_1, \ldots, \mathrm{Den}_\alpha \mathsf{t}_n))$ for each function symbol f,

(CTD$_{at}$) $\forall\alpha\forall \mathsf{t}_1 \cdots \forall \mathsf{t}_n\ (\mathrm{Sat}_\alpha \dot{R}(\mathsf{t}_1, \cdots, \mathsf{t}_n) \leftrightarrow R(\mathrm{Den}_\alpha \mathsf{t}_1, \cdots, \mathrm{Den}_\alpha \mathsf{t}_n))$ for each n-ary relation symbol R,

(CTD$_\neg$) $\forall\alpha\forall\phi\ (\mathrm{Sat}_\alpha \dot{\neg}\phi \leftrightarrow \neg \mathrm{Sat}_\alpha \phi)$,

(CTD$_\wedge$) $\forall\alpha\forall\phi\forall\psi\ (\mathrm{Sat}_\alpha \phi \dot{\wedge} \psi \leftrightarrow \mathrm{Sat}_\alpha \phi \wedge \mathrm{Sat}_\alpha \psi)$,

(CTD$_\forall$) $\forall\alpha\forall\phi\forall i\ (\mathrm{Sat}_\alpha \dot{\forall} \dot{\mathsf{v}}_i \phi \leftrightarrow \forall\beta(\forall j(j \neq i \rightarrow \alpha(j) = \beta(j)) \rightarrow \mathrm{Sat}_\beta \phi))$;

(V) Syntactic induction (schema):

(Ind$_S$) $\Phi(0_s) \wedge \forall k(\Phi(k) \rightarrow \Phi(\mathsf{S}_s k)) \rightarrow \forall k \Phi(k)$ where k is a variable of the syntax theory and Φ is any formula.

(VI) The axioms of PA_b are true:

(TrAx) $\forall\phi(\mathrm{Ax}^{\mathrm{s}}_{\mathrm{O}}\phi \rightarrow \mathrm{Sat}_\alpha \phi)$ where $\mathrm{Ax}^{\mathrm{s}}_{\mathrm{O}}$ is the formula canonically expressing the property (of syntactic objects) of being an axiom of PA_b.

Leigh and Nicolai (2013: 626) show that $\mathsf{PA}_b + \mathsf{PA}_s^{CT} \vdash \mathrm{Con}_s(\mathsf{PA}_b)$, where $\mathrm{Con}_s(\mathsf{PA}_b)$ is the *syntactic* consistency statement for PA_b, i.e. the sentence in the syntactic vocabulary $\mathscr{L}_{\mathsf{PA}_s}$ expressing the consistency of PA_b. But $\mathsf{PA}_b + \mathsf{PA}_s^{CT} \nvdash \mathrm{Con}_b(\mathsf{PA}_b)$, where $\mathrm{Con}_b(\mathsf{PA}_b)$ is the *arithmetical* consistency statement for PA_b, i.e. the coded sentence in the arithmetical vocabulary $\mathscr{L}_{\mathsf{PA}_b}$ expressing the consistency of PA_b. Indeed, $\mathsf{PA}_b + \mathsf{PA}_s^{CT}$ is conservative over PA_b (Leigh and Nicolai 2013: 627).

But Halbach (2011) argues that it is highly artificial to treat syntactic and arithmetical consistency sentences asymmetrically in this way:

> [T]he very strict separation of syntax and mathematics that facilitates the proof of conservativity just outlined is highly artificial. Although in informal metamathematics we do distinguish between syntactic and mathematical objects such as numbers and sets and the associated theories, we are usually happy to pass from the syntactic consistency statement, to its coded

> counterpart. [...] To obtain a setting that is more natural than [such a disentangled theory], one would have to add 'bridge' laws between [the base theory] and [the syntax theory], axioms that allow one to connect mathematical and syntactic objects. (Halbach 2011: 320)

Leigh and Nicolai implement Halbach's suggestion by adding 'coding axioms', employing a new primitive cross-type predicate C (for the syntactico-mathematical bridging relation):

(CodAx1) $\forall x(C(x, 0_s) \leftrightarrow x = 0_b)$;

(CodAx2) $\forall x \forall i(C(x, i) \rightarrow C(\mathsf{S}_b x, \mathsf{S}_s i))$;

(CodAx3) $\forall x \exists! i\, C(x, i)$.

Leigh and Nicolai claim that the addition of (CodAx1)–(CodAx3) (collectively, CodAx) "seems to offer a satisfactory picture of our informal metatheoretic discussion as characterized in Halbach" (2013: 628).

We think this is at best partially correct: although $\mathsf{PA}_b + \mathsf{PA}_s^{CT} + \mathsf{CodAx}$ captures some features of informal metamathematics, it fails to be coherently integrated: the coding axioms, underivable from $\mathsf{PA}_b + \mathsf{PA}_s$, cry out for a more fundamental justification than Leigh and Nicolai provide. In contrast, in the system DZ^{2CT} we shall introduce, versions of CodAx can be *derived*, rather than merely posited: this stronger system better reflects our informal metamathematical practice.

II. Epistemically Stable Theories

As we have presented it, the disentanglement programme focuses on *combinations* of theories: base theories plus theories of syntax, perhaps enriched with truth-theoretic apparatus. But which combinations of theories ought to be considered?

Much of the work in the programme proceeds from a technical stance—one motivated by the desire to prove theorems as strong as possible using minimal resources. Richard Kimberly Heck's result (2015: 457) that $I\Sigma_1^{CT}$ (Robinson arithmetic plus Σ_1 induction plus compositional truth axioms) is a combined syntax and truth theory capable of proving the syntactic consistency statement for finitely axiomatized base theories is an example of this approach. The result is striking, but it is difficult to come up with an autonomous philosophical justification for accepting induction only for Σ_1 formulas.

In contrast, we adopt a *philosophical stance*, concerned primarily with theories possessing an internally coherent motivation. To elucidate this notion further, we appeal

to the idea of a *foundational equivalence thesis* (FET). A *foundational stance* is an informal conception of a mathematical domain (e.g. the natural numbers, the universe of sets, syntactic objects) or mode of reasoning (e.g. constructive or finitistic proof), corresponding to a principled position in the philosophy of mathematics that one might coherently adopt. An FET is a thesis to the effect that a foundational stance is extensionally equivalent to a given formal mathematical theory.[9]

We are interested in FETs because they allow us to characterize a class of systems that can be regarded as *epistemically stable*. We take this notion from Walter Dean, who ascribes it to a system when "there exists a coherent rationale for accepting [it] which does not entail or otherwise oblige a theorist to accept statements which cannot be derived from [its] axioms" (2015: 53).

Of course, one might desire more from a foundational stance than mere coherence: a coherent stance might be impoverished relative to its intended domain or simply misguided. Nonetheless, if a formal system can be linked to a *prima facie* coherent foundational stance via an appropriate FET, it is reasonable to regard it as internally motivated in the required sense.

Some examples of FETs may help to clarify the notion:

> DEDEKIND'S THESIS (cf. Dedekind 1888): There is a philosophical conception of the natural numbers according to which they are the smallest structure containing an initial element 0 and closed under the successor relation. This informal conception is captured precisely by the formal system of second-order arithmetic (Z^2).
>
> ISAACSON'S THESIS (Isaacson 1987; cf. Dean 2015: 53–54): There is a distinction between, on the one hand, the "purely arithmetical" content of our conception of the natural numbers, and on the other, "higher-order" or "set-theoretic" content that is revealed only from an extra-arithmetical vantage point (Isaacson 1987: 147). The purely arithmetical truths about the natural numbers are captured precisely by the theorems of first-order PA. In other words, "If we are to give a proof of any true sentence of [$\mathscr{L}_{\mathsf{PA}}$] which is independent of PA then we will need to appeal to ideas that go beyond those that are required in understanding PA" (Smith 2008: 1).
>
> TAIT'S THESIS (Tait 1981; cf. Dean 2015: 50–52): Finitism, in the sense of Hilbert and Bernays (1934–39), is a conception or mode of reasoning about the natural numbers that does not regard them as a completed infinite totality. This position is captured precisely by the formal system of Primitive

[9] FETs are one example of the larger class of *informal equivalence theses*, such as the Church-Turing thesis—characteristic instances of Kreisel (1967)'s method of 'informal rigour'.

> Recursive Arithmetic (PRA). Slightly less roughly, (i) any finistically acceptable function is primitive recursive, and conversely any primitive recursive function is finistically acceptable; (ii) any proof within PRA is acceptable to the finitist, and conversely any finitistic proof corresponds to a proof in PRA with the same conclusion.

Tait does not claim that the finitist ought to endorse PRA itself. (After all, PRA is committed to the existence of (infinitely many) total functions on the natural numbers, which cannot be recognized by the finitist as objects.) Rather, like many FETs, the position is characterized externally. This illustrates a recurring point: in general it is not required for the truth of an FET that a proponent of the foundational stance in question be in a position to recognize it as true.

> FEFERMAN-SCHÜTTE THESIS (Kreisel 1960; Feferman 1964; Schütte 1965a, 1965b): Predicativism given the natural numbers is a conception of the natural numbers and sets thereof motivated by the vicious circle principle in the sense of Poincaré and Russell, according to which sets of natural numbers are acceptable insofar as they can be defined without quantifying over totalities to which they belong. This philosophical position is captured precisely by the formal system $\mathsf{RA}_{<\Gamma_0}$ of ramified analysis up to the Feferman-Schütte ordinal Γ_0.

Before proceeding further, however, let us consider an objection to the very idea of epistemic stability. On one influential view (Kreisel 1958, 1965; Feferman 1991), anyone who rationally endorses a reasonably strong theory $\mathbf{T}$ is thereby implicitly committed to extending it via reflection principles—at a minimum, the local reflection principle $\mathrm{LocRefl}(\mathbf{T})$, $\mathrm{Bew}_{\mathbf{T}}(\ulcorner\Phi\urcorner) \rightarrow \Phi$, stating that if Φ is provable in $\mathbf{T}$, then Φ.[10]

If so, a fully rational agent who accepts $\mathbf{T}$ ought to accept $\mathbf{T} + \mathrm{LocRefl}(\mathbf{T})$, $\mathbf{T} + \mathrm{LocRefl}(\mathbf{T}) + \mathrm{LocRefl}(\mathbf{T} + \mathrm{LocRefl}(\mathbf{T}))$, and so on—each of these theories strictly stronger than the last. If this conception holds, then (on a straightforward construal of the notion of epistemic stability), there can be no such thing as an epistemically stable theory of nontrivial strength.

There are three points to make in response. First, the claim that acceptance of a theory incurs implicit commitment to reflection can be challenged. As Dean (2015) has forcefully argued, the mathematical consequences of reflection are closely linked to various induction principles. But where an informal foundational stance explicitly

[10]Many versions of reflection principles have been discussed. Another is the uniform reflection principle $\mathrm{UniRefl}(\mathbf{T})$, $\forall x(\mathrm{Bew}_{\mathbf{T}}(\ulcorner\Phi(\dot{x})\urcorner) \rightarrow \Phi(x))$. Other candidates for implicit commitments include the consistency statement $\mathrm{Con}_{\mathbf{T}}$ or the Gödel sentence $\mathrm{G}_{\mathbf{T}}$. These are equivalent to Π_1 restrictions of both local and uniform reflection principles (Smoryński 1977: 846).

motivates a theory with restricted induction (as is the case for PRA according to Tait and PA according to Isaacson), imposing reflection amounts to begging the question against the foundational stance. In such cases, we should reject the demand that, in order to be epistemically stable, a theory should display closure under reflection.

Second, it is relevant that many FETs are characterized externally. The sense in which $\mathsf{RA}_{<\Gamma_0}$ is supposed to capture predicativism is not that the predicativist ought to adopt it as her overall theory; rather, each of its theorems could be accepted by the predicativist, and conversely every claim accepted by the predicativist can be proven within it. Thus, even if it were true that anyone who accepted $\mathsf{RA}_{<\Gamma_0}$ were obligated to accept additional claims, this obligation would not apply to the predicativist herself.

Third, even if the proponent of strong implicit commitment is correct that no non-trivial mathematical theory is *absolutely* epistemically stable, there is still a useful notion of *stability modulo reflective closure*. A theory which is unstable for reasons independent of reflection is, in a straightforward sense, pathological. The proponent of strong implicit commitment should accept that, even if there are no fully stable theories, at least some theories are non-pathological in the relevant sense. Thus there is a notion of relative stability, distinguishing the theories which have an intrinsic motivation from those which do not. Proponents of strong implicit commitment are thus invited to understand 'epistemic stability', in what follows, as 'epistemic stability modulo reflective closure'.

We conclude that, from the philosophical perspective, theories of foundational interest must be epistemically stable. If the FETs above are correct, then—as mathematical theories taken in isolation—Z^2, PA, PRA, and $\mathsf{RA}_{<\Gamma_0}$ are internally coherent frameworks worthy of study.

What impact does this picture have on the disentanglement programme? Taking the programme seriously requires considering theories not in isolation but in combination, for the objects of study will be joint theories consisting of a theory of syntax added to a base mathematical theory. In such a setting, the possibility arises that even if some mathematical theory and theory of syntax are epistemically stable when taken in isolation, they are nevertheless *jointly* unstable when paired with one another. We believe this possibility is realized; indeed, it affects the main disentangled theories proposed in the literature.

III. Joint Epistemic Stability

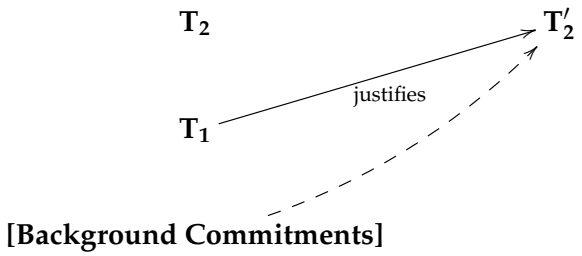
How might a collection of individually stable theories fail to be jointly stable? If the theories to be combined were, for example, an arithmetical theory and a physical theory of electromagnetism, it is hard to see how joint instability could arise. But the natural

numbers and syntactic objects are more intimately related. Even though disentanglement begins from the principle that syntax and arithmetic are distinct, the nature of the two domains places constraints on their interaction. It is no accident that, as Gödel showed, we can arithmetize syntax, and, as Quine (1946) showed, we can carry out arithmetic in a theory of strings and concatenation.

3.1 Two Kinds of Joint Instability

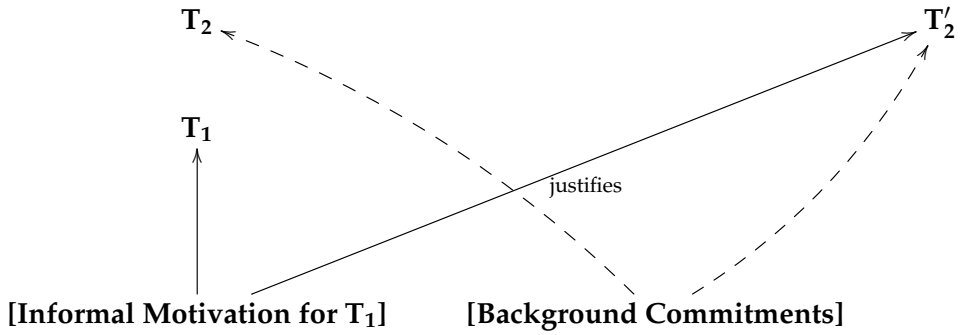
At an abstract level, there are at least two possible ways in which a combination of theories might exhibit joint instability. The first, *local joint instability*, arises when theories $\mathbf{T}_1$ and $\mathbf{T}_2$ are combined, but $\mathbf{T}_1$ (possibly together with background commitments) motivates a theory $\mathbf{T}_2'$ strictly stronger than $\mathbf{T}_2$.[11]

Fig 1. Local Joint Instability.



The second, *structural joint instability*, arises not when $\mathbf{T}_1$ and background commitments themselves lead beyond $\mathbf{T}_2$, but when $\mathbf{T}_1$ itself can only be motivated from a more general informal foundational stance—of the kind captured by an FET—which justifies a $\mathbf{T}_2'$ strictly stronger than $\mathbf{T}_2$.

Fig 2. Structural Joint Instability.



[11] The notion can be extended in the obvious way to combinations of more than two theories.

3.2 The Hilbert-Parsons Principle

The primary source of local joint instability we will discuss arises from the links between arithmetic and syntax. The idea can be sharpened by considering the following passage from Hilbert:

> [T]he objects [*Gegenstände*] of number theory are for me—in direct contrast to Dedekind and Frege—the signs themselves, whose shape [*Gestalt*] can be generally and certainly recognized by us—independently of space and time, of the special conditions of the production of the sign and of insignificant differences in the finished product. The solid philosophical attitude that I think is required for the grounding of pure mathematics—as well as for all scientific thought, understanding, and communication—is this: In the beginning was the sign. (Hilbert 1922: 163/Mancosu 1998: 202; cf. Hilbert and Bernays 1934–39: **1**: 20)

Here Hilbert defends a very strong claim: that the natural numbers *simply are* syntactic types. Few others have found this metaphysical thesis plausible. But it is worth asking why Hilbert found it attractive in the first place. The best explanation is that, even if Hilbert was mistaken in diagnosing the connection between syntax and arithmetic as arising from a syntactic metaphysics of number, the link nonetheless has epistemic purchase.

Further insight can be found in the work of Charles Parsons, who holds that we can learn arithmetical facts by consideration of syntax as follows. First, we learn about string-types (which he describes as "quasi-concrete" objects: abstract but possessing "an intrinsic relation to the concrete" tokens that instantiate them and "determined by their concrete embodiments" (2006: 33)) because we literally perceive them:

> [W]e stand in a perceptual relation to the expressions of this simple formal language [i.e. Hilbert's system of stroke-numerals]. The same reasons for talking of perception of expression-types with reference to natural language arise also with reference to this language. (Parsons 2006: 159–60)

Second, although we have no comparable perception of the natural numbers themselves, we can learn about them indirectly, using this perceptual knowledge of facts about stroke-types:

> [T]he system of strings [...] is still, in some sense, an intuitive model of arithmetic. [...] [I]t consists of objects of intuition in the sense that there is actual intuition of strings sufficiently early in the sequence and it is possible

> to draw some conclusions about an arbitrary string intuitively. [...] We can easily satisfy ourselves that it satisfies the Dedekind-Peano axioms. If we understand the strings as what is obtained from | by iterated application of the operation of adjoining one more, then it should be as evident that induction holds for them as that it holds for any structure characterized in this particular way. (Parsons 2006: 235)

We need not follow the details of Parsons's account; what is important is simply that there is a structural similarity between string-types and natural numbers, leading to the possibility of learning facts about arithmetic by reflecting on facts about syntactic objects. Parsons explains this structural similarity in terms of the fact that the system of strings satisfies the Dedekind-Peano axioms. In other words:

> HILBERT-PARSONS PRINCIPLE (HPP): We can learn arithmetical claims by learning syntactic claims, an ability which is explained by the fact that arithmetic can be given an (unintended) interpretation in terms of syntactic objects.

3.3 From HPP to Interpretability

These considerations can be used to motivate the claim that syntax should be *proof-theoretically relatively interpretable* within arithmetic.[12]

The argument from HPP to an explicit formal constraint in terms of proof-theoretic interpretability proceeds as follows. Suppose that the pair $\mathbf{T}_b + \mathbf{T}_s$ is epistemically stable. We have at least a partial grasp of the intended models $\mathscr{A}$ and $\mathscr{S}$ of arithmetical and syntactic objects. There is a subtheory of $\mathbf{T}_s$ (the theory of the null string and finite sequences of some arbitrary symbol) which we can reinterpret, along the lines set out above, through a mapping $\mu : \mathscr{S} \to \mathscr{A}$ taking the null string to zero and so on. The mapping μ yields characterizations of the operations of $\mathbf{T}_b$ in terms of $\mathbf{T}_s$; let † denote the induced function from formulas of $\mathscr{L}_{\mathbf{T}_s}$ to formulas of $\mathscr{L}_{\mathbf{T}_b}$. One who works within the appropriate foundational stance is justified in believing all the theorems of $\mathbf{T}_s$ and believing that † preserves truth. But if † took a theorem Φ of $\mathbf{T}_s$ to a nontheorem of $\mathbf{T}_b$, then $\mathbf{T}_b \cup \{\Phi^\dagger\}$ would be a proper extension of $\mathbf{T}_b$ justified by the combined theory, which is ruled out by epistemic stability. So if $\mathbf{T}_s \vdash \Phi$, then $\mathbf{T}_b \vdash \Phi^\dagger$. And since † clearly respects connectives and identity, this means that † is a relative interpretation of $\mathbf{T}_s$ in $\mathbf{T}_b$. We thus have:

[12] In the proof-theoretic sense, the interpretability of $\mathbf{T_1}$ within $\mathbf{T_2}$ means roughly that there is a function from the sentences in the language of $\mathbf{T_1}$ to sentences in the language of $\mathbf{T_2}$ such that theorems are mapped to theorems and the logical structure of complex sentences is preserved, modulo quantifier relativization (Lindström 2003: 96–98).

(HPP*) If $\mathbf{T}_b + \mathbf{T}_s$ is jointly epistemically stable, then $\mathbf{T}_s$ is relatively interpretable in $\mathbf{T}_b$.

3.4 HPP*, Deflationism, and Local Joint Instability

This constraint can be applied to some theories considered in the disentanglement literature. As noted above, Heck focusses on $I\Sigma_1^{CT}$ as an attractive theory of truth and syntax for the reason that it yields the consistency of any finitely axiomatized base theory to which it is added. But suppose $I\Sigma_1^{CT}$ is added to a relatively weak base theory incapable of interpreting it, such as Q or $I\Delta_0$.[13] Given HPP*, this combination of views is epistemically unstable, since someone who accepts $I\Sigma_1^{CT}$ can already move from Q or $I\Delta_0$ to the stronger base theory $I\Sigma_1$. Even waiving worries about the individual stability of $I\Sigma_1^{CT}$, there are strong constraints on the base theories with which it can paired.

We now turn to the theories considered by Leigh and Nicolai. The joint stability of $\mathsf{PA}_b + \mathsf{PA}_s$ is plausible. Given Isaacson's Thesis, PA_b is individually stable, and a syntactic analogue of Isaacson's Thesis can be formulated according to which PA_s is individually stable too. Since PA_b and PA_s are synonymous, they are mutually interpretable, and thus HPP* reveals no impediment to their joint stability.

But as appealing as $\mathsf{PA}_b + \mathsf{PA}_s$ may be, it is not a theory of any relevance to issues of truth, for instance the dispute between deflationists and their opponents, for it does not even contain a truth predicate. We are thus led to consider its most natural truth-theoretic extension $\mathsf{PA}_b + \mathsf{PA}_s^{CT}$.

Here, however, an argument against the joint stability of $\mathsf{PA}_b + \mathsf{PA}_s^{CT}$ can be made out via the HPP*. Since PA_s^{CT} proves $\mathrm{Con}_s(\mathsf{PA}_b)$ but $\mathsf{PA}_b + \mathsf{PA}_s^{CT}$ does not prove $\mathrm{Con}_b(\mathsf{PA}_b)$, the PA_b fragment on its own cannot interpret PA_s^{CT}. Thus, the theory is unstable according to HPP*.

Notice that the objection just made is distinct from the one raised by Halbach (2016: 306) and discussed in §1. Our objection is not that $\mathsf{PA}_b + \mathsf{PA}_s^{CT}$ is unnatural, or that it is unfaithful as a codification of our metamathematical practice; rather, our objection is that, given HPP*, it is an epistemically unstable combination of theories.

Despite first appearances, then, $\mathsf{PA}_b + \mathsf{PA}_s^{CT}$ cannot be embraced by the deflationist who wishes to sidestep the conservativeness objection. It is true that it conservatively extends PA_b. But it does so at the cost of sharply distinguishing its treatment of the *arithmetical* and *syntactic* consistency statements for PA_b in a way that leads to failures of interpretability and thus of the HPP*. Since this leads to an unacceptable epistemic instability, $\mathsf{PA}_b + \mathsf{PA}_s^{CT}$ should not be accepted as a final theory—by the deflationist or,

[13] See Hájek and Pudlák (1993: 108).

for that matter, by anyone else.

3.5 Structural Joint Instability and the Coding Axioms

Finally, we consider Leigh and Nicolai's preferred theory $\mathsf{PA}_b + \mathsf{PA}_s^{CT} + \mathsf{CodAx}$. Adding CodAx to the combined theory brings the strength of the base theory up to that of the syntax theory; there is thus no mismatch of interpretability strength. But HPP* is not the only possible cause of joint instability; as noted above, structural joint instability arises when the underlying foundational stance motivating one subtheory justifies an increase in strength in another subtheory. This is what we believe occurs in the case of the coding axioms.

Consider how the coding axioms might be philosophically motivated. We see three possibilities: intrinsically, truth-theoretically, and on higher-order grounds (in Isaacson's sense). In other words: CodAx might be well-motivated on its own, independently of $\mathsf{PA}_b + \mathsf{PA}_s^{CT}$; CodAx might be motivated on the basis of the conception of truth behind PA_s^{CT}; and CodAx might be motivated based on facts about the natural number structure that exceed what is captured in the first-order theories PA_b and PA_s.

We argue that only the last is plausible. But, if higher-order justification is taken seriously, it motivates both a stronger base theory and a stronger syntax theory than PA. Thus, $\mathsf{PA}_b + \mathsf{PA}_s^{CT} + \mathsf{CodAx}$ is epistemically unstable.

It would be desperately implausible to argue that CodAx possesses a motivation independently of some underlying conception of arithmetic. The coding axioms assert that there is a particular function coding up syntax in the base theory. But this is the kind of claim that, if true, cries out for explanation, and cannot be taken as brutely justified.

Nor can truth-theoretic considerations suffice. As already pointed out, such considerations can motivate $\mathsf{PA}_b + \mathsf{PA}_s^{CT}$ itself, but there is no way to use them to extend that theory. The addition of CodAx to $\mathsf{PA}_b + \mathsf{PA}_s^{CT}$ results in a non-conservative extension, so it requires some additional motivation.

In contrast, 'higher-order' considerations in Isaacson's sense seem perfectly suited for the job. Facts about the coding between arithmetic and syntax are, for Isaacson, paradigmatic examples of higher-order content:

> The key technique of Gödel's proof is the use of coding, the coding of syntactic relations and properties by properties and relations of natural numbers. At least in the case of Gödel sentences[...] the understanding of these sentences rests crucially on understanding this coding and our grasp of the situation being coded. The phenomenon of coding reveals fixed links between two situations or facts, one in the structure of arithmetic, the other in

> the realm of syntax of a formal system. These facts, and the link between them, are revealed by the description of the coding, but their existence is not dependent on being described. (Isaacson 1985: 214)

So, assuming Isaacson's Thesis and its syntactic analogue, the coding axioms are alien to the conception embodied by $\mathsf{PA}_b + \mathsf{PA}_s$ or even $\mathsf{PA}_b + \mathsf{PA}_s^{CT}$. Their justification relies not just on a 'local' appreciation of the individual structure of arithmetical or syntactic objects, but on the fact that the two domains exhibit an extremely strong form of structural similarity.

We thus see that Leigh and Nicolai's attempt to use the non-conservativeness of $\mathsf{PA}_b + \mathsf{PA}_s^{CT} + \mathsf{CodAx}$ over PA_b against the deflationist does not succeed. Although it is true that CodAx captures certain elements of informal metamathematical reasoning, the resulting theory is nevertheless epistemically unstable. The deflationist can appeal to this fact as a reason for resisting the expanded theory. But, as we shall show in the next section, this ultimately provides little comfort: the reasoning behind our argument for instability argument motivates a stronger disentangled truth and syntax theory, which we term DZ^{2CT}. We will go on to show that this theory also non-conservatively extends its arithmetical base theory, and is thus no more amenable to deflationism than $\mathsf{PA}_b + \mathsf{PA}_s^{CT} + \mathsf{CodAx}$.

IV. Double Second-Order Arithmetic

4.1 Dedekind's Thesis and DZ^2

We rejected $\mathsf{PA}_b + \mathsf{PA}_s^{CT} + \mathsf{CodAx}$ as unstable because the Coding Axioms appear to require a deeper justification. So we are led to ask: which conceptions of the natural numbers/syntax *are* rich enough to underwrite such claims of structural similarity, and which theories are motivated by these conceptions?

Dedekind's Thesis, the claim that the structure of the natural numbers is captured by Z^2, provides a natural answer. Again, there is an obvious syntactic analogue of Dedekind's Thesis which leads to a parallel claim for the structure of the syntactic domain.

We thus consider the theory that results from taking Z^2 as both the base theory and the syntax theory. We will call this theory DZ^2, for 'double' Z^2. Unlike $\mathsf{PA}_b + \mathsf{PA}_s^{CT} + \mathsf{CodAx}$, it and its truth-theoretical extension DZ^{2CT} are epistemically stable. Moreover, they have the resources to *demonstrate* complete structural similarity between the two domains—isomorphism—with no need for additional assumptions; the coding axioms are directly justified and need not be added in by hand.

We now set out DZ^2, *double second-order arithmetic*. In the terminology of §1, this theory is separated but not fully disentangled; due to the result of Corcoran et al. (1974) discussed there, no generality is lost.

We assume a standard second-order deductive system such as that in Shapiro (1991: 65–69), with each instance of the comprehension schema:

(CA) $\quad \exists \mathbf{X}^n \forall \mathbf{x}_1 \ldots \mathbf{x}_n (\mathbf{X}^n \mathbf{x}_1 \cdots \mathbf{x}_n \leftrightarrow \Phi)$ where $\mathbf{X}^n$ is not free in Φ.

We allow function constants but no quantification over functions. The signature of DZ^2 is $\{\mathsf{N}_b, \mathsf{N}_s, 0_b, 0_s, \mathsf{S}_b, \mathsf{S}_s\}$. Its proper axioms are:

(A1$_\xi$) $\quad \mathsf{N}_\xi 0_\xi$

(A2$_\xi$) $\quad \forall x (\mathsf{N}_\xi x \rightarrow \mathsf{S}_\xi x \neq 0_\xi)$

(A3$_\xi$) $\quad \forall x \forall y (\mathsf{N}_\xi x \wedge \mathsf{N}_\xi y \rightarrow (\mathsf{S}_\xi x = \mathsf{S}_\xi y \rightarrow x = y))$

(A4$_\xi$) $\quad \forall X (X 0_\xi \wedge \forall x (\mathsf{N}_\xi x \rightarrow (Xx \rightarrow X\mathsf{S}_\xi x)) \rightarrow \forall x (\mathsf{N}_\xi x \rightarrow Xx))$

(A5) $\quad \forall x \forall y (\mathsf{N}_b x \wedge \mathsf{N}_s y \rightarrow x \neq y)$

for $\xi \in \{b, s\}$.

Note that this is a one-sorted theory, unlike Leigh and Nicolai's $\mathsf{PA}_b + \mathsf{PA}_s$. In particular, second-order terms can apply to mixed collections (containing both mathematical and syntactic objects).

We write $\mathscr{L}_{\mathsf{Z}^2_b}$ and $\mathscr{L}_{\mathsf{Z}^2_s}$ for the languages with signatures $\{\mathsf{N}_b, 0_b, \mathsf{S}_b\}$ and $\{\mathsf{N}_s, 0_s, \mathsf{S}_s\}$, respectively, and we write Z^2_b and Z^2_s for the fragments of DZ^2 in $\mathscr{L}_{\mathsf{Z}^2_b}$ and $\mathscr{L}_{\mathsf{Z}^2_s}$. Given a formula Φ, we write $\Phi^{(b)}$ (resp. $\Phi^{(s)}$) for the result of relabelling the non-logical components with their b-analogues (resp. s-analogues).

4.2 Consistency and the Transfer Theorem

Regarding the stability of DZ^2, the main result of interest is the following:

Theorem 1 (Transfer Theorem). $\mathsf{DZ}^2 \vdash \Phi^{(b)} \leftrightarrow \Phi^{(s)}$ *for* $\Phi \in \mathsf{Sent}(\mathscr{L}_{\mathsf{DZ}^2})$.

This is a generalization of Leigh and Nicolai's (2013: 631) Corollary 3.17.[14] From the Transfer Theorem,

[14] In our notation, Leigh and Nicolai (2013: 361) establish that $\mathsf{PA}_b + \mathsf{PA}_s + \mathsf{CodAx} \vdash \Phi^{(b)} \leftrightarrow \Phi^{(s)}$ whenever Φ is a first-order sentence in the language $\mathscr{L}_{\mathsf{PA}_b + \mathsf{PA}_s}$. Our generalization applies to all sentences, not only first-order sentences. This is more than is required to establish Transfer of Consistency since, even though Z^2_b is a second-order theory, $\mathsf{Con}_b(\mathsf{Z}^2_b)$ is Π^0_1 in DZ^2; the extension to second-order sentences does not alter the structure of the proof, but is nevertheless desirable given that we work in a second-order setting.

Lemma 2 (Transfer of Consistency). $\mathsf{DZ}^2 \vdash \mathsf{Con}_b(\mathsf{Z}_b^2) \leftrightarrow \mathsf{Con}_s(\mathsf{Z}_b^2)$

follows immediately.

In order to prove the Transfer Theorem, we draw on a result from Väänänen and Wang (2015: 124). Second-order arithmetic is *internally categorical*, in that it proves the existence of an isomorphism between any two Z^2-structures, and thus between any Z_b^2-structure and any Z_s^2-structure (since the axioms of Z_b^2 and Z_s^2 differ only by a relabelling of constants):[15]

Lemma 3 (Internal Categoricity (Väänänen and Wang)). $\mathsf{DZ}^2 \vdash \exists f\, f : \langle \mathsf{N}_s, 0_s, \mathsf{S}_s\rangle \underset{\text{iso}}{\rightarrow} \langle \mathsf{N}_b, 0_b, \mathsf{S}_b\rangle$.

We sketch the proof: consider any Henkin model $\mathscr{M}$ satisfying (CA) such that $\mathscr{M} \vDash_H \mathsf{Z}_b^2$ and $\mathscr{M} \vDash_H \mathsf{Z}_s^2$, where $\vDash_H$ is the truth-in-a-model relation for Henkin models. We write D_1 and D_2 for the first- and second-order domains of $\mathscr{M}$. We proceed using a technique essentially equivalent to Frege's definition of the class of natural numbers: say that $g \in {}^{D_1}D_1$ is a *protomapping* if (1) $g(0_s) = 0_b$ and (2) $g(\mathsf{S}_s(x)) = \mathsf{S}_b(g(x))$. Let h be the function whose graph is given by $\bigcap\{g : g \text{ is a protomapping}\}$. Because h is a definable function, $h \in D_2$. It is easy to verify that h is an isomorphism between N_s and N_b; thus $\vDash_H \mathsf{DZ}^2 \rightarrow \exists f\, f : \langle \mathsf{N}_s, 0_s, \mathsf{S}_s\rangle \underset{\text{iso}}{\rightarrow} \langle \mathsf{N}_b, 0_b, \mathsf{S}_b\rangle$, so by completeness for Henkin models $\mathsf{DZ}^2 \vdash \exists f\, f : \langle \mathsf{N}_s, 0_s, \mathsf{S}_s\rangle \underset{\text{iso}}{\rightarrow} \langle \mathsf{N}_b, 0_b, \mathsf{S}_b\rangle$. The appeal to model-theoretic apparatus is inessential, for h can be defined directly inside DZ^2, and one instance of (CA) suffices for the existential claim.[16]

Henceforth, for convenience, we use h both for the function itself and as a meta-linguistic abbreviation for a second-order term in $\mathscr{L}_{\mathsf{DZ}^2}$ denoting it; $\hat{h}(A)$ functions similarly for $\{h(x) : x \in A\}$.

By construction, h provably satisfies the analogues of (CodAx1) and (CodAx2):

(IA1) $\qquad h(0_s) = 0_b$;

[15] We abuse notation somewhat in stating the theorem: in our second-order language, all functions are defined on the whole domain, but the behaviour of f on anything other than the extension of N_s is irrelevant. Likewise, we do not care about the behaviour of S_b or S_s outside the extension of N_b or N_s, respectively. For a general discussion of internal categoricity, see Button and Walsh (2018: 223–50).

[16] Examination of the proof reveals that only Π^1_1–CA is required for the internal categoricity result; see Simpson and Yokoyama (2013). We do not think that this detracts from the interest of Z^2 and DZ^2: in our view, Π^1_1–CA is internally epistemically unstable for reasons analogous to those which afflict $I\Sigma_1$. In a similar vein, one might understand PA as involving an *open-ended* induction schema (Parsons 2008: 269-70; McGee 1997; Lavine MS). Analogues of the coding axioms for a system with two copies of PA can then be derived, drawing on the formal similarity between open-ended schemata and the Π^1_1 fragment of the corresponding second-order theory. How to assess the epistemic stability of an open-ended schematic framework raises questions beyond the scope of this paper, since open-ended schemata are more akin to functors from the class of languages to the class of theories than simple collections of syntactic objects. We hope to consider this and similar issues in future work.

(IA2) $\quad \forall x \forall y (h(x) = y \to h(\mathsf{S}_s(x)) = \mathsf{S}_b(h(x)))$.

No analogue of (CodAx3) is required, for functionality is built in by definition.

We turn now to the proof of the Transfer Theorem.

Proof. We first establish a more general schematic claim:

$$\begin{aligned}\mathsf{DZ}^2 \vdash \forall X_1 \cdots \forall X_n \forall Y_1 \cdots \forall Y_n \forall x_1 \cdots \forall x_m \forall y_1 \cdots \forall y_m \\ (Y_1 = \hat{h}(X_1) \wedge \cdots \wedge Y_n = \hat{h}(X_n) \wedge y = h(x_1) \wedge \cdots \wedge y_m = h(x_m) \to \\ (\Phi^{(b)}(X_1, \ldots, X_n, x_1, \ldots x_m) \leftrightarrow \Phi^{(s)}(Y_1, \ldots, Y_n, y_1, \ldots y_m)))\end{aligned}$$

for $\Phi \in \mathsf{Fmla}(\mathscr{L}_{\mathsf{DZ}^2})$, which we abbreviate $\mathsf{DZ}^2 \vdash \Xi(\Phi^{(b)}(\vec{X}, \vec{x}), \Phi^{(s)}(\vec{Y}, \vec{y}))$ or simply $\mathsf{DZ}^2 \vdash \Xi(\Phi)$.

We show this by a metatheoretic induction on the complexity of Φ, adapting the proof strategy of Leigh and Nicolai's (2013: 631) Theorem 3.16. In the base case, where Φ is atomic, $\mathsf{DZ}^2 \vdash \Xi(\Phi)$ follows directly from (A1)-(A2) and the definition of the relabelling. The induction clauses for the sentential connectives are trivial. Now consider $\Phi = \forall z\, \Psi(z)$. If z is one of $\{\vec{x}\} \cup \{\vec{y}\}$, then it is bound, not free, in $\Phi^{(b)}$ and $\Phi^{(s)}$, and the satisfaction of the biconditional $\Phi^{(b)}(X_1, \ldots, X_n, x_1, \ldots x_m) \leftrightarrow \Phi^{(s)}(Y_1, \ldots, Y_n, y_1, \ldots y_m)$ does not depend on its value, but is already guaranteed by the induction hypothesis. So, without loss of generality, we can assume z is distinct from each of the x_i and y_i. By the induction hypothesis, we have $\mathsf{DZ}^2 \vdash \Xi(\Psi^{(b)}(\vec{X}, \vec{x}, z), \Psi^{(s)}(\vec{Y}, \vec{y}, z))$. But

$$\Xi(\Psi^{(b)}(\vec{X}, \vec{x}, z), \Psi^{(s)}(\vec{Y}, \vec{y}, z)) \to \Xi(\forall z\, \Psi^{(b)}(\vec{X}, \vec{x}), \forall z\, \Psi^{(s)}(\vec{Y}, \vec{y}))$$

follows from logic alone since z is free in Ψ; so $\mathsf{DZ}^2 \vdash \Xi(\forall z\, \Psi^{(b)}(\vec{X}, \vec{x}) \forall z\, \Psi^{(s)}(\vec{Y}, \vec{y}))$, i.e. $\mathsf{DZ}^2 \vdash \Xi(\Phi)$. A precisely analogous argument applies when $\Phi = \forall Z^n\, \Psi(Z^n)$. $\dashv$

4.3 The Stability of DZ^2 and DZ^{2CT}

We now argue that DZ^2 is a jointly stable theory. We have considered three ways in which a disentangled theory can fail to be stable: failures of (i) individual stability, (ii) local joint stability, and (iii) structural joint stability. How does DZ^2 fare on these grounds?

Note first that, given Dedekind's Thesis and its syntactic analogue, the individual components of DZ^2 are stable. The other two kinds of instability are more interesting.

The primary source of local joint instability discussed so far has arisen from the Hilbert-Parsons Principle, when the syntactic theory is not relatively interpretable within the arithmetical base theory. But for DZ^2, clearly no such worries arise. The

Transfer Theorem shows that the syntax theory can be interpreted within the base theory via the obvious interpretation mapping each part of the syntax language to its analogue in the base language.

What about other sources of local joint instability? It is hard to see how any could arise. The base theory can be interpreted in the same way within the syntax theory by the Transfer Theorem, and these interpretations are stable under any extension of the joint language: if background commitments motivate adding some third subtheory to DZ^2, the resulting theory will never prove a sentence in $\mathscr{L}_{\mathsf{Z}^2_b}$ without proving the corresponding sentence in $\mathscr{L}_{\mathsf{Z}^2_s}$ and vice versa.

With respect to structural joint instability, the key example we considered was the addition of a truth theory and coding axioms to Leigh and Nicolai's $\mathsf{PA}_b + \mathsf{PA}_s$. What is the analogous situation concerning DZ^2? We know from the Transfer Theorem that no interpretability failure can arise, but it remains at least possible that the extended theory could rely on an informal conception motivating a stronger base theory than Z^2_b. In the case of Leigh and Nicolai's theory the problems arose not from the truth theory but from the coding axioms, which embody a conception of arithmetic and syntax (and the structural relations between them) going beyond the Peano-Dedekind axioms.

We shall now introduce a truth theory, DZ^{2CT}, which extends DZ^2 with the natural compositional truth clauses for a second-order language. This truth theory can be autonomously motivated. Furthermore, in the appendix, we demonstrate that it is capable of *proving* the syntactic consistency statement for the base theory and thus, by Transfer of Consistency, the coded arithmetic consistency statement for the base theory in the base language. Thus DZ^{2CT} captures all the informal metamathematical reasoning we expect from adding a truth theory. Unlike Leigh and Nicolai's system, it does so autonomously, appealing to no supplemental axioms—such as CodAx—that require informal motivation beyond the commitments of DZ^2 and the truth theory. The spectre of structural joint instability is thus dispelled.

We now formally state DZ^{2CT}, the compositional theory for DZ^2. First fix a coding $\ulcorner\urcorner : \mathsf{Sent}(\mathscr{L}_{\mathsf{Z}^2_b}) \to \mathbb{N}_s$: terms for codes of $\mathscr{L}_{\mathsf{Z}^2_b}$ expressions will be terms of the *syntax language* $\mathscr{L}_{\mathsf{Z}^2_s}$. In order to make dealing with assignments tractable, we take advantage of the fact that pairing is definable in DZ^2. For each n-ary higher-order entity, we can define a function f_n coding it into a singulary second-order entity—i.e. a subset of the domain. We can also define a function g allowing us to simulate first-order quantification using a second-order entity by lifting each individual to its singleton. We thus have a denumerable collection $\langle g_1, g_2, \ldots, f^1_1, f^1_2, \ldots, f^2_1, f^2_2, \ldots\rangle$ going proxy for the values on an assignment of the variables $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{X}^1_1, \mathbf{X}^1_2, \ldots, \mathbf{X}^2_1, \mathbf{X}^2_2, \ldots$ (which are in turn represented in our syntax theory by codes we write as $\mathsf{v}_1, \mathsf{v}_2, \ldots, \mathsf{V}^1_1, \mathsf{V}^1_2, \ldots, \mathsf{v}^2_1, \mathsf{v}^2_2, \ldots$).

We can use additional coding devices to simulate a countable collection of subsets

of an infinite domain by a single subset. In this way we can use singulary second-order entities as proxies for assignments of values to all variables of the language (both first- and second-order), as well as a family of definable functions of each adicity allowing us to extract the value of a given variable from the assignment. We use α, β to range over assignments (in this coded sense of 'assignment'), and we write $[\![\mathsf{v}]\!]_\alpha$ (resp. $[\![\mathsf{V}]\!]_\alpha$) for the value of a given first- or second-order variable on α. We write $\alpha \overset{\mathsf{v}}{\sim} \beta$ to indicate that α differs from β only in the value assigned to v, $\alpha^{[[\![\mathsf{v}]\!]:=x]}$ for the assignment differing from α only by assigning x to v, and $\alpha^{[\mathsf{v}_1/\mathsf{v}_2]}$ for the assignment differing from α only by exchanging the values of v_1 and v_2; all of these notations are extended in the obvious way to higher-order variables and sequences of variables.

Further, as a matter of convenience, we extend $[\![]\!]_\alpha$ to all variables and all terms of the language in the obvious way: $[\![\bar{0}_b]\!]_\alpha = 0_b$, $[\![\dot{\mathsf{S}}\mathsf{t}]\!]_\alpha = \mathsf{S}[\![\mathsf{t}]\!]_\alpha$, and so forth. (The $[\![]\!]_\alpha$ expression is, of course, shorthand for a family of complex formulas picking out entities of various types, but in practice this will cause no confusion).

Finally, we introduce a new primitive cross-type predicate Sat which takes a singulary second-order entity and a syntactic entity: on the intended interpretation, $\mathsf{Sat}_\alpha\phi$ is true if and only if ϕ, the coded syntactic object, denotes a formula of the base language which is satisfied on the assignment coded up by α.

To obtain DZ^{2CT}, we add the following truth-theoretic axioms to DZ^2:

(T1) $\quad \forall\alpha\forall\mathsf{t}_1\forall\mathsf{t}_2(\mathsf{Sat}_\alpha(\mathsf{t}_1 \dot{=} \mathsf{t}_2) \leftrightarrow [\![\mathsf{t}_1]\!]_\alpha = [\![\mathsf{t}_2]\!]_\alpha)$

(T2) $\quad \forall\alpha\forall\mathsf{t}(\mathsf{Sat}_\alpha \dot{\mathsf{N}}_b\mathsf{t} \leftrightarrow \mathsf{N}_b[\![\mathsf{t}]\!]_\alpha)$

(T3) $\quad \forall\alpha\forall\mathsf{V}^n\forall\mathsf{t}_1\cdots\forall\mathsf{t}_n(\mathsf{Sat}_\alpha \dot{\mathsf{V}}^n\mathsf{t}_1\cdots\mathsf{t}_n \leftrightarrow [\![\mathsf{V}^n]\!]_\alpha[\![\mathsf{t}_1]\!]_\alpha\cdots[\![\mathsf{t}_n]\!]_\alpha)$

(T4) $\quad \forall\alpha\forall\phi(\mathsf{Sat}_\alpha \dot{\neg}\phi \leftrightarrow \neg\mathsf{Sat}_\alpha\phi)$

(T5) $\quad \forall\alpha\forall\phi\forall\psi(\mathsf{Sat}_\alpha(\phi \dot{\wedge} \psi) \leftrightarrow \mathsf{Sat}_\alpha\phi \wedge \mathsf{Sat}_\alpha\psi)$

(T6) $\quad \forall\alpha\forall\phi\forall\psi(\mathsf{Sat}_\alpha(\phi \dot{\vee} \psi) \leftrightarrow \mathsf{Sat}_\alpha\phi \vee \mathsf{Sat}_\alpha\psi)$

(T7) $\quad \forall\alpha\forall\phi\forall\mathsf{v}(\mathsf{Sat}_\alpha \dot{\forall}\mathsf{v}\phi \leftrightarrow (\forall\beta \overset{\mathsf{v}}{\sim} \alpha)\mathsf{Sat}_\beta\phi)$

(T8) $\quad \forall\alpha\forall\phi\forall\mathsf{V}^n(\mathsf{Sat}_\alpha \dot{\forall}\dot{\mathsf{V}}^n\phi \leftrightarrow (\forall\beta \overset{\mathsf{V}^n}{\sim} \alpha)\mathsf{Sat}_\beta\phi)$

Conventions for the use of $\forall\mathsf{t}$ and similar expressions correspond to the obvious disentangled analogues of those introduced for PA^{CT}. Note that (T3) and (T8) are schematic in n, as is appropriate for a polyadic theory.[17]

The crucial result, whose proof we defer to the Appendix is,

[17] In the Appendix, for technical purposes, we work with a simpler theory which obviates the need for this device.

Theorem 4. $\mathsf{DZ}^{2CT} \vdash \mathrm{Con}_s(\mathsf{Z}_b^2)$.

Thus, by the Transfer of Consistency Lemma, $\mathsf{DZ}^{2CT} \vdash \mathrm{Con}_b(\mathsf{Z}_b^2)$.

V. Concluding Discussion

We have introduced DZ^2 and its truth theoretic extension DZ^{2CT}, and argued that (unlike their main disentangled competitors) both are epistemically stable. In this final section we consider some objections and some potential implications of our results for deflationism, Dedekind's Thesis, and Isaacson's Thesis.

Our results make use of second-order arithmetic. It might be objected, therefore, that their interest is limited. The standard semantics for second-order theories is given within set theory: the second-order quantifiers are taken to range over the full powerset of the domain. In our case, the intended domain is countably infinite. But, the objection goes, it is problematic to suppose that we have a determinate conception of second-order quantification, since the determinacy of the powerset of the natural numbers (and syntax) is dubious. Of course, set-theoretic realists will find an appeal to $\mathscr{P}(\mathbb{N})$ unproblematic; there are, however, many positions in the philosophy of mathematics—predicativism, strong set-theoretic pluralism, and so forth—on which determinately quantifying over *every* subset of $\mathbb{N}$ is impossible. Has our argument any interest for proponents of those positions?

We maintain that it has. We have two main responses to the objection. First, nothing in our treatment of second-order logic made essential use of set theory. There is no reason why we could not work (e.g. in our coding of assignments) within a higher-order metalanguage. Second, and more fundamentally, there is no sense in which we presuppose the determinacy of second-order logic. All of the results we give are *theorems* of the relevant second-order theories, not merely semantic consequences. Internal Categoricity, for instance, which is used in the proof of the Transfer Theorem, is simply a theorem: thus it is valid not only on all standard interpretations of the second-order quantifiers but also on all Henkin interpretations (whereby the quantifiers range over some specified, possibly proper, subset of the full powerset of the domain). Unlike the standard semantics, there exist sound and complete proof procedures for Henkin semantics.

There is also another possible worry about the status of second-order resources in DZ^2. Our demonstration of the Transfer Theorem requires that second-order quantifiers range over higher-order entities whose extensions are *mixed* in that they include both mathematical and syntactic objects. Is this compatible with the basic idea behind disentanglement? One of the central motivations of disentanglement in the first-order setting, after all, was to separate out syntactic from purely mathematical instances of

the induction schema. In the second-order setting induction is an axiom, but we have the full panoply of syntactic, mathematical, and mixed instances of comprehension. Is it not unsurprising that $\mathrm{Con}_b(\mathsf{Z}_b^2)$ is provable? After all, $\mathrm{Con}_b(\mathsf{PA}_b)$ becomes provable in Leigh and Nicolai's system (2013: 627) once the induction schema is fully extended.

In our view, there is an important disanalogy between the induction schema in PA_b and the comprehension schema in DZ^2. In a second-order setting, (CA) is a logical principle: in particular, it is a natural complement of unrestricted existential generalization. If we lose (CA) for formulas with syntactic vocabulary, then we lose such principles as $\mathrm{Sent}(\ulcorner\Phi\urcorner) \vdash \exists X\ X\ulcorner\Phi\urcorner$; but surely, if we accept second-order logic at all, we should accept this inference. In contrast, in the first-order framework, mathematical induction is *contentual*, and adding new instances adds new subject matter: as Leigh and Nicolai note, extending induction "is somewhat unnatural, at least from our point of view, as the interaction between 'mathematical' and 'syntactic' schemas [...] was exactly what the setting with 'disentangled syntax' wanted to avoid" (2013: 628).

What are the implications of our discussion for Isaacson's and Dedekind's Theses? In our view, it lends support to both. Bridge principles such as Leigh and Nicolai's coding axioms are not derivable within $\mathsf{PA}_b + \mathsf{PA}_s$. Isaacson's Thesis offers a satisfying and elegant explanation of this fact: the coding axioms are paradigmatically 'higher order' propositions, whose justification relies on our appreciation of a structural similiarity between arithmetic and syntax. We take it that this lends some abductive support to Isaacson's Thesis. Finally, it is a mark in favour of Dedekind's Thesis that the coding axioms are derivable within DZ^2. To be sure, it cannot be the case that Dedekind's Thesis requires everything true in (our fullest conception of) the natural number structure to be provable from Z^2, for obvious Gödelian reasons. But in the case of DZ^2, the Transfer Principle does not represent any increase in consistency strength: if it were naturally justified by the conception behind DZ^2 but not provable from it, this would be a lacuna not directly explicable on Gödelian grounds, suggesting that DZ^2 in fact failed to capture our conception of the structure of the natural numbers and syntax.

What exactly is the upshot of our argument for deflationism? We've argued that DZ^{2CT} is a very natural theory of arithmetic, syntax, and truth to adopt; and furthermore, that it is hard to find a weaker disentangled truth theory that is not vulnerable to serious stability-related objections. If deflationist theories of truth and syntax must indeed be conservative over their arithmetical base theories, then this result spells trouble for deflationism. At the very least, it is incumbent upon the deflationist to propose an alternative that is both epistemically stable and conservative.

We conclude by gesturing at a final set of issues for further investigation. The conservativeness constraint with which we began can be understood in two different ways:

(WEAK CONSERVATIVENESS) Adding a theory of truth to *our best total truth-*

free theory must result in a conservative extension;

(STRONG CONSERVATIVENESS) Adding a theory of truth to *any natural fragment of our best total truth-free theory* (including, presumably, our best total theory of natural numbers and syntax) must result in a conservative extension.

Much of the literature has presupposed that the relevant constraint is something like Strong Conservativeness: otherwise, it would be irrelevant to focus upon theories of arithmetical truth, since arithmetic has no reasonable claim to being our best total mathematical theory, let alone our best total theory. If Strong Conservativeness holds, then our arguments above show that the disentanglement programme leaves deflationism in bad shape. But if, by contrast, Weak Conservativeness is the best way of understanding the constraint, then many issues remain to be explored. In particular: a full assessment of the conservativeness objection would require an investigation of adding a disentangled theory of truth and syntax to theories—for instance, various formulations of set theory— which have a plausible claim to being our best total mathematical theories. Perhaps this approach holds out a possible means of escape for the deflationist; perhaps not. But that is work for another day.

Appendix

This appendix outlines DZ^{2CT} and substantiates various claims about it made above. Our main aim is to show that $\mathsf{DZ}^{2CT} \vdash \mathrm{Con}_b(\mathsf{Z}^2_b)$. It follows, via results in §4, that DZ^{2CT} also proves $\mathrm{Con}_s(\mathsf{Z}^2_b)$, $\mathrm{Con}_b(\mathsf{Z}^2_s)$, and $\mathrm{Con}_s(\mathsf{Z}^2_s)$. The fact that DZ^{2CT} proves the consistency of the base and syntax theories is evidence in favour of its naturalness as a disentangled truth-theoretic framework.

We will work not in Z^{2CT} but in the monadic second-order system, $\hat{\mathsf{Z}}^{2CT}$. No generality is lost, since polyadic second-order quantification can be coded via a pairing function. Because this coding is primitive recursive, $\mathrm{Con}(\mathsf{Z}^2) \leftrightarrow \mathrm{Con}(\hat{\mathsf{Z}}^2)$ will be provable in a very weak base theory. So if $\hat{\mathsf{Z}}^{2CT} \vdash \mathrm{Con}(\hat{\mathsf{Z}}^2)$ then $\hat{\mathsf{Z}}^{2CT} \vdash \mathrm{Con}(\mathsf{Z}^2)$. But since $\hat{\mathsf{Z}}^{2CT}$ is a subtheory of Z^{2CT}, $\mathsf{Z}^{2CT} \vdash \mathrm{Con}(\mathsf{Z}^2)$ if $\hat{\mathsf{Z}}^{2CT} \vdash \mathrm{Con}(\hat{\mathsf{Z}}^2)$. Furthermore, Z^2_b can be relatively interpreted in Z^2, $\hat{\mathsf{Z}}^{2CT} \vdash \mathrm{Con}(\hat{\mathsf{Z}}^2)$ only if $\mathsf{DZ}^{2CT} \vdash \mathrm{Con}_b(\mathsf{Z}^2_b)$.

$\hat{\mathsf{Z}}^{2CT}$ comprises base axioms:

(A1) $\quad \forall x(\mathsf{S}x \neq 0)$

(A2) $\quad \forall x \forall y(\mathsf{S}x = \mathsf{S}y \rightarrow x = y)$

(A3) $\quad \forall X(X0 \wedge \forall x(Xx \rightarrow X\mathsf{S}x) \rightarrow \forall x Xx)$

together with truth-theoretic axioms:

(T1) $\forall\alpha\forall t_1\forall t_2(\mathrm{Sat}_\alpha(t_1 \doteq t_2) \leftrightarrow [\![t_1]\!]_\alpha = [\![t_2]\!]_\alpha)$

(T2) $\forall\alpha\forall v\forall t(\mathrm{Sat}_\alpha vt \leftrightarrow [\![v]\!]_\alpha[\![t]\!]_\alpha)$

(T3) $\forall\alpha\forall\phi(\mathrm{Sat}_\alpha \dot{\neg}\phi \leftrightarrow \neg\mathrm{Sat}_\alpha\phi)$

(T4) $\forall\alpha\forall\phi\forall\psi(\mathrm{Sat}_\alpha(\phi \dot{\wedge} \psi) \leftrightarrow \mathrm{Sat}_\alpha\phi \wedge \mathrm{Sat}_\alpha\psi)$

(T5) $\forall\alpha\forall\phi\forall\psi(\mathrm{Sat}_\alpha(\phi \dot{\vee} \psi) \leftrightarrow \mathrm{Sat}_\alpha\phi \vee \mathrm{Sat}_\alpha\psi)$

(T6) $\forall\alpha\forall\phi\forall v(\mathrm{Sat}_\alpha \dot{\forall} v\,\phi \leftrightarrow (\forall\beta \overset{v}{\sim} \alpha)\mathrm{Sat}_\beta\phi)$

(T7) $\forall\alpha\forall\phi\forall V(\mathrm{Sat}_\alpha \dot{\forall} V\,\phi \leftrightarrow (\forall\beta \overset{V}{\sim} \alpha)\mathrm{Sat}_\beta\phi)$

We define $T\phi$ as $\forall\alpha\mathrm{Sat}_\alpha\phi$. Note that, in a mild abuse of notation, this allows truth to be attributed to open formulas as well as sentences.

We will show that $\hat{\mathsf{Z}}^{2CT}$ proves the global reflection principle for $\hat{\mathsf{Z}}^2$:

Theorem 5. $\hat{\mathsf{Z}}^{2CT} \vdash \forall\phi(\mathrm{Bew}_{\hat{\mathsf{Z}}^2}\phi \rightarrow T\phi)$.

We follow the usual strategy of formalizing the 'semantic argument': all the axioms of the system are true; all its rules of inference are truth preserving; so all its theorems are true. For the sake of definiteness we use the deductive system in Shapiro (1991). To that end we prove the following six claims.

(Sem1) $\hat{\mathsf{Z}}^{2CT} \vdash \forall\phi(\mathrm{LogAx}\phi \rightarrow T\phi)$,

(Sem2) $\hat{\mathsf{Z}}^{2CT} \vdash \forall\phi(\mathrm{PropAx}\phi \rightarrow T\phi)$,

(Sem3) $\hat{\mathsf{Z}}^{2CT} \vdash \forall\phi(\mathrm{CompAx}\phi \rightarrow T\phi)$,

(Sem4) $\hat{\mathsf{Z}}^{2CT} \vdash \forall\phi\forall\psi((T(\phi \dot{\rightarrow} \psi) \wedge T\phi) \rightarrow T\psi)$,

(Sem5) $\hat{\mathsf{Z}}^{2CT} \vdash \forall\phi\forall\psi\forall v((T(\phi \dot{\rightarrow} \psi) \wedge \neg\mathrm{Free}(v, \phi)) \rightarrow T(\phi \dot{\rightarrow} \dot{\forall} v\,\psi))$,

(Sem6) $\hat{\mathsf{Z}}^{2CT} \vdash \forall\phi\forall\psi\forall V((T(\phi \dot{\rightarrow} \psi) \wedge \neg\mathrm{Free}(V, \phi)) \rightarrow T(\phi \dot{\rightarrow} \dot{\forall} V\,\psi))$.

Here LogAx expresses the property of being the code of an instance of a logical axiom, PropAx expresses the property of being the code of an instance of (A1)–(A3), and CompAx expresses the property of being a code of an instance of (CA). (Sem4) formalizes the truth-preservingness of modus ponens; similarly (Sem5) and (Sem6) for the rules of inference governing the quantifiers.

From (Sem1)-(Sem6), the required reflection principle will follow, and so too will $\mathrm{Con}_b(\mathsf{Z}^2_b)$.

A number of additional lemmata will be of use.

Lemma 6. (Disquotation Lemma) *If Φ has no free variables, then* $\hat{\mathsf{Z}}^{2CT} \vdash \Phi \leftrightarrow T^{\ulcorner}\Phi^{\urcorner}$.

Lemma 7. (Closure) *Let* $\mathrm{ucl}(^{\ulcorner}\Phi^{\urcorner})$ *be the code of* Φ*'s universal closure.* $\hat{\mathsf{Z}}^{2CT} \vdash \forall\phi(T\phi \leftrightarrow T\mathrm{ucl}(\phi))$.

Lemma 8. (Substitution of Provable Equivalents) $\hat{\mathsf{Z}}^{2CT} \vdash \forall\phi\forall\psi\forall\chi((\phi \leftrightarrow \psi) \rightarrow (T\chi \leftrightarrow T\chi^{\psi}/_{\phi}))$.

Lemma 9. (Variable-Swapping) *Let* $\tilde{\alpha}$ *be* $\alpha^{[\mathsf{v}_{a_1},\ldots,\mathsf{v}_{a_m},\mathsf{v}_{b_1},\ldots,\mathsf{v}_{b_n}/\mathsf{v}_{c_1},\ldots,\mathsf{v}_{c_m},\mathsf{v}_{d_1},\ldots,\mathsf{v}_{d_n}]}$. *Then* $\hat{\mathsf{Z}}^{2CT} \vdash \forall\phi\forall\alpha(\mathrm{Sat}_{\alpha}\phi \leftrightarrow \mathrm{Sat}_{\tilde{\alpha}}\phi^{\mathsf{v}_{a_1},\ldots,\mathsf{v}_{a_m},\mathsf{v}_{b_1},\ldots,\mathsf{v}_{b_n}/\mathsf{v}_{c_1},\ldots,\mathsf{v}_{c_m},\mathsf{v}_{d_1},\ldots,\mathsf{v}_{d_n}})$.

Proof. Disquotation follows by a simple induction in the metalanguage. Closure, Substitution of Provable Equivalents, and Variable-Swapping proceed by internal inductions. ⊣

We first show (Sem4)–(Sem6). (Sem4) follows straightforwardly from (T3), (T5), and the definition of the conditional. For (Sem5) and (Sem6), we need the following lemma, saying that if two assignments agree on all free variables in Φ, then they agree on Φ:

Lemma 10.

$$\hat{\mathsf{Z}}^{2CT} \vdash \forall\alpha\forall\beta\forall\phi\forall\mathsf{V}\forall\mathsf{v}(((\mathrm{Free}(\mathsf{v},\phi) \rightarrow [\![\mathsf{v}]\!]_{\alpha} = [\![\mathsf{v}]\!]_{\beta}) \wedge (\mathrm{Free}(\mathsf{V},\phi) \rightarrow [\![\mathsf{V}]\!]_{\alpha} = [\![\mathsf{V}]\!]_{\beta})) \rightarrow (\mathrm{Sat}_{\alpha}\phi \leftrightarrow \mathrm{Sat}_{\beta}\phi));$$

Proof. By an internal induction on complexity of formulas in $\hat{\mathsf{Z}}^{2CT}$. We write $\alpha \overset{\mathrm{fv}}{\sim} \beta$ if α and β agree on all free variables in the relevant formula. The base case is clear; it relies only on (T1), (T2), and the definition of $[\![]\!]$. For the induction step, suppose $\forall\psi\forall\alpha\forall\beta(\alpha \overset{\mathrm{fv}}{\sim} \beta \rightarrow (\mathrm{Sat}_{\alpha}\psi \leftrightarrow \mathrm{Sat}_{\beta}\psi)$ for all ψ of complexity $< n$ and that ϕ has complexity n. The only difficult cases are the quantifiers. We show the $\forall\mathsf{v}$ case; the $\forall\mathsf{V}$ case is similar.

We reason informally in $\hat{\mathsf{Z}}^{2CT}$. First, if v is not free in ψ, then $\forall\mathsf{v}\psi$ and ψ are equivalent, and so we are done. Thus assume v is free in ψ and suppose $\alpha \overset{\mathrm{fv}}{\sim} \beta$ and $\mathrm{Sat}_{\alpha}\forall\dot{\mathsf{v}}\psi$. Then for all $\alpha' \overset{\mathsf{v}}{\sim} \alpha$, $\mathrm{Sat}_{\alpha'}\psi$. Now suppose $\beta' \overset{\mathsf{v}}{\sim} \beta$. Then there is some $\alpha' \overset{\mathsf{v}}{\sim} \alpha$ such that $\beta' \overset{\mathrm{fv}}{\sim} \alpha'$. Now, by the IH, we have $\mathrm{Sat}_{\alpha'}\psi \leftrightarrow \mathrm{Sat}_{\beta'}\psi$. So $\mathrm{Sat}_{\beta'}\psi$. So for all β' such that $\beta' \overset{\mathsf{v}}{\sim} \beta$, $\mathrm{Sat}_{\beta'}\psi$. But then $\mathrm{Sat}_{\beta}\forall\dot{\mathsf{v}}\psi$. ⊣

Two immediate consequences of Lemma 10 are:

$$\forall\phi\forall\mathsf{v}(\neg\mathrm{Free}(\mathsf{v},\phi) \rightarrow (\forall\alpha\forall\beta(\alpha \overset{\mathsf{v}}{\sim} \beta \rightarrow (\mathrm{Sat}_{\alpha}\phi \leftrightarrow \mathrm{Sat}_{\beta}\phi)))$$

and

$$\forall\phi\forall\mathsf{V}(\neg\mathrm{Free}(\mathsf{V},\phi) \rightarrow (\forall\alpha\forall\beta(\alpha \overset{\mathsf{V}}{\sim} \beta \rightarrow (\mathrm{Sat}_{\alpha}\phi \leftrightarrow \mathrm{Sat}_{\beta}\phi)))$$

From these two claims, (Sem5) and (Sem6) follow by (T7) and (T8).

To show (Sem1) is more difficult and requires a formalized induction.

Lemma 11. $\hat{\mathsf{Z}}^{2CT} \vdash \forall\phi(\mathrm{LogAx}\phi \to T\phi)$.

Proof. These are straightforward; as an example, we show

$$\forall\psi\forall\mathsf{v}\forall\mathsf{t}(\mathrm{FreeFor}(\psi,\mathsf{t},\mathsf{v}) \to T(\forall\dot{\mathsf{v}}\phi\dot{\to}\phi^{\mathsf{t}}/_{\mathsf{v}})).$$

Working within $\hat{\mathsf{Z}}^{2CT}$, assume for reductio that, for some α, $\mathrm{Sat}_\alpha\forall\dot{\mathsf{v}}\phi$ but not $\mathrm{Sat}_\alpha\phi^{\mathsf{t}}/_{\mathsf{v}}$. So for all $\beta \overset{\mathsf{v}}{\sim} \alpha$, $\mathrm{Sat}_\beta\phi$. We define $\alpha' = \alpha^{[\mathsf{v}:=[\![\mathsf{t}]\!]_\alpha]}$. Since $\alpha' \overset{\mathsf{v}}{\sim} \alpha$, $\mathrm{Sat}_{\alpha'}\phi$; but, by the definition of $[\![]\!]_\alpha$ and the fact that t is free for v in ϕ, $\mathrm{Sat}_{\alpha'}\phi$ if and only if $\mathrm{Sat}_\alpha\phi^{\mathsf{t}}/_{\mathsf{v}}$. ⊣

We turn to (Sem2). In the setting with only monadic second-order quantification, the proper axioms of $\hat{\mathsf{Z}}^2$ are just (A1)–(A3).

Either x is free in (A3) or it is not; in the former case, we can apply Closure to reduce it to a closed sentence. Then, since there are only finitely many proper axioms, Disquotation yields the desired result.

The hardest case is (Sem3). In order to prove that *all* instances of the comprehension axiom are true (as opposed to each of the instances, which is immediate from Disquotation), we appeal to a single judiciously chosen instance of comprehension.

Lemma 12. $\hat{\mathsf{Z}}^{2CT} \vdash \forall\phi(\mathrm{CompAx}\phi \to T\phi)$.

Proof. We first define Sat^*_α as $\mathrm{Sat}_{\alpha[[\mathsf{V}]:=X,[\mathsf{v}]:=x]}$. Using (CA), we have

$$\hat{\mathsf{Z}}^{2CT} \vdash \exists X\forall x(Xx \leftrightarrow \mathrm{Sat}^*_\alpha\psi)$$

with both ψ and α free; generalizing yields

$$\hat{\mathsf{Z}}^{2CT} \vdash \forall\psi\forall\alpha\exists X\forall x(Xx \leftrightarrow \mathrm{Sat}^*_\alpha\psi).$$

Now, by Substitution of Provable Equivalents, we have

$$\hat{\mathsf{Z}}^{2CT} \vdash \forall\psi\forall\alpha\exists X\forall x(\mathrm{Sat}^*_\alpha\dot{\mathsf{V}}\mathsf{v} \leftrightarrow \mathrm{Sat}^*_\alpha\psi).$$

Applying satisfaction clauses for the connectives, we obtain

$$\hat{\mathsf{Z}}^{2CT} \vdash \forall\psi\forall\alpha\exists X\forall x\mathrm{Sat}^*_\alpha(\dot{\mathsf{V}}\mathsf{v}\dot{\leftrightarrow}\psi).$$

But note that $\forall x\ \mathrm{Sat}_{\alpha[[\mathsf{V}]:=X,[\mathsf{v}]:=x]}\chi$ is provably equivalent to $(\forall\beta \overset{\mathsf{v}}{\sim} \alpha)\mathrm{Sat}_{\beta[[\mathsf{V}]:=X]}\chi$, which in turn is provably equivalent, using (T6), to $\mathrm{Sat}_{\alpha[[\mathsf{V}]:=X]}\forall\dot{\mathsf{v}}\ \chi$. Using Substitution

of Provable equivalents again, we have

$$\hat{\mathsf{Z}}^{2CT} \vdash \forall\psi\forall\alpha\exists X\mathrm{Sat}_{\alpha^{\frown}[\ulcorner\mathsf{V}\urcorner:=X]}\dot{\forall}\dot{\mathsf{v}}(\dot{\mathsf{V}}\dot{\mathsf{v}}\dot{\leftrightarrow}\psi).$$

But $\exists X\mathrm{Sat}_{\alpha^{\frown}[\ulcorner\mathsf{V}\urcorner:=X]}\chi$ is provably equivalent to $(\exists\beta \overset{\mathsf{V}}{\sim} \alpha)\mathrm{Sat}_{\beta}\chi$, which in turn is provably equivalent to $\mathrm{Sat}_{\alpha}\dot{\exists}\dot{\mathsf{V}}\chi$ by (T7), definitions, and predicate logic. So a final application of Substitution of Provable Equivalents yields

$$\hat{\mathsf{Z}}^{2CT} \vdash \forall\psi\forall\alpha\mathrm{Sat}_{\alpha}\dot{\exists}\dot{\mathsf{V}}\dot{\forall}\dot{\mathsf{v}}(\dot{\mathsf{V}}\dot{\mathsf{v}}\dot{\leftrightarrow}\psi),$$

or, equivalently,

$$\hat{\mathsf{Z}}^{2CT} \vdash \forall\psi T(\dot{\exists}\dot{\mathsf{V}}\dot{\forall}\dot{\mathsf{v}}(\dot{\mathsf{V}}\dot{\mathsf{v}}\dot{\leftrightarrow}\psi)).$$

But this is the general form of an instance of CompAx; expanding definitions and applying clauses for the connectives, we get $\hat{\mathsf{Z}}^{2CT} \vdash \forall\phi(\mathrm{CompAx}\phi \rightarrow T\phi)$. ⊣

Works Cited

Addison, J.W., Leon Henkin, and Alfred Tarski, eds. (1965). *The Theory of Models: Proceedings of the 1963 International Symposium at Berkeley*. Amsterdam: North-Holland.

Barwise, Jon, ed. (1977). *Handbook of Mathematical Logic*. Amsterdam: North-Holland.

Button, Tim, and Sean Walsh (2018). *Philosophy and Model Theory*. Oxford: Oxford University Press.

Cieśliński, Cesary (2017). *The Epistemic Lightness of Truth: Deflationism and Its Logic*. Cambridge: Cambridge University Press.

Corcoran, John, William Frank, and Michael Maloney (1974). 'String Theory'. *Journal of Symbolic Logic* **39**: 625–37.

Cortois, Paul, ed. (1995). *The Many Problems of Realism*. Tilburg: Tilburg University Press.

Crossley, J.N., and Michael Dummett, eds. (1965). *Formal Systems and Recursive Functions*. Amsterdam: North-Holland.

de Bouvère, K.L. (1965a). 'Logical Synonymity'. *Indagationes Mathematicae* **22**: 622–29.

———————— (1965b). 'Synonymous Theories'. In Addison et al. (1965): 402–6.

Dean, Walter (2015). 'Arithmetical Reflection and the Provability of Soundness'. *Philosophia Mathematica* (3rd ser.) **23**: 31–64.

Dedekind, Richard (1888). *Was sind und was sollen die Zahlen?* Braunschweig: Vieweg und Sohn.

Feferman, Solomon (1964). 'Systems of Predicative Analysis'. *Journal of Symbolic Logic* **29**: 1–30.

——————— (1991). 'Reflecting on Incompleteness'. *Journal of Symbolic Logic* **56**: 1–49.

Field, Hartry (1999). 'Deflating the Conservativeness Argument'. *Journal of Philosophy* **96**: 533–40.

Fujimoto, Kentaro (MS). 'Deflationism beyond Arithmetic'. Forthcoming, *Synthese*.

Grzegorczyk, Andrzej (2005). 'Undecidability without Arithmetization'. *Studia Logica* **79**: 163–230.

Halbach, Volker (1999). 'Disquotationalism and Infinite Conjunctions'. *Mind* **108**: 1–22.

——————— (2001). 'How Innocent is Deflationism?' *Synthese* **126**: 167–94.

——————— (2011). *Axiomatic Theories of Truth*. Cambridge: Cambridge University Press.

——————— (2014). *Axiomatic Theories of Truth*. 2nd edn. Cambridge: Cambridge University Press.

Hájek, Pavel, and Petr Pudlák (1993). *Metamathematics of First-Order Arithmetic*. Berlin: Springer.

Heck, Richard Kimberly (2015). 'Consistency and the Theory of Truth'. *Review of Symbolic Logic* **8**: 424–66 (originally published under the name "Richard G. Heck, Jr").

——————— (2018). 'The Logical Strength of Compositional Principles'. *Notre Dame Journal of Formal Logic* **55**: 1–33 (originally published under the name "Richard G. Heck, Jr").

——————— (MS). 'The Strength of Truth Theories'. Typescript.

Hilbert, David (1922). 'Neubegründung der Mathematik: Erste Mitteilung'. *Abhandlungen aus dem Seminar der Hamburgischen Universität* **1**: 157–177. Rpt. in Hilbert (1932–35): **3**: 157–77; citations to reprint. Trans. as 'The New Grounding of Mathematics: First Report' in Mancosu (1987): 198–214.

——————— (1932–35). *Gesammelte Abhandlungen*. 3 vols. Berlin: Springer.

———————, and Paul Bernays (1934–39). *Grundlagen der Mathematik*. 2 vols. Berlin: Springer.

Horsten, Leon (1995). 'The Semantical Paradoxes, the Neutrality of Truth and the Neutrality of the Minimalist Theory of Truth'. In Cortois (1995): 173–87.

——————— (2011). *The Tarskian Turn. Deflationism and Axiomatic Truth*. Cambridge, Mass.: MIT Press.

Isaacson, Daniel (1987). 'Arithmetical Truth and Hidden Higher-order Concepts'. In Paris Logic Group (1987): 147–59.

Ketland, Jeffrey (1999). 'Deflationism and Tarski's Paradise'. *Mind* **108**: 69–94.

Kreisel, Georg (1958). 'Ordinal Logics and the Characterization of Informal Concepts of Proof'. In Todd (1958): 289–99.

——————— (1960). 'La prédicativité'. *Bulletin de la Société Mathématique de France* **88**: 371–91.

——————— (1965). 'Mathematical Logic'. In Saaty (1965): **3**: 95–195.

——————— (1967). 'Informal Rigour and Completeness Proofs'. In Lakatos (1967): 138–85.

Lakatos, Irme, ed. (1967). *Problems in the Philosophy of Mathematics*. Amsterdam: North-Holland.

Lavine, Shaughan (MS). 'Skolem Was Wrong'. Typescript.

Leigh, Graham, and Carlo Nicolai (2013). 'Axiomatic Truth, Syntax and Metatheoretic Reasoning'. *Review of Symbolic Logic* **6**: 613–36.

Lindström, Per (2003). *Aspects of Incompleteness*. 2nd edn. Cambridge: Cambridge University Press.

Mancosu, Paolo, ed. (1998). *From Brouwer to Hilbert: The Debate on the Foundations of Mathematics in the 1920s*. New York: Oxford University Press.

McGee, Vann (1990). *Truth, Vagueness, and Paradox: An Essay on the Logic of Truth*. Indianapolis: Hackett.

——————— (1997). 'How We Learn Mathematical Language'. *Philosophical Review* **106**: 35–68.

Nicolai, Carlo (2015). 'Deflationary Truth and the Ontology of Expressions'. *Synthese* **192**: 4031–55.

Paris Logic Group, eds. (1987). *Logic Colloquium '85: Proceedings of the Colloquium Held in Orsay, France, July 1985*. Amsterdam: North-Holland..

Parsons, Charles (2008). *Mathematical Thought and Its Objects*. Cambridge: Cambridge University Press.

Quine, W.V. (1946). 'Concatenation as a Basis for Arithmetic'. *Journal of Symbolic Logic* **11**: 105–14.

Saaty, T.R., ed. (1965). *Lectures on Modern Mathematics*. 3 vols. New York: Wiley.

Schütte, Kurt (1965a). 'Predicative Well-Orderings'. In Crossley and Dummett (1965): 280–303.

——————— (1965b). 'Eine Grenze für die Beweisbarkeit der Transfiniten Induktion in der verzweigten Typenlogik'. *Archiv für mathematischen Logik und Grundlagenforschung* **7**: 45–60.

Shapiro, Stewart (1991). *Foundations without Foundationalism: A Case for Second-Order Logic*. Oxford: Clarendon Press.

——————— (1998). 'Proof and Truth: Through Thick and Thin'. *Journal of Philosophy* **95**: 493–521.

Simpson, Stephen G. (2009). *Subsystems of Second Order Arithmetic*. 2nd edn. Cambridge: Cambridge University Press.

——————— and Yokoyama, Keita (2013). 'Reverse Mathematics and Peano Categoricity'. *Annals of Pure and Applied Logic* **164**: 284–93.

Smith, Peter (2008). 'Ancestral Arithmetic and Isaacson's Thesis'. *Analysis* **68**: 1–10.

——————— (2013). *An Introduction to Gödel's Theorems*. 2nd edn. Cambridge: Cambridge University Press.

Smoryński, Craig (1977). 'The Incompleteness Theorems'. In Barwise (1977): 821–66.

Švedjar, Vítězslav (2009). 'On Interpretability in the Theory of Concatenation'. *Notre Dame Journal of Formal Logic* **50**: 87–95.

Tait, W.W. (1981). 'Finitism'. *Journal of Philosophy* **78**: 524–46.

Tarski, Alfred (1935). 'Der Wahrheitsbegriff in den formalisierten Sprachen'. *Studia Philosophica* 1: 261–405. Trans. as 'The Concept of Truth in Formalized Languages' in Tarski (1983): 152–278.

——————— (1983). *Logic, Semantics, Metamathematics*. Trans. J.H. Woodger. Ed. John Corcoran. 2nd edn. Indianapolis: Hackett.

———————, Andrzej Mostowski, and Raphael M. Robinson (1953). *Undecidable Theories*. Amsterdam: North-Holland.

Todd, J.A., ed. (1958). *Proceedings of the International Congress of Mathematicians, 14–21 August 1958*. Cambridge: Cambrisge University Press.

Väänänen, Jouko, and Tong Wang (2015). 'Internal Categoricity in Arithmetic and Set Theory'. *Notre Dame Journal of Formal Logic* **56**: 121–34.

Visser, Albert (2009). 'Growing Commas: A Study of Sequentiality and Concatenation'. *Notre Dame Journal of Formal Logic* **50**: 61–85.

——————— and Harvey Friedman (2014). 'When Bi-Interpretability Implies Synonymity'. Logic Group Preprint Series (University of Utrecht) **320**.

Waxman, Daniel (2017). 'Deflationism, Arithmetic, and the Argument from Conservativeness'. *Mind* **126**: 429–63.