TWO-SORTED FREGE ARITHMETIC IS NOT CONSERVATIVE

STEPHEN MACKERETH
Department of Philosophy, University of Pittsburgh
and
JEREMY AVIGAD
Department of Philosophy, Carnegie Mellon University

Abstract. Neo-Fregean logicists claim that Hume's Principle (HP) may be taken as an implicit definition of cardinal number, true simply by fiat. A long-standing problem for neo-Fregean logicism is that HP is not deductively conservative over pure axiomatic second-order logic. This seems to preclude HP from being true by fiat. In this paper, we study Richard Kimberly Heck's Two-Sorted Frege Arithmetic (2FA), a variation on HP which has been thought to be deductively conservative over second-order logic. We show that it isn't. In fact, 2FA is not conservative over n-th order logic, for all $n \geq 2$. It follows that in the usual one-sorted setting, HP is not deductively Field-conservative over second- or higher-order logic.

§1. Introduction. Frege [10–12] sought to derive the theorems of arithmetic from nothing but basic logical laws and definitions. Such a derivation, called a *logicist derivation of arithmetic*, would provide the ultimate foundation for our arithmetical knowledge. It would justify the theorems of arithmetic once and for all by deriving them from principles that needed no justification—principles that were either self-evident ('basic logical laws') or true simply by stipulation ('definitions').

By his own lights, Frege did not manage to give a logicist derivation of arithmetic. But he did show how to derive a very powerful system of arithmetic from a single, natural principle, known as *Hume's Principle* (HP).[1] Informally, HP says, 'The number of Fs is equal to the number of Gs iff there is a one–one correspondence between the Fs and the Gs.' In second-order logic, HP is expressible as the universal closure of

$$\#F = \#G \leftrightarrow \exists R(F \approx_R G),$$

where # is an operator that combines with monadic second-order variables F, G to form terms of object type, and $F \approx_R G$ abbreviates the statement that R is a one–one

Received: June 15, 2021.
2020 *Mathematics Subject Classification*: 00A30, 03A05, 03B30, 03F35
Key words and phrases: Frege, logicism, Hume's Principle, neo-Fregeanism, neologicism, abstractionism, conservative extension, field-conservativeness, second-order arithmetic, second-order logic, stipulative definition.

[1] Following Boolos and Heck's reading of Frege. See [17, pp. 137–178] and [2]. For the history of this reading of Frege, see [23, p. 106, n. 5].

doi:10.1017/S1755020322000156

correspondence between the Fs and the Gs.[2] Then we have the following beautiful result:

THEOREM 1.1 (Frege's Theorem). *The theorems of second-order arithmetic are derivable in second-order logic from* HP *together with eliminative definitions of natural number, zero, and successor.*[3]

Neo-Fregean logicists, preeminently Hale and Wright [14], argue that Frege's Theorem already yields a logicist derivation of arithmetic. They claim that HP may be taken as an *implicit definition* of the operator # ('the number of') in purely logical terms.[4] Hale and Wright's notion of implicit definition is deeply controversial. For our purposes, the main point is that Hale and Wright conceive of implicit definitions as true simply by stipulation [14, p. 117]. Such definitions need no justification. They are true by fiat. So, if Hale and Wright are correct, Frege's Theorem does indeed yield a logicist derivation of arithmetic.

Not just anything can be stipulated to be true. We cannot establish any new 'substantive' truths by fiat. No one could have established by fiat that the Morning Star is the Evening Star. Accordingly, it is natural to think that any legitimate stipulative definition must meet the following requirement, known as conservativeness:

DEFINITION 1.2. *Let T be a theory in a formal language L. Let Δ be a definition of one new sign, and let L^+ be the language obtained by adding that new sign to L. Assume that deductive systems for L and L^+ have been specified. Then Δ is conservative over T iff any L-formula that is derivable$_{L^+}$ from $T + \Delta$ is already derivable$_L$ from T.*

Intuitively, a definition is conservative over our theory T just in case adding it to our theory does not yield any new theorems expressible entirely in old vocabulary. The definition does not settle any open questions that we already knew how to ask.

But HP is not conservative. More precisely, HP is not conservative over pure axiomatic second-order logic—which presumably ought to be the starting theory for aspiring logicists.[5] For HP proves a sentence DI in the language of pure second-order logic which says that the universe is Dedekind-infinite ('there is a one–one mapping from the universe into itself that is not onto'). But DI is not a theorem of pure second-order logic. So, it seems that HP cannot be a legitimate stipulative definition. Call this the *conservativeness problem for neo-Fregean logicism*.

The conservativeness problem is robust. Definitions that are conservative over pure second-order logic seem to be mathematically very weak, and hence unable to provide a foundation for arithmetic. Furthermore, adding more basic logical laws won't help

[2] That is, $F \approx_R G$ abbreviates $\forall x \forall y(Rxy \to (Fx \wedge Gy)) \wedge \forall x(Fx \to \exists!\, yRxy) \wedge \forall y(Gy \to \exists!\, xRxy)$.

[3] *Second-order arithmetic* (Z_2) is a powerful theory that seems capable of proving almost any ordinary mathematical theorem expressible in terms of countable mathematical objects and structures. By *derivable in second-order logic*, we mean derivable in Shapiro's deductive system D2 minus the axiom schema of choice [26, pp. 66–67]. Note that D2 includes full second-order comprehension.

[4] Hale and Wright make various other claims about the epistemological status of HP. For example, they claim that HP is analytic. But note that they explicitly base the analyticity claim on the claim that HP is a legitimate implicit definition [14, pp. 4, 12–14].

[5] By *pure* (*axiomatic*) *second-order logic*, we mean the deductive system described in footnote 3.

unless those laws suffice to prove DI. But it seems like a tall order to prove the existence of infinitely many objects from basic logical laws alone.

Hale and Wright respond to the conservativeness problem by denying that stipulative definitions must be conservative in the sense of Definition 1.2. Roughly speaking, they hold that stipulative definitions need only satisfy a modified conservativeness requirement, known as Field-conservativeness.[6] We set out to explore a different approach. Is it possible to find a variant of HP that is conservative in the standard deductive sense—the sense of Definition 1.2?

A promising direction is suggested by Heck's work on the Julius Caesar problem [15, 16]. Heck reconstrues Hume's Principle as introducing a new sort of singular term into the language. Call the reconstrued principle *two-sorted Hume's Principle* (2HP), and the theory that results from supplementing 2HP with logical axioms for the expanded language, *two-sorted Frege Arithmetic* (2FA). The theory 2FA interprets second-order arithmetic in the numerical sort. In particular, 2FA proves that the numerical universe is Dedekind-infinite. But there is no obvious witness to non-conservativeness, because the numerical sort is not part of the base language. Indeed, it has been claimed that 2FA is conservative over pure second-order logic [3, p. 237, n. 7].

In this paper we prove that 2FA is not conservative over pure second-order logic. In fact, we prove something stronger. Our strategy is based on the following little fact:

LEMMA 1.3. *Let T be a theory in a formal language L, and let A be any L-sentence. Suppose that a sentence Δ is not conservative over $T + A$. Then Δ is not conservative over T.*

Proof. Let φ be an L-sentence such that $T + A + \Delta \vdash \varphi$, but $T + A \nvdash \varphi$. By the Deduction Theorem, we have $T + \Delta \vdash A \to \varphi$, but $T \nvdash A \to \varphi$. □

In Section 7, we consider a theory w2FA that is much weaker than 2FA. We show that w2FA is non-conservative over pure second-order logic together with an axiom saying that the base universe is infinite.[7] In other words, even if we already know that there are infinitely many objects, w2FA tells us something new about them! Then from Lemma 1.3, it follows that w2FA, and hence 2FA, is non-conservative over pure second-order logic.

In Section 8, we show that for the weaker theory w2FA, the non-conservativeness vanishes if we strengthen the base theory in either of two natural ways. First, w2FA is conservative over third- or higher-order logic. Second, w2FA is conservative over second-order logic plus 'the base universe is finite'.

In Section 9, we present a different proof that 2FA is not conservative over pure second-order logic. This proof shows that 2FA remains non-conservative over the stronger base theories discussed in the previous section. Specifically, we show that 2FA is not conservative over second-order logic plus 'the base universe is finite', and the proof of this fact generalizes to third- and higher-order logic.

---

[6] See [14, pp. 133, 296–297, 319–320, 324–330]. Actually, Hale and Wright allow that some legitimate stipulative definitions may fail to be Field-conservative. However, in such cases, our entitlement to accept the definition 'cannot be purely stipulative' (p. 133). Also, on Hale and Wright's view, legitimate stipulative definitions must meet some other requirements besides Field-conservativeness; see [24, p. 450] for a nice summary.

[7] The axiom says: 'There is no well-ordering of the universe whose converse is also a well-ordering.' See the discussion of Stäckel-finiteness in Section 5.

In order to state and prove these results, we will need some preliminaries. In Section 2, we explain the logical setting for the paper: many-sorted axiomatic second-order logic. In Section 3, we explain how to construe Hume's Principle in a many-sorted setting, and we define the theories w2FA and 2FA. In Section 4, we present some background material on first- and second-order arithmetic. In Section 5, we show how to formalize some facts about well-orderings and finiteness in second-order logic. In Section 6, we discuss the Fraenkel model, which is the minimal infinite model of pure second-order logic.

In Sections 7–9, we prove the main results. Lastly, in Section 10, we connect our work to the literature on Field-conservativeness and related notions. Our main result implies that in a one-sorted setting, HP is neither deductively Field-conservative nor deductively Caesar-neutral conservative over second- or higher-order logic. This answers some open problems raised by Shapiro and Weir [27, p. 298], Fine [9, p. 192, n. 1], and Studd [30, p. 597]. We conclude by mentioning some open problems of our own.

§2. Many-sorted second-order logic. We work in axiomatic second-order logic with many sorts of singular terms and first-order variables. In this section we explain the logical framework in considerable generality.

In Section 2.1, we define 'sort'. In Sections 2.2 and 2.3, we define second-order languages $\mathcal{L}_J[K]$ for any nonempty set of object sorts J and any set of constant symbols K. We present deductive systems and general semantics for these languages. In Section 2.4, we define the two many-sorted second-order languages that will be central to the rest of the paper, called the *base language* $\mathcal{L} := \mathcal{L}_{\{0\}}[\varnothing]$ and the *expanded language* $\mathcal{L}^+ := \mathcal{L}_{\{0,n\}}[\{\#_0, \#_n\}]$.

2.1. Sorts. Let J be any nonempty set of symbols. These symbols are called *first-order sorts* or *object sorts*.

Let $\mathit{Sorts}^2(J)$ be the set of all tuples $\langle j_1, \dots, j_n\rangle$ with $n \geq 1$ and $j_1, \dots, j_n \in J$. These tuples are *second-order relation sorts* formed from J.

Let $\mathit{Sorts}^3(J)$ be the set of all tuples $\langle \tau_1, \dots, \tau_n\rangle$ with $n \geq 1$ and $\tau_1, \dots, \tau_n \in J \cup \mathit{Sorts}^2(J)$, and with at least one of $\tau_1, \dots, \tau_n$ belonging to $\mathit{Sorts}^2(J)$. These tuples are *third-order relation sorts* formed from J.

Let $\mathit{FnSorts}(J)$ be the set of all tuples $\langle \tau_1, \dots, \tau_n; \tau_{n+1}\rangle$ with $n \geq 1$ and $\tau_1, \dots, \tau_n, \tau_{n+1} \in J \cup \mathit{Sorts}^2(J)$. These tuples are *function sorts* formed from J.

Let $\mathit{Sorts}(J) = J \cup \mathit{Sorts}^2(J) \cup \mathit{Sorts}^3(J) \cup \mathit{FnSorts}(J)$.

Intuitively, $\langle \tau_1, \dots, \tau_n\rangle$ is the sort of n-ary relations with arguments of sorts $\tau_1, \dots, \tau_n$, while $\langle \tau_1, \dots, \tau_n; \tau_{n+1}\rangle$ is the sort of n-ary functions with arguments of sorts $\tau_1, \dots, \tau_n$ and values of sort τ_{n+1}.

EXAMPLE 2.4. *Suppose* $J = \{0, 1\}$. *Then* $\langle 1, 1, 0\rangle \in \mathit{Sorts}^2(J)$, $\langle 0, 1, \langle 0, 1\rangle\rangle \in \mathit{Sorts}^3(J)$, *and* $\langle\langle 0\rangle; 1\rangle \in \mathit{FnSorts}(J)$.

In the languages $\mathcal{L}_J[K]$, there will be no function variables and no third-order variables. We only allow variables of sorts $\tau \in J \cup \mathit{Sorts}^2(J)$. However, there may be constant symbols of any sort $\tau \in \mathit{Sorts}(J)$.

2.2. Languages without constant symbols. For any set of object sorts J, we define the second-order language $\mathcal{L}_J$ as follows:

(i) The alphabet of $\mathcal{L}_J$ contains variables $x^j, y^j, z^j, \dots$ for each object sort $j \in J$, and relation variables $X^\tau, Y^\tau, Z^\tau, \dots$ for each second-order sort $\tau \in \mathit{Sorts}^2(J)$.

https://doi.org/10.1017/S1755020322000156 Published online by Cambridge University Press

Table 1. *Deductive system for $\mathcal{L}_J$.*

Propositional logic	all tautologies	
1° quantification	$\forall x\varphi(x) \to \varphi(t)$	t is sub. for x
	$\forall x(\varphi \to \psi) \to (\forall x\varphi \to \forall x\psi)$	
	$\varphi \to \forall x\varphi$	x not free in φ
Identity	$x = x$	
	$x = y \to (\alpha \to \alpha')$	$(*)$
2° quantification	$\forall X\varphi(X) \to \varphi(T)$	T is sub. for X
	$\forall X(\varphi \to \psi) \to (\forall X\varphi \to \forall X\psi)$	
	$\varphi \to \forall X\varphi$	X not free in φ
Comprehension	$\exists X\forall\bar{x}(X\bar{x} \leftrightarrow \varphi(\bar{x}))$	X not free in φ
Rule of inference	from φ and $\varphi \to \psi$, infer ψ	

Let φ, ψ be any formulas of $\mathcal{L}_J$. Let x, y, X be variables, and t, T be terms. (Note that x, y, t must all be of the same sort. Likewise, X and T must be of the same sort.) Let $\varphi(t)$ be the result of substituting t for all free occurrences of x in φ. In $(*)$, let α be any atomic formula of $\mathcal{L}_J$, and let α' be any formula obtained from α by replacing zero or more occurrences of x with y. In Comprehension, we write $X\bar{x}$ to abbreviate $X^{\langle j_1,\ldots,j_n\rangle}x_1^{j_1}\cdots x_n^{j_n}$.

There are no nonlogical constant symbols. The logical constants are $\neg$, $\to$, $\forall$, $=$.

(ii) The terms of sort τ are the variables of sort τ, for each $\tau \in J \cup \mathit{Sorts}^2(J)$.

(iii) In atomic formulas, we require that the sorts match. More precisely, the atomic formulas are strings of the form $t_1^j = t_2^j$ and $T^{\langle j_1,\ldots,j_n\rangle}t_1^{j_1},\ldots,t_n^{j_n}$, where each t^j is a term of sort $j \in J$, and $T^{\langle j_1,\ldots,j_n\rangle}$ is a term of sort $\langle j_1,\ldots,j_n\rangle$.

(iv) If φ, ψ are formulas and x^j, X^τ are variables, then $\neg\varphi$, $\varphi \to \psi$, $\forall x^j\varphi$, $\forall X^\tau\varphi$ are also formulas.

The *deductive system* for $\mathcal{L}_J$ is essentially equivalent to Shapiro's D2 minus the axiom schema of choice [26, pp. 66–67]. Compare [8, pp. 112–113]. Its axioms are all closed universal generalizations of the formulas depicted in Table 1. For legibility, we suppress sorts. But note that x, y, and t must all be of the same sort, and X and T must be of the same sort. This requirement is induced by the formation rules of the language.

An $\mathcal{L}_J$*-prestructure* $\mathcal{M}$ is a collection of nonempty sets $\{M_\tau : \tau \in J \cup \mathit{Sorts}^2(J)\}$ such that $M_{\langle j_1,\ldots,j_n\rangle} \subseteq \mathcal{P}(M_{j_1} \times \cdots \times M_{j_n})$ for all $j_1,\ldots,j_n \in J$. Satisfaction and truth in $\mathcal{M}$ are defined inductively, taking variables of sort τ to range over domain M_τ.

A *general* $\mathcal{L}_J$*-structure* is an $\mathcal{L}_J$-prestructure in which the second-order comprehension axioms are satisfied. Our deductive system is sound and complete with respect to general $\mathcal{L}_J$-structures.

A *standard* $\mathcal{L}_J$*-structure* $\mathcal{M}$ is a general $\mathcal{L}_J$-structure in which $M_{\langle j_1,\ldots,j_n\rangle} = \mathcal{P}(M_{j_1} \times \cdots \times M_{j_n})$ for all $j_1,\ldots,j_n \in J$. So, a standard $\mathcal{L}_J$-structure is fully specified by its object domains $\{M_j : j \in J\}$. Our deductive system is sound but not complete with respect to standard structures.

2.3. Languages with constant symbols. We will now sketch how to add constant symbols to the languages $\mathcal{L}_J$.

For each $\tau \in \mathit{Sorts}(J)$, let K_τ be a set of new symbols, called *constant symbols*. Each constant symbol is assigned to a particular sort τ, and is classified as an object, relation, or function constant accordingly. Assume that the K_τ's are pairwise disjoint, or use superscripts to keep track of sorts. Let $K = \bigcup_{\tau \in \mathit{Sorts}(J)} K_\tau$.

Define the language $\mathcal{L}_J[K]$ as follows:

(i) The alphabet of $\mathcal{L}_J[K]$ is the alphabet of $\mathcal{L}_J$ expanded by K.
(ii) If $\tau \in J \cup \mathit{Sorts}^2(J)$, the atomic terms of sort τ are the variables x^τ and the constants in K_τ.
If $\tau \in \mathit{Sorts}^3(J)$, the atomic terms of sort τ are the constants in K_τ.
If $\tau = \langle \tau_1, \dots, \tau_n; \tau_{n+1} \rangle \in \mathit{FnSorts}(J)$, and $f^\tau \in K_\tau$, and $t_1^{\tau_1}, \dots, t_n^{\tau_n}$ are terms of the indicated sorts, then $f^\tau t_1^{\tau_1} \cdots t_n^{\tau_n}$ is a term of sort τ_{n+1}.
(iii) The atomic formulas are defined as in $\mathcal{L}_J$, except that we also allow atomic formulas of the form $T^\tau t_1^{\tau_1} \cdots t_n^{\tau_n}$ with $\tau = \langle \tau_1, \dots, \tau_n \rangle \in \mathit{Sorts}^3(J)$.
(iv) The inductive clauses generating the set of all formulas are unchanged.

The *deductive system* for $\mathcal{L}_J[K]$ is obtained from the deductive system for $\mathcal{L}_J$ by allowing φ, ψ to range over $\mathcal{L}_J[K]$-formulas, α to range over atomic $\mathcal{L}_J[K]$-formulas, and adding axioms of Extensionality analogous to the axioms of Identity.[8]

An $\mathcal{L}_J[K]$-*prestructure* $\mathcal{M} = (\mathcal{S}, I)$ consists of an $\mathcal{L}_J$-prestructure $\mathcal{S}$ together with an interpretation I of the constant symbols that meets the following three conditions:

(i) If c^j is an object constant of sort $j \in J$, then $I(c^j) \in M_j$.
(ii) If R^τ is a relation constant of sort $\tau = \langle \tau_1, \dots, \tau_n \rangle \in \mathit{Sorts}^2(J) \cup \mathit{Sorts}^3(J)$, then $I(R^\tau) \in \mathcal{P}(M_{\tau_1} \times \cdots \times M_{\tau_n})$.
(iii) If f^τ is a function constant of sort $\tau = \langle \tau_1, \dots, \tau_n; \tau_{n+1} \rangle \in \mathit{FnSorts}(J)$, then $I(f^\tau)$ is a function from $M_{\tau_1} \times \cdots \times M_{\tau_n}$ into $M_{\tau_{n+1}}$.

General and standard $\mathcal{L}_J[K]$-structures are defined analogously to $\mathcal{L}_J$-structures.

2.4. The languages $\mathcal{L}$ and $\mathcal{L}^+$. We now define the two languages that will be at the center of the rest of the paper.

DEFINITION 2.5. *The base language is* $\mathcal{L} := \mathcal{L}_{\{0\}}$.

DEFINITION 2.6. *The expanded language is* $\mathcal{L}^+ := \mathcal{L}_{\{0,n\}}[\{\#_0, \#_n\}]$, *where* $\#_0$ *and* $\#_n$ *are function constants of sorts* $\langle\langle 0\rangle; n\rangle$ *and* $\langle\langle n\rangle; n\rangle$, *respectively.*

The logical axioms for $\mathcal{L}$ and $\mathcal{L}^+$ will be denoted by $Ax_{\mathcal{L}}$ and $Ax_{\mathcal{L}^+}$, respectively.

Some notational conventions:

(i) We generally drop the superscripts 0, $\langle 0\rangle$, $\langle 0, 0\rangle$, ...
(ii) We generally write variables of sorts $\tau \in \{n\} \cup \mathit{Sorts}^2(\{n\})$ in boldface, and drop the superscripts n, $\langle n\rangle$, $\langle n, n\rangle$, $\cdots$.
(iii) When we write second-order relation superscripts, we drop the angle brackets and commas. For example, we write X^{n0} instead of $X^{\langle n,0\rangle}$.
(iv) We drop the subscripts from $\#_0$ and $\#_n$, writing # for both.
(v) Following Frege, we refer to monadic relations as *concepts*.

[8] Let X, Y be variables of sort $\tau \in \mathit{Sorts}^2(J)$. Let α be any atomic formula of $\mathcal{L}_J[K]$, and let α' be any formula obtained from α by replacing zero or more occurrences of X with Y. Then any closed universal generalization of $\forall \bar{x}(X\bar{x} \leftrightarrow Y\bar{x}) \to (\alpha \to \alpha')$ is an Extensionality axiom.

https://doi.org/10.1017/S1755020322000156 Published online by Cambridge University Press

§3. Heck's theory 2FA. Think of the base language $\mathcal{L}$ as our starting language, and $Ax_{\mathcal{L}}$ as our starting theory. Heck [15], [16, pp. 150–151] reconstrues Hume's Principle as introducing a new, numerical sort of object (sort n), together with a host of new second-order relation sorts. The operator # ('the number of') may be applied to a concept variable of either sort, yielding a singular term of the numerical sort.

DEFINITION 3.7. *Weak two-sorted Hume's Principle* (w2HP) *is the universal closure of*:

$$\#F^0 = \#G^0 \leftrightarrow \exists R^{00}(F^0 \approx_{R^{00}} G^0).$$

Here, $F^0 \approx_{R^{00}} G^0$ *abbreviates the statement that* R^{00} *is a one–one correspondence between* F^0 *and* G^0.

Intuitively, w2HP gives the criterion of identity for numbers belonging to base-sort concepts. It tells us how to count base-sort objects. But w2HP does not tell us how to count numbers. Since we do in fact count numbers, we are motivated to consider a stronger principle.

DEFINITION 3.8. *Two-sorted Hume's Principle* (2HP) *is the conjunction of the universal closures of the following three* $\mathcal{L}^+$*-formulas*:

$$\begin{aligned} \#F^0 = \#G^0 &\leftrightarrow \exists R^{00}(F^0 \approx_{R^{00}} G^0), \\ \#F^n = \#G^n &\leftrightarrow \exists R^{nn}(F^n \approx_{R^{nn}} G^n), \\ \#F^n = \#G^0 &\leftrightarrow \exists R^{n0}(F^n \approx_{R^{n0}} G^0). \end{aligned}$$

The first line is w2HP. The second line gives the criterion of identity for numbers belonging to numerical concepts. The third line gives the mixed criterion of identity, which tells us (e.g.) whether the number of Julio-Claudian emperors equals the number of prime numbers less than 12.

Using our superscript-dropping conventions, we may write 2HP as follows:

$$\begin{aligned} \#F = \#G &\leftrightarrow \exists R(F \approx_R G), \\ \#\mathbf{F} = \#\mathbf{G} &\leftrightarrow \exists \mathbf{R}(\mathbf{F} \approx_{\mathbf{R}} \mathbf{G}), \\ \#\mathbf{F} = \#G &\leftrightarrow \exists R^{n0}(\mathbf{F} \approx_{R^{n0}} G). \end{aligned}$$

DEFINITION 3.9. *Weak two-sorted Frege Arithmetic* (w2FA) *is the theory whose logical axioms are* $Ax_{\mathcal{L}^+}$ *and whose sole nonlogical axiom is* w2HP.[9] *In other words,*

$$\text{w2FA} = Ax_{\mathcal{L}^+} + \text{w2HP}.$$

DEFINITION 3.10. *Two-sorted Frege Arithmetic* (2FA) *is the theory whose logical axioms are* $Ax_{\mathcal{L}^+}$ *and whose sole nonlogical axiom is* 2HP. *In other words,*

$$\text{2FA} = Ax_{\mathcal{L}^+} + \text{2HP}.$$

Notice that the logical axioms of 2FA include full second-order comprehension for the expanded language. So, by Frege's Theorem, 2FA interprets second-order arithmetic in the numerical sort. It follows that 2FA proves a sentence which says that the numerical universe is Dedekind-infinite. But this is not a witness to non-conservativeness, because the numerical sort is not part of the base language.

[9] Beware: Linnebo [22] calls *this* theory 'Two-Sorted Frege Arithmetic'. We follow Heck's usage.

https://doi.org/10.1017/S1755020322000156 Published online by Cambridge University Press

Prima facie, it seems quite plausible that 2FA should be a conservative extension of $Ax_{\mathcal{L}}$.

§4. Arithmetic. We will study 2FA by comparing it with other, better-known systems of arithmetic. In Section 4.1, we describe the usual systems of first- and second-order arithmetic. In Section 4.2, we describe systems of arithmetic with no function symbols.

4.1. First- and second-order arithmetic. We begin with first-order arithmetic. For reference, see [13, pp. 12–13, 28–29].

Definition 4.11. *The language of Peano arithmetic, L_{PA}, is a classical first-order language with identity whose nonlogical vocabulary is $(0, S, \leq, +, \cdot)$. Here, 0 is a constant symbol, S is a unary function symbol, $\leq$ is a binary relation symbol, and $+, \cdot$ are binary function symbols.*

Definition 4.12. *Robinson arithmetic, Q, is the theory in L_{PA} with the following eight axioms*:

$$\begin{aligned}
&0 \neq Sx,\\
&Sx = Sy \to x = y,\\
&x \neq 0 \to \exists y(x = Sy),\\
&x + 0 = x,\\
&x + Sy = S(x + y),\\
&x \cdot 0 = 0,\\
&x \cdot Sy = (x \cdot y) + x,\\
&x \leq y \leftrightarrow \exists z(z + x = y).
\end{aligned}$$

Definition 4.13. *Peano arithmetic,* PA, *is the result of adding to Q the following axiom schema of induction*:

$$\varphi(0) \wedge \forall x(\varphi(x) \to \varphi(Sx)) \to \forall x \varphi(x),$$

where $\varphi(x)$ is any formula of L_{PA}.

We write $(\forall x \leq t)(\cdots)$ to abbreviate $\forall x(x \leq t \to \cdots)$, and similarly we write $(\exists x \leq t)(\cdots)$. The quantifiers occurring in these expressions are said to be *bounded*.

An L_{PA}-formula is called *bounded*, or Σ_0, if all quantifiers occurring in it are bounded.

An L_{PA}-formula is called Σ_n $(n \geq 0)$ if it consists of a string of n alternating unbounded quantifiers, the first of which is existential, followed by a bounded formula. That is, a Σ_n formula has the form $\exists x \forall y \exists z \forall w \cdots \theta$, where θ is bounded.

Definition 4.14. *The theory $I\Sigma_n$ $(n \geq 0)$ is the result of adding to Q the axiom schema of induction above, restricted to Σ_n formulas.*

We now turn our attention to second-order arithmetic. For reference, see [28, pp. 2–5].

Definition 4.15. *The language of second-order arithmetic, L_2, is a two-sorted language consisting of all the vocabulary of L_{PA}, together with denumerably many monadic second-order variables $X, Y, Z, \ldots$ and a second-order quantifier $\forall X$. The atomic formulas of L_2*

https://doi.org/10.1017/S1755020322000156 Published online by Cambridge University Press

include all strings of the form Xt, where t is a first-order term and X is a second-order variable.

The second-order variables of L_2 are usually called *set variables*, and the atomic formulas Xt are sometimes written $t \in X$. For our purposes, there is no difference between set variables and concept variables, and the predication relation $\in$ may be left implicit. Hence, L_2 may be regarded as an expansion of the monadic fragment of $\mathcal{L}$.

DEFINITION 4.16. *Second-order arithmetic, Z_2, is the theory in L_2 whose axioms are those of Q, together with the second-order induction axiom*

$$X0 \wedge \forall x(Xx \to X(Sx)) \to \forall x Xx$$

and the second-order comprehension scheme

$$\exists X \forall x(Xx \leftrightarrow \varphi(x))$$

for each formula φ of L_2 not containing X free. As usual, φ may contain parameters, i.e., free first- or second-order variables other than x.

4.2. First- and second-order arithmetic with no function symbols. In this section, we introduce an arithmetical language L' in which successor, addition, and multiplication are rendered as relations (which may be only partially defined) instead of functions. This allows us to define BA', a weak system of arithmetic that does not assume the existence of infinitely many natural numbers. The main point of the section is to state Lemma 4.23 and prove Lemmas 4.25 and 4.28. We will use these lemmas in Section 9 only, so feel free to skip this section and return to it later.

For reference, see [13, pp. 86–89, 233].

DEFINITION 4.17. *Let L' be the classical first-order language with identity whose nonlogical vocabulary is $(0, S, \leq, A, M)$. Here, 0 is a constant symbol, S and $\leq$ are binary relation symbols, and A and M are ternary relation symbols.*

An L'-formula is called *bounded′*, or Σ_0', if it contains only bounded quantifiers.

DEFINITION 4.18. *BA' is the theory in L' with the following axioms*:

1. *$\leq$ is a discrete linear order with least element* 0,
2. *Sxy iff y is the upper neighbor of x with respect to $\leq$,*
3. *Definitions of A and M*:

$$\begin{aligned}
&Ax0z \leftrightarrow z = x,\\
&Syy' \wedge Szz' \to (Axyz \leftrightarrow Axy'z'),\\
&Mx0z \leftrightarrow z = 0,\\
&Syy' \wedge Azxz' \to (Mxyz \leftrightarrow Mxy'z'),
\end{aligned}$$

4. *Commutativity and associativity of A and M, distributivity, monotonicity of addition, monotonicity of multiplication by a positive number, and $x \leq y \leftrightarrow (\exists u \leq y)Axuy$,*
5. *Induction scheme for Σ_0' formulas*:

$$\varphi(0) \wedge \forall x \forall y(\varphi(x) \wedge Sxy \to \varphi(y)) \to \forall x \varphi(x).$$

DEFINITION 4.19. *$I\Sigma_0'$ is the result of adding to BA' axioms saying that S, A, M define total functions, namely $\forall x \exists y Sxy$, etc.*

An L'-formula is called Σ'_n $(n \geq 0)$ if it consists of a string of n alternating unbounded quantifiers, the first of which is existential, followed by a bounded$'$ formula.

Definition 4.20. *The theory $I\Sigma'_n$ $(n \geq 0)$ is the result of adding to $I\Sigma'_0$ the axiom schema of induction above, extended to Σ'_n formulas.*

We now state some useful facts about BA' and its relatives.

Definition 4.21. *Let $\mathfrak{D}$ be the conjunction of the following three $(L_{\mathrm{PA}} \cup L')$-formulas:*

$$\begin{aligned} Sx = y &\leftrightarrow Sxy, \\ x + y = z &\leftrightarrow Axyz, \\ x \cdot y = z &\leftrightarrow Mxyz. \end{aligned}$$

For each $n \in \mathbb{N}$, let $x \doteq n$ abbreviate the L'-formula

$$(\exists u_1, \dots, u_{n-1} \leq x)(S0u_1 \wedge Su_1u_2 \wedge \cdots \wedge Su_{n-1}x).$$

Lemmas 4.22 and 4.23 tell us that the theories $I\Sigma_n$ and $I\Sigma'_n$ are in a strong sense equivalent.

Lemma 4.22. *Let $n \geq 0$. Then $I\Sigma'_n + \mathfrak{D} \vdash I\Sigma_n$, and conversely $I\Sigma_n + \mathfrak{D} \vdash I\Sigma'_n$.*

Lemma 4.23. *Let φ be a Σ_n formula with $n \geq 1$. Then there is a Σ'_n formula φ' with the same free variables as φ such that $I\Sigma'_n + \mathfrak{D} \vdash \varphi \leftrightarrow \varphi'$.*

For proof, see [13, pp. 88–89].[10]

Lemma 4.24. *$I\Sigma'_0$ and BA' prove the same bounded$'$ formulas.*

For proof, see [13, p. 233].

Lemma 4.25. *Let $\varphi(x_1, \dots, x_k)$ be a bounded$'$ formula, and let $a_1, \dots, a_k \in \mathbb{N}$ be such that $\mathbb{N} \vDash \varphi(a_1, \dots, a_k)$. Then*

$$BA' \vdash x_1 \doteq a_1 \wedge \cdots \wedge x_k \doteq a_k \to \varphi(x_1, \dots, x_k).$$

Proof. Let ψ be the L_{PA}-formula obtained from φ by replacing Sxy, $Axyz$, $Mxyz$ with $Sx = y$, $x + y = z$, $x \cdot y = z$ respectively. Observe that $\psi(S^{a_1}0, \dots, S^{a_k}0)$ is a true bounded sentence of L_{PA}.

Now we argue as follows:

$$\begin{aligned} I\Sigma_0 &\vdash \psi(S^{a_1}0, \dots, S^{a_k}0), \\ I\Sigma'_0 + \mathfrak{D} &\vdash \psi(S^{a_1}0, \dots, S^{a_k}0), \\ I\Sigma'_0 + \mathfrak{D} &\vdash x_1 \doteq a_1 \wedge \cdots \wedge x_k \doteq a_k \to \psi(x_1, \dots, x_k), \\ I\Sigma'_0 + \mathfrak{D} &\vdash x_1 \doteq a_1 \wedge \cdots \wedge x_k \doteq a_k \to \varphi(x_1, \dots, x_k), \\ I\Sigma'_0 &\vdash x_1 \doteq a_1 \wedge \cdots \wedge x_k \doteq a_k \to \varphi(x_1, \dots, x_k), \\ BA' &\vdash x_1 \doteq a_1 \wedge \cdots \wedge x_k \doteq a_k \to \varphi(x_1, \dots, x_k). \end{aligned}$$

[10] In Hájek and Pudlák's proof of I.2.88 (p. 88), the lower bound for y is incorrect. The proof can easily be fixed by replacing $\max \mathbf{x}$ with $\max \mathbf{x} + 2$. Compare V.5.1(1) (p. 362), where the correct bound is given.

The first line holds because $I\Sigma_0$ proves all true bounded sentences.[11] The second line follows by Lemma 4.22. Regarding the third line, it is easy to check that for each $n \in \mathbb{N}$,

$$I\Sigma_0' + \mathfrak{D} \vdash x = S^n 0 \leftrightarrow x \doteq n.$$

The fourth line follows by propositional logic, because φ and ψ differ only by applications of the equivalences in $\mathfrak{D}$. The fifth line follows because $I\Sigma_0' + \mathfrak{D}$ is conservative over $I\Sigma_0'$ for L'-formulas. The sixth line follows by Lemma 4.24. □

Lastly, we describe a system of second-order arithmetic without function symbols.

DEFINITION 4.26. *The language L_2' is just like L_2, but with the vocabulary of L' replacing the vocabulary of L_{PA}.*

DEFINITION 4.27. *Let Z_2' be the theory in L_2' whose axioms are those of $I\Sigma_0'$, plus the second-order induction axiom*

$$X0 \wedge \forall x \forall y (Xx \wedge Sxy \rightarrow Xy) \rightarrow \forall x Xx$$

and the second-order comprehension scheme for L_2'.

LEMMA 4.28. *Z_2 and Z_2' are mutually interpretable. Indeed, $Z_2' + \mathfrak{D} \vdash Z_2$, and conversely $Z_2 + \mathfrak{D} \vdash Z_2'$.*

Proof. We argue that $Z_2' + \mathfrak{D} \vdash Z_2$. The other direction is easy.

Observe that $(Z_2' + \mathfrak{D}) \vdash (I\Sigma_0' + \mathfrak{D}) \vdash I\Sigma_0 \vdash Q$. Furthermore, the two ways of formulating the second-order induction axiom are equivalent in the presence of $Sx = y \leftrightarrow Sxy$.

It remains to show that $Z_2' + \mathfrak{D}$ proves the second-order comprehension scheme for L_2. Take any L_2-formula φ. Let ψ be the formula obtained from φ by replacing each atomic predication Xt with $\exists z(Xz \wedge z = t)$, where z is a new variable. Then every non-atomic term in ψ occurs in an equation $t_1 = t_2$. These equations are L_{PA}-formulas. By Lemma 4.23, $Z_2' + \mathfrak{D}$ proves each L_{PA}-formula to be equivalent to an L'-formula. So, there is an L_2'-formula φ' such that $Z_2' + \mathfrak{D} \vdash \varphi \leftrightarrow \varphi'$. Now apply second-order comprehension to φ', and we are done. □

§5. Well-orderings and finiteness. In this section, we define 'well-ordering' in $\mathcal{L}$, and we note that $Ax_{\mathcal{L}}$ proves that all well-orderings are comparable (Lemma 5.29). Then we define the notion of Stäckel-finiteness and prove the important lemma of induction on finite concepts (Lemma 5.32). We will use these lemmas throughout the paper.

For simplicity, we work in $\mathcal{L}$. However, these notions can easily be extended to $\mathcal{L}^+$.

Let $\varnothing$ denote the empty concept. Let V denote the universal concept.

Let $Y \subseteq X$ abbreviate $\forall x(Yx \rightarrow Xx)$.

Let '(X, R) is a linear order' abbreviate the formula

$$\forall x \forall y (Rxy \wedge Ryx \rightarrow x = y) \wedge \forall x \forall y \forall z (Rxy \wedge Ryz \rightarrow Rxz) \\ \wedge \forall x \forall y (Xx \wedge Xy \leftrightarrow (Rxy \vee Ryx)).$$

In other words, (X, R) is a linear order just in case R is an antisymmetric, transitive, total relation on X.

[11] This is because $I\Sigma_0 \vdash Q$, and Q is Σ_1-complete [13, pp. 30–31, I.1.8–9].

Let '(X, R) is well-founded' abbreviate

$$\forall Y(Y \neq \varnothing \wedge Y \subseteq X \rightarrow \exists x(Yx \wedge \forall y(Yy \rightarrow Rxy))).$$

Say that (X, R) is a *well-ordering* if (X, R) is a well-founded linear order.

We say that two well-orderings $(X, \leq_X)$ and $(Y, \leq_Y)$ are *order-isomorphic*, denoted $(X, \leq_X) \simeq_o (Y, \leq_Y)$, just in case there is a bijection $f : X \rightarrow Y$ such that

$$\forall x \forall y(x \leq_X y \leftrightarrow f(x) \leq_Y f(y)).$$

Strictly speaking, we should represent f as a relation, but we will go on using functional notation informally.

If (X, R) is a well-ordering, let $X \upharpoonright a$ be the *initial segment* of (X, R) up to a, defined by

$$(X \upharpoonright a)x \;\leftrightarrow\; Xx \wedge Rxa.$$

We also regard $\varnothing$ as an initial segment of (X, R). An initial segment of (X, R) is *proper* if it is not equal to X.

Let $(X, \leq_X) <_o (Y, \leq_Y)$ abbreviate the statement that $(X, \leq_X)$ is order-isomorphic with a proper initial segment of $(Y, \leq_Y)$.

We borrow the next lemma from [7, p. 611].

Lemma 5.29 (Comparability of well-orderings). *It is provable from $Ax_{\mathcal{L}}$ that any two well-orderings $(X, \leq_X)$ and $(Y, \leq_Y)$ are comparable, in the sense that exactly one of the following holds*:

$$(X, \leq_X) <_o (Y, \leq_Y), \quad (X, \leq_X) \simeq_o (Y, \leq_Y), \quad (X, \leq_X) >_o (Y, \leq_Y).$$

Proof. Copy the usual set-theoretic proof [18, pp. 18–19]. □

We now define the notion of Stäckel-finiteness.

If R is a binary relation, let R^{-1} be the converse of R, defined by $R^{-1}xy \leftrightarrow Ryx$.

Definition 5.30. *Say that (X, R) is a double well-ordering if (X, R) and (X, R^{-1}) are both well-orderings.*

Say that X is Stäckel-finite, abbreviated $Fin(X)$, if X admits a double well-ordering. That is,

$$Fin(X) \iff_{\text{df}} \exists R((X, R) \textit{ is a double well-ordering}).$$

Remark 5.31. *The double well-ordering criterion is proposed as a definition of finiteness in* [29]. *The criterion is also discussed in* [34, 35]. *For historical remarks, see* [25].

Stäckel-finiteness is strictly stronger than Dedekind-finiteness, in the sense that

$$Ax_{\mathcal{L}} \vdash Fin(X) \rightarrow DFin(X),$$
$$Ax_{\mathcal{L}} \nvdash DFin(X) \rightarrow Fin(X),$$

where of course $DFin(X)$ abbreviates that X is Dedekind-finite. Indeed, $Fin(X) \rightarrow DFin(X)$ is a version of the pigeonhole principle. It is provable from $Ax_{\mathcal{L}}$ by induction on finite concepts (Lemma 5.32). On the other hand, the Fraenkel model (defined in Section 6) is a model of $DFin(V) + \neg Fin(V)$, witnessing that $Ax_{\mathcal{L}} \nvdash DFin(X) \rightarrow Fin(X)$.

Lastly, we show that $Ax_{\mathcal{L}}$ proves a principle of induction on Stäckel-finite concepts.

Let $X \cup \{a\}$ be the concept defined by

$$(X \cup \{a\})x \leftrightarrow (Xx \vee x = a).$$

LEMMA 5.32 (Induction on finite concepts). *Let $\varphi(X)$ be any formula of $\mathcal{L}$. Then $Ax_{\mathcal{L}}$ proves the universal closure of*

$$\varphi(\varnothing) \wedge \forall X \forall a(Fin(X) \wedge \varphi(X) \to \varphi(X \cup \{a\})) \to \forall X(Fin(X) \to \varphi(X)).$$

Proof. Assume the antecedent. Take any X such that $Fin(X)$. Fix a double well-ordering (X, R), and let Y be defined by $Yx \leftrightarrow (Xx \wedge \varphi(X \restriction x))$. It suffices to show that $Y = X$.

Suppose not. Since (X, R) is a well-ordering, there is an R-least y such that $Xy \wedge \neg Yy$. It is easy to see that y cannot be the R-least element of X. Since (X, R^{-1}) is a well-ordering, y has a unique (X, R)-predecessor, call it z. By the minimality of y, we have Yz, and hence $\varphi(X \restriction z)$. Also, it is easy to see that $Fin(X \restriction z)$. It follows that $\varphi((X \restriction z) \cup \{y\})$, which is to say $\varphi(X \restriction y)$. But this contradicts our choice of y. □

§6. The Fraenkel model. In this section, we define the Fraenkel model and show that it is a model of $Ax_{\mathcal{L}} + \neg Fin(V)$ (Lemmas 6.38 and 6.39). Then we show that the relations occurring in the Fraenkel model are exactly the sets definable by Boolean combinations of equalities with object parameters (Lemma 6.40). We will make good use of these facts in Section 7.

We remark that Lemma 6.40 implies that the Fraenkel model is the minimal infinite model of $Ax_{\mathcal{L}}$—i.e., it is a submodel of any infinite model of $Ax_{\mathcal{L}}$.

DEFINITION 6.33. *Let $A \subseteq \mathbb{N}^n$ and $E \subseteq \mathbb{N}$. We say that E is a support of A if every permutation $\pi : \mathbb{N} \to \mathbb{N}$ that fixes E pointwise fixes A setwise*:

$$(\forall e \in E)(\pi(e) = e) \implies \forall x_1, \dots, x_n((x_1, \dots, x_n) \in A \leftrightarrow (\pi(x_1), \dots, \pi(x_n)) \in A).$$

Using the notation $\pi(A) = \{(\pi(x_1), \dots, \pi(x_n)) \in \mathbb{N}^n : (x_1, \dots, x_n) \in A\}$, we can restate this property as follows: for every permutation $\pi : \mathbb{N} \to \mathbb{N}$,

$$(\forall e \in E)(\pi(e) = e) \implies \pi(A) = A.$$

DEFINITION 6.34. *A set $A \subseteq \mathbb{N}^n$ is symmetric if it has a finite support $E \subseteq \mathbb{N}$.*

DEFINITION 6.35. *The Fraenkel model is the $\mathcal{L}$-prestructure $\mathcal{M}$ whose object universe is $\mathbb{N}$, and whose n-ary relations are the symmetric subsets of $\mathbb{N}^n$. That is, writing M_n for $M_{\langle 0, \dots, 0 \rangle}$ (n zeroes),*

$$M_0 = \mathbb{N},$$
$$M_n = \{A \subseteq \mathbb{N}^n : A \text{ is symmetric}\}.$$

It is well known that $\mathcal{M}$ is a model of $Ax_{\mathcal{L}}$ (i.e., it is a general $\mathcal{L}$-structure) [32]. However, we are not aware of any English-language source that gives the proof. For the reader's convenience, we present the proof from [1] in the next two lemmas.

LEMMA 6.36. *If $A \subseteq \mathbb{N}^n$ is symmetric, and $\sigma : \mathbb{N} \to \mathbb{N}$ is any permutation, then $\sigma(A) \subseteq \mathbb{N}^n$ is also symmetric.*

https://doi.org/10.1017/S1755020322000156 Published online by Cambridge University Press

Proof. Let E be a support for A. We show that $\sigma^{-1}(E)$ is a support for $\sigma(A)$. Indeed, take any permutation $\pi : \mathbb{N} \to \mathbb{N}$ that fixes $\sigma^{-1}(E)$ pointwise. Then the permutation $\sigma^{-1}\pi\sigma : \mathbb{N} \to \mathbb{N}$ fixes E pointwise. So, $(\sigma^{-1}\pi\sigma)(A) = A$, and hence $\pi(\sigma(A)) = \sigma(A)$. □

COROLLARY 6.37. *Each relation domain M_n of the Fraenkel model is closed under the action (on $\mathbb{N}^n$) of permutations of $\mathbb{N}$.*

LEMMA 6.38. *The Fraenkel model is a model of $Ax_{\mathcal{L}}$.*

Proof. Let $\mathcal{M}$ be the prestructure defined above. We show that $\mathcal{M}$ satisfies Comprehension. Take any formula $\varphi(\bar{x}, \bar{b}, \bar{B})$ of $\mathcal{L}$, with free variables $\bar{x} = (x_1, \dots, x_n)$ and parameters $\bar{b} = (b_1, \dots, b_j)$ and $\bar{B} = (B_1, \dots, B_k)$ drawn from $\mathcal{M}$. Say that $A = \{\bar{a} \in \mathbb{N}^n : \mathcal{M} \vDash \varphi(\bar{a}, \bar{b}, \bar{B})\}$. We show that $A \in M_n$.

Since the relation parameters $\bar{B}$ are drawn from $\mathcal{M}$, each set B_i has a finite support E_i $(i = 1, \dots, k)$. Let $E = \{b_1, \dots, b_j\} \cup E_1 \cup \cdots \cup E_k$. Clearly, E is finite. We show that E is a support for A.

Take any permutation $\pi : \mathbb{N} \to \mathbb{N}$ that fixes E pointwise, and take any $\bar{a} = (a_1, \dots, a_n) \in \mathbb{N}^n$. We check that $\bar{a} \in A \iff \pi(\bar{a}) = (\pi(a_1), \dots, \pi(a_n)) \in A$. Indeed,

$$\begin{aligned} \bar{a} \in A &\iff \mathcal{M} \vDash \varphi(\bar{a}, \bar{b}, \bar{B}) \\ &\iff \mathcal{M} \vDash \varphi(\pi(\bar{a}), \pi(\bar{b}), \pi(\bar{B})) \\ &\iff \mathcal{M} \vDash \varphi(\pi(\bar{a}), \bar{b}, \bar{B}) \\ &\iff \pi(\bar{a}) \in A. \end{aligned}$$

(Notation: $\pi(\bar{b}) = (\pi(b_1), \dots, \pi(b_j))$ and $\pi(\bar{B}) = (\pi(B_1), \dots, \pi(B_k))$. By Lemma 6.36, each $\pi(B_i)$ is a parameter from $\mathcal{M}$.) The second step works because permuting everything uniformly doesn't change any truth-values relative to any variable-assignment. This is easily proved by induction on formulas. The third step works because π fixes E pointwise, hence fixes all the parameters. □

LEMMA 6.39. *The Fraenkel model is a model of $\neg Fin(V)$.*

Proof. In fact, we will prove something stronger: the Fraenkel model does not contain any linear ordering of the universe.

Consider any relation $R \subseteq \mathbb{N}^2$ with finite support $E \subseteq \mathbb{N}$. Suppose for sake of contradiction that R is a linear ordering of the universe. Since R is total, we may choose distinct $a, b \in \mathbb{N} \setminus E$ such that Rab. Let π be any permutation fixing E such that $\pi(a) = b$ and $\pi(b) = a$. Since E is a support of R, it follows that Rba. But this contradicts the assumption that R is antisymmetric.

So, $\mathcal{M}$ contains no linear ordering of the universe. It follows that $\mathcal{M}$ contains no double well-ordering of the universe, i.e., $\mathcal{M} \vDash \neg Fin(V)$. □

We close this section by giving a simple characterization of symmetric sets.

LEMMA 6.40. *Let $E \subseteq N$ be a finite set. A set $A \subseteq N$ is symmetric with support E iff A is definable by Boolean combinations of equalities with parameters from E.*

Proof. Define an equivalence relation $\sim_E$ on $\mathbb{N}^n$, as follows:

$$\begin{aligned} (a_1, \dots, a_n) \sim_E (b_1, \dots, b_n) \iff & [(\forall i, j \leq n)(a_i = a_j \leftrightarrow b_i = b_j) \wedge \\ & (\forall e \in E)(\forall i \leq n)(a_i = e \leftrightarrow b_i = e)]. \end{aligned}$$

https://doi.org/10.1017/S1755020322000156 Published online by Cambridge University Press

In words: $\bar{a} \sim_E \bar{b}$ iff $\bar{a}$ and $\bar{b}$ are n-tuples with the same pattern of identity and distinctness which agree on members of E. It is easy to see that $\sim_E$ really is an equivalence relation.

($\Longrightarrow$). Suppose $A \subseteq \mathbb{N}^n$ is symmetric with support E. Observe that A is a union of equivalence classes of $\sim_E$. Indeed, if $\bar{a} \sim_E \bar{b}$, then there is a permutation $\pi : \mathbb{N} \to \mathbb{N}$ fixing E such that $\pi(\bar{a}) = \bar{b}$.

Now, each equivalence class of $\sim_E$ is definable by a Boolean combination of equalities with parameters from E, of the following form:

$$\bigwedge_{\substack{i,j \leq n \\ i \neq j}} (\neg)\, x_i = x_j \ \wedge \ \bigwedge_{\substack{i \leq n \\ e \in E}} (\neg)\, x_i = e.$$

(The parenthesized negations may or may not be present in each conjunct.) Furthermore, $\sim_E$ has only finitely many equivalence classes, because there are only finitely many possible patterns of identity and distinctness among $x_1, \ldots, x_n$ and the members of E. Hence, A is definable by a disjunction of formulas like the one above.

($\Longleftarrow$). Suppose A is definable by a Boolean combination of equalities with parameters from E. We show that A is symmetric with support E.

Take any permutation $\pi : \mathbb{N} \to \mathbb{N}$ fixing E pointwise. That is, for all $x_i, x_j \in \mathbb{N}$ and $e \in E$,

$$\begin{aligned} x_i = x_j &\leftrightarrow \pi(x_i) = \pi(x_j), \\ x_i = e &\leftrightarrow \pi(x_i) = e. \end{aligned}$$

By induction on formulas, it is easy to see that $\mathbb{N} \vDash \varphi(\bar{x}, \bar{e}) \leftrightarrow \varphi(\pi(\bar{x}), \bar{e})$ for any Boolean combination of equalities $\varphi(\bar{x}, \bar{e})$. Since A is defined by some such Boolean combination, it follows that $\pi(A) = A$.

Since π was arbitrary, we conclude that A is symmetric with support E. □

§7. The non-conservativeness of w2FA. In this section, we prove Theorem 7.47, which says that w2FA is not conservative over $Ax_{\mathcal{L}} + \neg Fin(V)$.

Here is the main idea of the proof. We have seen that $Ax_{\mathcal{L}} + \neg Fin(V)$ has a model whose relations are easy to describe in finitary terms (Section 6). Hence, $Ax_{\mathcal{L}} + \neg Fin(V)$ is a fairly weak theory; in fact it is mutually interpretable with first-order Peano arithmetic. (To show that $Ax_{\mathcal{L}} + \neg Fin(V)$ interprets PA, the trick is to code arithmetical statements as statements about finite concepts.) On the other hand, adding w2FA to $Ax_{\mathcal{L}} + \neg Fin(V)$ results in a much stronger theory, one which proves that the numerical sort is Dedekind-infinite and hence interprets second-order arithmetic. Second-order arithmetic is not conservative over Peano arithmetic. By means of a carefully chosen interpretation, this non-conservativeness can be transferred to the theories of interest to us. For example, w2FA + $\neg Fin(V)$ proves the interpretation of a consistency statement for Peano arithmetic, while $Ax_{\mathcal{L}} + \neg Fin(V)$ does not.

Let $X \approx Y$ abbreviate that there is a bijection between X and Y, in which case we say that X and Y are *equinumerous* concepts.

If Ryz is a binary relation, let R_y be the concept defined by $R_y z \leftrightarrow Ryz$. (This is terrible notation, but we only use it in the following definition.)

Definition 7.41. *Define Succ, Leq, Add, Mult as follows*:

$$\begin{aligned} \mathit{Succ}(X, Y) &\iff \exists a(\neg Xa \ \wedge \ Y \approx X \cup \{a\}), \\ \mathit{Leq}(X, Y) &\iff \exists X'(X \approx X' \ \wedge \ X' \subseteq Y), \\ \mathit{Add}(X, Y, Z) &\iff \exists Y'(Y \approx Y' \ \wedge \ X \cap Y' = \varnothing \ \wedge \ X \cup Y' \approx Z), \\ \mathit{Mult}(X, Y, Z) &\iff \exists R[\forall y \forall z(Ryz \to (Yy \wedge Zz)) \wedge \forall y(Yy \to R_y \approx X) \\ &\qquad \wedge \forall z(Zz \to \exists!\, yRyz)]. \end{aligned}$$

In other words, $\mathit{Mult}(X, Y, Z)$ says that Z is equinumerous with the union of $|Y|$ disjoint copies of X.

Definition 7.42. *Define the translation $\alpha : L_2 \to \mathcal{L}^+$ as follows.*

Identify first-order variables of L_2 with base-sort concept variables of $\mathcal{L}^+$. Identify second-order variables of L_2 with numerical-sort concept variables of $\mathcal{L}^+$.

Relativize $\forall x$ to the formula $\mathit{Fin}(X)$.

Relativize $\forall X$ to the formula $\mathit{FinNums}(\mathbf{X}) := \forall \mathbf{y}(\mathbf{Xy} \to \exists Y[\mathbf{y} = \# Y \wedge \mathit{Fin}(Y)])$.

Translate predication and equality as follows:

$$\begin{aligned} (Xy)^\alpha &:= \mathbf{X}(\# Y), \\ (x = y)^\alpha &:= X \approx Y. \end{aligned}$$

Translate $0, S, \leq, +, \cdot$ as follows:

$$\begin{aligned} (x = 0)^\alpha &:= X = \varnothing, \\ (Sx = y)^\alpha &:= \mathit{Succ}(X, Y), \\ (x \leq y)^\alpha &:= \mathit{Leq}(X, Y), \\ (x + y = z)^\alpha &:= \mathit{Add}(X, Y, Z), \\ (x \cdot y = z)^\alpha &:= \mathit{Mult}(X, Y, Z). \end{aligned}$$

We may extend this translation to all L_2-formulas via the usual techniques for eliminating definite descriptions. For example, write $SSx = y$ as $\exists z(Sx = z \wedge Sz = y)$, and so on.

Lemma 7.43. *Restricted to L_{PA}-formulas, the translation $\alpha : L_2 \to \mathcal{L}^+$ is an interpretation of* PA *in $Ax_{\mathcal{L}} + \neg \mathit{Fin}(V)$.*

Proof. Note that if φ is an L_{PA}-formula, then φ^α is an $\mathcal{L}$-formula. We will show that $Ax_{\mathcal{L}} + \neg \mathit{Fin}(V)$ proves the α-translation of each axiom of PA, and also proves that *Succ*, *Add*, *Mult* define total functions (up to $\approx$).

First we prove that *Succ* defines a total function (up to $\approx$). In other words, we show that for any Stäckel-finite concepts X, Y, Z,

$$\begin{aligned} &\exists W(\mathit{Fin}(W) \wedge \mathit{Succ}(X, W)), \\ &\mathit{Succ}(X, Y) \wedge \mathit{Succ}(X, Z) \to Y \approx Z. \end{aligned}$$

We reason in $Ax_{\mathcal{L}} + \neg \mathit{Fin}(V)$. For the first claim, take any concept X such that $\mathit{Fin}(X)$. Then X is not V. So, there exists a such that $\neg Xa$. Then $\mathit{Succ}(X, X \cup \{a\})$, and it is easy to check that $\mathit{Fin}(X \cup \{a\})$. This gives us the first claim. The second claim is obtained simply by unpacking the definition of *Succ*.

We postpone the proofs that *Add* and *Mult* define total functions (up to $\approx$).

The α-translations of the axioms of Q can be expressed as follows (after eliminating definite descriptions in a convenient way). For any Stäckel-finite concepts X, Y, Z, Y', Z',

$$\begin{aligned}
&\neg Succ(X, \varnothing),\\
&Succ(X, Z) \wedge Succ(Y, Z) \rightarrow X \approx Y,\\
&Add(X, \varnothing, Z) \leftrightarrow Z \approx X,\\
&Succ(Y, Y') \rightarrow (Add(X, Y', Z') \leftrightarrow \exists Z[Fin(Z) \wedge Add(X, Y, Z) \wedge Succ(Z, Z')]),\\
&Mult(X, \varnothing, Z) \leftrightarrow Z = \varnothing,\\
&Succ(Y, Y') \rightarrow (Mult(X, Y', Z') \leftrightarrow \exists Z[Fin(Z) \wedge Mult(X, Y, Z) \wedge Add(Z, X, Z')]),\\
&Leq(X, Y) \leftrightarrow \exists Z(Fin(Z) \wedge Add(Z, X, Y)).
\end{aligned}$$

(We drop the third axiom of Q, since it is redundant in PA.) It is tedious but straightforward to check that all of these claims are provable from $Ax_{\mathcal{L}} + \neg Fin(V)$.

The previous step essentially provides us with recursive definitions of *Add* and *Mult*. Using these recursive definitions, it is then easy to prove that *Add* and *Mult* define total functions (up to $\approx$). For *Add*, we must show that for any Stäckel-finite concepts X, Y, Z, W,

$$\begin{aligned}
&\exists U(Fin(U) \wedge Add(X, Y, U)),\\
&Add(X, Y, Z) \wedge Add(X, Y, W) \rightarrow Z \approx W.
\end{aligned}$$

Both of these claims are provable by induction on the finite concept Y (Lemma 5.32), using the recursive definition of *Add*. The proof for *Mult* is similar.

Lastly, the α-translation of the induction scheme of PA follows from induction on finite concepts (Lemma 5.32 again). □

LEMMA 7.44. *The translation* $\alpha : L_2 \rightarrow \mathcal{L}^+$ *is an interpretation of* Z_2 *in* w2FA + $\neg Fin(V)$.

Proof. By Lemma 7.43, we already know that the α-translation is an interpretation of PA in $Ax_{\mathcal{L}} + \neg Fin(V)$, and hence in w2FA $+ \neg Fin(V)$. It remains to check that w2FA $+ \neg Fin(V)$ proves the α-translations of the second-order induction and comprehension axioms.

The translation of the second-order induction axiom is equivalent to

$$\mathbf{X}(\#\varnothing) \wedge \forall X\left(Fin(X) \wedge \mathbf{X}(\#X) \wedge Succ(X, Y) \rightarrow \mathbf{X}(\#Y)\right) \rightarrow \forall X\left(Fin(X) \rightarrow \mathbf{X}(\#X)\right).$$

This is easily proved by induction on finite concepts, generalized to $\mathcal{L}^+$-formulas. The generalization is proved in the same way as Lemma 5.32.

The comprehension scheme translates as follows:

$$\exists \mathbf{X}\left(FinNums(\mathbf{X}) \wedge \forall Y\left(Fin(Y) \rightarrow \left(\mathbf{X}(\#Y) \leftrightarrow \varphi^{\alpha}(Y)\right)\right)\right).$$

To prove this in w2FA $+ \neg Fin(V)$, apply comprehension (in $\mathcal{L}^+$) to the formula

$$\exists Y\left(\mathbf{x} = \#Y \wedge Fin(Y) \wedge \varphi^{\alpha}(Y)\right).$$

Then use w2FA and the fact that $\approx$ is a congruence with respect to $\varphi^{\alpha}(Y)$. □

We will now define a translation $\beta : \mathcal{L} \rightarrow L_{\mathrm{PA}}$ inspired by the Fraenkel model, and show that it is an interpretation of $Ax_{\mathcal{L}} + \neg Fin(V)$ in PA.

Fix primitive recursive encodings of finite sets and sequences as natural numbers. For finite sequences, this amounts to specifying the following functions in L_{PA}:

(i) for each $n \in \mathbb{N}$, a primitive recursive function $\langle x_1, \dots, x_n \rangle$, which codes this tuple as a single number,
(ii) primitive recursive functions $length(s)$ and $(s)_i$, which return the length and the i-th element of the finite sequence coded by s.

We identify finite sets and sequences with their codes. We use the letter E for finite sets, and the letter s for finite sequences.

Fix a primitive recursive Gödel numbering of L_{PA}-formulas. We identify formulas with their Gödel numbers. For each formula φ, let $\ulcorner \varphi \urcorner$ be a formal numeral that denotes (the Gödel number of) φ.

Next, we describe L_{PA}-formulas $BoolEq$, $BoolSat$, pad_n representing certain primitive recursive relations and functions.

Let $BoolEq(x, y, E)$ just in case: x is a Boolean combination of L_{PA}-equalities with exactly y free variables and with constant symbols drawn from $\{S^e 0 : e \in E\}$.

Let $BoolSat(x, s)$ just in case: x is a Boolean combination of L_{PA}-equalities that is satisfied when the i-th variable of L_{PA} is assigned the value $(s)_i$, for all $i \leq length(s)$. This is primitive recursive, because truth and satisfaction for bounded (Σ_0) formulas are primitive recursive notions.

For each $n \in \mathbb{N}$, let $pad_n(x_1, \dots, x_n, y_1, \dots, y_n) = s$ just in case: s is the shortest finite sequence whose x_i-th element is y_i (for all $1 \leq i \leq n$) and whose other elements are all zero.

Definition 7.45. *Define the translation $\beta : \mathcal{L} \to L_{\mathrm{PA}}$ as follows.*

Let the variables of L_{PA} and the object variables of $\mathcal{L}$ be enumerated by $v_1, v_2, v_3, \dots$.

Translate each object variable v_i of $\mathcal{L}$ by the even-numbered variable v_{2i}. Translate each relation variable X of $\mathcal{L}$ by a distinct odd-numbered variable $v_X \in \{v_1, v_3, v_5, \dots\}$. In the last clause, E is a fresh variable and n is the arity of X.

$$\begin{aligned}
(Xv_{i_1} \cdots v_{i_n})^\beta &:= BoolSat(v_X, pad_n(S^{i_1}0, \dots, S^{i_n}0, v_{2i_1}, \dots, v_{2i_n})).\\
(v_i = v_j)^\beta &:= v_{2i} = v_{2j}.\\
(\varphi \to \psi)^\beta &:= \varphi^\beta \to \psi^\beta.\\
(\neg\varphi)^\beta &:= \neg\varphi^\beta.\\
(\forall v_i\ \varphi)^\beta &:= \forall v_{2i}\ \varphi^\beta.\\
(\forall X\ \varphi)^\beta &:= \forall v_X(\exists E\ BoolEq(v_X, S^n 0, E) \to \varphi^\beta).
\end{aligned}$$

Lemma 7.46. *The translation $\beta : \mathcal{L} \to L_{\mathrm{PA}}$ is an interpretation of $Ax_{\mathcal{L}} + \neg Fin(V)$ in* PA.

Proof. It is easy to check that the β-translation of any non-comprehension axiom is a theorem of first-order logic, and hence is provable in PA.[12] It remains to show that PA proves the β-translation of each comprehension axiom, and also that PA proves $(\neg Fin(V))^\beta$.

[12] In general, PA does not prove the β-translation of $\forall X\varphi(X) \to \varphi(Y)$. However, it is not these formulas that are axioms of $\mathcal{L}$, but rather the *closed* universal generalizations of such formulas. And PA does prove the latter.

The idea is to formalize the proofs of Lemmas 6.38, 6.39, and 6.40 in PA. The main obstacle is that we defined symmetric sets $A \subseteq \mathbb{N}^n$ in terms of arbitrary permutations of $\mathbb{N}$, and it is not obvious how to formalize those in PA. But in fact we do not need arbitrary permutations. Say that a permutation $\pi : \mathbb{N} \to \mathbb{N}$ is *essentially finite* if $\pi(a) = a$ for all but finitely many $a \in \mathbb{N}$. If we go through Section 6, replacing 'permutation' with 'essentially finite permutation' everywhere, we get exactly the same model, and all the proofs still work.

We formalize Lemma 6.40 as follows. Say that an L_{PA}-formula $\varphi(\bar{x})$ is *symmetric with support E* just in case, for every essentially finite permutation π,

$$(\forall e \in E)(\pi(e) = e) \implies \forall \bar{x}(\varphi(\bar{x}) \leftrightarrow \varphi(\pi(\bar{x}))).$$

Then we prove a theorem scheme in PA which says: 'An L_{PA}-formula is symmetric iff there is a Boolean combination of equalities coextensive with it.' More precisely, let $\varphi(v_{i_1}, \dots, v_{i_n})$ be any L_{PA}-formula with exactly the free variables displayed. Then PA proves the following: $\varphi(v_{i_1}, \dots, v_{i_n})$ is symmetric with support E iff there exists y such that

$$BoolEq(y, S^n 0, E) \wedge \forall \bar{x}(BoolSat(y, pad_n(S^{i_1}0, \dots, S^{i_n}0, \bar{x})) \leftrightarrow \varphi(\bar{x})).$$

($\Longrightarrow$). We reason in PA. Suppose that $\varphi(v_{i_1}, \dots, v_{i_n})$ is symmetric with support E. Let $\psi_1, \dots, \psi_m$ be all possible disjunctions of formulas of the form

$$\bigwedge_{\substack{j,k \le n \\ j \ne k}} (\neg)\, v_{i_j} = v_{i_k} \ \wedge \ \bigwedge_{\substack{j \le n \\ e \in E}} (\neg)\, v_{i_j} = S^e 0,$$

where parenthesized negations may or may not be present. Argue that $\bar{x} \sim_E \bar{y} \to (\varphi(\bar{x}) \leftrightarrow \varphi(\bar{y}))$, and hence

$$\forall \bar{x}(\varphi(\bar{x}) \leftrightarrow \psi_1(\bar{x})) \vee \cdots \vee \forall \bar{x}(\varphi(\bar{x}) \leftrightarrow \psi_m(\bar{x})).$$

Then observe that $\psi_i(\bar{x}) \leftrightarrow BoolSat(\ulcorner \psi_i \urcorner, pad_n(S^{i_1}0, \dots, S^{i_n}0, \bar{x}))$, for each $1 \le i \le m$.[13] Reasoning by cases, we are done.

For the ($\Longleftarrow$) direction, copy the rest of the proof of Lemma 6.40.

Next, we formalize Lemma 6.38. We replace $\mathcal{M} \vDash \varphi$ ('$\mathcal{M}$ satisfies φ') with φ^β throughout. For each $\mathcal{L}$-formula $\varphi(\bar{x}, \bar{y}, \bar{Y})$ not containing X free, we wish to show that PA proves

$$(\forall \bar{y} \forall \bar{Y} \exists X \forall \bar{x}[X\bar{x} \leftrightarrow \varphi(\bar{x}, \bar{y}, \bar{Y})])^\beta.$$

This basically says: 'There is a Boolean combination of equalities coextensive with $\varphi(\bar{x}, \bar{y}, \bar{Y})^\beta$.' By the formalized version of Lemma 6.40, it suffices to prove in PA that $\varphi(\bar{x}, \bar{y}, \bar{Y})^\beta$ is a symmetric L_{PA}-formula. To do this, use induction on $\mathcal{L}$-formulas $\varphi(\bar{x}, \bar{X})$ to prove the following theorem scheme in PA:

$$\pi \text{ is an essentially finite permutation} \to (\forall \bar{x} \forall \bar{X}[\varphi(\bar{x}, \bar{X}) \leftrightarrow \varphi(\pi(\bar{x}), \pi(\bar{X}))])^\beta.$$

(This corresponds to our earlier observation that permuting everything uniformly doesn't change any truth-values in $\mathcal{M}$ relative to any variable-assignment.) Then copy the rest of the proof of Lemma 6.38.

In the same way, it is easy to formalize Lemma 6.39 in PA. □

[13] This theorem scheme is provable in PA. See [19, p. 125, theorem 9.13] or [13, p. 56, I.1.70].

We are now ready to prove the first main theorem of the paper.

THEOREM 7.47. w2FA *is not conservative over* $Ax_{\mathcal{L}} + \neg Fin(V)$.

Proof. Let Con_{PA} denote a standard consistency statement for PA. We claim that $(Con_{\mathrm{PA}})^{\alpha}$ is a witness to non-conservativeness. That is,

$$Ax_{\mathcal{L}} + \neg Fin(V) \nvdash (Con_{\mathrm{PA}})^{\alpha}, \tag{1}$$

$$\mathrm{w2FA} + \neg Fin(V) \vdash (Con_{\mathrm{PA}})^{\alpha}. \tag{2}$$

Proof of claim (1). Write $\rhd$ for 'interprets'. From Lemmas 7.43 and 7.46, we have

$$\mathrm{PA} \;\rhd^{\beta}\; Ax_{\mathcal{L}} + \neg Fin(V) \;\rhd^{\alpha}\; \mathrm{PA}.$$

Suppose for a contradiction that $Ax_{\mathcal{L}} + \neg Fin(V) \vdash (Con_{\mathrm{PA}})^{\alpha}$. Then $\mathrm{PA} \vdash ((Con_{\mathrm{PA}})^{\alpha})^{\beta}$, and hence $\mathrm{PA} \rhd^{\beta\circ\alpha} \mathrm{PA} + Con_{\mathrm{PA}}$. However, by a strong version of Gödel's second incompleteness theorem, $\mathrm{PA} \not\rhd (\mathrm{PA} + Con_{\mathrm{PA}})$.[14] Contradiction.

Proof of claim (2). It is well known that $Z_2 \vdash Con_{\mathrm{PA}}$. Hence, by Lemma 7.44,

$$\mathrm{w2FA} + \neg Fin(V) \vdash (Con_{\mathrm{PA}})^{\alpha}.$$

□

COROLLARY 7.48. w2FA *is not conservative over* $Ax_{\mathcal{L}}$.

For proof, see Lemma 1.3.

COROLLARY 7.49. 2FA *is not conservative over* $Ax_{\mathcal{L}}$.

§8. w2FA is conservative over stronger base theories. It is surprising that w2FA is not conservative over $Ax_{\mathcal{L}}$. However, the next two theorems establish some limits to the non-conservativeness of w2FA.

THEOREM 8.50. w2FA *is conservative over third-order logic.*

Proof. Let $\mathcal{L}^3$ be the third-order analog of the base language $\mathcal{L}$. Let $Ax_{\mathcal{L}^3}$ denote the axioms of the deductive system for $\mathcal{L}^3$, including full third-order comprehension in the base sort. Note that w2FA still only includes second-order comprehension for the numerical sort.

Take any $\mathcal{L}^3$-formula φ, and suppose that $\mathrm{w2FA} + Ax_{\mathcal{L}^3} \vdash \varphi$. We show that $Ax_{\mathcal{L}^3} \vdash \varphi$. Our strategy is to define an interpretation of w2FA in $Ax_{\mathcal{L}^3}$ that leaves $\mathcal{L}^3$-sentences fixed (up to renaming of bound variables). Under such an interpretation, any derivation of φ from $\mathrm{w2FA} + Ax_{\mathcal{L}^3}$ is transformed into a derivation of φ from $Ax_{\mathcal{L}^3}$. The idea is to interpret each cardinality $\#X$ as the concept X from whence it came, with numerical-sort equality being interpreted as equinumerosity.

First, we define a pre-translation from variables of $\mathcal{L}^3 \cup \mathcal{L}^+$ into variables of $\mathcal{L}^3$. Translate each variable of sort τ as a variable of sort τ^*, where

[14] See [13, pp. 191–192, III.4.7–8]. The notation '$T \supseteq I\Sigma_1$' is explained at (p. 150, III.1.10). Hájek and Pudlák generally assume that equality is interpreted as equality (p. 149, II.1.5(2)). However, it is easy to adapt the proof of III.4.7–8 so as to dispense with this assumption. See also [20, p. 76, theorem 1] for more details.

https://doi.org/10.1017/S1755020322000156 Published online by Cambridge University Press

$$0^* := 0,$$
$$n^* := \langle 0 \rangle,$$
$$\langle \tau_1, \ldots, \tau_k \rangle^* := \langle \tau_1^*, \ldots, \tau_k^* \rangle.$$

In other words, τ^* is obtained from τ by replacing each occurrence of n with $\langle 0 \rangle$.

Set up the pre-translation so that distinct variables of $\mathcal{L}^3 \cup \mathcal{L}^+$ are translated as distinct variables of $\mathcal{L}^3$. For example, let the base-sort concept variables be enumerated by $X_0, X_1, X_2, \ldots$, and the numerical-sort object variables by $\mathbf{v}_0, \mathbf{v}_1, \mathbf{v}_2, \ldots$. Then let the pre-translations be

$$X_i^* := X_{2i},$$
$$\mathbf{v}_i^* := X_{2i+1}.$$

Similarly for other sorts.

We now define the translation $* : \mathcal{L}^3 \cup \mathcal{L}^+ \to \mathcal{L}^3$. In the first and last lines, let $\tau = \langle \tau_1, \ldots, \tau_k \rangle$ be any second- or third-order sort. In the last line, $Cong_{\approx}((X^\tau)^*)$ is a metalinguistic abbreviation of the statement: '$\approx$ is a congruence for the relevant argument-places of $(X^\tau)^*$', where the sort τ determines which argument-places are relevant.

$$(X^\tau x_1^{\tau_1} \cdots x_k^{\tau_k})^* := (X^\tau)^*(x_1^{\tau_1})^* \cdots (x_k^{\tau_k})^*.$$
$$(x = y)^* := x^* = y^*.$$
$$(\mathbf{x} = \mathbf{y})^* := \mathbf{x}^* \approx \mathbf{y}^*.$$
$$(\mathbf{x} = \#X)^* := \mathbf{x}^* \approx X^*.$$
$$(\varphi \to \psi)^* := \varphi^* \to \psi^*.$$
$$(\neg\varphi)^* := \neg\varphi^*.$$
$$(\forall x\, \varphi)^* = \forall x^*\, \varphi^*.$$
$$(\forall \mathbf{x}\, \varphi)^* = \forall \mathbf{x}^*\, \varphi^*.$$
$$(\forall X^\tau\, \varphi)^* = \begin{cases} \forall (X^\tau)^*\, \varphi^*, & \text{if } \tau \in Sorts^3(\{0\}), \\ \forall (X^\tau)^*(Cong_{\approx}((X^\tau)^*) \to \varphi^*), & \text{else.} \end{cases}$$

It is easy to check that the $*$-translation of each axiom of w2FA is provable from $Ax_{\mathcal{L}^3}$. So, the translation works. □

To prove the next theorem, we need another little fact about conservativeness.

LEMMA 8.51. *Let T be a theory in a formal language L, and let A be any L-sentence. Suppose that a sentence Δ is conservative over $T + A$ and is also conservative over $T + \neg A$. Then Δ is conservative over T.*

Proof. Take any $\varphi \in L$, and suppose that $T + \Delta \vdash \varphi$. We show that $T \vdash \varphi$. Indeed

$$T + A + \Delta \vdash \varphi,$$
$$T + A \vdash \varphi,$$
$$T \vdash A \to \varphi.$$

By the same reasoning, we also have $T \vdash \neg A \to \varphi$. Hence, $T \vdash \varphi$. □

THEOREM 8.52. w2FA *is conservative over $Ax_{\mathcal{L}} + Fin(V)$.*

Proof. Let $|V| = 1$ abbreviate the formula $\forall x \forall y\ x = y$. By Lemma 8.51, we may divide into cases according to whether $|V| = 1$ or $|V| \neq 1$. The rest of the proof is contained in Lemmas 8.53 and 8.54. $\square$

LEMMA 8.53. w2FA *is conservative over* $Ax_{\mathcal{L}} + Fin(V) + |V| \neq 1$.

Proof. We follow the same strategy as in Theorem 8.50. That is, we show how to define an interpretation † of w2FA in $Ax_{\mathcal{L}} + Fin(V) + |V| \neq 1$ that leaves $\mathcal{L}$-sentences fixed (up to renaming of bound variables). The idea is to interpret cardinalities $\#X$ as pairs of base-sort objects. Specifically, we will fix distinct base-sort objects a and b, represent $\#(V \restriction x)$ as (x, a), and represent $\#\varnothing$ as (a, b).

First, we define a pre-translation from variables of $\mathcal{L}^+$ into variables of $\mathcal{L}$. Translate each variable of sort τ as a distinct variable or pair of variables of sort(s) $\tau^\dagger$, where

$$
\begin{aligned}
0^\dagger &:= 0, \\
n^\dagger &:= 0, 0, \\
\langle \tau_1, \dots, \tau_k \rangle^\dagger &:= \langle \tau_1^\dagger, \dots, \tau_k^\dagger \rangle.
\end{aligned}
$$

For example, $\langle n, 0, n \rangle^\dagger = \langle 0, 0, 0, 0, 0 \rangle$ and $\langle \langle n \rangle, n \rangle^\dagger = \langle \langle 0, 0 \rangle, 0, 0 \rangle$.

Set up the pre-translation so that no variable of $\mathcal{L}$ is ever used twice. For definiteness, let the base-sort object variables be enumerated by $v_0, v_1, v_2, \dots$, and the numerical-sort object variables by $\mathbf{v}_0, \mathbf{v}_1, \mathbf{v}_2, \dots$. Then let the pre-translations of the object variables be

$$
\begin{aligned}
v_i^\dagger &:= v_{3i}, \\
\mathbf{v}_i^\dagger &:= v_{3i+1} v_{3i+2}.
\end{aligned}
$$

Similarly for second-order variables.

Now we define the interpretation $\dagger : \mathcal{L}^+ \to \mathcal{L}$. Fix a well-ordering $\leq$ of V, and fix distinct base-sort objects $a \neq b$. In the first and last lines, let $\tau = \langle \tau_1, \dots, \tau_k \rangle$ be any second-order sort.

$$
\begin{aligned}
(X^\tau x_1^{\tau_1} \cdots x_k^{\tau_k})^\dagger &:= (X^\tau)^\dagger (x_1^{\tau_1})^\dagger \cdots (x_k^{\tau_k})^\dagger. \\
(v_i = v_j)^\dagger &:= v_{3i} = v_{3j}. \\
(\mathbf{v}_i = \mathbf{v}_j)^\dagger &:= v_{3i+1} = v_{3j+1} \wedge v_{3i+2} = v_{3j+2}. \\
(\mathbf{v}_i = \#X)^\dagger &:= (X \approx (V \restriction v_{3i+1}) \wedge v_{3i+2} = a) \vee (X = \varnothing \wedge v_{3i+1} = a \wedge v_{3i+2} = b). \\
(\varphi \to \psi)^\dagger &:= \varphi^\dagger \to \psi^\dagger. \\
(\neg \varphi)^\dagger &:= \neg \varphi^\dagger. \\
(\forall v_i\ \varphi)^\dagger &= \forall v_{3i}\ \varphi^\dagger. \\
(\forall \mathbf{v}_i\ \varphi)^\dagger &= \forall v_{3i+1} \forall v_{3i+2}\ \varphi^\dagger. \\
(\forall X^\tau\ \varphi)^\dagger &= \forall (X^\tau)^\dagger\ \varphi^\dagger.
\end{aligned}
$$

In order to justify the interpretation of #, we must check that for each base concept X, there is a unique initial segment of $(V, \leq)$ that is equinumerous with X. For the existence claim, recall that $Ax_{\mathcal{L}}$ proves that any two well-orderings are comparable (Lemma 5.29). In particular, $(X, \leq)$ is order-isomorphic with a segment of $(V, \leq)$, and hence X is equinumerous with that segment. For the uniqueness claim, use the pigeonhole principle (Remark 5.31).

Now it is easy to check that the †-translation of each axiom of w2FA is provable from $Ax_{\mathcal{L}} + Fin(V) + |V| \neq 1$. So, the interpretation works. □

LEMMA 8.54. w2FA *is conservative over* $Ax_{\mathcal{L}} + |V| = 1$.

Proof. Observe that $Ax_{\mathcal{L}} + |V| = 1$ is a categorical theory, and hence it is a complete theory. So, the only way that w2FA could be non-conservative over $Ax_{\mathcal{L}} + |V| = 1$ is if the combined theory w2FA $+ |V| = 1$ were inconsistent. But w2FA $+ |V| = 1$ is consistent: it has a model $\mathcal{M}$ with object domains $M_0 = \{a\}$ and $M_n = \{0, 1\}$ and with $I(\#)$ being the function mapping each base-sort concept to its cardinality. □

§9. The non-conservativeness of 2FA. In the previous section, we established some limits to the non-conservativeness of w2FA. In this section, we will show that 2FA is more deeply non-conservative than w2FA. The main result is Theorem 9.67, which says that 2FA is non-conservative over $Ax_{\mathcal{L}} + Fin(V)$. Our proof of this result can be generalized to show that 2FA is non-conservative over pure axiomatic n-th order logic for any $n \geq 2$, or even over simple type theory.

Roughly, the idea is to construct a Gödel sentence for $Ax_{\mathcal{L}} + Fin(V)$. By a variation on Gödel's first incompleteness theorem, $Ax_{\mathcal{L}} + Fin(V)$ does not prove its own Gödel sentence. On the other hand, 2FA $+ Fin(V)$ does prove the Gödel sentence, because it is a powerful theory: it interprets second-order arithmetic in the new sort (and it is smart enough to relate that arithmetic to the Gödel sentence expressed in $\mathcal{L}$).

But $Ax_{\mathcal{L}} + Fin(V)$ says that the universe is finite, so it cannot interpret Q. How, then, is it possible to pull off the Gödel argument? The trick is that $Ax_{\mathcal{L}} + Fin(V)$ has arbitrarily large models. If $Ax_{\mathcal{L}} + Fin(V)$ proved its own Gödel sentence, then any sufficiently large model would contain a witness to the paradoxical derivation, yielding a contradiction.

To implement this argument, it will be convenient to work with a definitional extension $T = Ax_{\mathcal{L} \cup L'} + Fin(V) + \Delta$, which we now describe.

DEFINITION 9.55. *Let* $\mathcal{L} \cup L' := \mathcal{L}_{\{0\}}[\{0, S, \leq, A, M\}]$.

We identify variables of L' with object variables of $\mathcal{L}$. Thus,

- 0 is a base object constant,
- S and $\leq$ are constants of sort $\langle 0, 0\rangle$,
- A and M are constants of sort $\langle 0, 0, 0\rangle$.

Let $Ax_{\mathcal{L}\cup L'}$ be the axioms of the deductive system for $\mathcal{L} \cup L'$.

DEFINITION 9.56. *Let* Δ *be the conjunction of the following* $(\mathcal{L} \cup L')$*-formulas*:

1. $(V, \leq)$ *is a double well-ordering with least element* 0,
2. Sxy *iff* y *is the upper neighbor of* x *with respect to* $\leq$,
3. *Definitions of* A *and* M:

$$\begin{aligned}
&Ax0z \leftrightarrow z = x,\\
&Syy' \wedge Szz' \rightarrow (Axyz \leftrightarrow Axy'z'),\\
&Mx0z \leftrightarrow z = 0,\\
&Syy' \wedge Azxz' \rightarrow (Mxyz \leftrightarrow Mxy'z').
\end{aligned}$$

DEFINITION 9.57. *Let* $T = Ax_{\mathcal{L}\cup L'} + Fin(V) + \Delta$.

https://doi.org/10.1017/S1755020322000156 Published online by Cambridge University Press

LEMMA 9.58. $T \vdash BA'$.

Proof. It is obvious that T proves the universal closures of the first three axioms of BA'. Furthermore, since $(V, \leq)$ is a well-ordering, we have induction for all $(\mathcal{L} \cup L')$-formulas. Using induction, it is easy to prove the universal closures of the remaining axioms of BA'. □

We will now describe the construction of the Gödel sentence of T.

Fix a Gödel numbering of $\mathcal{L} \cup L'$. We describe L_{PA}-formulas Der_T, $diag$ representing certain primitive recursive notions.

Let $Der_T(x, y)$ just in case: x is the Gödel number of a T-derivation of a formula with Gödel number y.

Let $diag(x) = y$ be a function with the following property: if n is the Gödel number of an $(\mathcal{L} \cup L')$-formula $\theta(y)$ with exactly the free variable y, then

$$diag(S^n 0) = diag(\ulcorner \theta(y) \urcorner) = \ulcorner \forall y (y \doteq n \to \theta(y)) \urcorner.$$

(The notation $y \doteq n$ is from Definition 4.21.) Note that $diag$ is modeled on the Gödel diagonal function: in essence, it substitutes into a formula its own Gödel number.

It is well known that recursive relations are Δ_1-definable in PA [13, p. 18, theorem 0.45]. So, we may choose Der_T and $diag$ so that $Der_T(x, diag(y))$ is a Σ_1 formula. By Lemma 4.23, there is an equivalent Σ_1' formula $\varphi(x, y)$ of L' such that, for any parameters $a, b \in \mathbb{N}$,

$$\mathbb{N} \vDash \varphi(a, b) \iff \mathbb{N} \vDash Der_T(S^a 0, diag(S^b 0)).$$

Let p be the Gödel number of $\forall x \neg \varphi(x, y)$. Then $diag(S^p 0) = diag(\ulcorner \forall x \neg \varphi(x, y) \urcorner) = \ulcorner G \urcorner$, where G is the following sentence:

$$G := \forall y (y \doteq p \to \forall x \neg \varphi(x, y)).$$

We say that G is the *Gödel sentence* of the theory T.

LEMMA 9.59. *The theory* $T = Ax_{\mathcal{L} \cup L'} + Fin(V) + \Delta$ *does not prove its own Gödel sentence* G.

Proof. Suppose for sake of contradiction that $T \vdash G$. Let d be the Gödel number of a derivation of G. Then we have

$$\begin{aligned} &\mathbb{N} \vDash Der_T(S^d 0, diag(S^p 0)), \\ &\mathbb{N} \vDash \varphi(d, p). \end{aligned}$$

Write $\varphi(x, y)$ as $\exists z \psi(x, y, z)$, where ψ is bounded$'$. Fix $r \in \mathbb{N}$ such that $\mathbb{N} \vDash \psi(d, p, r)$. By Lemma 4.25 and the Generalization Theorem,

$$BA' \vdash \forall x \forall y \forall z (x \doteq d \wedge y \doteq p \wedge z \doteq r \to \psi(x, y, z)).$$

By Lemma 9.58,

$$T \vdash \forall x \forall y \forall z (x \doteq d \wedge y \doteq p \wedge z \doteq r \to \psi(x, y, z)).$$

It follows that

$$T \vdash \exists x \exists y \exists z(x \doteq d \wedge y \doteq p \wedge z \doteq r) \rightarrow \exists y(y \doteq p \wedge \exists x \exists z \psi(x, y, z)),$$
$$T \vdash \exists x \exists y \exists z(x \doteq d \wedge y \doteq p \wedge z \doteq r) \rightarrow \neg G.$$

We assumed that $T \vdash G$. Hence,

$$T \vdash \neg \exists x \exists y \exists z(x \doteq d \wedge y \doteq p \wedge z \doteq r).$$

But T has arbitrarily large finite models. In particular, $\mathbb{N} \upharpoonright \max\{d, p, r\}$ is a model of T that satisfies $\exists x \exists y \exists z(x \doteq d \wedge y \doteq p \wedge z \doteq r)$. Contradiction. □

Let us now turn our attention to what is provable in the stronger theory 2FA + *Fin*(V).

LEMMA 9.60. 2FA *interprets* Z_2, *and hence* Z_2'.

The proof is an easy variation on Frege's Theorem.

It will be convenient to fix a particular interpretation of Z_2 and Z_2' in the numerical sort of 2FA.

DEFINITION 9.61. *Fix a translation* $\gamma : L_2' \rightarrow \mathcal{L}^+$ *which interprets* Z_2' *in the numerical sort of* 2FA. *The interpretants of the nonlogical vocabulary items of* L_2' *will be denoted by* $\mathbf{0}$, $\mathbf{S}$, $\leq$, $\mathbf{A}$, $\mathbf{M}$. *The universe of the interpretation is defined by the following formula* $\mathbb{N}(\mathbf{x})$:

$$\forall \mathbf{X}(\mathbf{X0} \wedge \forall \mathbf{y} \forall \mathbf{z}(\mathbf{Xy} \wedge \mathbf{Syz} \rightarrow \mathbf{Xz}) \rightarrow \mathbf{Xx}).$$

Object quantifiers are relativized to $\mathbb{N}(\mathbf{x})$. *Set quantifiers are relativized to* $\forall \mathbf{x}(\mathbf{Xx} \rightarrow \mathbb{N}(\mathbf{x}))$.

The interpretation of Z_2 *in* 2FA *is obtained by extending the* γ*-translation so as to interpret* $Z_2' + \mathfrak{D}$, *where* $\mathfrak{D}$ *consists of the definitions of* $S, +, \cdot$ *in terms of* S, A, M (*Definition* 4.21).

The next two lemmas show that 2FA + *Fin*(V) is smart enough to relate the arithmetic in its base sort (BA') with the arithmetic in its numerical sort (Z_2').

To ease clutter, we will often write '2FA + *Fin*$(V) \vdash \Delta \rightarrow \cdots$' when we really mean 2FA + *Fin*$(V) \vdash \forall(0, S, \leq, A, M)(\Delta \rightarrow \cdots)$.

LEMMA 9.62. 2FA + *Fin*(V) *proves that the base universe is order-isomorphic with an initial segment of the natural numbers in the numerical sort*:

$$2\mathrm{FA} + \mathit{Fin}(V) \vdash \Delta \rightarrow \exists \mathbf{a}\left((V, \leq) \simeq_o (\mathbb{N} \upharpoonright \mathbf{a}, \leq)\right).$$

Proof. We reason in 2FA + *Fin*(V). Fix $0, S, \leq, A, M$, and suppose Δ. Then $(V, \leq)$ is a double well-ordering. Further, it is easy to show that $(\mathbb{N}, \leq)$ is a well-ordering.

By the comparability of well-orderings (Lemma 5.29, generalized to $\mathcal{L}^+$), exactly one of the following holds:

$$(V, \leq) <_o (\mathbb{N}, \leq), \quad (V, \leq) \simeq_o (\mathbb{N}, \leq), \quad (V, \leq) >_o (\mathbb{N}, \leq).$$

We can rule out the latter two options, because they imply that the converse of $(\mathbb{N}, \leq)$ is a well-ordering, which it isn't. Hence, $(V, \leq) <_o (\mathbb{N}, \leq)$. This is what we wanted. □

For the next definition, fix $0, S, \leq, A, M$, and suppose Δ. Also fix $\mathbf{a}$ as in the statement of Lemma 9.62.

Definition 9.63. *Let $\delta : L' \to \mathcal{L}^+$ be a translation which is exactly like γ, except that object quantifiers are relativized to $\mathbb{N} \upharpoonright \mathbf{a}$, and we restrict the translation to first-order formulas. In other words,*

$$\begin{aligned}
(x = 0)^\delta &:= \mathbf{x} = \mathbf{0},\\
(Sxy)^\delta &:= \mathbf{Sxy},\\
(x \leq y)^\delta &:= \mathbf{x} \leq \mathbf{y},\\
(Axyz)^\delta &:= \mathbf{Axyz},\\
(Mxyz)^\delta &:= \mathbf{Mxyz},\\
(x = y)^\delta &:= \mathbf{x} = \mathbf{y},\\
(\varphi \to \psi)^\delta &:= \varphi^\delta \to \psi^\delta,\\
(\neg\varphi)^\delta &:= \neg\varphi^\delta,\\
(\forall x\, \varphi)^\delta &:= \forall \mathbf{x}((\mathbb{N} \upharpoonright \mathbf{a})\mathbf{x} \to \varphi^\delta).
\end{aligned}$$

Lemma 9.64. *For any formula φ of L',*

$$2\text{FA} + \mathit{Fin}(V) \vdash \Delta \to (\varphi \leftrightarrow \varphi^\delta).$$

Proof. We reason in $2\text{FA} + \mathit{Fin}(V)$. Fix $0, S, \leq, A, M$, and suppose Δ. By Lemma 9.62, there is an order-isomorphism $f : V \to \mathbb{N} \upharpoonright \mathbf{a}$. In other words, there is a bijection f such that $f(0) = \mathbf{0}$ and

$$x \leq y \;\leftrightarrow\; f(x) \leq f(y).$$

We wish to prove corresponding statements for the other atomic formulas of L', namely,

$$\begin{aligned}
Sxy &\;\leftrightarrow\; \mathbf{S}f(x)f(y),\\
Axyz &\;\leftrightarrow\; \mathbf{A}f(x)f(y)f(z),\\
Mxyz &\;\leftrightarrow\; \mathbf{M}f(x)f(y)f(z).
\end{aligned}$$

The first statement holds because S is definable in terms of $\leq$:

$$\begin{aligned}
Sxy &\;\leftrightarrow\; y \text{ is the upper neighbor of } x \text{ with respect to } \leq\\
&\;\leftrightarrow\; \forall z((x \leq z \wedge x \neq z) \leftrightarrow y \leq z)\\
&\;\leftrightarrow\; \forall z((f(x) \leq f(z) \wedge f(x) \neq f(z)) \leftrightarrow f(y) \leq f(z))\\
&\;\leftrightarrow\; \forall \mathbf{z}((\mathbb{N} \upharpoonright \mathbf{a})\mathbf{z} \to [(f(x) \leq \mathbf{z} \wedge f(x) \neq \mathbf{z}) \leftrightarrow f(y) \leq \mathbf{z}])\\
&\;\leftrightarrow\; f(y) \text{ is the upper neighbor of } f(x) \text{ with respect to } \leq\\
&\;\leftrightarrow\; \mathbf{S}f(x)f(y).
\end{aligned}$$

The second statement holds because A and $\mathbf{A}$ satisfy the same recursive definition along their respective well-orderings (Definition 4.18). So, by the recursion theorem, A and $\mathbf{A}$ are isomorphic. (If they are not isomorphic, then consider a counterexample where y is $\leq$-minimal and derive a contradiction.)

The third statement holds for the same reason: M and $\mathbf{M}$ satisfy the same recursive definition along their respective well-orderings.

By induction on formulas, $\varphi \leftrightarrow \varphi^\delta$ for every L'-formula φ. □

In the next two lemmas, we show that 2FA formalizes the proof of Lemma 9.59.

Let $\varphi(x, y)$, $\psi(x, y, z)$, and G be the L'-formulas from Lemma 9.59.

Let $\mathbf{p}$ be a term in the numerical sort of $\mathcal{L}^+$ that denotes the Gödel number of $\forall x \neg\varphi(x, y)$. In other words, $\mathbf{p} = \ulcorner \forall x \neg\varphi(x, y) \urcorner$.

Let $\tilde{G}$ be the following formula in the numerical sort of $\mathcal{L}^+$:

$$\tilde{G} := \forall \mathbf{x} \neg\varphi^{\gamma}(\mathbf{x}, \mathbf{p}).$$

Observe that $2\mathrm{FA} \vdash \tilde{G} \leftrightarrow G^{\gamma}$. (This is because we chose the interpretations of Z_2 and Z_2' to be compatible with one another. See Definition 9.61.) Intuitively, $\tilde{G}$ says: 'The Gödel sentence for T is not derivable in T.' In other words, $\tilde{G}$ formalizes the statement of Lemma 9.59.

It is well known that Z_2 formalizes Tarskian definitions of truth and satisfaction for L_{PA} [31, pp. 183–187]. In the same way, 2FA formalizes Tarskian definitions of truth and satisfaction for L' with respect to the standard model $\mathbb{N}$. Denote the truth predicate by $Tr_{\mathbb{N}}(\mathbf{x})$ and the satisfaction predicate by $Sat_{\mathbb{N}}(\mathbf{x}, \mathbf{y})$.

Lemma 9.65. *Let θ be an L'-formula whose free variables are among the first k free variables of L'. Then* 2FA *proves*

$$\forall x_1 \cdots \forall x_k (Sat_{\mathbb{N}}(\ulcorner \theta \urcorner, \langle x_1, \dots, x_k \rangle) \leftrightarrow \theta^{\gamma}(x_1, \dots, x_k)).$$

For proof, compare [31, pp. 186–187, proposition 18.12].

Lemma 9.66. $2\mathrm{FA} \vdash \tilde{G}$.

Proof (sketch). The idea is to formalize the proof of Lemma 9.59 in 2FA.

We reason in 2FA. Suppose $\neg\tilde{G}$. Then there exists $\mathbf{d}$ such that $\varphi^{\gamma}(\mathbf{d}, \mathbf{p})$. By Lemma 9.65, we have $Sat_{\mathbb{N}}(\ulcorner \varphi \urcorner, \langle \mathbf{d}, \mathbf{p} \rangle)$. Write $\varphi(x, y) = \exists z \psi(x, y, z)$. Unpacking the definition of $Sat_{\mathbb{N}}$, there exists $\mathbf{r}$ such that $Sat_{\mathbb{N}}(\ulcorner \psi \urcorner, \langle \mathbf{d}, \mathbf{p}, \mathbf{r} \rangle)$.

Formalize Lemma 4.25 to obtain

$$\exists \mathbf{x}\, Der_{BA'}(\mathbf{x}, \ulcorner x \doteq d \wedge y \doteq p \wedge z \doteq r \rightarrow \psi(x, y, z) \urcorner),$$

and so on, until we reach

$$\exists \mathbf{x}\, Der_T(\mathbf{x}, \ulcorner \neg\exists x \exists y \exists z (x \doteq d \wedge y \doteq p \wedge z \doteq r) \urcorner).$$

Let $\mathbf{m} = \max\{\mathbf{d}, \mathbf{p}, \mathbf{r}\}$. Argue that Der_T is sound with respect to the semantics $Tr_{\mathbb{N} \restriction \mathbf{m}}$, in the sense that

$$\forall \mathbf{y} (\exists \mathbf{x}\, Der_T(\mathbf{x}, \mathbf{y}) \rightarrow Tr_{\mathbb{N} \restriction \mathbf{m}}(\mathbf{y})).$$

Finally, check that $\neg Tr_{\mathbb{N} \restriction \mathbf{m}}(\ulcorner \neg\exists x \exists y \exists z (x \doteq d \wedge y \doteq p \wedge z \doteq r) \urcorner)$. Contradiction. □

We are finally ready to prove the second main theorem of the paper.

Theorem 9.67. 2FA *is not conservative over $Ax_{\mathcal{L}} + Fin(V)$.*

Proof. We establish the following witness to non-conservativeness:

$$Ax_{\mathcal{L}} + Fin(V) \nvdash \forall (0, S, \leq, A, M)(\Delta \rightarrow G), \tag{3}$$

$$2\mathrm{FA} + Fin(V) \vdash \forall (0, S, \leq, A, M)(\Delta \rightarrow G). \tag{4}$$

https://doi.org/10.1017/S1755020322000156 Published online by Cambridge University Press

Proof of claim (3). Suppose not. Then we have

$$\begin{aligned} Ax_{\mathcal{L}} + Fin(V) &\vdash \forall(0, S, \leq, A, M)(\Delta \to G), \\ Ax_{\mathcal{L}\cup L'} + Fin(V) &\vdash \forall(0, S, \leq, A, M)(\Delta \to G), \\ Ax_{\mathcal{L}\cup L'} + Fin(V) &\vdash \Delta \to G, \\ Ax_{\mathcal{L}\cup L'} + Fin(V) + \Delta &\vdash G. \end{aligned}$$

But this contradicts Lemma 9.59.

Proof of claim (4). We reason in 2FA $+$ $Fin(V)$. Fix $0, S, \leq, A, M$, and suppose Δ. Also fix **a** as in the statement of Lemma 9.62. We show G.

By Lemma 9.66, we have $\tilde{G}$. Then we reason as follows:

$$\tilde{G} \implies G^{\gamma} \implies G^{\delta} \implies G.$$

The first arrow holds because we set up the interpretations of Z_2 and Z_2' correctly (Definition 9.61). The second arrow holds by quantificational logic, using the fact that G is Π_1'. (The idea is that universal formulas are preserved when passing to a submodel.) The third arrow holds by Lemma 9.64. Hence, we obtain G. □

By Lemma 1.3, this gives us another proof that 2FA is non-conservative over $Ax_{\mathcal{L}}$. By the same argument, we have:

Corollary 9.68. 2FA *is not conservative over pure axiomatic n-th order logic, for any* $n \geq 2$.

Corollary 9.69. 2FA *is not conservative over simple type theory.*

§10. HP is not deductively Field-conservative.

As we noted in the introduction, Hale and Wright hold that legitimate stipulative definitions need not be conservative in the standard deductive sense. They need only be Field-conservative, i.e., conservative over 'previously recognized ontology' [14, p. 133].

An *abstraction principle* is a purported implicit definition of a new operator @ by means of a sentence of the form

$$@F = @G \leftrightarrow \varphi(F, G),$$

where $\varphi(F, G)$ is an equivalence relation. In the special case of abstraction principles, Hale and Wright [14, p. 319, n. 21] adopt a precise formulation of Field-conservativeness, which we now describe.

For any formula φ, let $\varphi^{A(x)}$ denote the relativization of φ to the formula $A(x)$.[15] For any theory T, let $T^{A(x)} = \{\varphi^{A(x)} : \varphi \in T\}$.

Definition 10.70. *Let T be a theory in a formal language L. Let Δ be an abstraction principle introducing the new operator @, and let $L^+ = L \cup \{@\}$. Then Δ is Field-conservative over T if for every L-formula φ,*

$$T^{\neg\exists F(x=@F)} + \Delta \vDash \varphi^{\neg\exists F(x=@F)} \implies T \vDash \varphi.$$

[15] If φ is a second-order formula, then $\varphi^{A(x)}$ is the formula obtained from φ by replacing first-order quantifiers $\forall x(\cdots)$ with $\forall x(A(x) \to \cdots)$, and replacing second-order quantifiers $\forall X(---)$ with $\forall X(\forall x_1 \cdots \forall x_k(Xx_1 \cdots x_k \to A(x_1) \wedge \cdots \wedge A(x_k)) \to ---)$.

https://doi.org/10.1017/S1755020322000156 Published online by Cambridge University Press

If L is a second- or higher-order language, then $\vDash$ denotes the consequence relation with respect to standard (full) semantics.

There are two differences between Field-conservativeness and standard deductive conservativeness. Firstly, Field-conservativeness involves relativizing some of the quantifiers to 'non-abstracts'. Secondly, Field-conservativeness is formulated semantically rather than deductively.

Hale and Wright's suggestion, then, is that abstraction principles need only be Field-conservative in order to be acceptable. Much of the neo-Fregean literature has followed Hale and Wright on this point, if only because there seemed to be no other way for the neo-Fregean project to get off the ground.[16]

Following [33, pp. 21–22], we may distinguish some notions closely related to Field-conservativeness. See [6, 33] for motivation and further discussion.

Definition 10.71. *Let L, L^+, T, Δ be as in Definition 10.70. Assume that deductive systems for L and L^+ have been specified. Let P (for 'previously recognized ontology') be a new unary predicate symbol. Then*:

1. *Δ is deductively Field-conservative over T iff for every L-formula φ,*
$$T^{\neg\exists F(x=@F)} + \Delta \vdash \varphi^{\neg\exists F(x=@F)} \implies T \vdash \varphi.$$
2. *Δ is Caesar-neutral conservative over T iff for every L-formula φ,*
$$T^P + \Delta \vDash \varphi^P \implies T \vDash \varphi.$$
3. *Δ is deductively Caesar-neutral conservative over T iff for every L-formula φ,*
$$T^P + \Delta \vdash \varphi^P \implies T \vdash \varphi.$$

Weir [33, p. 24, theorem 4.1] proved that HP is both Field-conservative and Caesar-neutral conservative over pure second-order logic. It has remained an open question whether HP satisfies the deductive analogue of either of these conditions. Our results imply that it does not.[17]

Theorem 10.72. HP *is not deductively Caesar-neutral conservative over pure axiomatic second-order logic.*

Proof. We proved that 2FA is not deductively conservative over pure axiomatic second-order logic $Ax_{\mathcal{L}}$ (Corollary 7.49). Let θ be an $\mathcal{L}$-sentence such that $2\mathrm{FA} \vdash \theta$ but $Ax_{\mathcal{L}} \nvdash \theta$. Let P be a new unary predicate symbol. It suffices to show that $Ax_{\mathcal{L}[\{\#,P\}]} + \mathrm{HP} \vdash \theta^P$.

---

16 Field-conservativeness and related notions have been extensively studied by Shapiro and Weir [27], Weir [33], Linnebo [21], Cook [5], Cook and Linnebo [6], and others. These authors, along with Fine [9] and Heck [16], do not require acceptable abstraction principles to be conservative in the standard deductive sense. (Note that many of these authors do not regard acceptable abstraction principles as stipulative definitions. Some of them conceive of acceptable abstraction principles as analytic, or 'epistemically innocent', or definitions of a non-stipulative variety, or philosophically significant in other ways.) On the other hand, Burgess [3, pp. 158–161] raises some doubts about giving up standard deductive conservativeness.

17 We are grateful to an anonymous referee who pointed this out to us.

Actually, we show that $Ax_{\mathcal{L}[\{\#,P\}]} + \mathrm{HP} + \exists x Px \vdash \theta^P$. Since P is supposed to stand for 'previously recognized ontology', the hypothesis $\exists x Px$ merely reflects the fact that classical logic requires a nonempty domain. In any case, we can absorb the extra hypothesis by replacing θ with $\exists x(x = x) \to \theta$.[18]

Let us define a translation $\heartsuit$ from our two-sorted language $\mathcal{L}^+ = \mathcal{L}_{\{0,n\}}[\{\#_0, \#_n\}]$ into the one-sorted language $\mathcal{L}[\{\#, P\}]$. The idea is to relativize base-sort quantifiers to P and relativize numerical-sort quantifiers to $Num(x) := \exists F(x = \#F)$.

First we define a pre-translation from variables of $\mathcal{L}^+$ into variables of $\mathcal{L}[\{\#, P\}]$. (Compare Theorem 8.50.) Translate each variable of sort τ as a variable of sort $\tau^\heartsuit$, where $\tau^\heartsuit$ is obtained from τ by replacing each occurrence of n with 0. Set up the pre-translation so that distinct variables of $\mathcal{L}^+$ are translated as distinct variables of $\mathcal{L}[\{\#, P\}]$.

We now define the translation $\heartsuit : \mathcal{L}^+ \to \mathcal{L}[\{\#, P\}]$. Let j be any object sort, and let $\tau = \langle j_1, \dots, j_k\rangle$ be any second-order sort. Let $Num(x) := \exists F(x = \#F)$. Let A_j be the relativization predicate for sort j:

$$A_j(x) := \begin{cases} Px, & \text{if } j = 0, \\ Num(x), & \text{if } j = n. \end{cases}$$

Then the translation runs as follows:

$$\begin{aligned}
(X^\tau x_1^{j_1} \cdots x_k^{j_k})^\heartsuit &:= (X^\tau)^\heartsuit (x_1^{j_1})^\heartsuit \cdots (x_k^{j_k})^\heartsuit, \\
(x^j = y^j)^\heartsuit &:= (x^j)^\heartsuit = (y^j)^\heartsuit, \\
(\#_0 X)^\heartsuit &:= \#(X^\heartsuit), \\
(\#_n \mathbf{X})^\heartsuit &:= \#(\mathbf{X}^\heartsuit), \\
(\varphi \to \psi)^\heartsuit &:= \varphi^\heartsuit \to \psi^\heartsuit, \\
(\neg\varphi)^\heartsuit &:= \neg\varphi^\heartsuit, \\
(\forall x\, \varphi)^\heartsuit &= \forall x^\heartsuit (P(x^\heartsuit) \to \varphi^\heartsuit), \\
(\forall \mathbf{x}\, \varphi)^\heartsuit &= \forall \mathbf{x}^\heartsuit (Num(\mathbf{x}^\heartsuit) \to \varphi^\heartsuit), \\
(\forall X^\tau\, \varphi)^\heartsuit &= \forall (X^\tau)^\heartsuit ((X^\tau)^\heartsuit \subseteq A_{j_1} \times \cdots \times A_{j_k} \to \varphi^\heartsuit).
\end{aligned}$$

In other words, predication and equality are translated as themselves, both $\#_0$ and $\#_n$ are translated as $\#$, and quantifiers are relativized to P and Num in the natural way.

We wish to show $Ax_{\mathcal{L}[\{\#,P\}]} + \mathrm{HP} + \exists x Px \vdash \theta^P$. We have $2\mathrm{FA} \vdash \theta$. Applying the -translation, we obtain

$$2\mathrm{FA}^\heartsuit + \exists x Px + \exists x Num(x) + \forall X((X \subseteq P \vee X \subseteq Num) \to Num(\#X)) \vdash \theta^\heartsuit.$$

(The extra hypotheses serve to make the assumptions of our two-sorted notation explicit.) Notice that $\theta^\heartsuit$ is just θ^P. So, it suffices to show that $Ax_{\mathcal{L}[\{\#,P\}]} + \mathrm{HP} + \exists x Px$ proves all of the following:

---

[18] We could have added the extra hypothesis to the definition of deductive Caesar-neutral conservativeness, so that it said: for every L-formula φ,

$$T^P + \Delta + \exists x Px \vdash \varphi^P \implies T \vdash \varphi.$$

It is easy to verify that this alternative definition is equivalent to ours.

https://doi.org/10.1017/S1755020322000156 Published online by Cambridge University Press

- $2\mathrm{FA}^{\heartsuit}$,
- $\exists x Px$,
- $\exists x Num(x)$,
- $\forall X((X \subseteq P \vee X \subseteq Num) \rightarrow Num(\#X))$.

The second, third, and fourth bullets are obvious. For the first bullet, we have $2\mathrm{FA}^{\heartsuit} = (Ax_{\mathcal{L}^+})^{\heartsuit} + 2\mathrm{HP}^{\heartsuit}$. Now, $(Ax_{\mathcal{L}^+})^{\heartsuit}$ merely consists of relativizations of the logical axioms $Ax_{\mathcal{L}[\{\#,P\}]}$. These relativizations are all provable from $Ax_{\mathcal{L}[\{\#,P\}]} + \exists x Px + \exists x Num(x)$.

Similarly, $2\mathrm{HP}^{\heartsuit}$ merely consists of relativizations of HP, such as

$$(\forall F, G \subseteq P)(\#F = \#G \leftrightarrow (\exists R \subseteq P \times P)(F \approx_R G)^{\heartsuit}).$$

These are all provable from $Ax_{\mathcal{L}[\{\#,P\}]} + \mathrm{HP} + \exists x Px + \exists x Num(x)$. The proof is complete. □

COROLLARY 10.73. HP *is not deductively Field-conservative over pure axiomatic second-order logic.*

Proof. Set $Px := \neg Num(x)$ in the proof of the previous theorem. □

The upshot is that it makes a very great difference for the neo-Fregean program whether conservativeness requirements are formulated deductively or semantically. There seems to be no deductive criterion of conservativeness on which HP, or any similar principle, is conservative. As a matter of fact, neo-Fregeans have tended to prefer semantic notions of conservativeness anyway. But it would be desirable to see more philosophical justification for the use of these semantic notions, given that the deductive alternatives simply don't work.[19]

By the way, just as our deductive conservativeness results in the two-sorted setting could easily be transferred to the one-sorted setting, so too, Weir's semantic conservativeness results for HP can easily be transferred to the two-sorted setting. Say that T_1 is *semantically conservative* over T_0 if every standard model of T_0 can be expanded to a standard model of T_1. Then we have the following result:

THEOREM 10.74. 2FA *is semantically conservative over* $Ax_{\mathcal{L}}$.

Proof. Our argument is a simple adaptation of [33, p. 24, theorem 4.1].

Take any standard $\mathcal{L}$-structure $\mathcal{M}$, with object domain M_0. We will show how to expand $\mathcal{M}$ to a standard $\mathcal{L}^+$-structure $\mathcal{N}$ that satisfies 2HP.

To specify $\mathcal{N}$, we have to specify object domains N_0 and N_n, and an interpretation I of the constant symbols $\#_0, \#_n$.

Set $N_0 = M_0$.

Set $N_i = \kappa \cup \{\kappa\}$, where κ is the least infinite cardinal such that $\kappa \geq |N_0|$.

Set $N_\tau = \mathcal{P}(N_{j_1} \times \cdots \times N_{j_m})$ for all other sorts $\tau = \langle j_1, \ldots, j_m \rangle$.

We claim that the cardinality of any base concept $A \in N_{\langle 0 \rangle}$ is a member of the numerical universe N_n. Indeed, take any $A \in N_{\langle 0 \rangle}$. Then

$$A \subseteq N_0 \implies |A| \leq |N_0| \leq \kappa \implies |A| \in N_n.$$

[19] See [14, p. 133, n. 32] and [33, pp. 22–24] for some philosophical discussion of the matter. Semantic notions of conservativeness have also been studied in the literature on truth [4].

Further, we claim that the cardinality of any numerical concept $A \in N_{\langle n\rangle}$ is a member of the numerical universe N_n. Indeed, take any $A \in N_{\langle n\rangle}$. Then

$$A \subseteq N_n \implies |A| \leq |N_n| = \kappa \implies |A| \in N_n.$$

Let $I(\#_0)$ be the function $N_{\langle 0\rangle} \to N_n$ which maps each concept to its cardinality. Let $I(\#_n)$ be the function $N_{\langle n\rangle} \to N_n$ which maps each concept to its cardinality. Then $\mathcal{N}$ is a standard $\mathcal{L}^+$-structure satisfying 2HP, and hence 2FA. □

We conclude with some interesting open problems.

Problem 10.75. *Is* HP *conservative over DI* $(= \neg DFin(V))$?

Problem 10.76. *Is* w2FA *conservative over* $Ax_{\mathcal{L}} + \neg DFin(V)$?

Problem 10.77. *Is* 2FA *conservative over* $Ax_{\mathcal{L}} + \neg DFin(V)$?

Problem 10.78. *Is* 2FA *conservative over axiomatic third-order logic* $+ \neg Fin(V)$?

§11. Acknowledgements. Thanks to Sophia Arbeiter, John Burgess, Sean Ebels-Duggan, Vivian Feldblyum, Warren Goldfarb, Anil Gupta, Alexander Johnstone, Daniel Kaplan, Benjamin Marschall, Thomas Ricketts, James Shaw, Sean Walsh, and Daniel Webber for helpful discussion. We are also grateful to Jouko Väänänen for providing us with a copy of Asser's book, and to two anonymous referees for their constructive feedback.

BIBLIOGRAPHY

[1] Asser, G. (1981). *Einführung in die mathematische Logik, Teil III: Prädikatenlogik höherer Stufe*. Thun: Verlag Harri Deutsch.

[2] Boolos, G., & Heck, R. K. (1998). Die Grundlagen der Arithmetik, §§82–3. In Schirn, M., editor. *The Philosophy of Mathematics Today*. Oxford: Clarendon Press, pp. 407–428. Reprinted in Boolos, G. (1998). *Logic, Logic, and Logic*. Cambridge, MA: Harvard University Press, pp. 315–341.

[3] Burgess, J. P. (2005). *Fixing Frege*. Princeton: Princeton University Press.

[4] Cieśliński, C. (2017). *The Epistemic Lightness of Truth: Deflationism and its Logic*. Cambridge: Cambridge University Press.

[5] Cook, R. T. (2012). Conservativeness, stability, and abstraction. *British Journal for the Philosophy of Science*, **63**(3), 673–696.

[6] Cook, R. T., & Linnebo, Ø. (2018). Cardinality and acceptable abstraction. *Notre Dame Journal of Formal Logic*, **59**(1), 61–74.

[7] Ebels-Duggan, S. C. (2021). Deductive cardinality results and nuisance-like principles. *Review of Symbolic Logic*, **14**(3), 592–623.

[8] Enderton, H. B. (2001). *A Mathematical Introduction to Logic* (second edition). San Diego: Academic Press.

[9] Fine, K. (2002). *The Limits of Abstraction*. Oxford: Oxford University Press.

[10] Frege, G. (1884). *Die Grundlagen der Arithmetik: Eine logisch mathematische Untersuchung über den Begriff der Zahl*. Breslau: W. Koebner.

[11] ———. (1893). *Grundgesetze der Arithmetik*, Vol. 1. Jena: H. Pohle.

[12] ———. (1903). *Grundgesetze der Arithmetik*, Vol. 2. Jena: H. Pohle.

[13] Hájek, P., & Pudlák, P. (1993). *Metamathematics of First-Order Arithmetic*. Berlin: Springer.

[14] Hale, B., & Wright, C. (2001). *The Reason's Proper Study*: *Essays towards a Neo-Fregean Philosophy of Mathematics*. New York–Oxford: Clarendon Press.

[15] Heck, R. K., editor. (1997). The Julius Caesar objection. In *Language*, *Thought*, *and Logic*: *Essays in Honour of Michael Dummett*. Oxford: Oxford University Press, pp. 273–308.

[16] ———. (2011). *Frege's Theorem*. Oxford: Oxford University Press.

[17] ———. (2012). *Reading Frege's Grundgesetze*. Oxford: Clarendon Press.

[18] Jech, T. (2002). *Set Theory* (third millennium edition). New York: Springer.

[19] Kaye, R. (1991). *Models of Peano Arithmetic*. Oxford: Clarendon Press.

[20] Lindström, P. (2017). *Aspects of Incompleteness*, Lecture Notes in Logic, Vol. **10**. Cambridge: Cambridge University Press.

[21] Linnebo, Ø. (2011). Some criteria for acceptable abstraction. *Notre Dame Journal of Formal Logic*, **52**(3), 331–338.

[22] ———. (2016). Impredicativity in the neo-Fregean program. In Ebert, P., & Rossberg, M., editors. *Abstractionism*: *Essays in Philosophy of Mathematics*. Oxford: Oxford University Press.

[23] MacBride, F. (2003). Speaking with shadows: A study of neo-logicism. *British Journal for the Philosophy of Science*, **54**(1), 103–163.

[24] MacFarlane, J. (2009). Double vision: Two questions about the neo-Fregean program. *Synthese*, **170**(3), 443–456.

[25] Parsons, C. (1987). Developing arithmetic in set theory without infinity: Some historical remarks. *History and Philosophy of Logic*, **8**(2), 201–213.

[26] Shapiro, S. (1991). *Foundations without Foundationalism*: *A Case for Second-Order Logic*. New York–Oxford: Clarendon Press.

[27] Shapiro, S., & Weir, A. (1999). New V, ZF and abstraction. *Philosophia Mathematica*, **7**(3), 293–321.

[28] Simpson, S. G. (2009). *Subsystems of Second Order Arithmetic*. Cambridge: Cambridge University Press.

[29] Stäckel, P. (1907). Zu H. Webers elementarer Mengenlehre. *Jahresbericht der Deutschen Mathematiker-Vereinigung*, **16**, 425–428.

[30] Studd, J. P. (2016). Abstraction reconceived. *British Journal for the Philosophy of Science*, **67**(2), 579–615.

[31] Takeuti, G. (1987). *Proof Theory* (second edition). Amsterdam: North-Holland.

[32] Väänänen, J. (2021). Second-order and higher-order logic. In Zalta, E. N., editor. *The Stanford Encyclopedia of Philosophy* (Fall 2021 edition). Stanford: Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/entries/logic-higher-order/

[33] Weir, A. (2003). Neo-Fregeanism: An embarrassment of riches. *Notre Dame Journal of Formal Logic*, **44**(1), 13–48.

[34] Zermelo, E. (1909a). Sur les ensembles finis et le principe de l'induction complete. *Acta Mathematica*, **32**, 185–193.

[35] ———. (1909b). Über die Grundlagen der Arithmetik. In Castelnuovo, G., editor. *Atti del IV Congresso Internazionale Dei Matematici*, Vol. 2. Rome: Tipografia della R. Accademia dei Lincei, pp. 8–11.

DEPARTMENT OF PHILOSOPHY
UNIVERSITY OF PITTSBURGH
PITTSBURGH, PA, USA
E-mail: sgm32@pitt.edu

DEPARTMENT OF PHILOSOPHY
CARNEGIE MELLON UNIVERSITY
PITTSBURGH, PA, USA
E-mail: avigad@cmu.edu

https://doi.org/10.1017/S1755020322000156 Published online by Cambridge University Press