The Review of Symbolic Logic
Volume 0, Number 0, Month 2013

Logicism, Interpretability, and Knowledge of Arithmetic

Sean Walsh
Department of Logic and Philosophy of Science, University of California, Irvine

Abstract. A crucial part of the contemporary interest in logicism in the philosophy of mathematics resides in its idea that arithmetical knowledge may be based on logical knowledge. Here an implementation of this idea is considered that holds that knowledge of arithmetical principles may be based on two things: (i) knowledge of logical principles and (ii) knowledge that the arithmetical principles are representable in the logical principles. The notions of representation considered here are related to theory-based and structure-based notions of representation from contemporary mathematical logic. It is argued that the theory-based versions of such logicism are either too liberal (the plethora problem) or are committed to intuitively incorrect closure conditions (the consistency problem). Structure-based versions must on the other hand respond to a charge of begging the question (the circularity problem) or explain how one may have a knowledge of structure in advance of a knowledge of axioms (the signature problem). This discussion is significant because it gives us a better idea of what a notion of representation must look like if it is to aid in realizing some of the traditional epistemic aims of logicism in the philosophy of mathematics.

Received February 2013
Acknowledgements: This is material coming out of my dissertation, and I would thus like to thank my advisors, Dr. Michael Detlefsen and Dr. Peter Cholak, as well as my teachers, Dr. Patricia Blanchette and Dr. Timothy Bays, for their guidance with this essay. I would also like to thank, for their invaluable comments, Andrew Arana, Sharon Berry, Sébastien Gandon, Martin Fischer, Øystein Linnebo, Christopher Porter, Iulian Toader, and the anonymous referees. I received valuable feedback when presenting this work at the fifth Ideals of Proof Fellows Seminar at the Ecole Normale Supérieure on September 8, 2009, at FregeFest on February 26, 2010 at the University of California, Irvine, at Michael Potter's seminar at the University of Cambridge on May 12, 2011, at Arché at the University of St. Andrews on May 31, 2011, at the conference *Actualité du logicisme* on June 16, 2011, at the Northern Institute of Philosophy at the University of Aberdeen on May 28, 2012, and at the Philosophy Department Colloquium at the University of California, Davis on March 15, 2013. Thanks in particular to: Kai Wehmeier, Michael Potter, Toby Meadows, Sébestian Gandon, Brice Halimi, Roy Cook, Aaron Cotnoir, and Aldo Antonelli. Finally, for my graduate studies wherein this work began and the post-doc wherein this work was completed, I would like also to acknowledge the generous financial support of the Philosophy Department at Notre Dame, the Mathematics Department at Notre Dame, the Ahtna Heritage Foundation, the Deutscher Akademischer Austausch Dienst, the Georg-August Universität Göttingen, the National Science Foundation (under NSF Grants 02-45167, EMSW21-RTG-03-53748, EMSW21-RTG-0739007, and DMS-0800198), Øystein Linnebo's European Research Council-funded project Plurals, Predicates, and Paradox, the Philosophy Department at Birkbeck, University of London, the Alexander von Humboldt Stiftung TransCoop Program, and the Ideals of Proof Project, which in turn was funded and supported by Agence Nationale de la Recherche, Université Paris Diderot – Paris 7, Université Nancy 2, Collège de France, and Notre Dame.

doi:10.1017/S1755020300000000

§1. Introduction Epistemic variants of logicism in the philosophy of mathematics contend that knowledge of arithmetical principles may be based on knowledge of logical principles. Here a principle is said to be "logical" if it is epistemically akin to modus ponens: it is apriori, it is analytic, etc. Much of the recent discussion of logicism has centered around Crispin Wright's arguments that Hume's Principle –which states that two properties have the same cardinality if and only if they can be one-one correlated with each other –is a logical principle in this sense.[1] However, Wright and other logicists are ultimately interested in Hume's Principle because they think that knowledge of it can account for our knowledge of arithmetical principles such as the Peano axioms.[2] While these axioms are crucial to contemporary mathematics, contemporary philosophers of mathematics have had relatively little to say about the epistemic status of the Peano axioms. For instance, among the Peano axioms is the Mathematical Induction Axiom, which says that if zero has a property and if $n + 1$ has this property whenever n does, then all natural numbers have this property. In his recent book Charles Parsons says of this principle: "Writers on the foundations of arithmetic have found it difficult to state in a convincing way why the principle of mathematical induction is evident" (Parsons (2008) p. 264). So part of what is distinctive about logicism is that it is one of the few contemporary accounts that explicitly addresses the question of the evidence for mathematical induction.[3]

Let us call the *Logicist Template* the following schematic claim: knowledge of arithmetical principles may be based on knowledge of logical principles and the knowledge that these arithmetical principles can be represented within the logical principles. This claim is schematic in that it presupposes some antecedently specified notion of what it is for one set of principles to be *represented* within another set of principles. In contemporary mathematical logic, there are a number of notions of representation –often called *interpretations* in this tradition –that differ from one another both in terms of what and how they represent. Some of these notions are theory-based, wherein the idea is that one theory is representable within another if provability within the represented theory is matched by provability within the representing theory. Others of these notions are structure-based, where the idea is that the represented structure is isomorphic to a structure definable in the representing structure. In §2, the essentials of these notions of representation are briefly reviewed.

Theory-based and structure-based versions of the Logicist Template are respectively articulated and evaluated in §3 and §4. In regard to the theory-based versions of the Logicist Template (§3), my thesis is that they cannot exert an appropriate amount of control over the variety and scope of the propositions that are represented. The evidence for this lies in the plethora and consistency problems, which respectively show that too much would be counted as knowledge by this view or

[1] For Wright's arguments, see Wright (1999) pp. 7-15, Hale & Wright (2001) pp. 308-320. See Appendix §6.1. for a formal definition of Hume's Principle.

[2] Indeed, Wright even says that "[. . .] nothing can be essentially involved in the epistemology of number theory that is not involved in an understanding, and knowledge of the truth of Hume's Principle" (Wright (1998a) p. 366, Hale & Wright (2001) p. 255). For a formal definition of the Peano Axioms, see Appendix §6.1.

[3] Other extant accounts of the evidence for mathematical induction include Shapiro (2000b) pp. 109 ff and Leitgeb (2009) §3 pp. 273 ff.

that incompatible propositions would each be counted as knowledge by this view. While the plethora problem has been previously noted in the secondary literature on logicism by Hochberg, Blanchette, and Heck, the consistency problem has not been so noted. It is important to note both because there are theory-based versions of the Logicist Template which are immune to the plethora problem but not to the consistency problem.

In §4, I turn to versions of the Logicist Template centered around structure-based notions of representation. I begin with a version in §4.1 which additionally requires that the arithmetical theory be known to be definable within the logical theory. While this version is able to overcome the plethora and consistency problems, there is an objection, due to Papert, Parsons, and Boolos, to the effect that this version of the Logicist Template is circular because one of the claims recording the definability is too conceptually close to the Mathematical Induction Axiom itself. My presentation of the circularity problem is distinctive in that the conceptual proximity is here rendered as a provable equivalence of the definability claim and the Mathematical Induction Axiom across background knowledge. As I discuss, one way to avoid this circularity problem involves contending that the definability claims are all known on the basis of a knowledge of the meaning of their constitutive arithmetical and logical terms. However, as I argue, it seems reasonable to think that some but not all of these definability claims are so known. Finally, in §4.2, I turn to a structure-based notion of representation that weakens the definability claim to an isomorphism claim. The chief difficulty here is understanding what knowledge of structure amounts to in advance of knowledge of axioms, and in particular I argue that this knowledge cannot discriminate between various rival formal languages or signatures.

My overall conclusion in this paper is that both the theory-based and structure-based versions of the Logicist Template face deep problems, and hence that hitherto no satisfactory version of the Logicist Template has been presented which can secure the inference from knowledge of logical principles such as Hume's Principle to knowledge of arithmetical principles such as the Peano axioms. Nonetheless, it's worth highlighting that there are other variants of logicism that remained untouched by argumentation adduced here, such as the idea that arithmetic is maximally applicable because it "[. . .] can be accounted for on the basis of our general knowledge of principles of reasoning discoverable in every domain of inquiry (Demopoulos & Clark (2005) p. 138) and the idea that arithmetic results from a fictionalist "encoding" of finite cardinality quantifiers (Hodes (1990) p. 350). Thus this paper is directed only toward the epistemic strand of logicism which takes up Frege's idea that the Mathematical Induction Axiom is "based on general logical laws" and Crispin Wright's idea that Hume's Principle gives us a way to "apprehend the truth" of the Peano axioms (Frege (1980) p. iv, §80 p. 93, §108 p. 118, Frege (1967) p. 104, Wright (1983) p. xiv, p. 131).

§2. Brief Overview of the Interpretability of Theories and Structures

The goal of this section is to provide some brief background on notions of representation or "interpretation," as they are called in mathematical logic. There are notions of interpretability for theories and notions of interpretability for structures, and whereas the former are centered around proof, the latter are centered around definability. Structures and theories are both relative to *formal languages* or *signatures*,

and these are simply specifications of a class of constant symbols, relation symbols, and function symbols. Given a signature, a structure then is simply a set along with distinguished constants, relations, and functions on this set corresponding to the symbols from the signature.[4] Likewise, given a signature, a theory is simply a collection of sentences in this signature (cf. Marker (2002) chapter 1 or Enderton (2001) chapters 1-2).

An illustrative example of one structure being interpretable in another is the complex numbers and the real numbers: roughly, the field of complex numbers is interpretable in the field of real numbers because the complex numbers can be taken to be pairs of real numbers. One may make this notion of intepretability precise by first recalling the definitions of "a definable set" and "isomorphism." If M is a structure, then a subset X of M^n is *definable* in M if there is a first-order formula $\varphi(x_1, \ldots, x_n)$, perhaps containing parameters from M, such that the set X contains a tuple of elements from M if and only if M models that this tuple satisfies the formula. Further, two structures in the same signature are said to be *isomorphic* if there is a structure-preserving one-one map from the one onto the other. More formally, suppose that M and N are two structures in the same signature. Then M and N are said to be *isomorphic*, and one writes $M \cong N$, if there is a map f from M onto N such that M models $\varphi(a_1, \ldots, a_n)$ if and only if N models $\varphi(f(a_1), \ldots, f(a_n))$ for every formula $\varphi(x_1, \ldots, x_n)$ in their shared signature and every tuple of elements $a_1, \ldots, a_n$ from M, i.e.: $M \models \varphi(a_1, \ldots, a_n)$ if and only if $N \models \varphi(f(a_1), \ldots, f(a_n))$.[5] Then one says that a structure M is *definable* in a structure M^* if the domain, constants, relations, and functions of M are definable in M^*. (Here, a function is said to be definable if its graph is definable, and a constant is said to be definable if the singleton consisting just of it is definable). Finally, one says that a structure M is *interpretable* in a structure M^* if it is isomorphic to a structure that is definable in M^*.

There are two natural modifications of this definition of interpretability, the first of which pertains to equivalence relations and the second of which pertains to many-sorted settings. Part of what is distinctive about the example of the complex numbers given above is that each complex number is determined by a unique pair of real numbers. In other famous examples of interpretability in mathematics, this uniqueness clause is not always satisfied. For instance, in the case of the real projective plane, each triple of real numbers determines a point of the real projective plane, but for instance $(1, 1, 1)$ and $(\text{-}1, \text{-}1, \text{-}1)$ determine the same point of the real projective plane. In general, the indicator for when two triples (a, b, c) and (x, y, z) of real numbers determine the same point of the real projective plane is the relation E of "being on the same line through the origin":

$$(a, b, c)E(x, y, z) \iff \exists\, \lambda \neq 0\, [a = \lambda x \;\&\; b = \lambda y \;\&\; c = \lambda z] \tag{1}$$

This relation E on the set $\mathbb{R}^3$ is an equivalence relation: it is reflexive, symmetric, and transitive. Given any equivalence relation E on any set X, an equivalence class

[4] Sometimes these assignments of set-theoretic entities to the elements of a formal signature are called *interpretations*. I eschew this terminology here so that this notion is not conflated with the notion of interpretation which I define in the next paragraphs and which is the subject of this paper.

[5] See Marker (2002) Definition 1.1.3 pp. 8-9 and the proof of Theorem 1.1.10 p. 13, or Enderton (2001) p. 94 and the Homomorphism Theorem part (c) p. 96.

of E is a set of the form $[a]_E = \{b \in X : aEb\}$ and the set of all equivalence classes is written X/E. Suppose that X and E are definable in M^*. Then a relation $R^* \subseteq (X/E)^n$ is said to be *definable* in M^* if there is an M^*-definable relation R such that (i) $R(a_1, \ldots, a_n)$ and $a_1Eb_1, \ldots, a_nEb_n$ implies $R(b_1, \ldots, b_n)$ and (ii) $R^* = \{([a_1]_E, \ldots, [a_n]_E) : M^* \models R(a_1, \ldots, a_n)\}$. Then M is a *quotient structure definable* in M^* if the domain of M is X/E for some M^*-definable set X and some M^*-definable equivalence relation E and the constants, relations and functions of M are likewise definable in M^*. Given the ubiquity of quotient structures in contemporary mathematics, it is common to revise the definition of interpretability provided in the above paragraph to the following more expansive notion: a structure M is *interpretable* in a structure M^* if it is isomorphic to a structure that is a quotient structure definable in M^*. This is the sense in which the real projective plane is interpretable in real space: each point of the real projective plane can be taken to be equivalence classes of triples of real numbers.

The second modification of the notion of interpretability concerns many-sorted structures, that is structures which consist of a multitude of different domains connected in various ways. For instance, it's natural to formalize Euclidean geometry in a two-sorted manner, wherein one has separate domains for the points and the lines as well as a primitive incidence relation connecting the two domains. Likewise, in simple type theory, one works with a domain of objects, along with a second domain for its properties, a third domain for properties of those properties, etc., along with membership or predication relations connecting these domains. In general, given a many-sorted signature equipped with a collection of sorts, a many-sorted structure M in this signature consists of domains M_s for each sort s, along with distinguished constants, relations, and functions on the various M_s corresponding to symbols from the signature. Given a finite sequence of sorts $s_1, \ldots, s_k$, a subset X of $M_{s_1} \times \cdots \times M_{s_k}$ is said to *definable* if there is a formula $\varphi(x_1, \ldots, x_k)$ where x_i has sort s_i such that the set X contains a tuple of elements from $M_{s_1} \times \cdots \times M_{s_k}$ if and only if M models that this tuple satisfies the formula. Finally, a many-sorted structure M is said to be definable in another M^* if the domains, the constants, the relations, and the functions of M are definable in M^*; and the notion of a definable quotient structure is defined similarly. Finally, one says that a many-sorted structure M is *interpretable* in a many-sorted structure M^* if it is isomorphic to a structure which is definable in M^* (or to a quotient structure which is definable in M^*).[6]

[6] Standard references on many-sorted logic include Manzano (1996) Chapter 6 and Enderton (2001) §4.3 and Ebbinghaus (1985) pp. 27 ff. It should be mentioned that there is an alternative definition of interpretability for many-sorted structures. This is related to the natural operation $M \mapsto \mathrm{Flat}(M)$ that associates each many-sorted structure M to a one-sorted structure $\mathrm{Flat}(M)$ obtained by adding new unary predicates U_s for each of the sorts s and taking the domain of $\mathrm{Flat}(M)$ to be the union of the M_s and interpreting U_s by M_s (cf. Monk (1976) §29.9 p. 484, Manzano (1996) §6.8.2 p. 258, Enderton (2001) §4.3 p. 297). The alternative definition of interpretability of structures would then read: M is interpretable in M^* precisely when $\mathrm{Flat}(M)$ is interpretable in $\mathrm{Flat}(M^*)$. This alternative definition trivially implies the original definition, but is not implied by it (cf. Hook (1985)). Here is an example of a many-sorted structures M and M^* such that M is interpretable in M^* but $\mathrm{Flat}(M)$ is not interpretable in $\mathrm{Flat}(M^*)$. Let M be a many-sorted structure with sorts t_k for $k > 0$ such that M_{t_k} has exactly one element, and let M^* be a saturated elementary extension

Whereas the key role in the interpretability of structures is played by the notion of definability, the key role in the interpretability of theories is played by the notion of provability. In particular, one says that a theory T is *interpretable* in a theory T^* if the primitives of the interpreted theory T can be translated into formulas of the interpreting theory T^* so that the translation φ^* of every theorem φ of T is a theorem of T^*. That is, the key idea is that *the translation of theorems are theorems*, i.e.: if $T \vdash \varphi$ then $T^* \vdash \varphi^*$. For instance, the Zermelo-Fraenkel axioms for set theory interpret the Peano axioms for arithmetic because one can associate the arithmetical primitive "being a natural number" with the set-theoretic formula "being a finite ordinal," and likewise one can associate "$x < y$" with "$x \in y$," and "$x = 0$" with "$x = \emptyset$." One can then verify that the translations of arithmetical theorems are set-theoretic theorems, where the translation is given compositionally, so that the translation of a conjunct is the conjunct of the translations, i.e.: $(\varphi \wedge \psi)^* \equiv (\varphi^* \wedge \psi^*)$. For instance, it is a theorem of Peano arithmetic that no natural number is less than zero, and it is likewise a theorem of Zermelo-Fraenkel set theory that no finite ordinal is contained in the empty set.[7]

Let us now briefly take note of how to incorporate equivalence relations and many-sortedness into the concept of the interpretability of theories. For instance, just as the theory of the complex numbers is interpretable in the theory of the real numbers, so one would like a way of saying that the theory of the real projective plane was interpretable in the theory of the real numbers. Since interpretability of theories involves translating the primitives of the interpreted theory into formulas of the interpreting theory, the natural thought is to simply view the identity symbol itself as yet another primitive of the interpreted theory. Since identity is an equivalence relation and the idea of interpretability is that translations of theorems are theorems, one will necessarily translate the identity symbol by a formula which

---

of the structure $(\omega, 0, S)$ (cf. Marker (2002) p. 138 for the definition of saturation and p. 116 for the associated concept of a type). Then by interpreting M_{t_k} by the singleton $\{S^k(0)\}$, one has that M is interpretable in M^*. Suppose that $\mathrm{Flat}(M)$ was interpretable in $\mathrm{Flat}(M^*)$, which is just M^* with an additional unary predicate symbol for the domain. In particular, suppose that $\mathrm{Flat}(M)$ was isomorphic to the structure M' which was definable in M^* (resp. a definable quotient structure in M^*). Then the type $p(\overline{v}) = \{M'(\overline{v}) \wedge \neg U_{t_k}(\overline{v}) : k > 0\}$ is finitely realized in M^* and hence realized in M^*. But then M' contains an element which is not in the interpretation of any of the predicates U_{t_k} and hence M' cannot be isomorphic to $\mathrm{Flat}(M)$. The advantage of using the original definition of interpretability of many-sorted structures given in the above text, as opposed to the alternative definition described in this footnote, is that the alternative definition has no obvious correlate in the setting of theories.

[7] For a more formal presentation of the notion of interpretability for theories, see Lindström (2003) pp. 96-97 or Hájek & Pudlák (1998) pp. 148-149 or Visser (2006) §2.2. There is a very natural correspondence between the notion of interpretability for structures and the notion of interpretability for theories, at least when the semantics for these theories has a completeness theorem. For, in this case, a theory T is interpretable in a theory T^* if and only if every model M^* of T^* uniformly defines a model M of T, where the sense of uniformly is that the same formulas are used each time. These formulas may be allowed to include parameters from a certain parameter-free definable class in M^*, so long as one stipulates that any choice of parameters from this class effects such a definition of a model of T (cf. Hájek & Pudlák (1998) Definition 1.4 p. 149, Visser (2006) §B.3, Hodges (1993) Remark 5 p. 215). This natural correspondence also holds in the many-sorted setting, and in particular for the model-based notion defined in the previous paragraph and the theory-based notion defined in the subsequent paragraph.

satisfies the axioms of an equivalence relation (when appropriately relativized to a formula that serves as the translation of the domain). So for instance, in the case of the real projective plane, one translates the identity symbol governing identity of points by associating it with a formula defining the equivalence relation E from equation (1). Finally, one defines interpretability between many-sorted theories by associating each sort s of the signature of the interpreted theory to some finite sequence of sorts $\sigma_s = (\sigma_s(1), \ldots, \sigma_s(\ell_s))$ from the signature of the interpreting theory. One then requires that n-ary relations $R(x_1, \ldots, x_n)$ from the signature of the interpreted theory, wherein x_i has sort s_i, be associated to formulas from the signature of the interpreting theory of the following form, where variable $x_{i,j}$ has sort $\sigma_{s_i}(j)$:

$$\varphi(x_{1,1}, \ldots, x_{1,\ell_{s_1}}, \ldots x_{n,1}, \ldots, x_{n,\ell_{s_n}}) \tag{2}$$

One requires similar "typing discipline" in the case of the translation of constant and function symbols. While notationally more complex, this is the natural modification to the many-sorted setting of the fundamental idea that interpretations involve associating primitives of the interpreted theory to formulas of the interpreting theory in such a way that the compositionally defined map from sentences of the former to sentences of the latter preserves theoremhood.[8]

It is instructive to contrast the interpretability of theories to the faithful interpretability of theories. A theory T is said to be *faithfully interpretable* in a theory T^* if T is interpretable in T^* so that translations of theorems are theorems *and* so that translations of non-theorems are non-theorems, i.e.: $T \vdash \varphi$ if and only if $T^* \vdash \varphi^*$. It turns out that there are many examples of interpretations which are not faithful interpretations. For instance, the interpretation of Peano arithmetic in Zermelo-Fraenkel set theory given above is not a faithful interpretation because Peano arithmetic doesn't prove its own consistency, whereas Zermelo-Fraenkel set theory does prove a formal arithmetical sentence expressive of the consistency of Peano arithmetic. Finally, we can define the mutual interpretability of theories and the mutual faithful interpretability of theories. In particular, two theories are said to be *mutually interpretable* if each interprets the other, and two theories are said to be *mutually faithfully interpretable* if each faithfully interprets the other. Just as faithful interpretability implies interpretability, so we have that mutual faithful interpretability implies mutual interpretability.

Finally, let us briefly describe two stronger notions of similarity of theories and structures that are studied often in mathematical logic, namely the notion

[8] If the signature of the interpreted and interpreting theories are finite and presented in a primitive recursive manner, then interpretations are automatically primitive recursive functions from the signature of the interpreted theory to formulas in the signature of the interpreting theory. This fact is sometimes used in applying interpretations in the settings of the incompleteness theorems, since it allows one to treat the interpretation function itself as a defined notion within the theories (cf. Lindström (2003) p. 97). However, in any infinite signature there is no longer any guarantee of interpretation functions being primitive recursive or even computable. This is relevant to many-sorted theories because the signature of any infinitely-sorted theory is automatically infinite. Hence, interpretations between infinitely-sorted theories are not guaranteed to be computationally tractable.

of biinterpretability and sentential equivalence.[9] Suppose that V and W are two classes of structures, perhaps in different signatures. Then V and W are said to be *biinterpretable* if four things happen:

(3) Every structure M from V uniformly defines a structure $\Gamma(M)$ from W,
(4) Every structure N from W uniformly defines a structure $\Delta(N)$ from V,
(5) For every structure M from V there is a uniformly M-definable bijection $f_M : \Delta(\Gamma(M)) \to M$ that induces an isomorphism $\overline{f}_M : \Delta(\Gamma(M)) \to M$,
(6) For every structure N from W there is a uniformly N-definable bijection $g_N : \Gamma(\Delta(N)) \to N$ that induces an isomorphism $\overline{g}_N : \Gamma(\Delta(N)) \to N$.

Further, V and W are said to be *sententially equivalent* if one has (3)-(4) as well as the following:

(5′) Every structure M from V is elementarily equivalent to $\Delta(\Gamma(M))$,
(6′) Every structure N from W is elementarily equivalent to $\Gamma(\Delta(N))$.

wherein two structures are said to be *elementarily equivalent* if they satisfy the same first order sentences. Hence, since being isomorphic implies being elementarily equivalent, one has that being biinterpretable implies being sententially equivalent. Now, the notion of uniformity invoked in the above clauses is the natural one: an interpretation of one structure within another is effected by a series of formulas, and so insisting on uniformity is just to insist that the formulas one uses do not vary from structure to structure. It's natural also to modify clauses (3)-(4) to allow $\Gamma(M)$ and $\Delta(N)$ to be definable quotient structures in M and N respectively, rather than mere definable structures. Further, the definition of sentential equivalence carries over directly to the many-sorted setting, but to define biinterpretability in the many-sorted setting one modifies (5)-(6) as follows to accommodate the different sorts:

(5″) For every structure M from V and every sort s there is a uniformly M-definable bijection $f_{M_s} : (\Delta(\Gamma(M)))_s \to M_s$ which together induce an isomorphism $\overline{f}_M : \Delta(\Gamma(M)) \to M$,
(6″) For every structure N from W and every sort s there is a uniformly N-definable bijection $g_{N_s} : (\Gamma(\Delta(N)))_s \to N_s$ which together induce an isomorphism $\overline{g}_N : \Gamma(\Delta(N)) \to N$.

Finally, one says that two structures M and M^* are *biinterpretable* (resp. *sententially equivalent*) if the classes of isomorphic copies $\{N : N \cong M\}$ and $\{N : N \cong M^*\}$ are biinterpretable (resp. sententially equivalent), and one says that two theories T and T^* are *biinterpretable* (resp. *sententially equivalent*) if the classes of models $\{M : M \models T\}$ and $\{M^* : M^* \models T^*\}$ are biinterpretable (resp. sententially equivalent). It's easy to see how these notions generalize that of mutual interpretability of structures: whereas mutual interpretability of M and M^* merely requires that M can represent M^* and vice-versa, biinterpretability and sentential equivalence additionally require that M has the resources to confirm the accuracy of its representation of M^*'s representation of it, and additionally that M^* has similar resources going in the other direction.

[9] For biinterpretability, see Ahlbrandt & Ziegler (1986) p. 67, Hodges (1993) p. 222, Nies (2007) pp. 333-334, Visser (2006) § 3.3 p. 295, Enayat et al. (2011) p. 61. For sentential equivalence, see Visser (2006) § 3.3 p. 295, Enayat et al. (2011) p. 62.

§3. Theory-Based Versions: the Plethora and Consistency Problems As we saw in the previous section, the interpretability of theories is the most basic theory-based notion of representation to be found in mathematical logic. Hence, it is natural to first consider a theory-based version of the Logicist Template centered around the interpretability of theories. This version contends that knowledge of an arithmetical theory such as the Peano axioms may be based on knowledge of a logical theory such as Hume's Principle and the knowledge that this arithmetical theory is interpretable in this logical theory. Of course, it is demonstrable that Hume's Principle interprets the Peano axioms, and this result is now called Frege's Theorem.[10] Hence, if one does not dispute our knowledge of Hume's Principle, then what is at issue here is whether *in general* principles may come to be known by interpreting them in known principles.

Prior to discussing this issue, let me briefly note one place where a prominent logicist seems to endorse something very similar to this theory-based version of the Logicist Template. The following is a passage which Wright repeats verbatim in two different essays:

> The neo-Fregean thesis about arithmetic is that a knowledge of its fundamental laws (essentially, the Dedekind-Peano axioms)– and hence of the existence of a range of objects which satisfy them– may be based on Hume's Principle as an explanation of the concept of cardinal number in general, and finite cardinal number in particular. More specifically, the thesis involves four ingredient claims: [¶] (i) that the vocabulary of higher-order logic plus the cardinality operator, octothorpe [#] or 'Nx: ...x...', provides a sufficient definitional basis for a statement of the basic laws of arithmetic; [¶] (ii) that when they are so stated, Hume's Principle provides for a derivation of those laws within higher-order logic [...] (Wright (1998b) p. 389, Wright (1999) p. 17, Hale & Wright (2001) pp. 256, 321).

It seems to me that the key idea expressed in these two roman numerals is that (i) there is a way of translating arithmetical primitives into formulas about cardinalities, and that (ii) all the axioms of Peano arithmetic become theorems of Hume's Principle when so translated. This, of course, implies that the translation of theorems of Peano arithmetic are theorems of Hume's Principle, which by definition is what it means for the Peano axioms to be interpretable in Hume's Principle. Hence, it seems that what Wright is here suggesting is that knowledge of the Peano axioms may be based on knowledge of Hume's Principle because the Peano axioms are interpretable in Hume's Principle.

One problem with this version of the Logicist Template, which I will call the *plethora problem*, has been voiced in different ways by Hochberg, Blanchette and Heck, among others.[11] The plethora problem stems from the fact that many theories

[10] See Wright (1983) pp. 154-169, Boolos (1996).

[11] Hochberg (1970) p. 396, Hochberg (1984) p. 321, Hochberg (1956) p. 119, Blanchette (1994) p. 95, Blanchette (2012) §4.3 pp. 84-85, and Heck (1999) p. 59, Heck (2000) p. 188, Heck (2011) p. 156, Heck (1997) p. 597, Cook (2007) p. 69, Heck (2011) p. 245, and Shapiro (2000a) §7 pp. 360-361, Cook (2007) §7 p. 250-251, Wright (2000) p. 323, Cook (2007) p. 261.

are interpretable in the Peano axioms. For instance, it is well-known from the work of Tarski that the complete first-order theory of the real and complex numbers are interpretable in the Peano axioms.[12] However, it would seem strange to suggest that these theories can come to be known by way of an interpretability result. For instance, some of the axioms of the complex numbers express the Fundamental Theorem of Algebra, which asserts that every non-zero polynomial with complex coefficients in one variable has a complex root. Now one might simply take the content of the Fundamental Theorem as epistemically basic, that is, as not derived from other more basic knowledge. But most mathematicians do not do this: rather, they require proof of this theorem, and the proofs of this theorem that they accept and teach to their students are all non-trivial, and typically require appeal to limits or to topological notions, each of which must be studied in its own right before one can begin to understand these proofs of the Fundamental Theorem of Algebra. It would seem counterintuitive to suggest that all of this could be circumvented by appeal to a comparatively elementary interpretability result. Hence, the *plethora problem* is that too much knowledge may be generated by the supposition that knowledge of one theory can be based on knowledge of a theory which interprets it.

One way to avoid the plethora problem is to strengthen the notion of interpretation in a philosophically well-motivated way, and one obvious idea is to focus on the strongest notion of interpretation that is provided by the traditional proof of Frege's Theorem. For, this proof actually establishes that the Peano axioms are faithfully interpretable in Hume's Principle, and moreover in such a way that the canonical singular terms of the Peano axioms are made inferentially indistinguishable with canonical singular terms of "applied arithmetic" coming from Hume's Principle. More precisely, consider the terms in the signature of the Peano axioms consisting of zero, the successor of zero, the successor of the successor of zero etc. These may be defined recursively (in the metalanguage) as follows:

$$\underline{0} = 0, \qquad \underline{n+1} = S(\underline{n}) \tag{3}$$

Correspondingly, consider the following terms $\overline{n}$ from the signature of Hume's Principle, where $\#$ is used to denote "the number of" or "the cardinality of":

$$\overline{0} = \#\{z : z \neq z\}, \qquad \overline{n+1} \equiv \#\{z : \bigvee_{\ell=0}^{n} z = \overline{\ell}\} \tag{4}$$

[12] See Appendix 6.1. for a formal statement of the Peano axioms. The easiest way to see that the complete first-order theory of the real and complex numbers are interpretable in the Peano axioms is to note three things. First, by the work of Tarski, the complete theories of the real and complex numbers are complete and recursively axiomatizable (cf. Marker (2002) Corollary 3.2.3 p. 85 and Corollary 3.3.16 p. 97). Second, by formalizing Henkin's proof of the completeness theorem, one can show that if the Peano axioms prove the consistency of a recursively axiomatizable theory, then they interpret that theory (cf. Lindström (2003) Theorem 4 p. 99 and Hájek & Pudlák (1998) Theorem 2.39 p. 169). Third, it is easy to show by a model construction within Peano arithmetic that the Peano axioms prove the consistency of the recursively axiomatizable fragments of the complete theories of the real and complex numbers (cf. Simpson (2009) Theorem II.9.4 p. 97 and Theorem II.9.7 p. 98).

These terms embody important principles of applied arithmetic because Hume's Principle allows one to prove the following, which has been called "Nq":

$$\forall\, F\ (\#F = \overline{n} \leftrightarrow \exists^{=n}\, x\ Fx) \tag{5}$$

Moreover, in their "pure" uses, the terms $\underline{n}$ are inferentially indistinguishable from $\overline{n}$, in that one can show that for every formula $\varphi(x_1, \ldots, x_k)$ in the signature of the Peano axioms and every $n_1, \ldots, n_k \geq 0$, one has that the Peano axioms prove $\varphi(\underline{n}_1, \ldots, \underline{n}_k)$ if and only if Hume's Principle proves that $\varphi^*(\overline{n}_1, \ldots, \overline{n}_k)$, where $\varphi \mapsto \varphi^*$ is the compositional map from formulas in the signature of the Peano axioms to formulas in the signature of Hume's Principle induced by the usual proof of Frege's Theorem. Of course, this inferential indistinguishability is a natural strengthening of the notion of faithful interpretations discussed in the previous section.

The applied character of the terms $\overline{n}$ expressed in Nq in equation (5) and the inferential indistinguishability of $\overline{n}$ and $\underline{n}$ has been thought by Wright and Hale to provide a crucial constraint on interpretations. Hale writes: "But why should we not regard $N^=$ [HUME'S PRINCIPLE] and [NQ], together with the definitions of the individual numerals [i.e. $\overline{n}$], as serving simultaneously to introduce the use of numerals and terms of the form $Nx : Fx$ [i.e. #F]?" (Hale (1987) p. 224). Indeed, it seems like these constraints do serve to circumvent the plethora problem. For, Wright and Hale have in effect suggested that knowledge is preserved under known interpretability in known premises provided that both of the following conditions hold: (i) the introduced singular terms and concepts are inferentially indistinguishable from a range of terms and concepts already available via the known premises, and (ii) the applied uses of the introduced concepts and singular terms are similarly made available via the known premises.[13] Further, this concern with applications is not idiosyncratic to the natural numbers, but extends to an account of our knowledge of other basic mathematical structures like the reals and Euclidean geometry. For instance, Hale writes that the "insistence that reals be defined as ratios of quantities derives from [the] belief that the *application* of reals as measures of quantities is *essential* to their very nature" (Hale (2000) p. 104). So Wright and Hale might respond to the plethora problem by suggesting that while the aforementioned work of Tarski provides for an interpretation of the theory of the real numbers inside Hume's Principle, it does nothing to secure the applications of real numbers.

However, while this response to the plethora problem seems well-motivated, there is another problem – which I call the *consistency problem* – which besets this and other theory-based versions of the Logicist Template. This problem is that these versions of the Logicist Template can provide us with knowledge of both a sentence and its negation. To illustrate this, let us call *the anti-Peano axioms* the Peano axioms but with the Mathematical Induction Axiom replaced by its *negation*. It then turns out that axioms containing the anti-Peano axioms are faithfully interpretable in Hume's Principle, and moreover in such a way that the terms $\overline{n}$

[13] One might distinguish between two different senses of application. First, Nq in equation (5) might be viewed as an application because the right-hand side does not contain any non-logical vocabulary. Second, Nq might be viewed as an application because of the functional character of the "number of" operator on the left-hand side. For discussion of this second aspect, see Wright (2000) p. 325, Cook (2007) p. 263.

and $\underline{n}$ continue to be inferentially indistinguishable. So as to not distract from the philosophical point, the proof of this result is deferred until Appendix 6.2. In short, the philosophical point here is this: if knowledge of the Peano axioms may be based on Hume's Principle because the Peano axioms are interpretable in Hume's Principle in such a way that the conditions (i)-(ii) from the previous paragraph are met, then if the anti-Peano axioms are likewise interpretable in Hume's Principle in a way that meets conditions (i)-(ii), then presumably knowledge of the anti-Peano axioms could be based on knowledge of Hume's Principle. However, presumably it is absurd to suggest that both the Mathematical Induction Axiom and its negation can be known.[14]

While examples of $T + \neg\varphi$ interpretable in $T + \varphi$ have *not* been previously discussed in the secondary literature on logicism, they and examples of mutually interpretable $T+\varphi$ and $T+\neg\varphi$ have been discussed in other parts of the philosophy of mathematics. For instance, Edward Nelson had the idea of characterizing a highly constructive theory of arithmetic, which he called "predicative arithmetic," as the collection of all those sentences φ of arithmetic such that $Q + \varphi$ was interpretable in Q, where Q denotes a weak base theory of arithmetic called Robsinson's Q (cf. Appendix §6.1. for the definition of this theory). Nelson then immediately noted and posed the problem of determining whether the conjunction of two sentences have this property whenever the two sentences themselves individually have this property (Nelson (1986) p. 63). But later work of Visser and Kalsbeek showed that both a sentence and its negation could have this property. As Kalsbeek notes, this results suggests that "the putative definition of predicative arithmetic is incoherent" (Kalsbeek (1989) p. 63, cf. Iwan (2000) p. 151, Buss (2006) p. 194).

In the philosophy of set theory, Peter Koellner has described a solution to the consistency problem, but there is no obvious analogue of Koellner's solution available to the logicist. Koellner draws attention to the Guaspari-Lindström theorem, which implies that two extensions of the Zermelo-Fraenkel axioms by finitely many new axioms in the same signature are mutually interpretable if and only if they prove exactly the same Π^0_1-sentences (cf. Lindström (2003) Theorem 6 pp. 103, 115). Here a Π^0_1-sentence is simply a sentence which begins with a universal quantifier over natural numbers and all of whose other quantifiers are bounded. Koellner's idea is that the Π^0_1-sentences are exactly the observational sentences, so that the Guispari-Lindström theorem implies that while two mutually interpretable set theories may disagree vastly about the nature of sets, they must of necessity have the same observational consequences (Koellner (2009) p. 98). However, there does not seem

[14] Another problem that ought not be confused with the consistency problem concerns Feferman's classical result that the Peano axioms plus the *negation* of the traditional consistency statement for these axioms is interpretable in these axioms themselves (cf. Lindström (2003) Theorem 8 p. 104, Feferman (1961) Theorem 6.6 p. 76). Considerations of space prevent me from fully discussing this problem, but suffice it to say that I do not view it as posing an incontrovertible problem for theory-based versions of the Logicist Template. It's perhaps also worth mentioning that the consistency problem is not the problem of multiple interpretations, such as the interpretation of the Peano axioms into set theory using the finite ordinals $0, 1, 2, \ldots, n, \ldots$ as opposed to the infinite ordinals $\omega + 0, \omega + 1, \omega + 2, \ldots, \omega + n, \ldots$. Rather, the consistency problem is that a given theory– like set theory– can interpret both $T + \varphi$ and $T + \neg\varphi$, so that interpreting a theory in a known theory like set theory cannot possibly be a way to acquire knowledge.

to be any analogue of Koellner's idea which is available to the logicist. For not only is the logicist interested in interpretations between theories in different signatures, but the sentences that the logicist is interested in are not Π^0_1-sentences, and so would not be covered by the Guaspari-Lindström theorem in the first place.[15]

In formal theories of truth, Martin Fischer has articulated a position which bears a certain family resemblance to the Logicist Template, and it seems that the consistency problem is also a problem for unqualified versions of Fischer's view. Fischer has suggested that if a theory of truth is mutually faithfully interpretable with its base theory of arithmetic, then the truth predicate and its arithmetical translate possess the same inferential role.[16] But, it turns out that there are examples of theories $T + \varphi$ and $T + \neg\varphi$ that are mutually faithfully interpretable.[17] It seems that these examples require qualification of Fischer's thesis. For, suppose that the only primitive appearing in T and φ is the relation symbol R, and that $T + \varphi$ is mutually faithfully interpretable in $T + \neg\varphi$ in such a way that R is translated by formula R^*. Then Fischer's suggestion would imply that R and R^* possess the same inferential role, and hence presumably that the sentence φ and its translation φ^* induced by translating R as R^* would possess the same meaning. Since $T + \neg\varphi$ proves φ^*, any agent who had reason to believe $T + \neg\varphi$ would thus have reason to believe φ^* and hence also its synonym φ. Thus taken in its most unqualified form, Fischer's suggestion would require us to impute having reason to believe $\varphi \wedge \neg\varphi$ to an agent merely in virtue of her belief in $T + \neg\varphi$.

The most promising avenue of response for both Fischer and the advocate of the Logicist Template is to seek out stronger theory-based notions of representation and to show that the consistency problem does not arise for these theories. One of the strongest theory-based notions is the notion of "definitional equivalence" or "logical synonymy" which is sometimes said to explicate the intuitive idea of two theories being mere "notational variants" of one another (Corcoran (1980) p. 232) or having the "same import" or being "various descriptions amount[ing] to the same thing" (de Bouvère (1965b) p. 622, cf. de Bouvère (1965a)). The motivating examples of this phenomena were: (i) taking $<$ as primitive in a linear order and defining $\leq$, as opposed to taking $\leq$ as primitive and defining $<$ (de Bouvère (1965b) p. 622), and (ii) taking meet and joint as primitive in a lattice and defining $\leq$-relation, as opposed to vice-versa (cf. Givant & Halmos (2009) p. 43, Corcoran (1980) p. 233). Given these examples, it should not be surprising that definitional equivalence implies the notion of biinterpretability defined at the close of the previous section (cf. Visser (2006) § 3.3). However, as we note in Appendix §6.3., the Peano axioms and Hume's Principle are not biinterpretable and hence not definitionally equivalent. This highlights one potential pitfall in pursuing this response to the consistency problem: as one develops more demanding theory-based notions of representation,

[15] Of course, if one isolated a surrogate of the natural numbers within the interpreted and interpreting theories, then one could attempt to formulate and prove versions of the theorem with respect to these surrogates. But then one would have to separately motivate the connection between the theorem and observational consequences.

[16] Fischer (2010) p. 367, cf. Horsten (2011) §7.6 pp. 94 ff.

[17] Lindström (2003) Exercise 6 (a) p. 91 and Theorem 14 p. 107.

the notion must not become so demanding as to be inapplicable to its intended philosophical targets.[18]

It ought to be mentioned that the shift from mere interpretability to stronger representational notions like definitional equivalence and biinterpretability will most likely be concomitant with a shift in one's reasons as to *why* knowledge is preserved under known representation in known principles. One might distinguish between at least two such reasons, which I call the *reliable mechanism* position and the *irrelevance of contingencies* position. The first position highlights the fact that interpretability is merely deduction coupled in a certain way with explicit definitions, and would suggest that just as both deduction and explicit definition are ways of extending knowledge, so interpretability is a means of extending knowledge. The best way to evaluate this position is not to demand any positive line of argumentation for it, just as we don't demand positive argumentation for why deduction may extend knowledge, but rather to try to assay whether this is indeed a reliable mechanism of knowledge extension: whether in general when we feed known inputs into this mechanism we obtain known outputs as a result. The plethora and consistency problems constitute two distinct ways in which such reliability fails when the notion of representation is taken to be notions like interpretability or faithful interpretability.

Alternatively, from the talk of "notational variants" and "various descriptions" one might extract the idea that there are various elements of formal renderings of mathematics into formal first-order theories that are largely underdetermined by the mathematics itself. For instance, in spite of our vast knowledge of the natural numbers, no one knows whether "the natural numbers have $<$ as opposed to $\leq$ as a primitive" is true or false. The *irrelevance of contingencies* position simply posits that such differences cannot make a difference to mathematical knowledge, so that a theory differing from a known theory only in these ways is itself known. As mentioned above, this position will not help the logicist who seeks to move from knowledge of Hume's Principle to knowledge of the Peano axioms, since Hume's Principle and the Peano axioms are not biinterpretable or definitionally equivalent (cf. §6.3.). However, this is not to say that every principle whose epistemic status is similar to Hume's Principle will likewise fail to be biinterpretable with the Peano axioms, and so this would be one theory-based way for the logicist to proceed that would not obviously fall prey to the plethora and consistency problems discussed in this section.

§4. Structure-Based Versions In this section, the aim is to set out and examine two versions of the Logicist Template that supplement theory-based notions of representability with additional contentions about the representability of arithmetical structure within logical structure –e.g. the representability of the natural numbers within models of Hume's Principle. The first structure-based notion

[18] Fischer also suggests (in personal correspondence) that one might consider a strengthening of mutual interpretability to sentential equivalence (cf. definition of this notion at the close of §2.). I do not know whether Hume's Principle and the Peano axioms are sententially equivalent. In § 6.3., it is shown however that Hume's Principle and the Peano axioms are not fully sententially equivalent, where "fullness" (as defined in § 6.2.) essentially requires that one not reinterpret the concept-object distinction.

of representability described in §4.1 is very traditional, and simply posits that the arithmetical structure is patently definable within the logical structure. One problem with this account is a circularity objection, to the effect that certain of the definability claims are trivially equivalent to the very arithmetical knowledge which the advocate of the Logicist Template seeks to secure in the first place. This naturally motivates a second structure-based notion of representability that merely requires that arithmetical structure be isomorphic to a structure definable in logical structure. However, in §4.2 I will argue that this second structure-based notion is to be rejected for reasons related to the manner in which it requires a knowledge of structures and their signatures in advance of knowledge of axioms.

Finally, before proceeding, it's worth remarking that the distinction between "theory-based" and "structure-based" notions of representation is more of a continuum than a binary opposition. The key idea in theory-based notions is that provability should be privileged, while the key idea in the structure-based notions is that definability should be privileged. However, the formal line between provability and definability is invariably muddled: for instance, all definability facts in a structure will be recorded as provability facts in its complete first-order theory, while more locally any specific definability fact about a structure could be formalized as yet another axiom of a first-order theory. The reason for separating the discussion of theory-based and structure-based notions of representation is simply that the epistemic hurdles seem different when we center matters around definability as opposed to provability. For instance, in the following subsection (§ 4.1.), we must ask questions about the epistemic status of definitions, questions which seem to go above and beyond knowledge of axioms like Hume's Principle.

4.1. First Version & The Circularity Problem To illustrate the first structure-based notion of representation that I want to consider in this paper, suppose that we prove in the usual way that the Peano axioms are interpretable in Hume's Principle, and suppose that we supplement this knowledge with the additional knowledge that the natural numbers are the kinds of cardinalities invoked in the interpretation, that the natural number zero is the specific cardinality invoked in the interpretation, and likewise for the other arithmetical operations like successor and addition. This supplementary knowledge would then be articulated in terms of the following biconditionals, wherein N, 0 and S refer respectively to the natural numbers, zero, and successor, and where N^*, Z^*, and S^* are their logical- or cardinality-theoretic correlates in the signature of Hume's Principle:

$$\forall\, x\, [N(x) \leftrightarrow N^*(x)] \quad \forall\, y\, [y = 0 \leftrightarrow Z^*(y)] \quad \forall\, y, z\, [S(y,z) \leftrightarrow [N(y)\ \&\ S^*(y,z)]] \tag{7}$$

There would of course be analogous biconditionals for addition, multiplication, and the less-than relation, but for the sake of simplicity we omit those here.

Now, traditionally, $Z^*(y)$ would be taken to say that y was the cardinality of some concept which had no members, and $S^*(y,z)$ would be taken to say that y was the cardinality of the Y's, z was the cardinality of the Z's, all the Y's were Z's, and there was exactly one Z which was not a Y. The traditional choice of $N^*(x)$ would be as follows, which intuitively says that x holds of all the concepts F which contain the ersatz of zero and are closed under the ersatz of successor:

$$N^*(x) \equiv \forall\, F\, [[(\forall\, y\, (Z^*(y) \rightarrow Fy))\ \&\ (\forall\, y, z\, (Fy\ \&\ S^*(y,z)) \rightarrow Fz)] \rightarrow F(x)] \tag{8}$$

For the sake of definiteness, let us call the three biconditionals from equation (7), with these traditional understandings of N^*, Z^*, and S^* in place, the *definability biconditionals*, and let us refer to them individually as the *domain biconditional*, *the zero biconditional*, and the *successor biconditional*.

The version of the Logicist Template which I want to consider in this section contends that knowledge of the Peano axioms could be based on knowledge of Hume's Principle because one knows Hume's Principle, the definability biconditionals, and that these two things together deductively imply the Peano axioms. This latter piece of knowledge is of course non-trivial, but it is something that falls naturally out of the usual proof of Frege's Theorem. This version of the Logicist Template is also more traditional than those considered in the last section. For, the definability biconditionals are specific instances of the "bridge laws" from traditional versions of intertheoretic reduction. Further, this version of the Logicist Template merely posits that knowledge is preserved under known implication from known premises, rather than under known interpretability in known premises.

But the decisive advantage of this version of the Logicist Template is that it avoids both the plethora and consistency problems. On the one hand, the plethora problem will be blocked because one would additionally need to know facts about how the represented structure is defined in the representing structure. For instance, one could infer from the Peano axioms to the Fundamental Theorem of Algebra according to this version of the Logicist Template only if one knew that complex numbers were certain kinds of sequences of natural numbers and that this in conjunction with the Peano axioms deductively implied that the first-order sentences expressive of the Fundamental Theorem of Algebra were true on the complex numbers. On the other hand, the consistency problem is blocked because the relevant notion of representation requires that theories are paired many-one with structures which render the axioms of the theory true. Given this, one simply can't have arithmetical $T + \varphi$ and $T + \neg\varphi$ being representable in the relevant sense in say, Hume's Principle. For, since both these theories are arithmetical, the representability in question would require that the arithmetical structure given by $(N, 0, S)$ model both φ and $\neg\varphi$, which is clearly impossible.

While this version of the Logicist Template has the advantage of being immune to the plethora and consistency problems, it must contend with an objection specific to the definability biconditionals. The objection is that this version of the Logicist Template is patently circular because some of the arithmetical knowledge in question is trivially equivalent to one of the definability biconditionals. More specifically, the relevant equivalence is that the Mathematical Induction Axiom is trivially equivalent to the domain biconditional against the background of the zero and successor biconditional and the supposition that zero is a natural number and that successors of natural numbers are natural numbers. So as to not distract from the discussion here, I defer the proof of this equivalence to Appendix §6.4. While variants of this circularity objection can be found in the writings of Papert, Parsons, and Boolos, rendering this objection in terms of such a provable equivalence is new. I believe this rendering has the advantage of making clear the modal force behind Papert and Parsons' concern that "[. . .] we must use the notion of natural number [. . .] to see the equivalence of the set-theoretical propositions and their number-theoretical correlates" (Parsons (1965) p. 198, Parsons (1983) p. 168) or that "one needs the principle of induction to justify the identifications even in extension of

arithmetical notions and the corresponding notions of the *Principa Mathematica*" (Papert (1960) p. 113).[19]

The circularity concern of Papert, Parsons, and Boolos has been occasionally discussed in recent decades, but never to my knowledge buttressed by any formal considerations such as the aforementioned provable equivalence. Steiner and Demopoulus suggest responding to the objection by suggesting that the agent who knows the definability biconditionals might be different than the agent who acquires knowledge of the Peano axioms, and that the corresponding sources and standards of justification might be different, and for instance perhaps pragmatic in character.[20] However, this response seems vulnerable to the plethora problem, depending on how easy it is for the second agent to provide metaproofs or how easy it is to meet the relevant pragmatic standards. More recently, Wright and Hale briefly discuss the objection that "logicism illicitly presupposes distinctively mathematical techniques– *par excellence*, mathematical induction– in, for instance, establishing the requisite correlations between basic mathematical theories and the logicist systems in which they are reconstructed" (Hale & Wright (2001) pp. 433-434). They continue: "The natural reply to that particular charge is that the alleged circularity wouldn't matter. What would be significant would be if in order for a thinker to follow through the logicist route to arithmetic, she would need to be competent in induction *in advance*, as an unreconstructed rule of inference" (Hale & Wright (2001) p. 434). It seems to me that the provable equivalence of the Mathematical Induction Axiom with the domain biconditional across the background of the other definability biconditionals makes a *primae facie* case that "in order for a thinker to follow through a particular logicist route, she needs to be competent in induction in advance." The only other more recent discussion of the circularity objection of which I am aware is the following remark of Heck, who writes: "This sort of [circularity] objection was originally pressed by Henri Poincare, but has in recent years been developed by Charles Parsons. Let me not discuss it further, however, except to say that it is bound up with concerns about impredicativity [...] (Heck (1999) p. 70). But it is not immediately clear to me how the circularity objection might be related to impredicativity. Further, as mentioned in Appendix § 6.1., if one lowers the levels of comprehension, one can't prove Frege's Theorem, and so the discussion of the significance of this theorem would not obviously get off the ground in a predicative setting.[21]

---

[19] There's obviously much that distinguishes Papert, Parsons, and Boolos. For instance, Papert is discussing the interpretations of arithmetic in *Principia Mathematica*, while Parsons and Boolos are discussing interpretations of arithmetic in set theory. It seems to me that their concern carries over to interpretations of arithmetic in Hume's Principle. However, perhaps responses to their concern could make use of the important distinctions between Hume's Principle, set theory, and *Principia Mathematica*. However, primarily for reasons of space, I shall not pursue such responses here.

[20] Cf. "[...] we have a metaproof [...]" Steiner (1975) p. 30, "[...] it is a wholly pragmatic question whether we should justify the adequacy of the definition [...]" Demopoulos (1994) p. 237)

[21] There is some recent work by Burgess (2005) p. 82 and Visser (2011) on how much one can do in a predicative setting. The types of arithmetic that one can recover in this setting are weak arithmetics like Robinson's Q as opposed to strong arithmetics like the Peano axioms which are the topic of this paper (cf. Appendix 6.1. for formal definitions

In my view, the best response to the circularity objection is to appeal to what we might call *the evidential similarity principle*: the impropriety of a trivial equivalence of premise with conclusion across background knowledge is significantly lessened when there is a similarity between the evidence for the background knowledge and the evidence for the premises. To illustrate this principle, consider the following example: no one thinks that it is objectionable to infer that all the samples have a given chemical property from careful measurements of the last sample and a prior knowledge that all the previous samples have the property *when* this prior knowledge is similarly based on earlier careful observations. This is the case even though the conclusion that all the samples have the chemical property is trivially equivalent to the premise that the last sample has the property, against the background of our prior knowledge. Moreover, a moment's reflection on this example suggests a natural explanation for why the evidential similarity principle holds. For, one might be of a mind that a trivial equivalence of conclusion with a premise across background knowledge is improper because a chief aim of argumentation is to remove doubt, and such trivial equivalence would indicate that any doubts one had about the conclusion would be automatically transferable to the premise. But one natural way to block this transfer of doubt is to note that the evidence one has for the premise is similar to the comparatively indubitable evidence that one has for the background knowledge.

So it seems open to the advocate of the Logicist Template to respond to the circularity objection by noting that there seems to be a deep similarity between our evidence for the zero- and successor-biconditionals and the domain biconditional. The idea would be that the circularity objection seems most threatening when we wrongly and artificially start out accepting the zero- and successor-biconditionals and then subsequently inquire after the domain biconditional. For, once we accept the zero and successor biconditionals, the line between successor S and ersatz successor S^* and zero 0 and ersatz zero Z^* becomes blurred, and accordingly the domain biconditional starts to have the "look" of mathematical induction. It is this "inductive character" which Boolos emphasized in his version of the circularity objection: "[. . .] if at the outset one doubted whether the natural numbers satisfied induction one will still doubt whether the natural numbers will be (isomorphic to) the objects satisfying the definition, precisely because of the inductive character of the definition " (Boolos (1984) p. 470, Boolos (1998) p. 371). But if we treat all of these biconditionals as premises rather than artificially treating some few of them as background knowledge, then the "inductive character" of the domain biconditional disappears, and it just becomes a long complicated sentence in a higher-order language.

So how should the advocate of the Logicist Template conceive of the common source for evidence for the definability biconditionals? As Benacerraf suggests, a natural idea would be that the logicist would argue that "the sentences of arithmetic, in their preanalytic senses, mean the same (or approximately the same) as their homonyms in the logicist system" (Benacerraf (1981) p. 20, Demopoulos (1995) p. 46). Or, as Dummett suggests, the logicist would seek "to settle the status of the arithmetical laws we already have, involving those arithmetical concepts

of these theories). Hence, for instance, one could not account for our knowledge of the Mathematical Induction Axiom using these results.

we already grasp; and, to do that, he must analyse those concepts and supply expressions of definitions long in use [...]. [...] if such definitions are to serve their purpose, they must surely be analytic ones" (Dummett (1996) p. 20). Thus the idea is that one reflects on the meanings of "zero," "successor," and "natural number," and realizes that they respectively mean the same things as the logicist correlates flanking the opposite sides of the definability biconditionals.

The chief difficulty with making good on this thought is that there are presumably patterns of use of the terms "zero," "successor," and "natural number," such that persons proficient in this use might habitually invoke the zero and successor biconditonal but not the domain biconditional. For instance, consider the activity of specifying exact numerosities, i.e. the activity of providing answers like "15" or "27" to questions of the form "How many F's are there?" In this activity, it seems plausible that competent speakers might invoke the zero and successor biconditionals but not the domain biconditional. For instance, we can easily see how to avail ourselves of the zero and successor biconditionals to verify that the number of dots on the screen is exactly 15 without making tacit or explicit appeal to the domain biconditional.[22] Further, there is a natural explanation for why we do not so appeal to the domain biconditional in this setting: for, in this activity, one simply does not need a uniform method to establish universal claims about natural numbers, and this is largely what the domain biconditional provides. This explanation must of course say something further about the larger aims that circumscribe our needs. Presumably the immediate aim here is one of compression of information, which presumably in turn facilitates communication, storage, and recall. In answering the question "How many F's are there?" with "15" as opposed to literally displaying the F's or an explicit bijection between them and some class of objects whose exact number is already known, a useful piece of information is rendered in a highly compact and easily conveyable form which can be communicated or stored in memory with a comparatively short sequence of words.

The points made here about exact numerosity seem to generalize to a broader class of related mathematical activities. For instance, our everyday inferential activity is such that we would readily assent to the claim that to deduce a conjunction it suffices to deduce each conjunct, whereas we would hesitate to assent to the claim that a deduction from a set of axioms is a finite sequence of steps such that each step is an axiom or follows from earlier steps by an *antecedently fixed* set of inference rules. One consideration that may serve to induce such hesitation is a moment's reflection on one traditional way of motivating the significance of the completeness theorems for propositional and predicate logic. For instance, Kleene writes of the former: "[...] we have listed eleven postulates for the propositional calculus [...]. Can we give a reason why we stop with just these? Might we with

[22] The inferential process to verify that the number of F's is exactly k would go like this. One selects arbitrarily one of the F's, which we might call n_1, and concludes on the basis of the zero and successor biconditionals that the number of objects equal to n_1 is equal to the natural number one. Relying on one's memory and visual faculties, one then selects another of the F's, which we might call n_2, and concludes on the basis of one's earlier knowledge and the successor biconditional that the number of objects equal to n_1 or n_2 is equal to the natural number two. One then continues in this way until one concludes that the number of objects equal to n_1, n_2, or ... n_k is equal to the natural number k.

advantage attempt to discover others which could be added to the list to give more provable formulas?" (Kleene (1952) p. 131). Another related example comes from geometry: our initial inculcation in Euclidean geometry is such that we know that an equilateral triangle may be constructed on a given line, whereas we would not know, at least on the same basis, that a geometrical construction is one that may be effected by finitely many operations with *only* lines and circles as opposed to other constructions.[23] As a final example, take calculation: part of our training in calculation shows us that our usual procedure for long division yields an effectively calculable operation, whereas it presumably does not so show us that an effective operation is one which may be obtained via finitely many applications of operations such as primitive recursion and minimization. So the thought here is that proficiency in these various activities (exact numerosity, deduction, geometric constructions, and calculating) does not require the invocation of *or* assent to "exclusion clauses," like the domain biconditional, that tell us that "that's all there is" to the relevant concept (natural number, proof, geometric constructibility, effectively calculable function).

This criticism of the appeal to epistemic analyticity in the philosophy of mathematics is different in character from Williamson's well-known recent critique of epistemic analyticity (Williamson (2007a), Williamson (2007b) Chapter 4). The easiest way to see this is to note that the logicist has a viable response to Williamson's objection that does not meet the above objection. Williamson employs alternative logics to construct counterexamples to the contention that knowledge of the meaning of certain paradigmatic analytic propositions like "All vixens are female foxes" suffices for knowledge of that proposition. His specific counterexamples involve supervaluational semantics or logics on which true universal hypotheses carry existential presuppositions. Advocates of such alternative logics will have reason to doubt "All F's are G" even in situations when F and G are known to be synonymous because of their doubt in the elementary logical truth "All F's are F". However, without thereby giving up on logicism, it may be conceded that this is an example of an agent who knows the meaning of the paradigmatic proposition without knowing the proposition itself. For, Williamson's example does not vitiate the contention that knowledge of the meaning of the paradigmatic proposition *plus knowledge of basic logical truths* entails knowledge of the paradigmatic proposition.[24] This is relevant because the logicist's entire plan is to reduce the problem of mathematical knowledge to the problem of logical knowledge, and hence it seems entirely apposite for the logicist to suppose that the mechanism of this reduction– namely epistemic analyticity– is itself facilitated by a good deal of logical knowledge. It's clear that this logicist response does not meet the criticism articulated in the previous para-

[23] See Bos (2001) esp. pp. 221 ff for a historical account of the vacillations in the concept of geometric constructibility in the early modern period.

[24] Williamson, unlike the advocate of alternative logics, thinks that we do know basic logical laws like "All F's are F." He writes: "For unless a radical form of scepticism holds, we know that every vixen is a vixen, even though Peter and Stephen do not" (Williamson (2007a) p. 28) and he likewise writes: "Other people just like Peter and Stephen except for having more logical insight do know *every vixen is a vixen*" (Williamson (2007b) p. 130). "Stephen" is the name Williamson gives to the advocate of supervaluational semantics, while "Peter" is the name Williamson gives to the theorist who has non-standard beliefs about presuppositions.

graphs: for, knowledge of further inference rules or basic logical laws of the predicate calculus won't undercut the thought that one may exhibit an understanding of a given mathematical domain without being in a position to assent to exclusion clauses like the domain biconditional.

4.2. Second Version & The Signature Problem Given the discussion in the previous section, it is thus natural to seek out a structure-based version of the Logicist Template that does not appeal to the definability biconditionals. Since the definability biconditionals express that arithmetical structure *is definable* in the logical structure (or cardinality-theoretic structure), it is natural to focus on the weaker claim that arithmetical structure *is isomorphic to a structure definable* in the logical structure. This condition is very natural, and one highlighted in accounts of ontological reduction by both Quine and Goodman respectively in "Ontological Reduction and the World of Numbers" and *The Structure of Appearances*.[25] Further, since Goodman was motivated by cases "[. . .] where the extensional coincidence of definiendum and definiens is far from evident" (Goodman (1951) p. 6), it's natural to hope that this condition will avoid the problems related to definability biconditionals discussed in the previous section.[26]

In §2, it was mentioned that to say that one structure is *interpretable* in a second structure is just to say that the first structure is isomorphic to a structure definable in the second structure. With this terminology in place, I can now state the final version of the Logicist Template to be considered in this paper. This version claims that one can know that the arithmetical theory is true of the arithmetical structure because (i) one knows that the arithmetical theory is interpretable in the logical theory, (ii) one knows that the arithmetical structure is interpretable in the logical structure, and (iii) one knows that the formulas used in the two interpretations are the same. This version of the Logicist Template thus demands a kind of pre-established harmony of interpretability of theories and structures, wherein one knows that the formulas which define some model of the arithmetical theory within an arbitrary model of the logical theory are the same as those which define an isomorphic copy of the arithmetical structure within the logical structure.

While this version of the Logicist Template is easily able to overcome the consistency problem, it is less obvious that it can overcome the plethora problem. Before turning to the latter, let us briefly describe the mechanism by which it avoids the consistency problem. This mechanism is an elementary consequence of the hybrid notion of representability described in the previous paragraph: namely, that it follows deductively from the hypotheses (i)-(iii) of the previous paragraph that the arithmetical theory is true of the arithmetical structure. For, since the logical structure models the logical theory, by (i) it defines a model M of the arithmetical theory. But by (ii), an isomorphic copy M^* of the arithmetical structure is definable in the logical structure. By (iii), one has that M and M^* are identical, so that M^*

[25] See Quine (1964) p. 215, Quine (1976) p. 218, Goodman (1951) p. 6, Polánski (2009) Definition 7, Hellman (1978) pp. 214, 218.

[26] I do not know of any logicist who has suggested or endorsed the second-structure based version discussed in this section, which is motivated by these ideas of Quine and Goodman. Despite this, it seems valuable to map out the conceptual space and to understand what other structure-based versions are available and what problems they might face.

also models the arithmetical theory. Since M^* is isomorphic to the arithmetical structure, it likewise follows that the arithmetical structure models the arithmetical theory. So this is the reason why hypotheses (i)-(iii) imply that the arithmetical theory is true of the arithmetical structure. Hence, were there two rival arithmetical theories $T+\varphi$ and $T+\neg\varphi$ which satisfied these hypotheses (i)-(iii), then one would have that *the* arithmetical structure satisfies both φ and $\neg\varphi$, which is impossible. This, in any case, is the elementary consideration which shows that this version of the Logicist Template is not afflicted by the consistency problem.

However, it is less than obvious whether this version of the Logicist Template is able to overcome the plethora problem. For, hypothesis (ii) of this version requires that one knows that arithmetical structure is interpretable in logical structure, in advance of knowing much else about the arithmetical structure, for instance whether it models the arithmetical theory. If this *kind of knowledge* is too easy to come by, so that it is too easy to know that this or that structure is interpretable in logical structure via antecedently specified formulas, then one might very well be concerned that the plethora problem would simply reemerge in this slightly more complex setting. So the advocate of this version of the Logicist Template has to walk the fine line between saying something about how it is known that arithmetical structure is interpretable in logical structure via specific formulas, without it thereby being the case that it is easy to know that a wide variety of structures are so interpretable.[27] While these reflections on the plethora problem do not generate crisp counterexamples to this version of the Logicist Template, they do serve to underscore the general opacity of the kind of knowledge of structure invoked by this version of the Logicist Template.

To get clearer on the nature of this knowledge of structure, it is helpful to bear in mind that we typically expect our foundational axiomatic theories of arithmetic, analysis, and set theory to effect two tasks in tandem: to record our entitlements and to describe a structure. That is, we expect such theories to both tell us what inferences are permitted in the relevant practice (arithmetic, analysis, set-theory), as well as to describe the subject-matter of the practice (the natural numbers, the real numbers, the cumulative hierarchy). By contrast, this version of the Logicist Template seeks to secure our entitlement to arithmetical axioms like the Peano axioms by presupposing that we have some independent grasp of arithmetical structure in advance of these axioms. One way to cast doubt on the nature of this presupposition is to ask whether the type of knowledge of structure implicated therein is able to discriminate between various relevant alternatives in advance of knowledge of arithmetical axioms. While this requirement is admittedly high, it is something that we are prone to expect of our logical knowledge in general: whatever the evidence is that vouches for modus ponens, presumably it is able to rule out this law holding with respect to propositions about middle-sized objects whilst failing in more esoteric settings. Thus if this version of the Logicist Template presupposed a type of knowledge that was unable to so discriminate, this would cast

[27] These reflections are similar to Hellman's reflections on whether Goodman's notion of ontological reduction is afflicted with a version of the plethora problem which Hellman dubbed "mathematicism": the problem that physical theory would be reducible to set theory (cf. Hellman (1978) §4 pp. 221 ff).

doubt on its capacity to vindicate the contention that our arithmetical knowledge is akin to our logical knowledge.

One way to develop this kind of objection is to focus in on the knowledge of *formal language* or *signature* that is implicated in the pre-axiomatic knowledge of arithmetical structure postulated by this version of the Logicist Template. In particular, this version of the Logicist Template clearly presupposes that one knows that the natural numbers are a structure in a specific signature, namely that of the Peano axioms, which includes addition and multiplication as well as zero and successor. However, there are other formal theories of arithmetic such as the Presburger axioms which involve only addition, zero and successor. But it is unclear that there is evidence for the natural numbers being a structure in the signature of the Peano axioms as opposed to the Presburger axioms that does not presuppose the very knowledge of the Peano axioms which this version of the Logicist Template seeks to secure. For instance, one good reason to think that the natural numbers are a structure in the signature of the Peano axioms is that one knows that there are infinitely many prime numbers on the basis of the Peano axioms, but that the set of prime numbers is not first-order definable in the signature of the first-order Presburger axioms.[28]

One response to this signature problem is to suggest that it is misleading to speak about *the* signature of the natural numbers, and that whatever the natural numbers are, they aren't something that comes equipped with a signature. This thought has a certain appeal to it and might be viewed as an extension of the received wisdom that attitudes like knowledge apply more properly to propositions than to sentences.[29] In the development of his brand of structuralism, Resnik likewise suggested that structures qua basic mathematical objects ought not be tied to specific signatures (cf. Resnik (1981) p. 535, cf. Resnik (1997) pp. 207-208), and he later suggested a way to handle this problem by defining a certain equivalence relation on structures. In particular, Resnik says that two structures are equivalent if they have the same domain and if the constants, relations, and functions of the one are first-order definable in the other.[30] One concern is that this is too fine-grained an equivalence relation for the purposes of this structure-based version of the Logicist Template. For instance, no two of the following first-order structures are pairwise equivalent in Resnik's sense, wherein N denotes the natural numbers:

$$(N, S), \quad (N, +), \quad (N, +, \times) \tag{9}$$

For, the even numbers are first-order definable in the second structure, whilst any first-order definable subset of the first structure is finite or cofinite (cf. Marker (2002) Exercise 3.4.3 p. 104). Likewise, multiplication is not first-order definable in the second structure, since if we could define multiplication, then we could define

[28] This non-definability results follows from the well-known quantifier elimination results for first-order Presburger arithmetic. See Marker (2002) pp. 81 ff.

[29] Further, one who was sympathetic to the irrelevance of contingencies position (described at the close of §3.) might obviously be sympathetic to this thought.

[30] See Resnik (1981) p. 536, Resnik (1997) pp. 208-209, noting that the key notion of "pattern occurrence" is explicitly cast in terms of definability at Resnik (1981) p. 533, Resnik (1997) p. 205. It's perhaps worth noting that Resnik's notion of equivalence of structures implies that the complete theories of these structures are definitionally equivalent (cf. end of §3. for discussion of and references on this notion).

the set of primes. Hence, if one modified the structure-based version of the Logicist Template so that it was phrased in terms of equivalence classes of structures, then one could still ask why the natural numbers are a structure in the equivalence class of the third structure and not in the equivalence class of the second structure.[31]

§5. Conclusion I want to close by contrasting the nature of the challenges which I have presented for the theory-based and structure-based versions of the Logicist Template. A highly abstract version of the Logicist Template may be presented in terms of the following valid argument:

> *Base Premise*: The logical principles are known.
> *Representability Premise*: It is known that the arithmetical principles are representable in the logical principles.
> *Preservation Premise*: For all principles P and P^*, if principles P^* are known, and it is known that P is representable in P^*, then principles P are known.
> *Conclusion*: The arithmetical principles are known.

This presentation of the Logicist Template obscures many important things –the role of the epistemic agent, the quality or grade of knowledge –but it serves to compactly introduce various contrasts. For, expressed in these terms, it seems fair to say that most of the recent literature on logicism has focused on the Base Premise, i.e., the epistemic status of Hume's Principle.

My aim has rather been to draw attention to the manner in which the truth of the Representability and Preservation Premises vary with the choice of the notion of representation. For instance, as described in §3, the theory-based notion of representation given by interpretability of theories renders the Representability Premise patently true –in this case it just follows directly from Frege's Theorem. But this notion of representation seems to make the Preservation Premise susceptible to counterexamples coming from both the plethora problem and the consistency problem. Likewise, while the structure-based versions seem to fare better with respect to some of these problems, it seems difficult to sustain the Representability Premise in this case. For, as described in §4.1, the circularity problem seemingly requires evidence that the domain biconditional is known on the basis of a knowledge of meaning, evidence which does not seem forthcoming. Likewise, as discussed in §4.2, securing the Representability Premise with respect to an isomorphism-based notion of representation seems to put one in the position of claiming to have knowledge of structures and signatures in advance of knowledge of axioms. The challenge is thus to find a notion of representation which renders both the Representability and

[31] If one moves from first-order structures to second-order structures with added standard second-order parts, then it seems that Resnik's equivalence relation becomes too coarse. For instance, with these second-order resources, the real and complex fields will be able to define the natural numbers, whereas one might have initially thought that the real and complex fields should be geometric rather than arithmetic in character. To take another example, most functions from natural numbers to natural numbers that come up in ordinary mathematics are definable in the natural numbers with these second-order resources, so that Resnik's criterion would then have that the addition of all these functions to this structure does not move one out of this equivalence class. Perhaps there is some middle-ground inbetween first-order and standard second-order on which Resnik's definition would be neither too coarse nor too fine for the purposes of importation into this logicist project.

Preservation Premise true, or at least a notion of representation which does not succumb to the specific problems described here.

§6. Appendix

6.1. Formal Definitions of Hume's Principle and The Peano Axioms The purpose of this brief appendix is merely to present formal definitions of Hume's Principle and the Peano axioms, that were introduced informally in §1, as well as to explain some notation related to these theories and their models that will be relevant for the proofs presented in the subsequent appendices. Formally, Hume's Principle is a sentence in an expansion of a second-order logic by a function symbol # from unary properties of first-order objects to first-order objects. That is, the idea is that if F is a unary property of first-order objects, then $\#F$ is itself a first-order object. The notion of a "one-one correspondence" can be formally captured with the idea of a bijection. A map $f : F \to G$ is a *bijection* if it is injective and surjective. The map $f : F \to G$ is *injective* if $f(x) = f(x')$ implies $x = x'$ for all x, x' from F, while the map $f : F \to G$ is *surjective* if for every y in G there is x in F such that $f(x) = y$. Hence, formally, Hume's Principle is the following sentence:

$$\forall\, F, G \;\; \#F = \#G \leftrightarrow \;\exists \text{ bijection } f : F \to G \tag{10}$$

So, as the right-hand side of Hume's Principle makes clear, the ambient logic of Hume's Principle is second-order logic. Moreover, natural renderings of the right-hand side of Hume's Principle suggest that we countenance at least two distinct kinds of second-order entities: unary or one-place properties of objects like F and G, as well as binary or two-place properties like the graph of the function f. So it seems most natural to formalize Hume's Principle in a second-order logic that has a distinct sort or type for n-place properties of objects for each natural number $n \geq 1$.

So formalized, Hume's Principle would thus be a theory in a many-sorted logic with a countably infinite number of distinct sorts (cf. discussion of many-sorted logic in §2.). However, there is an obvious two-sorted alternative formalization available. In particular, one could posit an injection $(x, y) \mapsto \langle x, y\rangle$ on pairs of objects, a so-called "pairing function". Using this, one could then define an injection $(x, y, z) \mapsto \langle x, y, z\rangle \equiv \langle x, \langle y, z\rangle\rangle$ on triples of objects. In this way, one could treat all n-ary properties of objects as unary properties of objects consisting of "codes" for n-tuples of objects, with the coding given by $(x_1, \ldots, x_n) \mapsto \langle x_1, \ldots, x_n\rangle$. This difference in formalization makes no difference to the results about Hume's Principle mentioned in this paper.

Since Hume's Principle has sorts designed to model second-order entities, there are several well-known alternative semantics available. These semantics differ on the issue of whether one requires that the second-order quantifiers for n-ary relations range over the entirety of the powerset $P(M^n)$ of the n-th Cartesian product M^n of the domain M of the first-order quantifiers. The semantics that insists on this requirement is called the *full* or *standard* semantics, while the semantics that does not insist on this requirement is called the *Henkin* semantics. None of the important differences between these semantics matter for the results mentioned in this paper. However, for the results mentioned in this paper, it is important that the second-order logic is assumed to have the full comprehension schema. This schema says that to each n-ary formula there corresponds an n-ary relation such that the relation

is predicated of all and only those n-tuples of objects of which the formula holds. The full comprehension schema is needed because one of the theorems discussed here, namely Frege's Theorem, does not hold if a more restricted version of the comprehension schema is used.

Let us now turn from Hume's Principle to the Peano axioms. The Peano axioms are given by the following axioms, called the axioms of *Robinson's Q*

(Q1) $Sx \neq 0$ (Q2) $Sx = Sy \rightarrow x = y$ (Q3) $x \neq 0 \rightarrow \exists\, w\, x = Sw$
(Q4) $x + 0 = x$ (Q5) $x + Sy = S(x + y)$ (Q6) $x \cdot 0 = 0$
(Q7) $x \cdot Sy = x \cdot y + x$ (Q8) $x \leq y \leftrightarrow \exists\, z\, x + z = y$

and by the Mathematical Induction Axiom:

$$\forall\, F\, [F(0)\ \&\ (\forall\, y, z\, (F(y)\ \&\ S(y) = z) \rightarrow F(z))] \rightarrow [\forall\, x\, F(x)] \tag{11}$$

The Mathematical Induction Axiom is obviously a second-order principle, since it begins with a universal quantifier over properties of objects. Hence, what I am describing in this paper as "the Peano axioms" is second-order Peano arithmetic, as described and studied in e.g. Simpson (2009). This is to be distinguished from first-order Peano arithmetic as studied in Hájek & Pudlák (1998), in which one remains within first-order logic and wherein the Mathematical Induction Axiom is replaced by the following infinite schema of formulas, where φ is a first-order formula:

$$[\varphi(0)\ \&\ (\forall\, y, z\, (\varphi(y)\ \&\ S(y) = z) \rightarrow \varphi(z))] \rightarrow [\forall\, x\, \varphi(x)] \tag{12}$$

Against the background of the full comprehension schema for second-order logic, the Mathematical Induction Axiom in equation (11) is equivalent to a version of the schema in equation (12) in which φ is allowed to be a second-order formula. Hence, the Peano axioms can be viewed as the result of generalizing the axiom schema of first-order Peano arithmetic to the second-order setting.

There are two issues concerning the formalization of the Peano axioms that are specific to logicism. First, in traditional formalizations of the Peano axioms as in Simpson (2009), the Peano axioms are only formalized in a two-sorted logic, wherein one sort is for first-order objects and another sort is for second-order unary properties of objects. This is because the Peano axioms have a natural pairing functions on objects (cf. Simpson (2009) p. 66). Given that we are formalizing Hume's Principle in a many-sorted logic with sorts for n-ary properties of objects for $n \geq 1$, it seems natural to suppose that we are also formalizing the Peano axioms in this fashion. Since the Peano axioms have this pairing function built-in, anything that can be proven with the two-sorted version of the Peano axioms can be proven in the infinitary-sorted version of the Peano axioms (under the obvious translations), and vice-versa. The second issue pertaining to logicism and the formalization of the Peano axioms is that in the discussion in §4.1. and §6.1., the successor operation is written as a binary relation $S(y, z)$ instead of a unary function $S(y) = z$. This is because logicism traditionally aims to establish such arithmetical truths like "every natural number has a successor," which is presupposed by the functional notation.

Finally, let us first introduce some notation for models of Hume's Principle and the Peano axioms which we will use in the subsequent appendices. Models of Hume's Principle will be written as follows:

$$\widehat{M} = (M, S_1[M], S_2[M], \ldots, \in_1[M], \in_2[M], \ldots, \#) \tag{13}$$

wherein M is a non-empty set serving as the interpretation of the first-order objects, and $S_k[M]$ is a non-empty set serving as the interpretation of the k-ary properties of objects, and $\in_k[M]$ is a $(k+1)$-ary relation serving as the interpretation of the relation of a k-tuple of objects $a_1, \ldots, a_n$ satisfying a k-ary property of objects R, and written in the object language of $\widehat{M}$ as $R(a_1, \ldots, a_n)$. If one says that a structure as in equation (13) is a model of Hume's Principle, then it is presumed that it satisfies Hume's Principle (10) and the comprehension schema. Further, let us say that the model $\widehat{M}$ in equation (13) is *normal* if $S_k[M] \subseteq P(M^k)$ and $\in_k[M]$ is interpreted as the membership relation from the ambient set theory, so that

$$\widehat{M} \models R(a_1, \ldots, a_k) \Longleftrightarrow (a_1, \ldots, a_k) \in R \tag{14}$$

In a normal structure $\widehat{M}$ one may omit the relations $\in_k[M]$ from the specification of $\widehat{M}$ in equation (13). Further, every structure $\widehat{M'}$ isomorphic to a normal structure $\widehat{M}$, and so if one is only concerned with structures up to isomorphism, then one may presume that every structure is normal. Finally, let us say explicitly what is meant by the cardinality of these many-sorted structures (cf. Manzano (1996) p. 231 and Ebbinghaus (1985) pp. 32, 64):

$$\text{cardinality of } \widehat{M} = \left|\widehat{M}\right| = \sup\{|M|, |S_k[M]| : k \geq 1\} \tag{15}$$

Note that this notion of cardinality is from the ambient set theory which serves as the metatheory, and hence is in general distinct from the notion of cardinality axiomatized by the #-operator from Hume's Principle. Finally, let us write models of the Peano axioms as follows:

$$\widehat{N} = (N, S_1[N], S_2[N], \ldots, \in_1[N], \in_2[N], \ldots, 0_N, S_N, +_N, \times_N, \leq_N) \tag{16}$$

wherein the conventions on N and $S_k[N]$ and $\in_k[N]$ are exactly as above and where the items $0_N, S_N, +_N, \times_N, \leq_N$ serve as the denotation in the model $\widehat{N}$ of the non-logical items $0, S, +, \times, \leq$ from the signature Robinson's Q and the Peano axioms described above circa equation (11). Likewise, cardinality of $\widehat{N}$ is defined exactly as the cardinality of $\widehat{M}$ as in equation (15).

6.2. Formal Result Related to Consistency Problem The aim of this brief appendix is to prove a result discussed in §3., namely there are axioms containing the anti-Peano axioms that are faithfully interpretable within Hume's Principle, and moreover in such a way that the terms $\overline{n}$ and $\underline{n}$ are inferentially indistinguishable and the application constraint Nq from equation (5) is met. Let us first introduce some terminology on the types of interpretations between theories that shall be useful in establishing this result. First, when the signatures of both theories are extensions of that of second-order logic that do not include any new sorts, it's natural to focus on interpretations that translate objects by certain kinds of objects N, unary concepts by subconcepts of N, binary concepts by subconcepts of $N \times N$, etc. Let us agree to call such interpretations *full*, and let us empoy a similar terminology for models. So for instance, if $\widehat{M}$ and $\widehat{M'}$ have signatures that are extensions of that of second-order logic that do not include any new sorts, then we say that $\widehat{M'}$ is *fully* definable in $\widehat{M}$ if the objects in $\widehat{M'}$ are defined as a class of objects N in $\widehat{M}$, the concepts of $\widehat{M'}$ are defined as concepts in $\widehat{M}$ which are subconcepts of N, etc. If both $\widehat{M}$ and $\widehat{M'}$ are normal (cf. Appendix § 6.1. for the definition of normality), then $\widehat{M'}$ is fully definable in $\widehat{M}$ if M' is equal to

an $\widehat{M}$-definable subset N of M and $S_n(M') = S_n(M) \cap P(N^n)$. Finally, it will be useful in what follows to focus on certain kinds of interpretations of extensions of Robinson's Q in Hume's Principle. In particular, if T is an extension of Robinson's Q in the signature of the Peano axioms, then let's say that an interpretation of T in Hume's Principle is *orderly* if whenever $\widehat{M}$ is a model of Hume's Principle and if $\widehat{N}$ is the model of T induced by the interpretation, then for all all m from N one has that $m = \#\{k \in N : k < m\}$. For instance, the usual proof of Frege's Theorem establishes that there is a full and orderly interpretation of the Peano axioms in Hume's Principle.

The other important concept that is needed is that of relative categoricity. Suppose that T is an extension of Robinson's Q in the signature of the Peano axioms. Then let's say that T is *relatively categorical* if whenever $\widehat{N}$ is a model of T and another model $\widehat{N}'$ of T is fully definable in $\widehat{N}$, then $\widehat{N}$ and $\widehat{N}'$ are isomorphic, and moreover an isomorphism $g : \widehat{N} \to \widehat{N}'$ is definable in the original structure $\widehat{N}$ using the same parameters as the definition of $\widehat{N}'$ in $\widehat{N}$. The usual proof of the Dedekind Categoricity Theorem shows that the Peano axioms are relatively categorical (cf. Parsons (2008) pp. 281 ff or Shapiro (1991) pp. 82 ff). Recall from §3. that the anti-Peano axioms are simply the Peano axioms but with the Mathematical Induction Axiom replaced by its negation. That is, since the Peano axioms are Robinson's Q plus the Mathematical Induction Axiom Θ and the full comprehension schema, the anti-Peano axioms are Robinson's Q plus $\neg\Theta$ and the full comprehension schema. Finally, *the anti-Peano* axioms* are the anti-Peano axioms plus the supposition that there is exactly one object ∞ such that $S(\infty) = \infty$ and that ∞ is greater than all the other numbers and ∞ added to anything is ∞, and that $\infty \cdot 0 = 0$ and $\infty \cdot x = \infty$ for all $x \neq 0$, and that the Mathematical Induction Axiom holds on the restricted domain of the numbers $< \infty$. The anti-Peano* axioms have a simple model whose first-order part is $\{0, 1, 2, \ldots\} \cup \{\omega\}$ where ω serves as the interpretation of ∞. Just as the Peano axioms are relatively categorical, so the anti-Peano*-axioms are relatively categorical, essentially because they merely result from the relatively categorical Peano axioms by the addition of a single "infinite number." Further, just as there is a full and orderly interpretation of the Peano axioms in Hume's Principle, so there is a full and orderly interpretation of the anti-Peano*-axioms in Hume's Principle. Finally, just as Hume's Principle is fully interpretable in the Peano axioms, so Hume's Principle is fully interpretable in the anti-Peano* axioms. The desired result then follows directly from these remarks and the below theorem:

THEOREM 6.1. *For each* $n \geq 0$

$$\mathrm{HP}^2 \vdash \forall\, F\ (\#F = \overline{n} \leftrightarrow \exists^{=n}\, x\ Fx) \tag{17}$$

wherein HP^2 *is an abbreviation for Hume's Principle plus full comprehension. Further, suppose that* T *is a relatively categorical extension of Robinson's* Q *and that* I *is an interpretation of* T *in* HP^2 *that is full and orderly and that there is a full interpretation* B *of* HP^2 *into* T*. Then for all formulas* $\varphi(x_1, \ldots, x_k)$ *in the signature of Robinson's* Q *and all* $n_1, \ldots, n_k \geq 0$*, we have the following:*

$$T \vdash \varphi(\underline{n}_1, \ldots, \underline{n}_k) \Longleftrightarrow \mathrm{HP}^2 \vdash \varphi^I(\overline{n}_1, \ldots, \overline{n}_k) \tag{18}$$

wherein $\varphi \mapsto \varphi^I$ *is the map from formulas in the signature of the Peano axioms to formulas in the signature of Hume's Principle that is compositionally induced by the*

interpretation I, and wherein the terms $\underline{n}$ and $\overline{n}$ are as defined in equations (3)-(4) from §3.

Proof. So suppose that I is an interpretation of T into HP^2 that is full and orderly. First let us note that due to the relative categoricity of T, we have that I is actually a faithful interpretation of T into HP^2. For, suppose that $\mathrm{HP}^2 \vdash \varphi^I$ but that there is some model $\widehat{N}$ of $T + \neg\varphi$. Let $\widehat{M}$ be the model of HP^2 induced by the action of interpretation B. Let $\widehat{N}'$ be the model of T induced the action of interpretation I on $\widehat{M}$, so that $\widehat{N}' \models \varphi$. Since both B and I are full, we have that $\widehat{N}'$ is fully definable in $\widehat{N}$. Hence, by relative categoricity of T, it follows that $\widehat{N}'$ is isomorphic to $\widehat{N}$, so that we have reached a contradiction, since $\widehat{N}'$ models φ while $\widehat{N}$ models $\neg\varphi$. Hence, in fact I is a faithful interpretation of T into HP^2.

The other concept that we need to appeal to is the notion of a skolemisation of a theory. Let's briefly recall the following standard definition of this notion, where here we follow Hodges (1993) pp. 88 ff. Suppose that H is a theory in a signature L. Then there is an expansion of L_* of the same cardinality as L and an L_*-theory H_* extending H, called *a skolemisation of* H, such that (i) every model of H can be expanded to a model of H_* and (ii) such that for every L_* formula $\varphi(x_1, \ldots, x_n, y)$ there is an L_*-term $\tau(x_1, \ldots, x_n)$ such that H_* proves

$$\forall\, x_1, \ldots, x_n\, [(\exists\, y\, \varphi(x_1, \ldots, x_n, y)) \rightarrow (\varphi(x_1, \ldots, x_n, \tau(x_1, \ldots, x_n)))] \tag{19}$$

Note that a skolemisation H_* of H is a conservative extension of H for L-sentences. This follows from clause (i). For, suppose that H_* proves an L-sentence φ but H does not prove φ. Then there is a model of $H + \neg\varphi$. Then clause (i) requires that we can expand this model to a model of H_* and hence $H_* + \varphi$.

Now we proceed to the proof of the theorem. Let HP^2_* be a skolemisation of HP^2. By the results of the two previous paragraphs, we have we have that for all sentences φ in the signature of T:

$$T \vdash \varphi \Longleftrightarrow \mathrm{HP}^2_* \vdash \varphi^I \tag{20}$$

Since I is a full interpretation of T in HP^2, let $N(x)$ be a formula in the signature of HP^2 which serves as the domain of the interpretation and let $\psi_n(x)$ be the formula $x = \underline{n}$ in the signature of T. Then for all $n \geq 0$ since T proves $\exists! x\ \psi_n(x)$, we have that HP^2 proves $\exists! x\ (N(x) \wedge \psi^I_n(x))$. Since HP^2_* is a skolemization of HP^2, there is a term τ^n_I in the signature of HP^2_* such that HP^2_* proves $N(\tau^n_I) \wedge \psi^I_n(\tau^n_I))$. Let $n_1, \ldots, n_k \geq 0$. Let ψ be the sentence $\varphi(\underline{n}_1, \ldots, \underline{n}_k)$. Then by equation (20), one has

$$T \vdash \psi \iff T \vdash\ \exists\, x_1, \ldots, x_k\, [(\bigwedge_{i=1}^{k} \psi_{n_i}(x_i)) \wedge \varphi(x_1, \ldots, x_k)] \tag{21}$$

$$\iff \mathrm{HP}^2 \vdash \exists\, x_1, \ldots, x_k\, [(\bigwedge_{i=1}^{k} N(x_i) \wedge \psi^I_{n_i}(x_i)) \wedge \varphi^I(x_1, \ldots, x_k)] \tag{22}$$

$$\iff \mathrm{HP}^2_* \vdash \varphi^I(\tau^{n_1}_I, \ldots, \tau^{n_k}_I) \tag{23}$$

where the bottom-to-top direction of last biconditional follows from skolemizations being conservative extensions of the theories which they skolemize. So we have established that if $\varphi(x_1, \ldots, x_k)$ is a formula in the signature of T and $n_1, \ldots, n_k \geq$

0, then

$$T \vdash \varphi(\underline{n}_1, \ldots, \underline{n}_k) \iff \mathrm{HP}^2_* \vdash \varphi^I(\tau^1_I, \ldots, \tau^k_I) \tag{24}$$

Now fix $n \geq 0$ in the metatheory. Then since T extends Robinson's Q, we have the following (cf. Hájek & Pudlák (1998) Theorem I.1.6.(4) p. 30):

$$T \vdash \forall\, x\, [x < \underline{n} \leftrightarrow (\bigvee_{\ell=0}^{n-1} x = \underline{\ell})] \tag{25}$$

Let $\psi_<(x, y)$ be a formula in the signature of HP^2 which serves as the interpretation of the less-than relation for I. Then by an application of equation (24), we have

$$\mathrm{HP}^2_* \vdash \forall\, x\, N(x) \rightarrow [\psi_<(x, \tau^n_I) \leftrightarrow (\bigvee_{\ell=0}^{n-1} x = \tau^\ell_I)] \tag{26}$$

Suppose that F is a concept such that Fz if and only if $\bigvee_{\ell=0}^{n-1} z = \tau^\ell_I$. Then $F \subseteq N$ and Fz if and only if $N(z) \wedge \psi_<(z, \tau^n_I)$. Then since the interpretation I is orderly we have:

$$\#F = \#\{z \in N : \psi_<(z, \tau^n_I)\} = \tau^n_I \tag{27}$$

So we have just shown that

$$\mathrm{HP}^2_* \vdash \forall\, F\, [(\forall\, z\, (Fz \leftrightarrow \bigvee_{\ell=0}^{n-1} z = \tau^\ell_I)) \rightarrow \#F = \tau^n_I] \tag{28}$$

Now we argue that this implies

$$\mathrm{HP}^2_* \vdash \forall\, F\, (\#F = \tau^n_I \leftrightarrow \exists^{=n}\, x\, Fx) \tag{29}$$

For, suppose that $\#F = \tau^n_I$. Let G be a concept such that Gz if and only if $\bigvee_{\ell=0}^{n-1} z = \tau^\ell_I$. Then by equation (28), one has that $\#G = \tau^n_I = \#F$. Hence, there is a bijection between G and F. By equation (24) one has that all the τ^ℓ_I for $0 \leq \ell \leq n-1$ are distinct. Thus $\exists^{=n} x\, Gx$ and hence $\exists^{=n} x\, Fx$. Conversely, suppose that $\exists^{=n} x\, Fx$. Let G be a concept such that Gz if and only if $\bigvee_{\ell=0}^{n-1} z = \tau^\ell_I$. Then by equation (28), one has that $\#G = \tau^n_I$. By equation (24) one has that all the τ^ℓ_I for $0 \leq \ell \leq n-1$ are distinct. Hence $\exists^{=n}\, x\, Gx$. Since $\exists^{=n}\, x\, Fx$ and $\exists^{=n}\, x\, Gx$, choose a bijection between F and G. Then we have $\#F = \#G = \tau^n_I$. So we have just finished verifying equation (29).

Now, suppose that $\{\rho^n : n \geq 0\}$ is a another sequence of terms in the signature of HP^2_* which satisfies:

$$\mathrm{HP}^2_* \vdash \forall\, F\, (\#F = \rho^n \leftrightarrow \exists^{=n}\, x\, Fx) \tag{30}$$

We argue that HP^2_* proves that $\tau^n_I = \rho^n$ for all $n \geq 0$. Fix $n \geq 0$, and work within a model $\widehat{M}$ of HP^2_*. Choose any concept F with exactly n elements, such as Fz iff $\bigvee_{\ell=0}^{n-1} z = \tau^\ell_I$. This has exactly n elements since by equation (24) one has that all the τ^ℓ_I for $0 \leq \ell \leq n-1$ are distinct. Then since $\{\tau^n_I : n \geq 0\}$ satisfies equation (29) and $\{\rho^n : n \geq 0\}$ satisfies equation (30), we have that $\#F = \rho^n$ and $\#F = \tau^n_I$ and hence $\rho^n = \tau^n_I$.

By this argument and equation (24) and the fact that HP^2_* is conservative over HP^2 for sentences in the signature of HP^2, it suffices to show the following:

$$\mathrm{HP}^2_* \vdash [\forall\, F\ (\#F = \overline{n} \leftrightarrow \exists^{=n}\, x\ Fx)] \wedge [\bigwedge_{i<j\leq n} \overline{i} \neq \overline{j}] \tag{31}$$

wherein we adopt the convention that $(\bigwedge_{i<j\leq 0} \overline{i} \neq \overline{j})$ is equivalent to a tautology. For the case $n = 0$, clearly this holds by our convention and since Hume's Principle implies that $\#F = \#\emptyset$ iff there are no F's. Suppose that the result holds for n, and let us attempt to show it for $n+1$. Let's first note that $\mathrm{HP}^2_* \vdash \bigwedge_{i<n+1} \overline{i} \neq \overline{n+1}$. For, suppose not. Then letting $F = \{z : \bigvee_{\ell=0}^{n} z = \overline{\ell}\}$ we have $\#F = \overline{n+1}$ by the definition of the term $\overline{n+1}$ and hence $\#F = \overline{n+1} = \overline{i}$ for some $i < n+1$. Then by induction hypothesis for i, we have that $\exists^{=i}\ z\ Fz$. But this contradicts the induction hypothesis that $\mathrm{HP}^2_* \vdash \bigwedge_{i<j\leq n} \overline{i} \neq \overline{j}$. Suppose now that

$$\#F = \overline{n+1} = \#\{z : \bigvee_{\ell=0}^{n} z = \overline{\ell}\} \tag{32}$$

Since by the induction hypothesis we have that $\mathrm{HP}^2_* \vdash \bigwedge_{i<j\leq n} \overline{i} \neq \overline{j}$, it follows that $\exists^{=n+1}\ z\ Fz$. Conversely, suppose that $\exists^{=n+1}\ z\ Fz$. Choose z such that Fz and let $G = F \setminus \{z\}$. Then $\exists^{=n} z\ Gz$ and hence by induction hypothesis $\#G = \overline{n} = \#\{z : \bigvee_{\ell=0}^{n-1} z = \overline{\ell}\}$. Since $\overline{n}$ is distinct from the $\overline{\ell}$ for $\ell < n$ by induction hypothesis, it follows that $\#F = \#\{z : \bigvee_{\ell=0}^{n} z = \overline{\ell}\} = \overline{n+1}$. □

6.3. Hume's Principle and the Peano Axioms are neither Biinterpretable nor Fully Sententially Equivalent In this section, it is shown that Hume's Principle and the Peano axioms are not biinterpretable and that they are not fully sententially equivalent. These results were mentioned respectively at the close of §3. and in Footnote 18 of §3. Further, recall that the notions of biinterpretability and sentential equivalence were defined respectively in equations (3)-(6) and (3)-(4)-(5′)-(6′) at the close of §2. Finally, it should be emphasized that while the full semantics are used in the course of these proofs, the notions of biinterpretability and full sentential equivalence are purely proof-theoretic notions. So the use of the full semantics here is similar to how one might use these semantics to illustrate that a sentence of second-order logic was not a deductive validity. Of course, the use of the full semantics raises the general question of how rich or strong the metatheory needs to be to establish these non-interpretability results.

So now let us prove that Hume's Principle and the Peano axioms are not biinterpretable. Suppose, for the sake of contradiction that Hume's Principle and the Peano axioms were biinterpretable. Then consider the following normal model of Hume's Principle (where, recall, the notion of normality was defined at the close of § 6.1.):

$$\widehat{M} = (M, S_1[M], S_2[M], \ldots, |\cdot|) \tag{33}$$

wherein $M = |P(\omega)| + 1$ and $S_k[M] = P(M^k)$ and $|\cdot| : P(M) \to M$ is simply the cardinality function of the ambient set theory. In the definition of M, the addition is ordinal addition from the ambient set theory. This has the consequence that if X is a member of $P(M)$ then $|X|$ is a member of $|P(\omega)| + 1 = M$. Note that the cardinality of $\widehat{M}$ is $|P(P(\omega))|$. Now the supposition of biinterpretability (cf.

equations (3)-(6) of §2.) entails that $\widehat{M}$ defines a model $\widehat{N}$ of the Peano axioms, that $\widehat{N}$ defines a model $\widehat{M'}$ of Hume's Principle, and that $\widehat{M}$ and $\widehat{M'}$ are isomorphic. Now, using methods familiar from the traditional proof of Frege's Theorem note that the standard model of second-order arithmetic is definable in $\widehat{M}$. Hence, using the isomorphism from $\widehat{M}$ to $\widehat{M'}$, we have that an isomorphic copy $\widehat{C}$ of the standard model of second-order arithmetic definable in $\widehat{M'}$ and hence in $\widehat{N}$. By primitive recursion in $\widehat{N}$ (cf. Simpson (2009) p. 69), there is an $\widehat{N}$-definable function $\iota : N \to C$ such that

$$\iota(0_N) = 0_C \qquad \iota(S_N(n)) = S_C(\iota(n)) \tag{34}$$

Using induction in $\widehat{N}$, along with the fact that both $\widehat{N}$ and $\widehat{C}$ are models of the axioms of Robinson's Q, one has that $\iota : N \to C$ is an injection, so that N is countable. From this it follows that the cardinality of $\widehat{N}$ is $\leq |P(\omega)|$ (cf. equation (15) for the definition of the cardinality of a many-sorted structure). But then we reach a contradiction. For, the cardinality of $\widehat{M'}$ is $|P(P(\omega))|$, while the cardinality of $\widehat{N}$ is $\leq |P(\omega)|$. Hence, $\widehat{M'}$ cannot possibly be definable in $\widehat{N}$.

Now we show that Hume's Principle and the Peano Axioms are not fully sententially equivalent. Recall that the notion of sentential equivalence was defined at the close of §2. and the notion of full interpretations was defined at the opening of § 6.1. To say that two theories are fully sententially equivalent is just to say that they are sententially equivalent via full interpretations. Now, let κ be the least infinite cardinal such that $\kappa = \aleph_\kappa$. This implies that there is a bijection $\iota : \{|X| : X \subseteq \kappa\} \to \kappa$. Define $\# : P(\kappa) \to \kappa$ to be $\#(X) = \iota(|X|)$, so that $\# : P(\kappa) \to \kappa$ is a surjection. Then consider the following normal model of Hume's Principle:

$$\widehat{M} = (M, S_1[M], S_2[M], \ldots, \#) \tag{35}$$

wherein $M = \kappa$ and $S_k[M] = P(M^k)$. Note that the cardinality of $\widehat{M}$ is $|P(\kappa)|$. Now the supposition of sentential equivalence entails that $\widehat{M}$ defines a model $\widehat{N}$ of the Peano axioms, that $\widehat{N}$ defines a model $\widehat{M'}$ of Hume's Principle, and that $\widehat{M}$ and $\widehat{M'}$ are elementary equivalent. Further, the supposition of fullness implies that $N \subseteq M$ and $S_k[N] = P(N^k)$ and $M' \subseteq N$ and $S_k[M'] = P((M')^k)$. Further, let us write the "number of" function of $\widehat{M'}$ as $\#'$. Now, note by construction we have the following, where Φ_ω is a sentence in the signature of second-order logic which is true on full models if and only if their first-order domain is countably infinite:

$$\widehat{M} \models (\forall\, x\, \exists\, F\, \#(F) = x)\ \&\ \neg\Phi_\omega \tag{36}$$

Since $\widehat{M}$ and $\widehat{M'}$ are elementarily equivalent, we have that

$$\widehat{M'} \models (\forall\, x\, \exists\, F\, \#'(F) = x)\ \&\ \neg\Phi_\omega \tag{37}$$

Let $\lambda = |M'|$. Since $\widehat{M'} \models \neg\Phi_\omega$, we have that $\lambda > \omega$. We claim that $\lambda = \kappa$. Choose a bijection $\pi : \lambda \to M'$ and the induced bijection $\overline{\pi} : P(\lambda) \to P(M')$ given by $\overline{\pi}(X) = \{\pi(x) : x \in X\}$. Consider the map $\iota' : \{|X| : X \subseteq \lambda\} \to \lambda$ given by $\iota'(|X|) = \pi^{-1}(\#'(\overline{\pi}(X)))$. By construction ι' is an injection and by the previous equation, ι' is a surjection, so that ι' is a bijection. Hence we have

$$|\{|X| : X \subseteq \lambda\}| = \lambda \tag{38}$$

Since $\lambda > \omega$ we have that

$$|\{|X| > \omega : X \subseteq \lambda\}| = \lambda \tag{39}$$

Then we claim that $\lambda = \aleph_\lambda$. For, suppose not. Then $\lambda \leq \aleph_\alpha$ for some least $\alpha < \lambda$. Consider the injection $j : \{|X| > \omega : X \subseteq \lambda\} \to (\alpha+1)$ satisfying $|X| = \aleph_{j(|X|)}$. But this is a contradiction, since it implies that $\lambda \leq |\alpha|$ and we know that $|\alpha| < \lambda$. So indeed $\lambda = \aleph_\lambda$. By the definition of κ as the smallest infinite cardinal satisfying $\kappa = \aleph_\kappa$, we indeed have that $\lambda = \kappa$. So choose a bijection $\theta : M' \to M$ and extend to a bijection $\overline{\theta} : P(M') \to P(M)$. Then define a bijection $\pi' : M' \to M$ by

$$\pi'(\#'(F)) = \#(\overline{\theta}(F)) \tag{40}$$

This function is well-defined and injective because:

$$\widehat{M'} \models \#'(F) = \#'(G) \Leftrightarrow |F| = |G| \Leftrightarrow \left|\overline{\theta}(F)\right| = \left|\overline{\theta}(G)\right| \Leftrightarrow \widehat{M} \models \#(\overline{\theta}(F)) = \#(\overline{\theta}(G)) \tag{41}$$

It has domain M' because of equation (37), and it is surjective because of equation (36). Further to show that $\overline{\pi}' : \widehat{M'} \to \widehat{M}$ is an isomorphism it suffices to show that

$$\#(\overline{\pi}'(F)) = \pi'(\#'(F)) \tag{42}$$

But we show this by arguing for the identity from right-to-left

$$\pi'(\#'(F)) = \#(\overline{\theta}(F)) = \#(F) = \#(\overline{\pi}'(F)) \tag{43}$$

Hence, in fact $\widehat{M}$ and $\widehat{M'}$ are isomorphic. Now we may argue as in the proof of biinterpretability.

6.4. Provable Equivalence of Mathematical Induction and Domain Biconditional So here we prove an equivalence mentioned in §4.1, namely that the Mathematical Induction Axiom is equivalent to the domain biconditional against the background of the zero and successor biconditionals and the supposition that zero is a natural number and that successors of natural numbers are natural numbers. This latter supposition can be rendered symbolically as:

$$N(0) \;\&\; (\forall\, y, z\; (N(y) \;\&\; S(y, z)) \to N(z)) \tag{44}$$

Further, for ease of verification, we reproduce here the definability biconditionals, which again we respectively call the domain biconditional, the zero biconditional, and the successor biconditional:

$$\forall\, x\; [N(x) \leftrightarrow N^*(x)] \qquad \forall\, y\; [y = 0 \leftrightarrow Z^*(y)] \qquad \forall\, y, z\; [S(y, z) \leftrightarrow [N(y) \;\&\; S^*(y, z)]] \tag{7}$$

For the purposes of the proof, the exact specifications of Z^* and S^* are immaterial. However, the proof depends crucially on the traditional specification of N^*, which again we repeat here for the ease of verification:

$$N^*(x) \equiv \forall\, F\; [(\forall\, y\; Z^*(y) \to Fy) \;\&\; (\forall\, y, z\; (Fy \;\&\; S^*(y, z)) \to Fz)] \to F(x) \tag{8}$$

Again, the way to read this equation is to note that the variable x appears on the right-hand-side only in the last consequent, so that $N^*(x)$ says that x has all properties F which satisfy a certain condition, namely, containing the ersatz of zero and being closed under the ersatz of successor. Further, we can formulate the

Mathematical Induction Axiom as:

$$\forall\, F\, [F(0)\, \&\, (\forall\, y, w\, (F(y)\, \&\, S(y,w) \rightarrow F(w))] \rightarrow [\forall\, x\, (N(x) \rightarrow F(x))] \qquad (45)$$

This differs from usual textbook presentations of the Mathematical Induction Axiom, such as we gave in Appendix 6.1., in that successor is treated as a relation instead of a function. This is because in the context of the definability biconditionals, one does not want to assume that the successor operation is total, since part of what the logicist seeks to do is to certify the logical credentials of this arithmetical knowledge. However, nothing in the below proof depends on this, and the reader can easily verify that everything goes through if the relational expression $S(y,z)$ is systematically replaced with the functional expression $S(y) = z$.

Now first let us note that our background knowledge suffices to straightforwardly deduce the right-to-left direction of the domain biconditional. For, suppose that $N^*(a)$ holds, and let us attempt to show that $N(a)$ holds. Since $N^*(a)$ holds, by substituting N for the variable F in equation (8), we see that it suffices to show that

$$[(\forall\, y\; Z^*(y) \rightarrow Ny)\, \&\, (\forall\, y, z\, (Ny\, \&\, S^*(y,z)) \rightarrow Nz)] \qquad (46)$$

To verify this, first suppose that $Z^*(y)$. Then by the right-to-left direction of zero biconditional, it follows that $y = 0$ and thus $N(y)$ by equation (44). To complete the verification, suppose that $Ny\, \&\, S^*(y,z)$. Then by the right-to-left direction of the successor biconditional, it follows that $S(y,z)$ and thus $N(z)$ by equation (44). It follows that we have finished verifying (46). So this is why the background knowledge implies the right-to-left direction of the domain biconditional, which we may abbreviate symbolically by $N^* \subseteq N$.

Now it remains to show that against this background knowledge, we have that the Mathematical Induction Axiom is equivalent to the domain biconditional. So first suppose that the Mathematical Induction Axiom holds. By the result of the previous paragraph, it suffices to show the left-to-right direction of the domain biconditional, or in symbols $N \subseteq N^*$. Now, the next step in the proof is to substitute N^* for F in the Mathematical Induction Axiom in equation (45). Formally, what justifies this procedure is an instance of the full comprehension schema, which would say that there is a property corresponding to the formula N^*. As mentioned in Appendix §6.1., we're assuming the full comprehension schema here since there are simply fewer results like Frege's Theorem if one restricts the comprehension schema. Now, substituting N^* for F in the Mathematical Induction Axiom in equation (45) in this manner, we see that it suffices to verify:

$$[N^*(0)\, \&\, (\forall\, y, w\, (N^*(y)\, \&\, S(y,w))) \rightarrow N^*(w))] \qquad (47)$$

To verify this, we first show that $N^*(0)$. By definition of N^* in equation (8), this means we have to verify:

$$\forall\, F\, [(\forall\, y\; Z^*(y) \rightarrow Fy)\, \&\, (\forall\, y, z\, (Fy\, \&\, S^*(y,z)) \rightarrow Fz)] \rightarrow F(0) \qquad (48)$$

So suppose that F satisfies the antecedent of this equation. By the left-to-right direction of the zero biconditional, it follows that $Z^*(0)$ and thus $F(0)$, which is exactly the consequent of this equation. So we have shown that $N^*(0)$. Now suppose that $N^*(y)\, \&\, S(y,w)$, and we try to show that $N^*(w)$. By definition of N^*

in equation (8), it again suffices to show that

$$\forall F\,[(\forall y\, Z^*(y) \rightarrow Fy)\ \&\ (\forall y, z\,(Fy\ \&\ S^*(y,z)) \rightarrow Fz)] \rightarrow F(w) \qquad (49)$$

So suppose that F satisfies the antecedent of this equation. Since $N^*(y)$ holds, we see from the definition in equation (8) that $F(y)$. Since $S(y,w)$, it follows from the left-to-right direction of the successor biconditional that $S^*(y,w)$. Since $F(y)$ and $S^*(y,w)$, it follows that $F(w)$ since F satisfies the antecedent of equation (49). Thus, we have succeeded in verifying equation (47), so that we may indeed conclude that $N \subseteq N^*$. Combining this with our earlier knowledge that $N^* \subseteq N$, we have that N and N^* are coextensive, which is exactly what the domain biconditional expresses.

Finally we want to show that from the background knowledge and the domain biconditional, we can deduce the Mathematical Induction Axiom in equation (45). So suppose that F is such that

$$[F(0)\ \&\ (\forall y, w\, F(y)\ \&\ S(y,w) \rightarrow F(w))] \qquad (50)$$

We must show that all the natural numbers have property F, or symbolically that $N \subseteq F$. Let G be the conjunction of F and N. Thus, it suffices to show that $N \subseteq G$. So suppose that $N(a)$. Then $N^*(a)$ by the left-to-right direction of the domain biconditional. By substituting G for the variable F in the definition of N^* in equation (8), to conclude that $G(a)$, it suffices to show that

$$[(\forall y\, Z^*(y) \rightarrow Gy)\ \&\ (\forall y, z\,(Gy\ \&\ S^*(y,z)) \rightarrow Gz)] \qquad (51)$$

Suppose then that $Z^*(y)$. By the right-to-left direction of the zero biconditional, it follows that $y = 0$ and thus the hypothesis of $F(0)$ from equation (50) implies $F(y)$. Likewise, since $y = 0$ and we have $N(0)$ from equation (44), we thus have $N(y)$. Since $F(y)$ and $N(y)$, one has $G(y)$. To finish the verification of equation (51), suppose that $Gy\ \&\ S^*(y,z)$. Then $Ny\ \&\ S^*(y,z)$. By the right-to-left direction of the successor biconditional, one has that $S(y,z)$. Since $G(y)$ one also has $F(y)\ \&\ S(y,z)$. But from the hypothesis in equation (50) we can conclude that $F(z)$. Further, from equation (44) we have that $Ny\ \&\ S(y,z)$ implies $N(z)$, so that it follows that $G(z)$, which finishes the verification of equation (51).

Bibliography

Ahlbrandt, G., & Ziegler, M. (1986). Quasi-Finitely Axiomatizable Totally Categorical Theories. Annals of Pure and Applied Logic **30**(1), 63–82.

Benacerraf, P. (1981). Frege: The Last Logicist. Midwest Studies in Philosophy **6**, 17–35. Reprinted in Demopoulos (1995).

Blanchette, P. A. (1994). Frege's Reduction. History and Philosophy of Logic **15**, 85–103.

Blanchette, P. A. (2012). Frege's Conception of Logic. Oxford: Oxford University Press.

Boolos, G. (1984). The Justification of Mathematical Induction. Proceedings of the Biennial Meeting of the Philosophy of Science Association **2**, 469–475. Reprinted in Boolos (1998).

Boolos, G. (1996). On the Proof of Frege's Theorem. In Benacerraf and His Critics, pp. 143–159. Blackwell. Edited by Adam Morton and Stephen P. Stich. Reprinted in Boolos (1998).

Boolos, G. (1998). Logic, Logic, and Logic. Cambridge, MA: Harvard University Press. Edited by Richard Jeffrey.

Bos, H. J. M. (2001). Redefining Geometrical Exactness: Descartes' Transformation of the Early Modern Concept of Construction. Sources and Studies in the History of Mathematics and Physical Sciences. New York: Springer.

Burgess, J. P. (2005). Fixing Frege. Princeton Monographs in Philosophy. Princeton: Princeton University Press.

Buss, S. R. (2006). Nelson's Work on Logic and Foundations and Other Reflections on the Foundations of Mathematics. In Faris, W. G., editor, Diffusion, Quantum Theory, and Radically Elementary Mathematics, Volume 47 of Mathematical Notes, pp. 183–208. Princeton, NJ: Princeton University Press.

Cook, R. T., editor (2007). The Arché Papers on the Mathematics of Abstraction, Volume 71 of The Western Ontario Series in Philosophy of Science. Berlin: Springer.

Corcoran, J. (1980). On Definitional Equivalence and Related Topics. History and Philosophy of Logic **1**, 231–234.

de Bouvère, K. (1965a). Logical Synonymity. Indagationes Mathematicae **27**, 622–629.

de Bouvère, K. (1965b). Synonymous Theories. In The Theory of Models, pp. 402–406. North-Holland. Edited by J. W. Addison et. al.

Demopoulos, W. (1994). Frege and the Rigorization of Analysis. Journal of Philosophical Logic **23**, 225–246. Reprinted in Demopoulos (1995).

Demopoulos, W., editor (1995). Frege's Philosophy of Mathematics. Cambridge: Harvard University Press.

Demopoulos, W., & Clark, P. (2005). The Logicism of Frege, Dedekind, and Russell. In Shapiro, S., editor, The Oxford Handbook of Philosophy of Mathematics and Logic, pp. 129–165. Oxford University Press.

Dummett, M. (1996). Frege and the Paradox of Analysis. In Frege and Other Philosophers, pp. 17–52. Oxford: Oxford University Press.

Ebbinghaus, H.-D. (1985). Extended Logics: The General Framework. In Barwise, J. & Feferman, S., editors, Model-Theoretic Logics, Perspectives in Mathematical Logic, pp. 25–76. New York: Springer.

Enayat, A., Schmerl, J. H., & Visser, A. (2011). ω-Models of Finite Set Theory. In Kennedy, J. & Kossak, R., editors, Set Theory, Arithmetic, and Foundations of Mathematics: Theorems, Philosophies, Volume 36 of Lecture Notes in Logic, pp. 43–65. La Jolla, CA: Association for Symbolic Logic.

Enderton, H. B. (2001). A Mathematical Introduction to Logic (Second ed.). Burlington: Harcourt.

Feferman, S. (1960/1961). Arithmetization of Metamathematics in a General Setting. Fundamenta Mathematicae **49**, 35–92.

Fischer, M. (2010). Deflationism and Reducibility. In Tadeusz Czarnecki, Katarzyna Kijania-Placek, O. P. & Wolenski, J., editors, The Analytical Way. Proceedings of the 6th European Congress of Analytic Philosophy, pp. 357–369. London: College Publications.

Frege, G. (1967). Kleine Schriften. Hildesheim: Olms. Edited by Ignacio Angelelli.

Frege, G. (1980). The Foundations of Arithmetic: A Logico-Mathematical Enquiry into the Concept of Number (second ed.). Evanston: Northwestern University Press. Translated by John Langshaw Austin.

Givant, S., & Halmos, P. (2009). Introduction to Boolean algebras. Undergraduate Texts in Mathematics. New York: Springer.

Goodman, N. (1951). The Structure of Appearance. Cambridge: Harvard University Press.

Hájek, P., & Pudlák, P. (1998). Metamathematics of First-Order Arithmetic. Perspectives in Mathematical Logic. Berlin: Springer.

Hale, B. (1987). Abstract Objects. Oxford: Basil Blackwell.

Hale, B. (2000). Reals by Abstraction. Philosophia Mathematica **8**(3), 100–123. Reprinted in Hale & Wright (2001), Cook (2007).

Hale, B., & Wright, C. (2001). The Reason's Proper Study. Oxford: Oxford University Press.

Heck, Jr., R. G. (1997). Finitude and Hume's principle. Journal of Philosophical Logic **26**(6), 589–617. Reprinted in Cook (2007) and with additional postscript in Heck (2011).

Heck, Jr., R. G. (1999). Frege's Theorem: An Introduction. The Harvard Review of Philosophy **7**, 56–73.

Heck, Jr., R. G. (2000). Cardinality, Counting, and Equinumerosity. Notre Dame Journal of Formal Logic **41**(3), 187–209. Reprinted in Heck (2011).

Heck, Jr., R. G. (2011). Frege's Theorem. Oxford: Oxford University Press.

Hellman, G. (1978). Accuracy and Actuality. Erkenntnis **12**, 209–228.

Hochberg, H. (1956). Peano, Russell, and Logicism. Analysis **16**(5), 118–120.

Hochberg, H. (1970). Russell's Reduction of Arithmetic to Logic. In Klemke, E., editor, Essays on Bertrand Russell, pp. 396–415. Urbana: University of Illinois Press. Reprinted in Hochberg (1984).

Hochberg, H. (1984). Logic, Ontology, and Language: Essays on Truth and Reality. München: Analytica (Philosophia Verlag).

Hodes, H. (1990). Where Do The Natural Numbers Come From? Synthese **84**(3), 347–407.

Hodges, W. (1993). Model Theory, Volume 42 of Encyclopedia of Mathematics and its Applications. Cambridge: Cambridge University Press.

Hook, J. L. (1985). A Note on Interpretations of Many-Sorted Theories. The Journal of Symbolic Logic **50**(2), 372–374.

Horsten, L. (2011). The Tarskian Turn. Cambridge: The MIT Press.

Iwan, S. (2000). On the Untenability of Nelson's Predicativism. Erkenntnis **53**(1-2), 147–154.

Kalsbeek, M. (1989). An Orey Sentence for Predicative Arithmetic. Unpublished. Master's Thesis, Institute for Language, Logic, and Information. ITLI Prepublication Series X-89-01.

Kleene, S. C. (1952). Introduction to Metamathematics, Volume 1 of Bibliotheca Mathematica. Amsterdam: North-Holland.

Koellner, P. (2009). Truth in Mathematics: The Question of Pluralism. In New Waves in the Philosophy of Mathematics, pp. 80–116. Palmgrave. Edited by Otávio Bueno and Øystein Linnebo.

Leitgeb, H. (2009). On Formal and Informal Provability. In New Waves in the Philosophy of Mathematics, pp. 263–299. New York: Palmgrave. Edited by Otávio Bueno and Øystein Linnebo.

Lindström, P. (2003). Aspects of Incompleteness (second ed.), Volume 10 of Lecture Notes in Logic. Urbana, IL: Association for Symbolic Logic.

Manzano, M. (1996). Extensions of First Order Logic, Volume 19 of Cambridge Tracts in Theoretical Computer Science. Cambridge: Cambridge University Press.

Marker, D. (2002). Model Theory: An Introduction, Volume 217 of Graduate Texts in Mathematics. New York: Springer-Verlag.

Monk, J. D. (1976). Mathematical Logic. Number 37 in Graduate Texts in Mathematics. New York: Springer-Verlag.

Nelson, E. (1986). Predicative Arithmetic, Volume 32 of Mathematical Notes. Princeton, NJ: Princeton University Press.

Nies, A. (2007). Describing groups. Bulletin of Symbolic Logic **13**(3), 305–339.

Papert, S. (1960). Sur le réductionniseme logique. In Gréco, P., Grize, J.-B., Papert, S., & Piaget, J., editors, Problèmes de la construction du nombre, Volume 11 of Etudes d'épistémologie génétique, pp. 97–116. Paris: Presses Universitaires de France.

Parsons, C. (1965). Frege's Theory of Number. In Black, M., editor, Philosophy in America, pp. 180–203. Ithaca: Cornell University Press. Reprinted with a new postscript in Parsons (1983), Demopoulos (1995).

Parsons, C. (1983). Mathematics in Philosophy: Selected Essays. Ithaca: Cornell University Press.

Parsons, C. (2008). Mathematical Thought and Its Objects. Cambridge: Harvard University Press.

Polánski, M. (2009). Goodman's Extensional Isomorphism and Syntactical Interpretations. Theoria. An International Journal for Theory, History and Foundations of Science **65**, 203–211.

Quine, W. (1964). Ontological Reduction and the World of Numbers. The Journal of Philosophy **61**(7), 209–216. Reprinted in Quine (1976).

Quine, W. (1976). Ways of Paradox and Other Essays. New York: Random House.

Resnik, M. D. (1981). Mathematics as a Science of Patterns: Ontology and Reference. Nous **15**(4), 529–550.

Resnik, M. D. (1997). Mathematics as a Science of Patterns. Oxford: Clarendon.

Shapiro, S. (1991). Foundations without Foundationalism: A Case for Second-Order Logic, Volume 17 of Oxford Logic Guides. New York: The Clarendon Press.

Shapiro, S. (2000a). Frege Meets Dedekind: A Neo-Logicist Treatment of Real Analysis. Notre Dame Journal of Formal Logic **41**(4), 335–364. Reprinted in Cook (2007).

Shapiro, S. (2000b). Philosophy of Mathematics: Structure and Ontology. Oxford: Oxford University Press.

Simpson, S. G. (2009). Subsystems of Second Order Arithmetic (second ed.). Cambridge: Cambridge University Press.

Steiner, M. (1975). Mathematical Knowledge. Ithaca: Cornell University Press.

Visser, A. (2006). Categories of Theories and Interpretations. In Enayat, A., Kalantari, I., & Moniri, M., editors, Logic in Tehran, Volume 26 of Lecture Notes in Logic, pp. 284–341. La Jolla: Association for Symbolic Logic.

Visser, A. (2011). Hume's Principle, Beginnings. Review of Symbolic Logic **4**(1), 114–129.

Williamson, T. (2007a). Conceptual Truth. Aristotelian Society, Supplementary Volume **80**(1), 1–41.

Williamson, T. (2007b). The Philosophy of Philosophy. Blackwell: Blackwell Publishers.

Wright, C. (1983). Frege's Conception of Numbers as Objects, Volume 2 of Scots Philosophical Monographs. Aberdeen: Aberdeen University Press.

Wright, C. (1998a). On the Harmless Impredictavity of $N^{=}$ (Hume's Principle). In Philosophy of Mathematics Today, pp. 393–368. Oxford: Clarendon Press. Edited by Matthias Schirn. Reprinted in Hale & Wright (2001).

Wright, C. (1998b). Response to Dummett. In Philosophy of Mathematics Today, pp. 389–405. Oxford: Clarendon Press. Reprinted in Hale & Wright (2001).

Wright, C. (1999). Is Hume's Principle Analytic? Notre Dame Journal of Formal Logic **40**(1), 6–30. Reprinted in Hale & Wright (2001) and Cook (2007).

Wright, C. (2000). Neo-Fregean Foundations for Real Analysis: Some Reflections on Frege's Constraint. Notre Dame Journal of Formal Logic **41**(4), 317–334. Reprinted in Cook (2007).