Stable and Unstable Theories of Truth and Syntax

Beau Madison Mount
Universität Konstanz
beau.mount@uni-konstanz.de

Daniel Waxman
National University of Singapore
danielwaxman@nus.edu.sg

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

The systematic formal study of theories of truth derives from Alfred Tarski's paper 'The Concept of Truth in Formalized Languages' (1935). But, in one important respect, contemporary work on truth has strayed from Tarski's approach, which sharply distinguishes the mathematical base theory under analysis from the theory of syntax whose objects are the strings of the base language. The standard approach develops a theory of truth over a single mathematical theory—typically an arithmetical system of some sort—whose objects play a dual role, both as mathematical entities and as surrogates for syntactic entities. Syntax is simulated within arithmetic by coding, using techniques familiar from Gödel.[1] Despite the mathematical elegance of coding syntax within arithmetic, there are philosophical reasons to attend to distinctions collapsed by the contemporary approach.

First, a theory of truth is presumably committed to truth-bearers—in the Tarskian tradition, sentences. But if so, this commitment should be reflected faithfully in our truth theories, rather than concealed by way of coding techniques (however technically useful they are).

[1] See McGee (1990), Horsten (2011), and Halbach (2014) on truth theories of this type.

Mind, Vol. 00 . 0 . 2021
doi:10.1093/mind/fzaa034

Furthermore, for some purposes it is important to make fine-grained evaluations of the resources involved in various combinations of base and truth theories. For instance, deflationism is often understood as requiring that principles governing truth are conservative over appropriate base theories. But adding standard compositional theories of truth to arithmetic (and extending induction to formulas involving the truth predicate) results in a non-conservative extension (Horsten 1995; Shapiro 1998; Ketland 1999; Halbach 1999, 2001). In response, some authors have distinguished between instances of induction whose motivation derives from *arithmetic* and those whose motivation derives from *syntax* (Field 1999; Waxman 2017). However, in order to make such distinctions cleanly, arithmetic and syntax must be distinguished in a manner closer to Tarski's original approach.

Recently, a number of authors have explored *disentangled* systems, in which base and syntax theories are clearly separated (Heck MS, 2015, 2018; Leigh and Nicolai 2013; Nicolai 2015, 2016; and Fujimoto 2019). Our aim in this paper is to discuss a novel set of philosophical issues arising within this setting. We argue, drawing on an idea due to Walter Dean (2015), that any adequate total theory must be *epistemically stable*: well-motivated on a basis that motivates no strictly stronger framework. Epistemic stability is especially interesting in the disentangled setting, for an individually stable base theory can be combined with an individually stable syntax theory in such a way that the result is nevertheless jointly unstable. We argue that some systems discussed in the literature fail on precisely these grounds. As an alternative, we develop a powerful disentangled system we call *double second-order arithmetic* (DZ^2). DZ^2 is not only epistemically stable but, when combined with a compositional theory of truth, provides a philosophically natural and metamathematically fruitful setting for studying the interaction of arithmetic, truth, and syntax.

1. Disentangling Truth and Syntax

Before introducing the notion of a disentangled truth theory, we sketch a version of the usual, entangled approach as a point of comparison.

We take first-order Peano arithmetic (PA) as a truth-free starting point to which a truth predicate is subsequently added. PA plays a double role: as a mathematical theory (about the natural numbers) and, by way of a Gödel coding, as a theory of syntax. The language of Peano arithmetic, $\mathscr{L}_{\mathrm{PA}}$, contains a constant 0, a unary function symbol S, and binary function 0 symbols $+$ and $\times$, with the obvious intended interpretations. The axioms of PA are:

(PA1) $\forall x \forall y (Sx = Sy \rightarrow x = y)$,

(PA2) $\neg \exists x \;\; Sx = 0$,

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

(PA4) $\forall x \forall y \;\; x + Sy = S(x + y)$,

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

(PA6) $\forall x \forall y \;\; x \times Sy = (x \times y) + x$,

(PA7) $\Phi(0) \wedge \forall x (\Phi(x) \rightarrow \Phi(Sx)) \rightarrow \forall x \Phi(x)$ where $\Phi(x)$ is a formula of $\mathscr{L}_{\mathrm{PA}}$.

We use capital Φ, Ψ, etc. as schematic letters in our metalanguage for formulas of the language under consideration; lower-case ϕ, ψ, etc. are reserved for variables ranging over codes of formulas, as described below.

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

We focus on the full *compositional theory of truth*, PA^{CT}, whose axioms specify that truth distributes over the logical connectives and whose arithmetical induction schema includes formulas containing T.[2] PA^{CT} results from PA by replacing (PA7) with

(PA7′) $\Phi(0) \wedge \forall x (\Phi(x) \rightarrow \Phi(Sx)) \rightarrow \forall x \Phi(x)$, where $\Phi(x)$ is a formula of $\mathscr{L}^T_{\mathrm{PA}}$

and adding:

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

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

[2] We use the following notational conventions: $\forall \phi \Phi$ abbreviates $\forall x (\mathrm{Sent}_{\mathrm{PA}}(x) \rightarrow \Phi[x/\phi])$, where $\mathrm{Sent}_{\mathrm{PA}}$ expresses the property of being the code of a sentence of $\mathscr{L}_{\mathrm{PA}}$; $\forall \mathrm{t} \cdots$ and $\forall \mathrm{v} \cdots$ function 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 code of a sentence, the code of its negation. The functor $^\circ$ abbreviates an expression taking the code of a closed term to its value; for each number n, $\overline{n}$ denotes the numeral representing n; and $\phi \frac{\mathrm{t}_1}{\mathrm{t}_2}$ 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 (see, e.g., Smith 2013), so the displayed expressions can be viewed as metalinguistic abbreviations.

$$(\mathrm{CT3})\quad \forall\phi\forall\psi(T(\phi \underset{\cdot}{\wedge} \psi) \leftrightarrow T\phi \wedge T\psi),$$

$$(\mathrm{CT4})\quad \forall\phi\forall \mathrm{v}(T(\underset{\cdot}{\forall}\, \mathrm{v}\phi) \leftrightarrow \forall \mathrm{t}\, T\phi\tfrac{\mathrm{t}}{\mathrm{v}}).$$

In many ways, PA^{CT} is an appealing theory: it allows the formalization of much ordinary metamathematical reasoning, such as the proof that, since the axioms of PA are true and the rules of inference preserve truth, all theorems of PA are true. It follows that PA^{CT} is not a conservative extension of PA; in particular, PA^{CT} proves Con(PA), the canonical arithmetized consistency statement for PA.

We now turn to the disentangled setting, where, in contrast to the standard approach, the base theory and syntax theory are separated: formulated in different languages, to be interpreted as ranging over distinct sorts of objects.

Given a fixed alphabet, there are two natural ways to provide an autonomous syntax theory. One takes *concatenation* as primitive: the theory contains constants for each member of the alphabet and a function for concatenation of strings. The other takes *adjunction* as primitive: for each symbol in the alphabet, the theory has a function for the operation of appending that symbol to a string.

Despite the fact that syntax can be developed in either of these ways, various technical results show that these approaches are in a precise sense formally equivalent.

First, given a fixed alphabet, second-order versions of the concatenation-based theory and the adjunction-based theory are synonymous (Corcoran et al. 1974). This implies that there is a translation which systematically turns any proof in one of the theories into a proof of the same result (suitably translated into the language of the other theory).[3] Second, theories with alphabets of different sizes are also synonymous. The upshot of these facts is that it is possible to work in a particular version of syntax theory, with a fixed alphabet of any given size, without losing any mathematical generality. Third, one such theory—the second-order adjunction-based theory with a single alphabet-symbol—is, up to choice of labelling, identical to second-order arithmetic (Z^2).

[3] As Friedman and Visser put it (2014, p. 1), synonymity is 'the strictest notion of sameness of theories except strict identity of signature and set of theorems'. While the synonymity results of Corcoran et al. apply only to second-order versions of the theories in question, similar but slightly weaker results carry over to first-order formulations (Švedjar 2009, p. 89; Visser 2009).

For these reasons, many authors working within the disentanglement programme (Leigh and Nicolai 2013, Nicolai 2015, Heck 2015) speak as if their syntax theory is formulated in a disjoint 'copy' of the language of arithmetic. As Richard Kimberly Heck puts it, one can think of the language of the syntax theory as 'the language of arithmetic written in boldface' (2015, p. 451).

This manoeuvre is technically convenient: arithmetical theories have been extensively studied, and it is useful to be able to import results wholesale. We shall follow this way of speaking, but we interpret the syntactic domains of the theories we present as ranging over genuinely syntactic entities. Readers uncomfortable with interpreting (e.g.) PA as a one-symbol syntax theory are invited to view such theories as placeholders for ones with larger alphabets and a full complement of primitive syntactic operations: the results of Corcoran et al. (1974) and others guarantee that the entire argument could be carried out in that framework.

We consider the system introduced by Graham Leigh and Carlo Nicolai (2013), which we call $\mathrm{PA}_b + \mathrm{PA}_s^{CT}$.[4] It is a three-sorted theory; (i) entities of sort b are natural numbers; (ii) entities of sort s are syntactic objects; (iii) entities of sort m are 'mixed'—sequences of sort-b objects serving as assignments of values to variables (where variables are syntactic entities).[5]

The theory of sort-b objects is PA, which we call PA_b; its primitives are $0_b, S_b, +_b$, and $\times_b$. The theory of sort-s objects is again PA—we call this PA_s, with primitives $0_s, S_s, +_s$, and $\times_s$.[6] Sort-m objects are handled by introducing three additional primitive notions: $\alpha[i]$ returns the value of the ith variable, v_i, on the assignment α; $\mathrm{Den}_\alpha \mathrm{t}$ 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.

[4] We have modified Leigh and Nicolai's notation for consistency.

[5] We understand the syntactic component of Leigh and Nicolai's theory as referring to genuinely syntactic objects, not to a mere duplicate copy of the arithmetical objects. Little turns on this, however; on the alternative reading, similar arguments can be made through an additional layer of coding.

[6] Interpreted syntactically, 0_s denotes the null string, S_s represents adjunction of a single symbol, and $+_s$ represents concatenation. $\times_s$ is the string-expansion operation that stands to concatenation as multiplication stands to addition. Following Hilbert and Bernays, it can be seen as a 'kind of replacement' (Parsons 2008, p. 255) where each symbol in a string is replaced by a number of copies (corresponding to the length of the other multiplicand). It is admittedly a less natural syntactic operation than concatenation, but, as above, readers are invited to treat PA_s as a placeholder for an alternative syntax theory.

Leigh and Nicolai's system includes:

(I) All axioms of PA (for objects of sort b);

(II) All axioms of PA (for objects of sort s);

(II) Axiom for sequences:

(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) $\forall\alpha\forall i\ \mathrm{Den}_\alpha \underset{.}{\mathrm{v}}_i=\alpha[i]$,

(CTD_c) $\forall\alpha\forall i\ \mathrm{Den}_\alpha \underset{.}{\mathrm{c}}_i=c_i$ for each constant-symbol c,

(CTD_f) $\forall\alpha\forall \mathrm{t}_1\cdots\forall \mathrm{t}_n(\mathrm{Den}_\alpha \underset{.}{f}(\mathrm{t}_1,\ldots,\mathrm{t}_n)=f(\mathrm{Den}_\alpha \mathrm{t}_1,\ldots,$ $\mathrm{Den}_\alpha \mathrm{t}_n))$ for each function symbol f,

(CTD_{at}) $\forall\alpha\forall \mathrm{t}_1\cdots\forall \mathrm{t}_n(\mathrm{Sat}_\alpha \underset{.}{R}(\mathrm{t}_1,\ldots,\mathrm{t}_n)\leftrightarrow R(\mathrm{Den}_\alpha \mathrm{t}_1,\ldots,$ $\mathrm{Den}_\alpha \mathrm{t}_n))$ for each n-ary relation symbol R,

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

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

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

(V) Syntactic induction (schema):

(Ind_S) $\Phi(0_s)\wedge\forall k(\Phi(k)\rightarrow\Phi(S_s k))\rightarrow\forall k\Phi(k)$ where k is a variable of the syntax theory and Φ is any formula.

(VI) An axiom stating that the axioms of PA_b are true:

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

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

As Volker Halbach argues, however, 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 2014, p. 306)

Using a suggestion due to Jeffrey Ketland, Leigh and Nicolai implement Halbach's idea by adding 'coding axioms', employing a new primitive cross-type predicate C (for the relation connecting mathematical and syntactic objects):

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

(CodAx2) $\forall x \forall i(C(x, i) \rightarrow C(S_b x, S_s i))$;

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

Here x ranges over sort-b objects, i over sort-s objects. In essence, these axioms (collectively, CodAx) state that there exists an isomorphism between the syntactic objects and the mathematical objects. Leigh and Nicolai claim that once CodAx have been added, the resulting theory provides 'a satisfactory picture of our informal metatheoretic discussion as characterized in Halbach' (2013, p. 628).

We think this is at best partially correct: although $\mathrm{PA}_b + \mathrm{PA}_s^{CT} +$ CodAx captures some features of informal metamathematics, it fails to be coherently integrated: the coding axioms, underivable from $\mathrm{PA}_b + \mathrm{PA}_s$, 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.

2. Epistemically Stable Theories

As presented, 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?

Most existing work proceeds from a *technical* perspective, motivated by the desire to prove theorems as strong as possible using minimal resources. An example of this approach is Heck's result (2015, p. 457) that $I\Sigma_1^{CT}$ (Robinson arithmetic plus Σ_1 induction plus compositional truth) proves the syntactic consistency statement for finitely axiomatized base theories. The result is striking, but it is difficult to come up with an autonomous philosophical justification for accepting induction only for Σ_1 formulas.[7] In contrast, we adopt a *philosophical* perspective, concerned primarily with theories possessing an internally coherent motivation. To elucidate this notion further, we appeal to the idea of a *foundational equivalence thesis*. 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. A foundational equivalence thesis is a thesis to the effect that a foundational stance is extensionally equivalent to a given formal mathematical theory.[8]

Foundational equivalence theses allow us to characterize a class of systems that can be regarded as *epistemically stable*. We take this notion from Dean, who ascribes epistemic stability 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, p. 53).

'Acceptance', here, means *non-instrumental* acceptance as an interpreted theory about an intended class of objects. There are interesting further questions whether theories can be *instrumentally* justified by being 'reduced' in some sense (perhaps proof-theoretic reduction or interpretability) to an antecedently justified theory, as some contend (Feferman 2000, Hofweber 2000), and if so, whether this justifies *acceptance* (as opposed to some other, more instrumental attitude)

[7] Perhaps $I\Sigma_1$ can be justified on finitist grounds, in virtue of the proof-theoretic reduction to PRA effected by Parsons's Theorem (Parsons 1970). However, any such justification would be instrumental, in the sense discussed below. There is also the further hurdle of finitistically justifying the truth-theoretic component of $I\Sigma_1^{CT}$, since it attributes truth to sentences which (on their intended interpretation) quantify over infinite domains. Thanks to a referee for raising this issue.

[8] For further discussion of Foundational equivalence theses in a different context, see Waxman (MS). These are examples of the larger class of *informal equivalence theses*, such as the Church-Turing thesis—characteristic instances of Kreisel (1967)'s method of 'informal rigour'.

towards the reduced theory. Here, however, we shall not assume that if $\mathbf{T}_1$ is interpretable within or proof-theoretically reducible to $\mathbf{T}_2$, then a rationale for accepting $\mathbf{T}_2$ carries over to $\mathbf{T}_1$—particularly when $\mathbf{T}_1$ and $\mathbf{T}_2$ concern different domains of objects. In §4, we argue for a special case of such a connection where numbers and syntactic objects are concerned, but we do so on independent grounds, not as a result of a general claim about interpretability.

From a technical perspective, many theories, including some that are epistemically unstable, are worth studying. But from a philosophical perspective, there are compelling reasons to focus upon epistemically stable theories: at least *prima facie*, agents who accept an unstable theory while failing to accept some stable extension of it are irrational, since they fail to accept all that is justified by their underlying rationale.

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 foundational equivalence thesis it is reasonable to regard it as internally well-motivated.

Some examples of foundational equivalence theses 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 function. This informal conception is captured precisely by the formal system of second-order arithmetic (Z^2).[9]
>
> Isaacson's Thesis (Isaacson 1987; cf. Dean 2015, pp. 53–54): There is a distinction between the 'purely arithmetical' content of our conception of the natural numbers and 'higher-order' content that is revealed only from an extra-arithmetical vantage point (Isaacson 1987, p. 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 [$\mathscr{L}_{\mathrm{PA}}$] which is independent of PA then we will need to appeal to ideas

[9] There is an issue whether second-order theories should be conceived model-theoretically (with either a full or Henkin interpretation of the second-order quantifiers) or as deductive systems. It would be anachronistic to attribute any such distinction to Dedekind (who worked, more or less, in informal set theory). As we only use the deductive system, our claims are compatible with either approach.

that go beyond those that are required in understanding PA' (Smith 2008, p. 1).

> Tait's Thesis (Tait 1981; cf. Dean 2015, pp. 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 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 finitistically acceptable, 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 as such by the finitist.) Rather, like many foundational equivalence theses, the position is characterized externally. This illustrates a key point: it is not required for the truth of a foundational equivalence thesis that a proponent of the foundational stance in question be in a position to recognize it as true.[10]

> Feferman-Schütte Thesis (Kreisel 1960b; Feferman 1964; Schütte 1965a, 1965b): Predicativism 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 $\mathrm{RA}_{<\Gamma_0}$ 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 1960a, 1965; Feferman 1991), anyone who rationally endorses a theory **T** is thereby implicitly committed to extending it via reflection

[10] Plausibly neither finitists nor predicativists could even in principle recognize the truth of a foundational equivalence thesis characterizing their informal stances. As Hellman (2004, p. 299) puts it, any attempt to characterize predicativity 'implicitly transcends predicativity in the very formulation of the limitative theses', and arguably the same goes for finitism. However, it is not clear whether these considerations can be turned into a completely general argument to the effect that proponents of a foundational stance cannot accept a foundational equivalence thesis characterizing it.

principles—at a minimum, the local reflection principle $\mathrm{Rfn}(\mathbf{T}), \mathrm{Bew}_{\mathbf{T}}(\ulcorner\Phi\urcorner)\to\Phi$, stating that if Φ is provable in $\mathbf{T}$, then Φ.[11]

If so, a rational agent who accepts $\mathbf{T}$ ought to accept $\mathbf{T}+\mathrm{Rfn}(\mathbf{T}), \mathbf{T}+\mathrm{Rfn}(\mathbf{T})+\mathrm{Rfn}(\mathbf{T}+\mathrm{Rfn}(\mathbf{T}))$, and so on—each of these theories strictly stronger than the last. If this view is correct, there can be no such thing as an epistemically stable theory.

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, p. 40) has noted, in some cases adding reflection is equivalent to adding new instances of induction; thus, if an informal foundational stance motivates a theory with restricted induction, imposing reflection would be question-begging against it. As this response is not directly relevant to the theories we shall consider, we do not develop it further.

Second, as noted above, foundational equivalence theses ordinarily characterize positions *externally*. The sense in which $\mathrm{RA}_{<\Gamma_0}$ 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 proved within it. Thus, even if anyone who explicitly adopted $\mathrm{RA}_{<\Gamma_0}$ were obligated to accept additional claims, this obligation would not apply to the predicativist herself.

Third, there is still a useful notion of *stability modulo reflective closure* that classifies theories in a way that carves at the joints. The proponent of strong implicit commitment should accept that, even if there are no fully stable theories, at least some theories are privileged as epistemically appropriate starting points for reflection. Proponents of strong implicit commitment are thus invited to understand ‘epistemic stability’, in what follows, as ‘epistemic stability modulo reflective closure’.[12]

We conclude that, from the philosophical perspective, theories of foundational interest must be epistemically stable. If the foundational

[11] For this constraint to apply, the theory must be strong enough to formulate the reflection principle or the agent must accept an additional syntax theory capable of formulating it. Many versions of reflection principles have been discussed: another is the uniform reflection principle $\mathrm{RFN}(\mathbf{T}), \forall x(\mathrm{Bew}_{\mathbf{T}}(\ulcorner\Phi(\underset{\cdot}{x})\urcorner)\to\Phi(x))$. Other candidates for implicit commitments include the consistency statement $\mathrm{Con}(\mathbf{T})$ or the Gödel sentence $\mathrm{G}_{\mathbf{T}}$.

[12] As a referee points out, it is not entirely clear how to give a precise formal definition of ‘reflective closure’, since subtle issues concerning ordinal notations arise (Feferman 1962). For our purposes, all we need is the idea that some theories serve as starting points for reflection, however exactly this is cashed out.

equivalence theses above are correct, then—as mathematical theories taken in isolation—Z^2, PA, PRA, and $\mathrm{RA}_{<\Gamma_0}$ are internally coherent frameworks worthy of study.

However, taking the disentanglement programme seriously requires considering not just mathematical theories in isolation, but combinations of base and syntax theories. 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 contend that this possibility is realized; indeed, it affects the main disentangled theories proposed in the literature.

3. 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.

3.1 The Hilbert-Parsons Principle

We first consider *local joint instability*—the instability that 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 $\mathbf{T}_2$.

The primary source of local joint instability we will discuss arises from the links between arithmetic and syntax. The idea can be motivated 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 [...] The solid philosophical attitude that I think is required [...] is this: In the beginning was the sign. (Hilbert 1922, p. 163/202; cf. Hilbert and Bernays 1934–39, vol. I, p. 20)

Here Hilbert defends a very strong claim: the natural numbers *simply are* syntactic types. Few others have found this metaphysical thesis appealing. But it is worth asking why Hilbert found it attractive in the first place. A plausible 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. He focuses on a simple case of syntax, inspired by Hilbert, based on an alphabet with a stroke | as its sole symbol. First, we learn about string-types (which he describes (2008, p. 33) as 'quasi-concrete' objects: abstract but 'determined by their concrete embodiments') by literally perceiving them:

> [W]e stand in a perceptual relation to the expressions of this simple formal language. The same reasons for talking of perception of expression-types with reference to natural language arise also with reference to this language. (Parsons 2008, pp. 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 2008, p. 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. In more detail, the process works as follows. Suppose we have a relatively clear apprehension of the string-types of our language (whether explained in Parsons's fashion or otherwise). Although any realistic language will be based on an alphabet with many characters, we can restrict attention to the strings formed from repeated adjunction of a single arbitrarily chosen symbol (which will play the same role as

Hilbert's stroke |). We can apprehend that there is a map from string-types of this kind to the natural numbers (taking the null string to zero, adjunction by the given symbol to successorship, and so on). Because we are in a position to see that this map is structure-preserving, we can learn about the structure of the natural numbers by considering the structure of string-types. Furthermore, we can semantically ascend: the map induces a translation between sentences in the single-symbol fragment of the language of syntax and (at least some) sentences in the language of arithmetic. We know that a translatable arithmetical sentence is true if and only if its syntactic translation is, enabling us to *reconceptualize* arithmetical claims as true of syntactic objects.

This yields what we term the Hilbert–Parsons Principle:

> 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 reconceptualized as true of strings, by way of a natural mapping between (a subclass of) syntactic and arithmetical objects.

3.2 From HPP to Interpretability

In HPP, 'reconceptualizing as true' is an informal notion, meaning merely that we can see that certain arithmetical claims are true because they can be reconstrued as true claims about syntactic objects. But this informal thesis motivates a *formal* proof-theoretic claim—that syntax should be *relatively interpretable* within arithmetic—as follows.[13]

Suppose 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 $\mathbf{T}'_s$ of $\mathbf{T}_s$ (the theory of the null string and finite sequences of some arbitrary symbol, with concatenation and perhaps other basic syntactic operations) which we can reinterpret, along the lines set out above, through a mapping $\mu : S \rightarrow A$ taking the null string to zero and so on. Because we have a definable arithmetical operation corresponding to every primitive syntactic operation on the one-symbol fragment, there is a function † from formulas of $\mathscr{L}_{\mathbf{T}'_s}$ to formulas of $\mathscr{L}_{\mathbf{T}_b}$. (This function need not

[13] 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, pp. 96–98).

be assumed to be surjective: for all we know, there may be arithmetical primitives without definable syntactic correlates.)

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 the following:

> (**Interpretability Constraint**) If $\mathbf{T}_b + \mathbf{T}_s$ is jointly epistemically stable, then the one-symbol fragment $\mathbf{T}'_s$ of $\mathbf{T}_s$ is relatively interpretable in $\mathbf{T}_b$.[14]

3.3 The Interpretability Constraint and Local Joint Instability

The Interpretability Constraint affects some theories discussed in the literature. As noted, Heck focusses on $[\mathrm{I}\Sigma_1]_s^{CT}$, i.e. $\mathrm{I}\Sigma_1$ as a syntax theory with compositional truth axioms, which they think can be added to an arbitrary base theory to yield an appealing theory of syntax and truth built upon it. However, we doubt that this invariably results in a stable theory. Notice that if $\mathbf{T}_b + [\mathrm{I}\Sigma_1]_s^{CT}$ is stable, presumably so too must be the truth-free theory $\mathbf{T}_b + [\mathrm{I}\Sigma_1]_s$. But, at least in some cases, this will run afoul of the Interpretability Constraint. Suppose $[\mathrm{I}\Sigma_1]_s$ is added to a relatively weak base theory incapable of interpreting it, such as Q_b or $[\mathrm{I}\Delta_0]_b$ (Hájek and Pudlak 1993, p. 391). Given the Interpretability Constraint, this results in an unstable combination.

[14] The Interpretability Constraint should be distinguished from:

> (**Reverse Interpretability Constraint**) If $\mathbf{T}_b + \mathbf{T}_s$ is jointly epistemically stable, then $\mathbf{T}_b$ is relatively interpretable in the one-symbol fragment $\mathbf{T}'_s$ of $\mathbf{T}_s$.

We remain neutral about the Reverse Interpretability Constraint here. Since, as mentioned above, † need not be surjective, it is not forced upon us by the HPP. There are weak systems of syntax—corresponding to Samuel Buss's S^1_2 (Buss 1986) or to PRA—which can perhaps be given independent combinatorial motivations. (See Nicolai 2016, pp. 97–103 and 114–17, for relevant results and discussion.) Since the HPP need not represent our only mode of insight into the structure of the natural numbers, perhaps such a system could be stably combined with a stronger arithmetical theory that had a different kind of justification. Thanks to a referee for pressing us to clarify here.

We now turn to the theories considered by Leigh and Nicolai. It is plausible that $\mathrm{PA}_b + \mathrm{PA}_s$ is jointly stable: 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 the Interpretability Constraint reveals no impediment to their joint stability.

As appealing as $\mathrm{PA}_b + \mathrm{PA}_s$ may be, it is inadequate to formalize informal metamathematical practice, since it contains no truth predicate, whereas informal metamathematics considers principles (e.g. global reflection) essentially involving the notion of truth. We are thus led to consider the theory's natural truth-theoretic extension, $\mathrm{PA}_b + \mathrm{PA}_s^{CT}$, resulting from the addition of compositional truth axioms.

The problem, however, is that adding compositional truth yields a failure of the Interpretability Constraint. Since $\mathrm{PA}_b + \mathrm{PA}_s^{CT}$ proves $\mathrm{Con}_s(\mathrm{PA}_b)$ but does not prove $\mathrm{Con}_b(\mathrm{PA}_b)$, the PA_b fragment on its own cannot interpret the restriction of PA_s^{CT} to the syntax language. Thus, the Interpretability Constraint diagnoses the theory as unstable.[15]

3.4 The Coding Axioms and Structural Joint Instability

Finally, we consider Leigh and Nicolai's preferred theory $\mathrm{PA}_b + \mathrm{PA}_s^{CT} + \mathrm{CodAx}$. 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 the kind of local joint instability addressed by the Interpretability Constraint is not the only possible cause of joint instability. We will argue that Leigh and Nicolai's treatment of the coding axioms manifests what we will call *structural joint instability*. Abstractly characterized, this arises for a pair of theories $\mathbf{T}_1$ and $\mathbf{T}_2$ when the only foundational stances which motivate $\mathbf{T}_1 + \mathbf{T}_2$ themselves motivate some $(\mathbf{T}_1 + \mathbf{T}_2)'$ strictly stronger than $\mathbf{T}_1 + \mathbf{T}_2$.

We believe this kind of instability afflicts Leigh and Nicolai's theory $\mathrm{PA}_b + \mathrm{PA}_s^{CT} + \mathrm{CodAx}$. On our view, any informal stance motivating

[15] Notice that the objection just made is distinct from the one raised by Halbach (2014, p. 306) and discussed in §2. Our objection is not that $\mathrm{PA}_b + \mathrm{PA}_s^{CT}$ is unnatural, or that it is unfaithful as a codification of our metamathematical practice; rather, our objection is that its syntactic component outstrips its arithmetical component, and thus, given the Interpretability Constraint, it is an epistemically unstable combination.

CodAx would also motivate theories stronger than $\mathrm{PA}_b + \mathrm{PA}_s^{CT}$. In particular, the only reasonable candidates for stances motivating CodAx are higher-order in Isaacson's sense; but any such stance itself motivates stronger theories than PA_b and PA_s.

To see this, consider how the coding axioms might be philosophically motivated. We see three *prima facie* possibilities: (i) on their own, independently of $\mathrm{PA}_b + \mathrm{PA}_s^{CT}$; (ii) on the basis of the conception of truth motivating PA_s^{CT}; or (iii) on the basis of 'higher-order' facts about the natural number structure (in Isaacson's sense) exceeding what is captured by the first-order theories PA_b and PA_s.

We argue that only (iii) is plausible. It would be somewhat desperate 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 a brute fact. Nor can truth-theoretic considerations suffice. As already pointed out, such considerations can motivate $\mathrm{PA}_b + \mathrm{PA}_s^{CT}$ itself, but there is no way to use them to extend that theory. The addition of CodAx to $\mathrm{PA}_b + \mathrm{PA}_s^{CT}$ 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 1987, pp. 158–59)

Thus, assuming Isaacson's Thesis and its syntactic analogue, the coding axioms are alien to the conception embodied by $\mathrm{PA}_b + \mathrm{PA}_s$ or even $\mathrm{PA}_b + \mathrm{PA}_s^{CT}$. Their justification relies not just on a 'local' appreciation of the individual structure of arithmetical or syntactic objects, but on the 'link between them'—the fact that the two domains are isomorphic. But reasoning about isomorphisms in general requires resources that transcend PA, by allowing us to talk about structures rather than just individual numbers. A theory capable of this will involve second-order quantification, some degree of set theory, or something else 'higher-order' in Isaacson's sense—and will thus transcend PA_b and PA_s. Therefore, the stance behind CodAx leads beyond the theories to which Leigh and Nicolai join it—rendering $\mathrm{PA}_b + \mathrm{PA}_s^{CT}$ + CodAx subject to structural joint instability.

So we are led to ask: which conceptions of the natural numbers and 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 a fuller conception of the natural numbers is captured by Z^2, provides a natural answer. Again, there is an obvious syntactic analogue of Dedekind's Thesis leading to a parallel claim for the syntactic domain.

We thus consider the theory that results from taking Z^2 as both the base theory and the syntax theory (interpreted as a theory of the natural numbers in the former case, and a one-symbol syntax theory in the latter). We call this theory DZ^2, for 'double' Z^2. Unlike $\mathrm{PA}_b + \mathrm{PA}_s^{CT}$ + CodAx, it and its truth-theoretical extension DZ^{2CT} are epistemically stable. Moreover, they have the resources to *demonstrate* isomorphism between the two domains with no need for additional assumptions; the coding axioms are directly justified and need not be added in by hand.

4. Double Second-Order Arithmetic

4.1 Dedekind's Thesis and DZ^2

We now set out DZ^2. We assume a standard second-order deductive system such as in Shapiro (1991, pp. 65–69), with each instance of the comprehension schema:

$$\text{(CA)} \quad \exists X^n \forall x_1 \ldots x_n (X^n x_1 \cdots x_n \leftrightarrow \Phi) \text{ where } X^n \text{ is not free in } \Phi.$$

We allow function constants but no quantification over functions. The signature of DZ^2 is $\{N_b, N_s, 0_b, 0_s, S_b, S_s\}$. Its proper axioms are, for $\xi \in \{b, s\}$:

$$(\mathrm{A1}_\xi) \quad N_\xi 0_\xi$$

$$(\mathrm{A2}_\xi) \quad \forall x (N_\xi x \rightarrow N_\xi (S_\xi x))$$

$$(\mathrm{A3}_\xi) \quad \forall x (N_\xi x \rightarrow S_\xi x \neq 0_\xi)$$

(A4$_\xi$) $\forall x \forall y (N_\xi x \wedge N_\xi y \rightarrow (S_\xi x = S_\xi y \rightarrow x = y))$

(A5$_\xi$) $\forall X (X 0_\xi \wedge \forall x (N_\xi x \rightarrow (Xx \rightarrow X S_\xi x)) \rightarrow \forall x (N_\xi x \rightarrow Xx))$

(A6) $\forall x \forall y (N_b x \wedge N_s y \rightarrow x \neq y)$

On the intended interpretation, $0_b, S_b 0_b, S_b S_b 0_b, \ldots$ denote the natural numbers 0, 1, 2, and so on; $0_s, S_s 0_s, S_s S_s 0_s, \ldots$ denote the null string, the string |, the string ||, and so on, where | is some fixed symbol.

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

We write $\mathscr{L}_{\mathrm{Z}^2_b}$ and $\mathscr{L}_{\mathrm{Z}^2_s}$ for the languages with signatures $\{N_b, 0_b, S_b\}$ and $\{N_s, 0_s, S_s\}$ respectively, and we write Z^2_b and Z^2_s for the fragments of DZ^2 in $\mathscr{L}_{\mathrm{Z}^2_b}$ and $\mathscr{L}_{\mathrm{Z}^2_s}$. 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:[16]

Theorem (Transfer). $\mathrm{DZ}^2 \vdash \Phi^{(b)} \leftrightarrow \Phi^{(s)}$ where Φ is a sentence of $\mathscr{L}_{\mathrm{DZ}^2}$.

We will use this result to argue that DZ^2 is a jointly stable theory. We 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?

Given Dedekind's Thesis and its syntactic analogue, the individual components of DZ^2 are stable. The two kinds of joint instability are more interesting.

The primary examples of local joint instability discussed so far have 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 natural mapping.

It is hard to see how joint instability could arise on other grounds. The Transfer Theorem also provides an interpretation in the other direction, and the equivalence it guarantees (for sentences in $\mathscr{L}_{\mathrm{Z}^2_{\mathrm{b}}}$ and

[16] For a proof, see the appendix.

$\mathscr{L}_{\mathrm{Z}^2_{\mathrm{s}}}$) will still hold if we add new axioms: the resulting system will never prove a sentence in $\mathscr{L}_{\mathrm{Z}^2_{\mathrm{b}}}$ without proving the corresponding sentence in $\mathscr{L}_{\mathrm{Z}^2_{\mathrm{s}}}$ and vice versa.

Turning to structural joint instability, the key example we considered was the addition of a truth theory and coding axioms to Leigh and Nicolai's $\mathrm{PA}_b + \mathrm{PA}_s$. 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 might 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.

To show that no such issue arises here, we introduce a truth theory, DZ^{2CT}, extending DZ^2 with compositional truth clauses for a second-order language. We demonstrate (in the appendix) that it is capable of *proving* the syntactic consistency statement for the base theory and thus, by the Transfer Theorem, 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—which require motivation going 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}. First fix a coding $\ulcorner\urcorner$ from the expressions of $\mathscr{L}_{\mathrm{Z}^2_b}$ to the syntactic objects. 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 taking it to a proxy monadic 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 $\langle g_1, g_2, \ldots, f^1_1, f^1_2, \ldots, f^2_1, f^2_2, \ldots\rangle$ as proxy for the values on an assignment of the variables $x_1, x_2, \ldots, X^1_1, X^1_2, \ldots, X^2_1, X^2_2, \ldots$ (which are in turn represented in our syntax theory by terms we write as $\mathrm{v}_1, \mathrm{v}_2, \ldots, \mathrm{V}^1_1, \mathrm{V}^1_2, \ldots, \mathrm{V}^2_1, \mathrm{V}^2_2, \ldots$).

We can use additional proxy functions to simulate a countable collection of subsets of an infinite domain by a single subset. In this way we can use a monadic second-order entity as a proxy variable assignment, from which the value of any given variable can be extracted (by a family of definable functions). We use α, β to range over assignments (in this proxy sense of 'assignment'), and we write $[\![\mathrm{v}]\!]_\alpha$ (resp. $[\![\mathrm{V}]\!]_\alpha$) for the value of a given first- or second-order variable on α. We write $\alpha \overset{\mathrm{v}}{\sim} \beta$ to indicate that α differs from β only in the value assigned to $\mathrm{v}, \alpha^{[[\![\mathrm{v}]\!]:=x]}$ for the assignment differing from α only by assigning x to v, and $\alpha^{[\mathrm{v}_1/\mathrm{v}_2]}$ 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 terms of the language.

Finally, we introduce a new primitive cross-type predicate Sat which takes a monadic second-order entity and a syntactic entity: on the intended interpretation, $\mathrm{Sat}_\alpha\phi$ holds if and only if ϕ, the coded syntactic object, denotes a formula of the base language satisfied on assignment α.[17]

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

(T1) $\forall\alpha\forall\mathrm{t}_1\forall\mathrm{t}_2(\mathrm{Sat}_\alpha(\mathrm{t}_1 \underset{\cdot}{=} \mathrm{t}_2) \leftrightarrow [\![\mathrm{t}_1]\!]_\alpha = [\![\mathrm{t}_2]\!]_\alpha)$

(T2) $\forall\alpha\forall\mathrm{t}(\mathrm{Sat}_\alpha \underset{\cdot}{N}_b\mathrm{t} \leftrightarrow N_b[\![\mathrm{t}]\!]_\alpha)$

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

(T4) $\forall\alpha\forall\phi(\mathrm{Sat}_\alpha \underset{\cdot}{\neg}\, \phi \leftrightarrow \neg\mathrm{Sat}_\alpha\phi)$

(T5) $\forall\alpha\forall\phi\forall\psi(\mathrm{Sat}_\alpha(\phi \underset{\cdot}{\wedge} \psi) \leftrightarrow \mathrm{Sat}_\alpha\phi \wedge \mathrm{Sat}_\alpha\psi)$

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

(T7) $\forall\alpha\forall\phi\forall\mathrm{V}^n(\mathrm{Sat}_\alpha \underset{\cdot}{\forall}\, \mathrm{V}^n\phi \leftrightarrow (\forall\beta \overset{\mathrm{V}^n}{\sim} \alpha)\mathrm{Sat}_\beta\phi)$

Conventions for the use of $\forall\mathrm{t}$ and similar expressions correspond to the obvious disentangled analogues of those introduced for PA^{CT}. Note that (T3) and (T7) are schematic in n, as is appropriate for a polyadic theory.[18]

The main result governing DZ^{2CT} is that it proves the syntactic consistency sentence for its base theory:

[17] The $[\![]\!]_\alpha$ expression is shorthand for a family of formulas picking out entities of various types, but in practice this will cause no confusion. To avoid notational clutter, we treat Sat as incorporating a tacit application operation $\mathrm{T}, \mathrm{t}_1, \ldots, \mathrm{t}_n \mapsto \ulcorner \mathrm{T}\mathrm{t}_1 \cdots \mathrm{t}_n \urcorner$ in the atomic case.

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

Theorem (Non-Conservativeness). $\mathrm{DZ}^{2CT} \vdash \mathrm{Con}_s(\mathrm{Z}^2_b)$.

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

For these reasons, DZ^2 and its truth-theoretic extension DZ^{2CT} are philosophically appealing theories within which to carry out the disentanglement programme. As the Non-Conservativeness Theorem demonstrates, they are sufficient to capture much informal metamathematical reasoning, with their syntax and base theories naturally moving in step. As we have argued, the individual components of these theories are well-motivated; furthermore, the results just mentioned show that, unlike their main disentangled competitors, no obvious source of joint instability arises from their interaction.

5. Concluding Discussion

In this final section we consider some objections and some potential implications of our results for Dedekind's Thesis, Isaacson's Thesis, and deflationism.

First, 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, according to the objection, it is problematic to suppose that we have a determinate conception of second-order quantification, since the determinacy of the powerset operation is dubious. Of course, set-theoretic realists will find an appeal to the powerset of $\mathbb{N}$ 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 within a higher-order metalanguage throughout. 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. They are thus valid not only on all standard interpretations of the second-order quantifiers but also on all Henkin interpretations (where the quantifiers range over some, possibly proper, subset of the full powerset of the domain); unlike the standard semantics, there exist sound and complete proof procedures for Henkin semantics. Any remaining scruples about second-order logic can be assuaged by viewing the second-order theories as, in effect, two-sorted first-order theories, along the lines of Simpson (2009).

In addition, there is 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 $\mathrm{Con}_b(\mathrm{Z}^2_b)$ is provable? After all, $\mathrm{Con}_b(\mathrm{PA}_b)$ becomes provable in Leigh and Nicolai's system (2013, p. 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. (CA) is equivalent to a rule of substitution, roughly according to which arbitrary formulas can be substituted for free second-order variables in demonstrated claims.[19] 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, p. 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 $\mathrm{PA}_b + \mathrm{PA}_s$. 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. It is also a mark in favour of Dedekind's Thesis that the coding axioms are derivable within

[19] See Boolos (1985, pp. 334–38) for a precise statement of the rule of substitution and proof of the equivalence.

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 a natural conception of natural numbers and syntax.

Let us close by briefly discussing the impact of our discussion on deflationism. To recall the main issue: deflationists are, allegedly, committed to the conservativeness of truth and syntax over mathematical base theories. Given that it is philosophically natural to move to disentangled theories, it is interesting to consider the implications for deflationism. At first, $\mathrm{PA}_b + \mathrm{PA}_s^{CT}$ might appear to be an appealing disentangled theory to adopt, since it is conservative over PA_b. However, Leigh and Nicolai argue, for reasons discussed in §2, that the theory is not so appealing after all, because it 'cannot capture our common metatheoretic practice' (2013, p. 635). In its place, they tentatively propose $\mathrm{PA}_b + \mathrm{PA}_s^{CT} + \mathrm{CodAx}$, which they think better captures metatheoretic practice. However, this theory is non-conservative over PA_b, and thus once again inhospitable to the deflationist. Our discussion takes the dialectic one stage further: as we argued in §4.5, $\mathrm{PA}_b + \mathrm{PA}_s^{CT} +$ CodAx exhibits structural joint instability, and for that reason should not be accepted by the deflationist (or, for that matter, anyone else).

In our discussion, we introduced DZ^{2CT}, which we advocate as an epistemically stable alternative. But this theory is not friendly to the deflationist either. For as we showed in §4, DZ^{2CT} is a non-conservative extension over the truth-free base theory. This result puts some pressure on deflationists: to the extent that a conservativeness constraint holds, it is incumbent upon them to propose an alternative disentangled truth theory, suitable for metamathematical reasoning, which is both epistemically stable and conservative over its base theory.

Whether this is a genuine problem for deflationists depends on a final set of issues, which we merely gesture at for further investigation. The conservativeness constraint to which deflationism is putatively committed 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, theories of arithmetical truth would be irrelevant, 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.[20] Whether this provides a means of escape for the deflationist, however, is a question that we cannot pursue here.[21]

Appendix

1. Proof of the Transfer Theorem

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

The Transfer Theorem generalizes Leigh and Nicolai’s (2013, p. 631) Corollary 3.17.[22] An immediate implication, appealed to in §4, is that $\mathrm{DZ}^2 \vdash \mathrm{Con}_b(\mathrm{Z}^2_b) \leftrightarrow \mathrm{Con}_s(\mathrm{Z}^2_b)$.

In order to prove the Transfer Theorem, we draw on a result from Väänänen and Wang (2015, p. 124). Second-order arithmetic is *internally categorical*, i.e. it proves that any two Z^2-structures are

[20] See Fujimoto (2019, pp. 1060–68) for recent work in this direction.

[21] Thanks to Kentaro Fujimoto, Carlo Nicolai, Lavinia Picollo, Oliver Tatton-Brown, Jared Warren, Philip Welch, Jack Woods, Jiji Zhang, audiences at New College, Oxford and the Humboldt-Universität Berlin, two anonymous referees for *Mind*, and, above all, Volker Halbach.

[22] In our notation, Leigh and Nicolai establish that $\mathrm{PA}_b + \mathrm{PA}_s + \mathrm{CodAx} \vdash \Phi^{(b)} \leftrightarrow \Phi^{(s)}$ whenever Φ is a first-order sentence in the language $\mathscr{L}_{\mathrm{PA}_b+\mathrm{PA}_s}$. 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, $\mathrm{Con}_b(\mathrm{Z}^2_b)$ 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.

isomorphic, as are any Z^2_b-structure and any Z^2_s-structure (since the axioms of Z^2_b and Z^2_s differ only by a relabelling of constants):[23]

Theorem (Internal Categoricity) (Väänänen and Wang). $DZ^2 \vdash \exists f \; f : \langle N_s, 0_s, S_s \rangle \underset{\text{iso}}{\to} \langle N_b, 0_b, S_b \rangle$.

We will appeal to the fact that, in the course of their proof, Väänänen and Wang provide a recipe for defining a function h which provably satisfies the analogues of the coding axioms (CodAx1)–(CodAx3):

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

(IA2) $\forall x \forall y (h(x) = y \to h(S_s(x)) = S_b(h(x)))$.

No analogue of (CodAx3) is required, for functionality is built in by definition.[24]

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

We turn now to the proof of the Transfer Theorem:

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

Proof. We first establish the more general schematic claim:

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

for Φ a formula of $\mathscr{L}_{DZ^2}$. We abbreviate this as $DZ^2 \vdash \Xi(\Phi^{(b)}(\vec{X}, \vec{x}), \Phi^{(s)}(\vec{Y}, \vec{y}))$ or simply $DZ^2 \vdash \Xi(\Phi)$.

We show this by a metatheoretic induction on the complexity of Φ, adapting the proof strategy of Leigh and Nicolai's (2013, p. 631) Theorem 3.16. In the base case, where Φ is atomic, $DZ^2 \vdash \Xi(\Phi)$ follows directly from (IA1)–(IA2) and the definition of the relabelling. The

[23] 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, pp. 223–50). In fact, the proof can be carried out in the much weaker system WKL_0; see Simpson and Yokoyama (2013). We do not think that this detracts from the interest of Z^2 and DZ^2: in our view, weak frameworks such as WKL_0 lack convincing philosophical motivations and face serious internal stability problems.

[24] All the proof requires is that h be defined on $\mathbb{N}_s$; we take the function to be total, with arbitrary values elsewhere.

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 $\mathrm{DZ}^2 \vdash \Xi(\Psi^{(b)}(\vec{X}, \vec{x}, z), \Psi^{(s)}(\vec{Y}, \vec{y}, z))$. 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 $\mathrm{DZ}^2 \vdash \Xi(\forall z \Psi^{(b)}(\vec{X}, \vec{x}), \forall z \Psi^{(s)}(\vec{Y}, \vec{y}))$, i.e. $\mathrm{DZ}^2 \vdash \Xi(\Phi)$. A precisely analogous argument applies when $\Phi = \forall Z^n \Psi(Z^n)$. □

2. Proof of the Non-Conservativeness Theorem

Here we show that $\mathrm{DZ}^{2CT} \vdash \mathrm{Con}_b(\mathrm{Z}_b^{\pm 2})$. It follows, via results in §4, that DZ^{2CT} also proves $\mathrm{Con}_s(\mathrm{Z}_b^2), \mathrm{Con}_b(\mathrm{Z}_s^2)$, and $\mathrm{Con}_s(\mathrm{Z}_s^2)$. 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.

It suffices to consider ordinary second-order arithmetic with the (standard, entangled) compositional truth axioms (Z^{2CT}) and the corresponding monadic second-order system, $\widehat{\mathrm{Z}}^{2CT}$. The use of $\widehat{\mathrm{Z}}^{2CT}$ simplifies the proofs, but no generality is lost, since polyadic second-order quantification can be coded via a pairing function. Because this coding is primitive recursive, $\mathrm{Con}(\mathrm{Z}^2) \leftrightarrow \mathrm{Con}(\widehat{\mathrm{Z}}^2)$ will be provable in a very weak base theory. So if $\widehat{\mathrm{Z}}^{2CT} \vdash \mathrm{Con}(\widehat{\mathrm{Z}}^2)$ then $\widehat{\mathrm{Z}}^{2CT} \vdash \mathrm{Con}(\mathrm{Z}^2)$. But since $\widehat{\mathrm{Z}}^{2CT}$ is a subtheory of $\mathrm{Z}^{2CT}, \mathrm{Z}^{2CT} \vdash \mathrm{Con}(\mathrm{Z}^2)$ if $\widehat{\mathrm{Z}}^{2CT} \vdash \mathrm{Con}(\widehat{\mathrm{Z}}^2)$. Furthermore, since Z_b^2 can be relatively interpreted in $\mathrm{Z}^2, \widehat{\mathrm{Z}}^{2CT} \vdash \mathrm{Con}(\widehat{\mathrm{Z}}^2)$ only if $\mathrm{DZ}^{2CT} \vdash \mathrm{Con}_b(\mathrm{Z}_b^2)$.

$\widehat{\mathrm{Z}}^{2CT}$ comprises base axioms:

(A_M1) $\forall x(Sx \neq 0)$

(A_M2) $\forall x \forall y(Sx = Sy \to x = y)$

(A_M3) $\forall X(X0 \wedge \forall x(Xx \to XSx) \to \forall x Xx)$

together with truth-theoretic axioms:

$$(\mathrm{T}_{M}1)\ \ \forall\alpha\forall \mathrm{t}_1\forall \mathrm{t}_2(\mathrm{Sat}_\alpha(\mathrm{t}_1 \underset{\cdot}{=} \mathrm{t}_2) \leftrightarrow [\![\mathrm{t}_1]\!]_\alpha = [\![\mathrm{t}_2]\!]_\alpha)$$

$$(\mathrm{T}_{M}2)\ \ \forall\alpha\forall \mathrm{V}\forall \mathrm{t}(\mathrm{Sat}_\alpha \mathrm{Vt} \leftrightarrow [\![\mathrm{V}]\!]_\alpha[\![\mathrm{t}]\!]_\alpha)$$

$$(\mathrm{T}_{M}3)\ \ \forall\alpha\forall\phi(\mathrm{Sat}_\alpha \underset{\cdot}{\neg}\, \phi \leftrightarrow \neg\mathrm{Sat}_\alpha\phi)$$

$$(\mathrm{T}_{M}4)\ \ \forall\alpha\forall\phi\forall\psi(\mathrm{Sat}_\alpha(\phi \underset{\cdot}{\wedge} \psi) \leftrightarrow \mathrm{Sat}_\alpha\phi \wedge \mathrm{Sat}_\alpha\psi)$$

$$(\mathrm{T}_{M}5)\ \ \forall\alpha\forall\phi\forall \mathrm{v}(\mathrm{Sat}_\alpha \underset{\cdot}{\forall}\, \mathrm{v}\phi \leftrightarrow (\forall\beta \overset{\mathrm{v}}{\sim} \alpha)\mathrm{Sat}_\beta\phi)$$

$$(\mathrm{T}_{M}6)\ \ \forall\alpha\forall\phi\forall \mathrm{V}(\mathrm{Sat}_\alpha \underset{\cdot}{\forall}\, \mathrm{V}\phi \leftrightarrow (\forall\beta \overset{\mathrm{V}}{\sim} \alpha)\mathrm{Sat}_\beta\phi)$$

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 $\widehat{\mathrm{Z}}^{2CT}$ proves the global reflection principle for $\widehat{\mathrm{Z}}^{2}$:

Theorem 1. $\widehat{Z}^{2CT} \vdash \forall\phi(\mathrm{Bew}_{\widehat{\mathrm{Z}}^2}\phi \rightarrow T\phi)$.

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. To that end we prove the following six claims.

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

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

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

(Sem4) $\widehat{\mathrm{Z}}^{2CT} \vdash \forall\phi\forall\psi((T(\phi \underset{\cdot}{\rightarrow} \psi) \wedge T\phi) \rightarrow T\psi)$,

(Sem5) $\widehat{\mathrm{Z}}^{2CT} \vdash \forall\phi\forall\psi\forall \mathrm{v}((T(\phi \underset{\cdot}{\rightarrow} \psi) \wedge \neg\mathrm{Free}(\mathrm{v}, \phi)) \rightarrow T(\phi \underset{\cdot}{\rightarrow} \underset{\cdot}{\forall}\, \mathrm{v}\psi))$,

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

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 (A_M1)–(A_M3), and CompAx expresses the property of being a code of an instance of (CA). (Sem^4) formalizes the truth-preservingness of modus ponens; similarly (Sem^5) and (Sem^6) for the rules of inference governing the quantifiers.

From (Sem1)–(Sem6), the required reflection principle will follow, and so too will $\text{Con}_b(Z_b^2)$.

A number of additional lemmata will be of use.

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

Lemma 3. (Closure) *Let* $\text{ucl}(\ulcorner\Phi\urcorner)$ *be the code of* Φ*'s universal closure.* $\widehat{Z}^{2CT} \vdash \forall\phi(T\phi \leftrightarrow T\text{ucl}(\phi))$.

Lemma 4. (Substitution of Provable Equivalents) $\widehat{Z}^{2CT} \vdash \forall\phi\forall\psi\forall\chi$ $((T\phi \leftrightarrow T\psi)\rightarrow(T\chi \leftrightarrow T\chi\psi/\phi)$.

Lemma 5. (Variable-Swapping) *Let* $\tilde{\alpha}$ *be* $\alpha\,[\text{V}_{a_1},\ldots,\text{V}_{a_m},\text{v}_{b_1},\ldots,\text{v}_{b_n}/\text{V}_{c_1},\ldots,\text{V}_{c_m},\text{v}_{d_1},\ldots,\text{v}_{d_n}]$. *Then* $\widehat{Z}^{2CT} \vdash \forall\phi\forall\alpha$ $(\text{Sat}_\alpha\phi \leftrightarrow \text{Sat}_{\tilde{\alpha}}\phi\text{V}_{a_1},\ldots,\text{V}_{a_m},\text{v}_{b_1},\ldots,\text{v}_{b_n}/\text{V}_{c_1},\ldots,\text{V}_{c_m},\text{v}_{d_1},\ldots,\text{v}_{d_n})$.

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 (T_M3), (T_M4), 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, they agree on Φ:

Lemma 6.

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

Proof. By an internal induction on complexity of formulas in $\widehat{Z}^{2CT}$. We write $\alpha\overset{\text{fv}}{\sim}\beta$ if α and β agree on all free variables in the relevant formula. The base case is clear; it relies only on (T_M1), (T_M2), and the definition of $[\![]\!]$. For the induction step, suppose $\forall\psi\forall\alpha\forall\beta(\alpha\overset{\text{fv}}{\sim}\beta\rightarrow(\text{Sat}_\alpha\psi \leftrightarrow \text{Sat}_\beta\psi)$ for all ψ of complexity $< n$ and that ϕ has complexity *n*. The only difficult cases are the quantifiers. We show the $\forall$v case; the $\forall$V case is similar.

We reason informally in $\widehat{Z}^{2CT}$. First, if v is not free in ψ, then $\forall v\psi$ and ψ are equivalent, and so we are done. Thus assume v is free in ψ and suppose $\alpha \overset{fv}{\sim} \beta$ and $\mathrm{Sat}_\alpha \underset{\cdot}{\forall}\, v\psi$. Then for all $\alpha' \overset{v}{\sim} \alpha, \mathrm{Sat}_{\alpha'}\psi$. Now suppose $\beta' \overset{v}{\sim} \beta$. Then there is some $\alpha' \overset{v}{\sim} \alpha$ such that $\beta' \overset{fv}{\sim} \alpha'$. 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 $\beta' \overset{v}{\sim} \beta, \mathrm{Sat}_{\beta'}\psi$. But then $\mathrm{Sat}_\beta \underset{\cdot}{\forall}\, v\psi$. □

Two immediate consequences of Lemma 6 are:

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

and

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

From these two claims, (Sem5) and (Sem6) follow by ($T_M\pm5$) and (T_M6).

To show (Sem1) requires a formalized induction.

Lemma 7. $\widehat{Z}^{2CT} \vdash \forall\phi(\mathrm{LogAx}\phi \to T\phi)$.

Proof. The various cases are fairly straightforward; as an example, we show

$$\forall\psi\forall v\forall t(\mathrm{FreeFor}(\psi, t, v) \to T(\underset{\cdot}{\forall}\, v\phi \underset{\cdot}{\to} \phi\frac{t}{v})).$$

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

We turn to (Sem2). In the setting with only monadic second-order quantification, the proper axioms of $\widehat{Z}^2$ are just (A_M1)–(A_M3).

Either *x* is free in (A_M3) 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 8. $\widehat{Z}^{2CT} \vdash \forall\phi(\mathrm{CompAx}\phi \to T\phi)$.

Proof. As an instance of (CA), we have

$$\widehat{Z}^{2CT} \vdash \exists X \forall x(Xx \leftrightarrow \mathrm{Sat}_{\alpha[\llbracket \mathrm{v} \rrbracket := x]}\psi)$$

with both ψ and α free. Generalizing yields

$$\widehat{Z}^{2CT} \vdash \forall\psi\forall\alpha\exists X\forall x(Xx \leftrightarrow \mathrm{Sat}_{\alpha[\llbracket \mathrm{v} \rrbracket := x]}\psi).$$

From this and the second corollary to Lemma 6, we have

$$\widehat{Z}^{2CT} \vdash \forall\psi\forall\alpha[\neg\mathrm{Free}(\mathrm{V},\psi) \to \exists X\forall x(Xx \leftrightarrow \mathrm{Sat}_{\alpha[\llbracket \mathrm{V} \rrbracket := X, \llbracket \mathrm{v} \rrbracket := x]}\psi)].$$

For brevity, define Sat^*_α as $\mathrm{Sat}_{\alpha[\llbracket \mathrm{V} \rrbracket := X, \llbracket \mathrm{v} \rrbracket := x]}$. Now, Xx is provably equivalent to $\mathrm{Sat}^*_\alpha \mathrm{Vv}$, so by Substitution of Provable Equivalents, we have

$$\widehat{Z}^{2CT} \vdash \forall\psi\forall\alpha[\neg\mathrm{Free}(\mathrm{V},\psi) \to \exists X\forall x(\mathrm{Sat}^*_\alpha \mathrm{Vv} \leftrightarrow \mathrm{Sat}^*_\alpha\psi)].$$

Applying satisfaction clauses for the connectives, we obtain

$$\widehat{Z}^{2CT} \vdash \forall\psi\forall\alpha[\neg\mathrm{Free}(\mathrm{V},\psi) \to \exists X\forall x\mathrm{Sat}^*_\alpha(\mathrm{Vv} \underset{\cdot}{\leftrightarrow} \psi)].$$

But note that $\forall x\mathrm{Sat}_{\alpha[\llbracket \mathrm{V} \rrbracket := X, \llbracket \mathrm{v} \rrbracket := x]}\chi$ is provably equivalent to $(\forall\beta \overset{\mathrm{v}}{\sim} \alpha)\mathrm{Sat}_{\beta[\llbracket \mathrm{V} \rrbracket := X]}\chi$, which in turn is provably equivalent, using (T_M6), to $\mathrm{Sat}_{\alpha[\llbracket \mathrm{V} \rrbracket := X]} \underset{\cdot}{\forall} \mathrm{v}\chi$. Using Substitution of Provable Equivalents again, we have

$$\widehat{Z}^{2CT} \vdash \forall\psi\forall\alpha[\neg\mathrm{Free}(\mathrm{V},\psi) \to \exists X\mathrm{Sat}_{\alpha[\llbracket \mathrm{V} \rrbracket := X]} \underset{\cdot}{\forall} \mathrm{v}(\mathrm{Vv} \underset{\cdot}{\leftrightarrow} \psi).]$$

But $\exists X\mathrm{Sat}_{\alpha[\llbracket \mathrm{V} \rrbracket := X]}\chi$ is provably equivalent to $(\exists\beta \overset{\mathrm{V}}{\sim} \alpha)\mathrm{Sat}_\beta\chi$, which in turn is provably equivalent to $\mathrm{Sat}_\alpha \underset{\cdot}{\exists} \mathrm{V}\chi$ by (T_M6), definitions, and predicate logic. So a final application of Substitution of Provable Equivalents yields

$$\widehat{Z}^{2CT} \vdash \forall\psi\forall\alpha[\neg\mathrm{Free}(\mathrm{V},\psi) \to \mathrm{Sat}_\alpha(\underset{\cdot}{\exists} \mathrm{V} \underset{\cdot}{\forall} \mathrm{v}(\mathrm{Vv} \underset{\cdot}{\leftrightarrow} \psi))],$$

or, equivalently,

$$\widehat{Z}^{2CT} \vdash \forall\psi[\neg\mathrm{Free}(\mathrm{V},\psi) \to T(\underset{\cdot}{\exists} \mathrm{V} \underset{\cdot}{\forall} \mathrm{v}(\mathrm{Vv} \underset{\cdot}{\leftrightarrow} \psi))].$$

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

References

Boolos, George 1985, ‘Reading the *Begriffsschrift*’, *Mind*, 94, pp. 331–44.

Buss, Samuel 1986, *Bounded Arithmetic* (Naples: Bibliopolis).

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

Corcoran, John, William Frank, and Michael Maloney 1974, ‘String Theory’, *Journal of Symbolic Logic* 39, pp. 625–37.

de Bouvère, K. L. 1965a, ‘Logical Synonymity’, *Indagationes Mathematicae* 22, pp. 622–29.

——1965b, ‘Synonymous Theories’, in J. W. Addison, Leon Henkin, and Alfred Tarski (eds.), *The Theory of Models: Proceedings of the 1963 International Symposium at Berkeley* (Amsterdam: North-Holland), pp. 402–6.

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

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

Feferman, Solomon 1962, ‘Transfinite Recursive Progressions of Axiomatic Theories’. *Journal of Symbolic Logic* 27, pp. 259–316.

——1964, ‘Systems of Predicative Analysis’, *Journal of Symbolic Logic* 29, pp. 1–30.

——1991, ‘Reflecting on Incompleteness’, *Journal of Symbolic Logic* 56, pp. 1–49.

Field, Hartry 1999, ‘Deflating the Conservativeness Argument’, *Journal of Philosophy* 96, pp. 533–40.

Fujimoto, Kentaro 2019, ‘Deflationism beyond Arithmetic’, *Synthese* 196, pp. 1045–69.

Grzegorczyk, Andrzej 2005, ‘Undecidability without Arithmetization’, *Studia Logica* 79, pp. 163–230.

Halbach, Volker 1999, ‘Disquotationalism and Infinite Conjunctions’, *Mind* 108, pp. 1–22.

——2001, ‘How Innocent is Deflationism?’ *Synthese* 126, pp. 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, pp. 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, pp. 1–33 (originally published under the name ‘Richard G. Heck, Jr.’).

Heck, Richard Kimberly MS, ‘The Strength of Truth Theories’, typescript available at <https://perma.cc/5K6J-CFMD>.

Hellman, Geoffrey 2004, ‘Predicativism as a Philosophical Position’, *Revue internationale de philosophie* 229, pp. 295–312.

Hilbert, David 1922, ‘Neubegründung der Mathematik: Erste Mitteilung’, *Abhandlungen aus dem Seminar der Hamburgischen Universität* 1, pp. 157–77; rpt. in his *Gesammelte Abhandlungen* (3 vols.; Berlin: Springer, 1932–35), vol. III, pp. 157–77; trans. as ‘The New Grounding of Mathematics: First Report’ in *From Brouwer to Hilbert: The Debate on the Foundations of Mathematics in the 1920s*, ed. Paolo Mancosu (New York: Oxford University Press, 1998), pp. 198–214.

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

Hofweber, Thomas 2000, ‘Proof-Theoretic Reduction as a Philosopher’s Tool’, *Erkenntnis* 53, pp. 127–46.

Horsten, Leon 1995, ‘The Semantical Paradoxes, the Neutrality of Truth and the Neutrality of the Minimalist Theory of Truth’, in Paul Cortois (ed.), *The Many Problems of Realism* (Tilburg: Tilburg University Press), pp. 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 (eds.), *Logic Colloquium ’85: Proceedings of the Colloquium Held in Orsay, France, July* 1985 (Amsterdam: North-Holland), pp. 147–59.

Ketland, Jeffrey 1999, ‘Deflationism and Tarski’s Paradise’, *Mind* 108, pp. 69–94.

Kreisel, Georg 1960a, ‘Ordinal Logics and the Characterization of Informal Concepts of Proof’, in J. A. Todd (ed.), *Proceedings of the International Congress of Mathematicians,* 14–21 *August* 1958 (Cambridge: Cambridge University Press), pp. 289–99.

——1960b, ‘La prédicativité’, *Bulletin de la Société Mathématique de France 88*, pp. 371–91.

——1965, ‘Mathematical Logic’, in T. R. Saaty (ed.), *Lectures on Modern Mathematics* (3 vols; New York: Wiley), vol. III, pp. 95–195.

——1967, 'Informal Rigour and Completeness Proofs', in Irme Lakatos (ed.), *Problems in the Philosophy of Mathematics* (Amsterdam: North-Holland), pp. 138–85.

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

Lindström, Per 2003, *Aspects of Incompleteness*, 2nd edn. (Cambridge: Cambridge 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, pp. 35–68.

Nicolai, Carlo 2015, 'Deflationary Truth and the Ontology of Expressions', *Synthese* 92, pp. 4031–55.

——2016, 'A Note on Typed Truth and Consistency Assertions', *Journal of Philosophical Logic* 45, pp. 89–119.

Parsons, Charles 1970, 'On a Number-Theoretic Choice Schema and Its Relation to Induction', in John Myhill, Akiko Kino, and R. E. Vesley (eds.), *Intuitionism and Proof Theory* (Amsterdam: North-Holland), pp. 459–73.

——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, pp. 105–14.

Schütte, Kurt 1965a, 'Predicative Well-Orderings', in J. N. Crossley and Michael Dummett (eds.), *Formal Systems and Recursive Functions* (Amsterdam: North-Holland), pp. 280–303.

——1965b, 'Eine Grenze für die Beweisbarkeit der Transfiniten Induktion in der verzweigten Typenlogik', *Archiv für mathematischen Logik und Grundlagenforschung* 7, pp. 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, pp. 493–521.

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

Simpson, Stephen G., and Keita Yokoyama 2013, 'Reverse Mathematics and Peano Categoricity', *Annals of Pure and Applied Logic* 164, pp. 284–93.

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

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

Smoryński, Craig 1977, 'The Incompleteness Theorems', in Jon Barwise (ed.), *Handbook of Mathematical Logic* (Amsterdam: North-Holland), pp. 821–66.

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

Tait, W. W. 1981, 'Finitism', *Journal of Philosophy* 78, pp. 524–46.

Tarski, Alfred 1935. 'Der Wahrheitsbegriff in den formalisierten Sprachen', *Studia Philosophica* 1, pp. 261–405, trans. J. H. Woodger as 'The Concept of Truth in Formalized Languages' in Alfred Tarski, *Logic, Semantics, Metamathematics*, 2nd edn., ed. John Corcoran (Indianapolis: Hackett), pp. 152–278.

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

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

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

Visser, Albert, 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, pp. 429–63.

——, MS: 'Did Gentzen Prove the Consistency of Arithmetic?' Typescript.