Notre Dame Journal of Formal Logic
Volume 40, Number 1, Winter 1999

# Is Hume's Principle Analytic?

CRISPIN WRIGHT

**Abstract** One recent 'neologicist' claim is that what has come to be known as "Frege's Theorem"—the result that Hume's Principle, plus second-order logic, suffices for a proof of the Dedekind-Peano postulate—reinstates Frege's contention that arithmetic is analytic. This claim naturally depends upon the analyticity of Hume's Principle itself. The present paper reviews five misgivings that developed in various of George Boolos's writings. It observes that each of them really concerns not 'analyticity' but either the truth of Hume's Principle or our entitlement to accept it and reviews possible neologicist replies. A two-part Appendix explores recent developments of the fifth of Boolos's objections—the problem of Bad Company—and outlines a proof of the principle $N^q$, an important part of the defense of the claim that what follows from Hume's Principle is not merely a theory which allows of interpretation as arithmetic but arithmetic itself.

***1*** It was George Boolos who, following Frege's somewhat charitable lead at *Grundlagen* §63, first gave the name, "Hume's Principle," to the constitutive principle for identity of cardinal number: that the number of *F*s is the same as the number of *G*s just in case there exists a one-to-one correlation between the *F*s and the *G*s. The interest—if indeed any—of the question whether the principle is analytic is wholly consequential on what has come to be known as *Frege's Theorem*: the proof, prefigured in *Grundlagen* §§82–83 [5] and worked out in some detail in Wright [21][1] that second-order logic plus Hume's Principle as sole additional axiom suffices for a derivation of second-order arithmetic—or, more cautiously, for the derivation of a theory which allows of interpretation as second-order arithmetic. (Actually I think the caution is unnecessary—more of that later.) Analyticity, whatever exactly it is, is presumably transmissible across logical consequence. If second-order consequence is indeed a species of logical consequence, the analyticity of Hume's Principle would ensure the analyticity of arithmetic—at least, provided it really is second-order arithmetic, and not just a theory which merely allows interpretation as such, which is a second-order consequence of Hume's Principle. What significance that finding would have would then depend, of course, on the significance of the notion of analyticity

*Received April 16, 1998*

itself. Later I shall suggest that the most important issues here are ones which are formulable without recourse to the notion of analyticity at all—so that much of the debate between Boolos and me could have finessed the title question.

Boolos wrote that "having to discuss whether Hume's Principle is analytic is rather like having to consider whether hydrogen sulphide is dephlogisticated"—a question formulated, I suppose he meant, in a discredited theoretical vocabulary ([4], p. 247). That would be consistent, of course, with there being a good question nearby of which that was merely a theoretically unfortunate expression; it would also be consistent with there being enough sense to the theoretically unfortunate question to allow of a negative answer in any case. I myself do not believe that when the dust settles on analytical philosophy's first century, our successors will find that the notion of analyticity *was* discredited by any of the well-known assaults. In particular, the two core lines of attack in "Two Dogmas of Empiricism," namely, that the notion resists all noncircular explanation and that no statement participating in general empirical theory can be immune to revision, set an impossible—Socratic—standard for conceptual integrity and confuse analyticity with indefeasible certainty, respectively. What is undeniable though is that the status and provenance of analytic truths, and the cognate class of a priori necessary truths, would have to be a lot clearer than philosophers have so far managed to make them before a positive answer to our title question could be justified and shown to have the sort of significance which early analytical philosophy would have accorded to it.

Boolos thought the situation was of the second kind: that the question is theoretically flawed but allows of well-motivated—though less than "knock-down"—arguments for a negative answer. To the best of my knowledge—I am drawing just on three of his papers (Boolos [2], [3], and [1]) which are reprinted in the excellent Demopoulos collection [7], plus his ipsonymous paper in Heck's volume for Dummett (Boolos [4] in [8])—he proffered exactly five such arguments. In what follows I shall briefly explore how a character I shall call the neo-Fregean might respond to each of these arguments. Each is interesting, some are very searching, but—if I am right—none does irreparable damage.

## *2*

***2.1 The ontological concern*** The ontological concern is epitomized in the following passage:

> I want to suggest that Hume's Principle is to be likened to 'the present King of France is a royal' in that we have no analytic guarantee that for every value of '*F*', there is an object that the open definite singular description, 'the number belonging to *F*' denotes. . . . Our present difficulty is this: just how do we know, what kind of guarantee do we have, why should we believe, that there is a function that maps concepts to objects in the way that the denotation of octothorpe [that is, '#', Boolos's symbol for the numerical operator] does if HP is true? . . . do we have any analytic guarantee that there is a function that works in the appropriate manner?[2]

The basic thought is that Hume's Principle *says too much* to be an analytic truth. As normally conceived, analytic truth must hold in any *possible* domain. On a (purportedly) more relaxed conception, some analytic truths are allowed to hold in any

*nonempty* domain. But how can a principle which entails—indeed, is strictly stronger than is necessary to entail—that there are infinitely many objects—indeed infinitely many objects of a special sort—possibly count as analytic?

Here is the neo-Fregean reply. There is, to be sure, a good sense in which whatever is entailed by certain principles together with truths of logic may be regarded as entailed by those principles alone. In this sense it is undeniable that Hume's Principle does entail the existence of infinitely many objects—at least if second-order consequence is a species of entailment. But the *manner* of the entailment is important. Hume's Principle is a second-order universally quantified biconditional. As such, we are not going to be able to elicit the existence of any objects at all out of it save by appropriate input into (instances of) its right-hand side. Thus we get the number zero by taking the instance of Hume's Principle,

$$Nx : x \neq x = Nx : x \neq x \longleftrightarrow x \neq x \quad 1 \approx 1 \quad x \neq x, \tag{1}$$

together with its right-hand side as a minor premise. Compare the fashion in which we derive the direction of the line $\underline{a}$ from an instantiation of Frege's illustrative equivalence for directions:

(DE) $$D\underline{a} = D\underline{a} \longleftrightarrow \underline{a}//\underline{a}$$

together with its right-hand side as a minor premise: the necessary truth, *modulo* the existence of line $\underline{a}$, that that line is parallel to itself. Sure, in the case of zero the minor premise

$$x \neq x \quad 1 \approx 1 \quad x \neq x \tag{2}$$

can be established in second-order logic. So the existence of zero follows from this truth of logic, together with Hume's Principle. *If*, accordingly, the latter can be regarded as, in all relevant respects, having a status akin to that of a *definition*, then the existence of zero is a consequence of logic and definitions. But that was exactly the classical account of analyticity: the analytical truths were to be those which follow from logic and definitions. So the existence of zero would be an analytic truth. And now with that in the bag, as it were, nothing stands in the way of regarding

$$x = 0 \quad 1 \approx 1 \quad x = 0 \tag{3}$$

as also an analytic truth, since it follows in second-order logic given only that there is such a thing as zero. But that is the right-hand side for the application of Hume's Principle which, following Frege, we use to obtain the number one. So its existence is also analytic. We may now proceed in similar fashion to obtain each of the finite cardinals from putatively analytic premises, in second-order logic. Our result is thus not quite—when done this way—that it is analytic that there is an infinity of finite cardinals, but rather that of each of the finite cardinals, it is analytic that it exists. Doubtless this will be equally offensive to the traditional understanding of analyticity—the (as nearly as possible) *existentially neutral* understanding of analyticity—called forth in the above quotation from Boolos. But my point now is simply that, for the reasons just sketched, *that* understanding of analyticity had to be in jeopardy all along provided there is a starting chance that Hume's Principle has an epistemic status relevantly similar to that of a definition.

In sum, on the classical account of analyticity, the analytical truths are those which follow from logic and definitions. So if the existence of zero, one, and so on follows from logic plus Hume's principle, then provided the latter has a status relevantly similar to that of a definition, it will be analytic, on the classical account, that $n$ exists, for each finite cardinal $n$. The idea which standardly accompanies the classical conception, that—with perhaps a very few, modest exceptions—existential claims can never be analytically true, is thus potentially in tension with the classical conception. If Hume's Principle has a status not relevantly different from that of a definition, then we learn that the classical conception will not marry with this standardly accompanying idea.

The core of the neo-Fregean stance is that Hume's Principle *does* have such a status: that it may be seen as an explanation of the concept of cardinal number in general, covering the finite cardinals as a special case. Boolos asks, "If numbers are supposed to be identical if and only if the concepts they are numbers of are equinumerous, what guarantee have we that every concept *has* a number?" ([4], p. 253). Earlier he suggested, in the passage quoted, that there is no such guarantee—or anyway no "analytic guarantee"—proposing a parallel between the principle and the statement, "The present King of France is a royal"—something which is analytically true, *modulo* its existential presupposition. This is also Field's position ([12], [13]) in his critical notice of Wright [21]. But I think this seemingly sane and reserved position is unstable.

Consider the case of direction again. How do we know there are any objects which behave in the way that the referents of direction terms ought to behave, given their introduction by the direction equivalence (DE); that is, given that they are identical just in case the associated lines are parallel and distinct just in case they are not? Shouldn't we just say that *provided* there are such things as directions in the first place, that will be the condition for their identity and distinctness? Well, if this were the right view of the matter, there could be no objection to making the presupposition explicit. The following principle would then count as *absolutely* analytic: that for any lines $\underline{a}$ and $\underline{b}$,

$$((\exists x)(\exists y)(x = D\underline{a} \,\&\, y = D\underline{b})) \longrightarrow (D\underline{a} = D\underline{b} \longleftrightarrow \underline{a}//\underline{b}) \qquad (4)$$

But think: how are we to understand the antecedent of this? The condition for its truth must now incorporate some unreconstructed idea of what it is for contexts of the form, '$p = D\underline{a}$' and '$q = D\underline{b}$' to be true—unreconstructed because Field and Boolos have just rejected the proposed sufficient conditions for the truth of such contexts, where '$p$' and '$q$' are, respectively, direction terms, incorporated in DE. However, no other such sufficient condition has been proposed. So, if we side with Field and Boolos, we do not have the slightest idea, actually, of what satisfaction of the antecedent of the supposedly more modest and reserved formulation could consist in.

True, the reserved formulation could be made to raise an intelligible issue if relativized to an *antecedently given* domain of quantification—the issue would be whether any of the objects thereby already recognized, perhaps certain equivalence classes, are appropriately identified and distinguished in the light of relations of parallelism among lines. But Frege, remember, was trying to address the question how we *come by and justify* the conception of a domain of abstracta in the first place. If it is insisted that abstraction principles *always* stand in need of justification by ref-

erence to an antecedently given domain of entities, that is just to presuppose—not argue—that they are useless in that project. And it is so far to offer no alternative conception of how the project might be accomplished. The neo-Fregean contention, by contrast, is that, under the right conditions, such principles are available to fix the truth-conditions of contexts of identity for a certain kind of thing and thereby—given appropriate input on their right-hand sides—to contribute toward determining that, and how it is possible for us to know that things of that kind exist.

Boolos's question "If numbers are supposed to be identical if and only if the concepts they are numbers of are equinumerous, what guarantee have we that every concept *has* a number?," raises a doubt in the way he presumably wished to do only if it is granted that the existence of numbers is a *further* fact, something which the (mere) equinumerosity of concepts may leave unresolved. But the neo-Fregean's intention in laying down Hume's Principle as an explanation is so to fix the concept of cardinal number that the equinumerosity of concepts $F$ and $G$ is *itself* to be necessary and sufficient, without further ado, for the identity of the number of $F$s with the number of $G$s, so that nothing more is required for the existence of those numbers beyond the equinumerosity of the concepts. This idea is discussed more fully in the early sections of Wright [22] and in Hale [14]. The key idea is that an instance of the left-hand side of an abstraction principle is meant to embody a *reconceptualization* of the type of state of affairs depicted on the right. Here is not the place to pursue this crucial idea further. My point is merely that Boolos's question either ignores this aspect of the neo-Fregean position or assumes it is ill-conceived.

***2.2 The epistemological concern*** A recurrent element in Boolos's misgivings about Hume's Principle concerns its proof-theoretic strength—more accurately, the strength of the system which results from its addition to axiomatic second-order logic. In part this concern relates to the ontological issues just reviewed. But there is a separate strand, nicely captured by a passage toward the end of "Is Hume's Principle Analytic?" Boolos was the first to show that second-order logic plus Hume's Principle is equi-interpretable with second-order arithmetic, and hence that each is consistent if the other is.[3] But he was not himself inclined to take that result as settling the question of the consistency of Hume's Principle. He writes:

> (it is *not* neurotic to think) we don't *know* that second-order arithmetic . . . is consistent. Do we really know that some hotshot Russell of the 23rd Century won't do for us what Russell did for Frege? The usual argument by which we think we can convince ourselves that analysis is consistent—"Consider the power set of the set of natural numbers . . . "—is flagrantly circular. . . . Uncertain as we are whether Frege arithmetic is consistent, how can we (dare to) call HP analytic? ([4], pp. 259–60)

Now, I do not myself know whether disclaiming knowledge of the consistency of Frege arithmetic is neurotic or not. But we must surely look askance at the presupposition of the concluding question, which arguably—as did Quine—confuses analyticity and certainty, or anyway insists that certainty is a precondition for *warranted* analyticity claims. That seems to me a great mistake. There is nothing incoherent in the idea that we can be *defeasibly* justified in believing or claiming to know that a proposition is true which, if true, is analytic. The neo-Fregean claim, remember, is

that Hume's Principle serves as an explanation of the concept of cardinal number. *If* it harbors some subtle inconsistency, then, of course, it fails as such an explanation—just as Basic Law V failed as an explanation of a coherent notion of set. But we can surely be fairly confident—though by all means with our eyes open—that Hume's Principle is successful in that regard, and correspondingly confident that it enjoys the kind of truth possessed by any successful implicit definition—and hence is analytic in whatever may be the attendant sense.

***2.3 The concern about the universal number*** The construction of the finite cardinals on the basis of Hume's Principle relies entirely on the legitimacy of applying the numerical operator to some necessarily empty concept at the first stage, the concept *not self-identical* being the standard choice. On the face of it there should accordingly be no obstacle to applying the operator to the *complement* of any such concept, so arriving at the universal number, *anti-zero*—the number of absolutely everything that there is. Certainly Hume's Principle as standardly formulated poses no obstacle to such an application. As Boolos puts it,

> As there is a number, zero, of things that are non self-identical, so, on the account of number we have been considering, there must be a number of things that are self-identical. the number of all the things that there are". ([4], p. 260)

Now, Hume's Principle can be no less dubious than any of its consequences, one of which is the claim then that there is such a number. But

> the worry is this: *is* there such a number as anti-zero? According to [ZF] there is no cardinal number that is the number of all the sets there are. The worry is that the theory of number we have been considering, Frege arithmetic, is incompatible with Zermelo-Frankel set theory plus standard definitions . . . one who seriously believes that [HP is an analytic truth] has to be bothered by the incompatibility of the consequence of Frege arithmetic that there is such a number as anti-zero with the claim made by ZF plus standard definitions . . . that there is no such number. (ibid.)

This objection, Boolos wrote, although it "may at first appear to be dismissible as silly or trivial," is "perhaps the most serious of all."

It is certainly an arresting objection about which there is a good deal to say. Clearly, there would be great discomfort in regarding any principle as analytically true if the cost of doing so was regarding Zermelo-Frankel set theory as *analytically false*. A first rejoinder would be that any such upshot would depend on cross-identification of the referents of terms in Frege arithmetic and terms in Zermelo-Frankel set theory—the "standard definitions" to which Boolos alludes. Who said numbers like anti-zero had to be *sets*, after all? However, the more general worry underlying Boolos's point—the worry about the coherence of Hume's Principle with standard set theory—need not depend on such cross-identification. Grant the plausible principle (to which I return below) that there is a determinate number of *F*s just provided that the *F*s compose a *set*. Zermelo-Frankel set theory implies that there is no set of all sets. So it would follow that there is no number of sets. Yet for all

we have so far seen, the property *set* lies within the range of the second-order quantifiers in Hume's Principle and the usual proof, via the reflexivity of equinumerosity, should therefore serve to establish, to the contrary, that there is such a number. So there would seem to be a collision with Zermelo-Frankel set theory in any case, whether or not anti-zero is identified with a set.

However, I think there is good reason to expect a principled and satisfying response to this general trend of objection. Consider the direction equivalence DE again. The reflexivity of the relation, . . . *is parallel to* . . . , ensures in the presence of DE that $\underline{a}$ has a direction, no matter what straight line $\underline{a}$ may be. But the question arises: what of the implications of DE for the case where $\underline{a}$ and $\underline{b}$ fail to be parallel because they are *not even lines*, as, for example, my hat fails to be parallel to my shoe. We might have been tempted to allow that the *D*-operator is totally defined—to allow that every object, without restriction, has a direction: in the case of an object which fails to be parallel to anything else because it is merely not a line, this would then be a direction that nothing else has. But a moment's reflection shows that is not an option: if the failure of parallelism between my hat and my shoe is down to the unsuitability of either object to be parallel to anything, then by the same token they are not *self*-parallel, and DE provides no incentive to regard either as having a direction at all. Moral: Just as not every object is suitable to determine a direction, so we should not assume without further ado that every concept—every entity an expression for which is an admissible substituend for the bound occurrences of the predicate letters in Hume's Principle—is such as to determine a number.

That is only a first step, of course. What is wanted for the exorcism of anti-zero is nothing less than grounds for affirming that whereas the concept, *not self-identical*, or any other self-contradictory concept, *is* a suitable case for application of the numerical operator, its complement is not. Here are two independent such lines of thought.

The first line is directed specifically at anti-zero. To accept Frege's insight that statements of number are higher-level—that they state things of concepts—is quite consistent with the familiar observation that a restriction is needed which he does not draw. The basic case in which the question How many *F*s are there? makes sense—or at least has a determinate answer—is that of a special class of substitutions for '*F*': what are sometimes called *count nouns*, or expressions for *sortal concepts*. While it is by no means the work of a moment to make this notion sharp, the usual intuitive understanding is that a sortal concept is one associated both with a criterion of application—a distinction between the things to which it applies and those to which it does not—*and* a criterion of identity: some principle determining the truth values of contexts of the form, 'X is *the same F* as Y'. 'Tree', 'person', 'city', 'river', 'number', 'set', 'time', 'place', are all, in at least certain uses, sortal concepts in the intended sense. By contrast, 'red', 'composed of gold', 'large'—in general, purely qualitative predicates, predicates of constitution, and attributive adjectives—although syntactically admissible substituends for occurrences of the predicate letters in higher order logic, are not. Call the latter class of expressions *mere* predicables. Where *F* is a mere predicable, then, the suggestion is that the question How many *F*s are there? is *ceteris paribus* deficient in sense and "the number of *F*s," accordingly, has no determinate reference.

It is easy to see that 'is self-identical' is a mere predicable. For reflect that—

prescinding from any cases of vagueness—mere predicables *do* nevertheless subserve determinate questions of cardinal number when their scope is restricted to that of some specific sortal concept: thus there can be a determinate number of *red apples in the bowl*, of *gold rings in the jeweler's window*, and of *large women at the reception*. So if 'self-identical' *were* a sortal concept, it should follow that there can be determinate numbers of red self-identicals in the bowl, golden self-identicals in the jeweler's window, and large self-identicals at the reception. However, since '*F* and self-identical' is equivalent to '*F*', it follows that there can be no such determinate number wherever there is no determinate number of *F*s. So, self-identity is not a sortal concept. If we take it that, save where *F* is assured an empty extension on purely logical grounds,[4] only sortal concepts, and concepts formed by restricting a mere predicable to a sortal concept, have cardinal numbers, it follows that there is no universal number.

To be sure, this first consideration will, of course, not engage the question whether we may properly conceive of a number of all *ordinals*, or all *cardinals*, or all *sets*—in general, cases where we are concerned with the results of applying the numerical operator to concepts which *are* (presumably) sortal but "dangerously" big. And as we saw, a variant of Boolos's objection, that there is a potential clash of Hume's Principle with Zermelo-Frankel set theory, does equally arise in those cases. However, a principled objection to the idea that there should be determinate numbers associated with *these* concepts may be expected to issue from the second line of thought, which concerns the tantalizing notion of *indefinite extensibility*.

As noted a little while ago, it seems natural and well motivated to suppose that the *F*s should have a determinate cardinal number just when they compose a *set*. But a long tradition in foundational studies would argue that sethood cannot be the right way to conceive of Frege's intentionally all-inclusive domain of objects: that Cantor's paradox shows, in effect, that there can be no universal set—no absolutely all-embracing totality which is subject, for example, to the operations and principles that provide for the proof of Cantor's theorem. That is not the same as saying that unrestricted first-order quantification is illegitimate—a concession which would, of course, be fatal to Frege's whole project. The point is rather that the objects that lie in the range of such unrestricted quantification compose not a determinate totality, but one that is, in the phrase coined by Dummett, "indefinitely extensible"[5]—a totality of such a sort that any attempt to view it as a determinate collection of objects will merely subserve the specification of new objects which ought, intuitively, to lie within the totality but cannot, on pain of contradiction, be supposed to do so. I do not know how best to sharpen this idea, still less how its best account might show that Dummett is right, both to suggest that the proof-theory of quantification over indefinitely extensible totalities should be uniformly intuitionistic and that the fundamental classical mathematical domains, such as those of the natural numbers, or the reals, should also be regarded as indefinitely extensible. But Dummett could be wrong about both those points and still be emphasizing an important insight concerning certain very large totalities—ordinal number, cardinal number, set, and indeed "absolutely everything." If there is anything at all in the notion of an indefinitely extensible totality—and there are signs that the issue is now being taken up in productive ways (see Clark [6], Oliver [17], and Shapiro [19])—one principled restriction on Hume's

Principle will surely be that *F* and *G not* be associated with such totalities. So that is a second definite program for understanding how, in particular, *not self-identical* might determine a cardinal number even though *self-identical* does not. Indeed, when the range of both individual and higher-order variables is unrestricted, the complement of any determinate finite concept is presumably *always* an indefinitely extensible totality.

***2.4 The concern about surplus content*** This is the objection I find it hardest to be sure I properly understand. Here is one of Boolos's expressions of it:

> It is known that Hume's Principle does not follow . . . from the conjunction of two of its strong consequences: . . . that nothing precedes zero and that *precedes* is a one-one relation. If HP is analytic, then it is strictly stronger . . . than some of its strong consequences. It's also known that arithmetic follows from these two statements alone . . . faced with these results, how can we really want to call HP analytic? ([4], p. 249)

The objection is developed and endorsed by Richard Heck in recent work [16] and I shall rely on his interpretation of it. Heck emphasizes that there is a long conceptual leap involved in advancing to the concept of cardinal number enshrined in Hume's Principle in full generality for one whose previous acquaintance with cardinal number—a pre-Cantorian as it were—is restricted to finite arithmetic and its applications. The length of the leap is reflected in the results about the proof-theoretic strength of various systems, including Fregean Arithmetic—that is, Hume's Principle plus second-order logic—second-order Peano Arithmetic and certain intermediaries which, building on work of Boolos, Heck demonstrates ([16], Section 4). Here is his conclusion:

> HP, conceptual truth or not, cannot be what underlies our knowledge of arithmetic. And no amount of reflection on the nature of arithmetical thought could ever convince one of HP, nor even of the coherence of the concept of cardinality of which it is purportedly analytic. Granted, any rationalist project of this sort will have to invoke a distinction between the 'order of discovery' and the 'order of justification'. But the objection is *not* that Hume's Principle is not known by ordinary speakers, nor that there was a time when the truths of arithmetic were known, but HP was not. It is that, even if HP is thought of as 'defining' or 'introducing' or 'explaining' our present concept of cardinality, the conceptual resources required if one is so much as to recognise the coherence of this concept (let alone HP's truth) vastly outstrip the conceptual resources employed in arithmetical reasoning. Wright's version of logicism is therefore untenable. ([16], pp. 597–98)

Heck goes on to consider whether some version of Hume's Principle restricted to finite concepts might be resistant to the particular objection, that is, whether such a version might be appreciable as a correct digest of its constitutive principles by one possessed just of the conceptual resources deployed in finite cardinal arithmetic and its applications. That is an interesting question, on which he offers interesting formal and informal reflection. But I have a prior difficulty in seeing that the original objection, concerning the conceptual excess of Hume's Principle over second-order Peano Arithmetic, does any serious damage to any contention that the neo-Fregean should want to make. Grant that a recognition of the truth of Hume's Principle cannot

be based purely on analytical reflection upon the concepts and principles employed in *finite* arithmetic. The question, however, surely concerned the reverse direction of things: it was whether access *to* those concepts and validation *of* those principles could be achieved via Hume's Principle, and whether Hume's Principle might in its own right enjoy a kind of conceptual status that would make that result interesting. The latter is, in effect, exactly the question raised by our title. But no particular view of it can be motivated merely by the reflection that the conceptual resources involved in Hume's Principle, insofar as an extension of the notion of cardinal number to the infinite case is involved, considerably exceed those involved in ordinary arithmetical competence.

Moreover, it is unclear how anyone wishing to demonstrate the analyticity of arithmetic could clear-headedly acquiesce in the rules of debate implicit in Heck's discussion. Those rules require that one canvass some principle which is supposedly analytic of ordinary arithmetical concepts in the precise sense that it could be recognized by reflection as systematizing those ordinary concepts and their proof theory. But, of course, an axiom could, in *that* sense, be analytic of a thoroughly synthetic theory, and itself as synthetic as that theory. (There might be a single such axiom which could be reflectively recognized as systematizing exactly Euclidean geometry). To be sure, it *is* a necessary condition of the success of the neo-Fregean project that the relevant principle does more than generate a theory within which arithmetic can be interpreted—there has to be a tighter conceptual relationship than that. But it is no necessary condition for the satisfaction of this necessary condition that there be no conceptual surplus of the axiom over the theory. And it is no sufficient condition of the analyticity of such an axiom that there be none; for again, a reflectively correct digest of a synthetic theory will be itself synthetic.

***2.5 The concern about bad company*** Boolos's final objection is perhaps the most interesting and challenging of all. It begins with the excellent observation that there are *close analogues* of Hume's Principle, specifically, principles taking the form of second-order abstractions, linking the obtaining of an (second-order logically definable) equivalence relation on concepts to the identity condition for certain associated objects, which are *self-consistent* (that is, the systems consisting of second-order logic plus one of these principles are, arguably, consistent) yet which are *inconsistent* with Hume's Principle. A nice example is what I have elsewhere called the *Nuisance Principle* (NP). The *nuisance* associated with the concept $F$ is the same as the nuisance associated with the concept $G$ just in case the *symmetric difference* between $F$ and $G$—the range of things which are either $F$ or $G$ but not both—is finite. Straightforward model-theoretic reasoning leads to the conclusion that any universe in which NP is satisfied must be a finite one.[6] But it is, apparently, a self-consistent principle—it does have finite models. If Hume's Principle is analytic, then NP is *analytically false*. But with what right could we make that claim—isn't the analogy between the two principles near enough perfect?

This challenge—there dubbed the 'Bad Company' objection—is treated in some detail in my [22] on which Boolos's "Is Hume's Principle Analytic?" was commentary. My suggestion in that paper was that the first step to disarming it is a deployment of (something very close to) Field's notion of conservativeness. A principle, or

set of principles, is conservative with respect to a given theory when, roughly, its addition to that theory results in no new theorems about the old ontology.[7] One could hope that Hume's Principle will be conservative with respect to any theory for which second-order Peano Arithmetic is conservative, (that is, one would hope, any theory whatever). By contrast, the consistent augmentation of any theory, $T$, by NP will result in a theory of which it is a consequence that all categories of the original ontology of $T$ are at most *finitely* instantiated. No pure definition could permissibly have that effect. So, no merely conceptual-explanatory principle—no principle whose role, as that of abstractions is supposed to be, is merely to fix the truth-conditions of a range of contexts featuring a new kind of singular-term forming operator and is otherwise to be as close as possible to that of a pure definition—can permissibly have it either. Since it has consequences for the size of extensions of concepts which are quite unrelated to that which it purportedly serves to introduce, NP thus cannot be viewed as such a conceptual-explanatory principle. Moreover, any abstraction principle which clashes with Hume's Principle by requiring the finitude of any domain in which it is to hold will be in like case. And indeed any abstraction principle which places an *upper bound*, finite or infinite, on the size of the universe will be nonconservative with respect to some consistent theory of things other than the abstracts it concerns.

The particular analogy is therefore broken: Hume's Principle, there is undefeated reason to hope, is conservative with respect to every consistent theory concerning things other than its own special ontology—the cardinal numbers. (That is, note, a kind of *weak* analyticity: if there were a possible world in which Hume's Principle failed, it would have to be by dint of its misrepresentation of the nature of the cardinals in that world.) NP and its kin, by contrast, come short by this constraint.

An abstraction is acceptable only if it is conservative with respect to every consistent theory whose ontology does not include its proper abstracts. It is a *logical* abstraction just in case its abstractive relation is definable in higher-order logic. The company kept by Hume's Principle is thus, we may presume, that of conservative, logical abstractions. But are *these* all Good Companions? Recent critics of neo-Fregeanism have observed that they are not, so that the fifth concern extends beyond the point that Boolos himself took it to. I pursue the matter in Appendix A.

***3*** It should now be apparent why I suggested earlier that my debate with Boolos could as well have proceeded, near enough, without recourse to the notion of analyticity. The point is simply that each of Boolos's objections is, in effect, independent of the problematical aspects of that notion: what was really bothering him was not whether Hume's Principle is analytic, but whether it is *true*, and whether and how we might be warranted in regarding it as being so. Thus, without any really significant loss, the five points of concern might be formulated as:

1. With what right do we regard ourselves as warranted in accepting a principle with such rich ontological implications—how do we know that there is any function which behaves as the referent of octothorpe must?
2. What warrant do we have for confidence that the strong theory—Fregean Arithmetic—to which Hume's Principle gives rise is a consistent theory?
3. Is not its inconsistency with Zermelo-Frankel set theory (plus standard definitions) a strong ground for doubting the truth of Hume's Principle?

4. What warrant is there for accepting a principle which is supposed to provide a foundation for arithmetic yet has so much surplus content over arithmetic?
5. With what right do we accept a principle which seems to be on all fours with other consistent principles which are inconsistent with it?

These are all good concerns, and I hope I have indicated, point by point, something of the direction in which the neo-Fregean should try to launch respective responses to them. The crucial point remains that the notion of analyticity is not required to formulate the concerns. What is really at stake, rather, is the nature of our *entitlement* to Hume's Principle.

A worked-out account of the notion of analyticity, in all its varieties, might well provide an answer to the question. The answer the neo-Fregean wants to give is not hostage to the provision of such an account. Let me rapidly recapitulate that answer. 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 a priori 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:[8]

1. that the vocabulary of higher-order logic plus the cardinality operator, octothorpe or '$Nx : \ldots x \ldots$', provides a sufficient definitional basis for a statement of the basic laws of arithmetic;
2. that when they are so stated, Hume's Principle provides for a derivation of those laws within higher-order logic;
3. that someone who understood a higher-order language to which the cardinality operator was to be added would learn, on being told that Hume's Principle governs the meaning of that operator, all that it is necessary to know in order to construe any of the new statements that would then be formulable;
4. finally and crucially, that Hume's Principle may be laid down *without significant epistemological obligation*: that it may simply be stipulated as an explanation of the meaning of statements of numerical identity, and that—beyond the issue of the satisfaction of the truth-conditions it thereby lays down for such statements—no competent demand arises for an independent assurance that there *are* objects whose conditions of identity are as it stipulates.

The first and third of these claims concern the epistemology of the *meaning* of arithmetical statements, while the second and fourth concern the recognition of their *truth*. With which of them would Boolos disagree? Even with a qualification I will come to in a minute, I think he had no quarrel with the first; nor, of course, with the second, which is just the point proved by Frege's Theorem. And, to accept just these two claims, of course, is already to acknowledge a substantial Fregean achievement: the analytical reduction of the primitive vocabulary of arithmetic to a base that contains just one nonlogical expression, the cardinality operator; and a demonstration that, on that basis, the fundamental laws of arithmetic can be reduced to just one: Hume's Principle itself.

The qualification concerning the first claim concerns the interpretation of the phrase "sufficient definitional basis." No question, of course, but that Frege shows how to define *expressions* which comport themselves like those for successor, zero,

and the predicate 'natural number', thus enabling the formulation of a theory which *allows of interpretation* as Peano Arithmetic. But—as we remarked right at the start—it is one thing to define expressions which, at least in pure arithmetical contexts, behave as though they express those various notions, another to define those notions themselves. And, it is the latter point, of course, that is wanted if Hume's Principle is to be recognized as sufficient for a theory which not merely allows of pure arithmetical interpretation but to all intents and purposes *is* pure arithmetic. How is the stronger point to be made good?

Well, I imagine it will be granted that to define the distinctively arithmetical concepts is so to define a range of expressions that the use thereby laid down for those expressions is indistinguishable from that of expressions which do indeed express those concepts. The interpretability of Peano Arithmetic within Fregean Arithmetic ensures that has already been accomplished as far as all *pure* arithmetical uses are concerned. So any doubt on the point has to concern whether the definition of the arithmetical primitives which Frege offers, based on Hume's Principle and logical notions, are adequate to the ordinary *applications* of arithmetic. Did Frege succeed in showing how the concepts of arithmetic, as understood both in their pure and applied uses, can be understood simply on the basis of second-order logic and the numerical operator, as constrained by Hume's Principle, or could someone fully understand the entirety of the construction without having the slightest inkling of the ordinary meaning of arithmetical claims?

The matter needs more detail than I will offer here, but I think it is clear that Frege did succeed in the more ambitious task, and a crucial first step in seeing that he did so is to realize that Hume's Principle provides for the proof of a very important principle, dubbed $N^q$ by Hale, to the effect that for each numeral, '$n_f$,' defined in Frege's way, we can establish that

$$n_f = Nx : Fx \longleftrightarrow \text{ there are exactly } n \ F\text{s} \ ,$$

where the second occurrence of '$n$' is schematic for the occurrence of an arabic numeral as ordinarily understood.[9] It follows that each Fregean numeral has exactly the meaning in application which it ought to have. That seems to me sufficient to ensure that Hume's Principle itself enforces the interpretation of Fregean Arithmetic as genuine arithmetic and not merely a theory which can be interpreted as such.

If this is right, then the key philosophical issues must concern the third and fourth claims. The importance of the third claim derives from the consideration that Hume's Principle is not, properly speaking, an eliminative definition—it allows the construction of uses of the numerical operator which it does not in turn provide the resources eliminatively to define. Its claim to serve as an explanatory basis for arithmetic must therefore depend on its ability somehow to explain such uses in a nonstrictly definitional fashion. Arguing the point requires stratifying occurrences of the numerical operator in sentences of Fregean Arithmetic according to the degree of complexity of the embedding context, and making a quasi-inductive case: first, that a certain range of basic uses are unproblematic, and second, that at every subsequent stage, the type of occurrence distinctive of that stage may be understood on the basis of an understanding of the mode of occurrence exemplified at the immediately preceding stage. There are some complications with this; I have tried to work through the point in some

detail elsewhere[10] and will not repeat the detail here. For what it is worth, it is Dummett rather than Boolos, who has been the most vociferous opponent of the third neo-Fregean claim.

It is the fourth claim—the claim that Hume's Principle can be laid down as an explanatory stipulation, without further epistemological obligation—which seems to me to be the heart of the issue. Boolos was indeed uncomfortable with this claim, suspecting that more had been smuggled into the notion of *explanation* in this setting than was consistent with the seeming modesty of the explanatory thesis. But I do not feel that I have understood his reservations very well. If nominalism is a misconception—if it is possible to know of abstract entities and their properties at all—then it has to be because we have so fixed the use of statements involving reference to and quantification over such entities as to bring the obtaining of their truth conditions somehow within our powers of recognition. And, whatever this fixing consisted in, it has to have been something we did by way of *determination of meaning*, and it should therefore have involved no epistemological obligations which are not involved in the construction of concepts and the determination of meanings generally. I really do not see why the fashion in which Hume's Principle—if it indeed succeeds in doing so—determines the truth conditions of statements which configure the cardinality operator with second-order logical concepts, should be epistemologically any more problematical than any definition or other form of stipulation whose effect is to fix the truth-conditions of statements containing a targeted (type of) term. It is of course—always—another question whether those truth conditions are satisfied: something which a definition, without supplementary considerations, is powerless to determine. But a good abstraction principle always determines very explicitly what those supplementary considerations are to be—you have only to look at its right-hand side. If there are good reservations about this way of looking at Hume's Principle, I do not think that they have yet been compellingly formulated.

Whatever the ultimate assessment of that issue may prove to be, it is my hope that the foregoing overview of Boolos's misgivings about the analyticity of Hume's Principle may serve as a reminder of two things: first, (we owe it to Frege to recognize) that there is still an unresolved debate to be had about the viability of something that is, in all essential respects, a Fregean philosophy of arithmetic and real and functional analysis;[11] second, that the progress made in the modern debate is owing in very considerable measure to George's brilliant and unique articles on the issues.

## *Appendix*

***A Conservativeness and modesty*** In [20], Shapiro and Weir observe that there are pairs of abstractions which result by various kinds of selection for $\varphi$ in

(D) $\quad (\forall F)(\forall G)(\Sigma F = \Sigma G \longleftrightarrow (\varphi F \,\&\, \varphi G) \vee (\forall x)Fx \longleftrightarrow Gx)$[12]

which are jointly unsatisfiable yet which are presumably conservative in the germane sense. For instance, take $\varphi$ respectively as 'is the size of the universe and some limit inaccessible' and 'is the size of the universe and some successor inaccessible'. (The neo-Fregean should resist any tendency to impatience at the rarefied character of the example. These notions are definable in higher-order logic.) Any instance of schema

(D) entails that some $F$ is $\varphi$. So the two indicated abstractions respectively entail that the universe is limit-inaccessible sized and that it is successor-inaccessible sized. It cannot be both. Yet neither implication places any overall bound on the size of the universe—so these abstractions do not involve the kind of nonconservativeness which NP entrained. Still, they cannot both be in good standing. And, if either is not, then it seems that neither should be. But, by what (well-motivated) principle might they be excluded? What virtue does Hume's principle have which they lack?

What is intuitively salient about any D-schematic abstraction (henceforward "Distraction"[13]) is that, the entailment notwithstanding, it provides no motive to *believe* that there is a concept which falls under its particular selection for '$\varphi$'—the result is obtained merely by exploitation of the embedded antinomy. For on the assumption of

$$(\forall F) - (\varphi F)$$

any Distraction entails Basic Law V:

$$(\forall F)(\forall G)(\Sigma F = \Sigma G \longleftrightarrow (\forall x)(Fx \longleftrightarrow Gx))$$

and thereby Russell's Paradox. Such abstractions thus have no more bearing on the *truth* of the relevant '$(\exists F)\varphi F$' than instances of the following schema have:

$$(\forall F)F \text{ is } \varphi\text{-terological} \longleftrightarrow F \text{ does not apply to itself or } \varphi F$$

which likewise, on the assumption of

$$(\forall F) - \varphi F$$

entail the well-known Heterological paradox:

$$(\forall F)F \text{ is heterological} \longleftrightarrow F \text{ does not apply to itself.}$$

Again, we can select for '$\varphi F$' that $F$ is the size of the universe and some limit inaccessible, or the size of the universe and some successor inaccessible, or that $F$ applies to God, or the Devil, . . . and proceed to infer that the universe is limit-inaccessible in cardinality, or successor inaccessible, or that God, or the Devil, exists. It is long familiar how Liar-family paradoxes can occur not merely in contexts of self-contained aporia but may be exploited to yield unmotivated a priori resolutions of intuitively unrelated issues. The Cretan and the Curry Paradox are the best known examples of the latter. The schema for $\varphi$-terologicality, and Distractions as a class, merely provide two more.

This perspective offers the option of a 'holding' response to the Shapiro/Weir objection: "You persuade me," the neo-Fregean may say, "that the general idea that a concept may be defined by stipulation of its satisfaction-conditions is somehow confounded by the possibility of pairwise incompatible yet consistent instances of the rubric for $\varphi$-terologicality and I will concede that the neo-Fregean conception of an abstraction principle is put in similar difficulties by conservative yet pairwise incompatible instances of (D)." This response is dialectically strong. Who would suppose that roguish cases such as "heterological" and instances of $\varphi$-terologicality somehow

show that we may no longer in good intellectual conscience regard the general run of definitions of the form

$$X \text{ is } F \quad \text{if and only if} \quad \ldots X \ldots,$$

as successful in fixing concepts? But then, someone who had *no other* objection to the claim of Fregean abstractions to play the role of truth-condition fixers for the kinds of context that feature on their left-hand sides, should not be fazed by roguish instances of (D).[14]

It is only a holding response, however. It refurbishes one's confidence that it has to be possible to draw the distinction which the neo-Fregean needs, but it does not draw it. The fact remains that just as a general explanation is owing of which are the pukka definitions of satisfaction-conditions and which may be dismissed as rogues, so we still need a characterization of which are the good abstractions and which are the (conservative but still) bad Distractions. In [22], motivated in part by the desire to legitimate Boolos's axiom New V:

$$(\forall F)(\forall G)(\Sigma F = \Sigma G \longleftrightarrow (\text{Big}(F) \,\&\, \text{Big} \vee (\forall x)Fx \longleftrightarrow Gx))$$

(where $F$ is *Big* just if it has a bijection with self-identity), I ventured an additional conservativeness constraint which would be tolerant of at least some instances of schema (D), but would reject the majority. Roughly, it was that those consequences of such an abstraction which follow by exploitation of its "paradoxical component" have to be in 'independent good standing'. I shall here attempt briefly to clarify and assess this proposal.

Distractions entail conditionals of the form:

$$-(\exists F)(\varphi F) \longrightarrow (\forall F)(\forall G)(\Sigma F = \Sigma G \longleftrightarrow (\forall x)(Fx \longleftrightarrow Gx))$$

The immediate intent of the proposed constraint is that anything derivable by the *reductio* of the antecedent of such a conditional afforded by its paradoxical consequent should be in independent good standing. New V fares well by this proposal: that there is a concept which is Big should presumably be a result in 'independent good standing' however that idea is filled out—for that self-identity itself is Big follows from the definition of 'Big' in second-order logic.

Of course, *any* abstraction will entail some such conditional. So the proposed constraint is quite general. How does Hume's principle fare by it? Well enough, presumably, though in a different way. We may, for instance, obtain a relevant conditional by selecting 'at least countably infinite' for $\varphi$. But this time the resources required to make good the consequences of the denial of the antecedent are afforded not just by second-order logic, but by Hume's Principle itself, via its independent proof of the infinity of the number series. Indeed, it is just because it independently entails that denial that we are able to show that Hume's Principle entails the selected conditional in the first place. By contrast, the kinds of roguish Distraction illustrated presumably fail the test. The only resources they have to show, for example, that the universe is limit-inaccessible, or successor inaccessible, or whatever, are those furnished by the inconsistency of Basic Law V and the consequent modus tollens on the relevant conditional.

So, an abstraction is good only if any entailed conditional whose consequent is Basic Law V (or, therefore, any other inconsistency) is such that all further consequences which can be obtained by discharging the antecedent are in independent good standing, as may be attested by their derivation in pure higher-order logic (like the case of New V) or their independent derivability from the abstraction in question (like the case of Hume's Principle). But this is unclear in a crucial respect: What is the relevant sense of 'independent derivability'? Clearly it would not be in keeping with the intended constraint if there were merely some collateral derivation of just the same suspect kind. The 'independent derivation' must be *bona fide*, must not proceed by "paradox-exploitative" means, as I expressed the matter. But what does that mean? In particular, how might it be characterized so as not to outlaw any proof by reductio ad absurdum?

One possible response—the one I offered in [22]— was that a relevantly narrow sense of "paradox-exploitative" may be captured by reinvoking the previous (Fieldian) notion of conservativeness in the following way: a derivation from a conservative abstraction is paradox-exploitative just if there is a representation of its form of which any instance is valid and of which some instance amounts to a proof of the nonconservativeness of another abstraction. For instance, the derivation of the successor-inaccessibility of the universe from the Distraction canvassed above is paradox-exploitative because it may be schematized under a valid form of which another instance is a derivation, from the appropriately corresponding Distraction, that the universe contains exactly 144 objects. The only Distractions which are good are those which are both conservative and such that any of their consequences which may validly be derived by paradox-exploitative means, in the stipulated sense, may also validly be derived by non-paradox-exploitative means. Otherwise put, the second conservativeness constraint is that the paradox-exploitative derivations from an abstraction have to be conservative with respect to the results obtainable from it non-paradox-exploitatively.

That was the essence of my previous proposal. In practice, its application would work like this. We would be defeasibly entitled to accept any (presumably) conservative abstraction, *A*, from which we had so far been able to construct no paradox-exploitative derivation—no proof of a valid form of which another instance demonstrated the nonconservativeness of another abstraction. But once we had such a derivation, it would then be inadmissible to accept *A* until we had found another non-paradox-exploitative derivation from it of the same conclusion: a formally valid derivation of which, so far as we could tell, no other instance was a proof of the nonconservativeness of another abstraction.

That is apt to seem uneasily complex and less clearly motivated than one would wish. And one might worry about its reliance on our ability to judge non-paradox-exploitative derivations. However, the play with 'paradox-exploitation' and its characterization in terms of nonconservativeness may now seem inessential. The basic idea was that some abstractions—the Distractions and some others—are at the service of noncogent proofs. We can tolerate this in particular cases so long as such proofs are matched by cogent ones of the same things. The natural—surely correct—objection to the derivation of, say, the successor inaccessibility of the universe from the appropriate Distraction is that it is unconvincing because "You could just as well prove the

opposite—or anything—like that," where "like that" means: by laying down a different (presumably consistent) Distraction and *reasoning in just the same way*. So, a natural thought would be that we should ban those distractions—or abstractions generally—some of whose consequences are such as to deserve that complaint. That would suggest the following stipulation: that an abstraction $A$ is unacceptable, at least pro tempore, if every proof it has yielded of some consequence $C$ is such that, schematized so that any instance of it is valid, some other (conservative) abstraction yields a proof of the same form of something inconsistent with $C$.

But there are still a number of salient concerns. First, it is not clear that any purpose is served by the continuing insistence on derivations of a given valid form. Why not just say that pairwise incompatible but individually conservative abstractions are ruled out—however the incompatibility is demonstrated—and have done with it? For think: if each such pair can be shown to be incompatible by proofs of a given single form, then the more complex formulation of the constraint is unnecessary; but if some pair cannot—if no derivation of $C$ from $A$ is of a valid form shared by some derivation of not-$C$ from $A^*$,—then there will still be pairwise incompatible but conservative abstractions which survive the new test. So there will still be Bad Company, for which some further treatment will need to be devised. Besides, how is the proposed constraint meant to be applied to *semantical* (model-theoretic) demonstrations of consequence—for which, of course, in the case of higher-order abstractions, there need be no effectively locating a corresponding derivation in higher-order logic?

This whole direction was stimulated by the desire to save some 'good' Distractions, *par excellence* New V. It is therefore germane that, as Shapiro has since observed, New V itself is in any case nonconservative—specifically that it entails that the universe can be well ordered, and hence that the nonabstracts can.[15] This result, to be sure, does not show that there is nothing to be gained from attempting to refine the second conservativeness constraint of [22]—that it has no point. But it should occasion a rethink of the motivation for the general direction.

I think there is something else amiss with the rogue Distractions—something which the second proposed constraint may well indirectly approximate but does not bring out with sufficient clarity. Start from the point that definitions proper should be innocent of substantive implications for the universe over which they range. Abstractions cannot in general match that, since in conjunction with logically (or other forms of metaphysically) necessary input, they may carry substantive implications for the abstracts whose concept they serve to introduce and hence—since those abstracts will be viewed, at least by neo-Fregeans, as full-fledged participants in the universe—at least some substantive implications for the universe as a whole. But to the extent that it is proposed to regard them as meaning-constituting stipulations, and hence as approximating definitions as nearly as possible, the character and scope of such implications needs to be curtailed. In brief: the requirement has to be that the only implications they may permissibly carry for the, as it were, enlarged universe in which their own abstracts participate must originate in what they imply—whether proof- or model-theoretically—about the abstracts they specifically concern. Hume's principle, for instance, implies of any object whatever that it participates in an at least countably infinite universe; but it carries that implication only via its entailment of the infinity of the cardinal numbers.

This is a different requirement to Field conservativeness. A non-Field conservative abstraction—one that, as we put it intuitively above, entails new results about the prior ontology— may, of course, violate it. But it is possible for an abstraction to be *immodest*— for it to carry implications for other objects in the universe which cannot be shown to originate in implications it carries for its own proper abstracts—without thereby being demonstrably nonconservative. Consider the Limit-inaccessible Distraction again. As noted, this entails that the universe is limit-inaccessibly sized. But because, unlike NP, it places no upper bound on the size of the universe, it is not nonconservative in the way that NP is—it limits the extension of no other concept. And if it is nonconservative in any other way, we have yet to see how. However, it *is* immodest. For its requirement that the universe have a certain kind of cardinality does not originate in any requirement that it imposes on its own abstracts.

It is easy to overlook the force of "originate" here. The Limit-inaccessible Distraction, to stay with that example, entails that any finite concept is non-$\varphi$. So it will allow singleton concepts to generate 'well-behaved' abstracts—abstracts whose identity and distinctness is governed by ordinary extensionality—of which there should therefore be no fewer than there are objects in the universe.[16] Thus this particular Distraction will indeed entail that its own abstracts are limit-inaccessible in number, from which the limit inaccessibility of the universe follows.[17] But—this is the crux—the result about the abstracts is not needed for the proof of the limit-inaccessibility of the universe. The Distraction provides no way of recognizing the limit-inaccessibility of the universe which goes via a priori recognition of what it entails about its own proper abstracts. Rather the inference is the other way about: the proof that the Distraction entails that result about the universe as a whole is needed in order to obtain the result about its own abstracts. That is immodesty.

Conservativeness constrains the *kind of consequences* which an acceptable abstraction is allowed to have: it is not allowable that there be any claim exclusively concerning the nonabstracts which was previously unprovable but which the abstraction, coupled with previous theory—now explicitly restricted to the previous ontology—enables us to prove. Modesty, by contrast, constrains the *kind of ground* which an acceptable abstraction can provide for consequences, not per se nonconservative, about the ontology of a theory in which that abstraction participates: such consequences must be grounded in what it requires of its own proper abstracts. But although the two constraints may seem different in character in this way, they are aspects of a single point *au fond*. Remember that the role of a legitimate abstraction, as I have repeatedly stressed, is merely to fix the truth conditions of a class of contexts featuring a novel term-forming operator. It cannot have more than that role and yet retain the epistemically undemanding character of a meaning-stipulation. Logical abstractions, to be sure, are so designed that, consistently with their playing just this role, logical resources may enable us to show that there are abstracts of the kind they concern and to establish things about them. But, no abstraction can be deemed to discharge the intended limited role successfully if, in conjunction with some consistent theory, it carries implications for the combined ontology which cannot be shown to derive from implications it has for its own abstracts. Nonconservativeness is (normally) one graphic way of failing that test. But, if even a conservative abstraction entails some conclusion about the combined ontology which cannot be justified by

reference to what it entails about its own abstracts, then knowledge of the truth of the abstraction cannot be founded in stipulation. Such an abstraction implicitly claims something about the world which might—for all we have shown to the contrary—be justified by reference to what it entails about its own abstracts; that is why we cannot accuse it of nonconservativeness. But equally, so long as we have no such justification, we have no defense against the suggestion that the abstraction is known only if we know that the world must be that way *in any case*, whether or not the abstracts themselves make it so. That would seem to demand knowledge about how the world would be even if the abstracts did *not* make it so. And that in turn is a substantial piece of collateral information which, by being prerequisite if we are to claim to be justified in laying down the abstraction in the first place, gives the lie to any claim that the abstraction is justified merely as a meaning-stipulation.

In sum, an abstraction is modest if its addition to any theory with which it is consistent results in no consequences (whether proof- or model-theoretically established) for the ontology of the combined theory which cannot be justified by reference to its consequences for its own abstracts. And again, *justification* is the crucial point: an abstraction may fail this constraint even though every consequence it has for the ontology of a combined theory may be seen to follow from things it entails about its proper abstracts; in particular, it will not count if, as in the case of the Limit-inaccessible Distraction, a consequence for the combined ontology is needed as a lemma in the proof that the abstracts have a property from which that very consequence follows.

Further clarification is needed of several matters: what kinds of proof should count in favor of the modesty of an abstraction—what it is to show that an abstraction independently carries certain implications for its own abstracts; whether the modesty constraint is effective against the general run of pairwise incompatible but (presumptively) conservative abstractions illustrated by Shapiro and Weir; what other constraints on Good Companions may be properly motivated. At the time of writing, these are largely open issues. I hope to return to them in future work.

## *B Proof of the Principle,* $\mathbf{N^q}$

***B.1 Stage-setting*** We assume the standard recursive definitions of the numerically definite quantifiers:

$$(\exists_0 x)Fx \longleftrightarrow (\forall x) - Fx,$$

$$(\exists_{n+1} x)Fx \longleftrightarrow (\exists x)(Fx \,\&\, (\exists_n y)(Fy \,\&\, y \neq x)),$$

and let '$n_f$' abbreviate Frege's definiens for $n$. Define '$Pxy$' (immediate predecession) as:

$$(\exists F)(\exists w)(Fw \,\&\, y = Nv{:}\,Fv \,\&\, x = Nz{:}[Fz \,\&\, z \neq w]).$$

Define 'Nat($x$)' ($x$ is a natural number) as:

$$x = 0 \vee P^*0x,$$

where '$P^*xy$' expresses ancestral predecession. Let '$(\exists R)(F\ \text{1-1}_R\ G)$' express that there is a one-one correspondence between $F$ and $G$.

We take three lemmas from the proof of the Peano axioms from HP outlined in the concluding section of *Frege's Conception* (numbering as there assigned).

**Lemma 51** $(\forall x)(\mathrm{Nat}(x) \longrightarrow x = Ny : [\mathrm{Nat}(y)\ \&\ P^*yx]$—*every natural number is the number of its ancestral predecessors.*

**Lemma 52** $(\forall x)(\mathrm{Nat}(x) \longrightarrow -P^*xx)$—*no natural number ancestrally precedes itself.*

**Lemma 5121** $(\forall x)(\forall y)(\mathrm{Nat}(x)\ \&\ \mathrm{Nat}(y) \longrightarrow (Pxy \rightarrow (\forall z)(\mathrm{Nat}(z)\ \&\ (P^*zx \vee z = x) \longleftrightarrow (\mathrm{Nat}(z)\ \&\ P^*zy)))$—*if one natural number immediately precedes another, then the natural numbers which ancestrally precede the second are precisely the first and those which ancestrally precede the first.*

Finally, recall that Frege's 0 is $Nx : x \neq x$ and that each successive $n+1_f$ is $Nx : [x = 0 \vee \cdots \vee x = n_f]$. Each of these objects qualifies as a natural number in the light of the above definition of $\mathrm{Nat}(x)$.

*Proof:* $0_f$ qualifies by stipulation; $n+1_f$ qualifies if $n_f$ does—take '$F$' in the definition of '$Pxy$' as '$[x = 0 \vee \cdots \vee x = n_f]$' and '$w$' as '$n_f$' to show that $P(n_f, n+1_f)$; then reflect that $Pxy \rightarrow P^*xy$ and that $P^*xy$ is transitive (*Frege's Conception*, Lemmas 3 and 4, respectively). □

### *B.2 Proof of* $\mathbf{N^q}$ *for Frege's natural numbers*

Induction Base: To show

$$Nx : Fx = 0_f \longleftrightarrow (\exists_0 x)Fx,$$

it suffices to reflect that the left-hand side holds just if $(\exists R)(Fx\ \text{1-1}_R\ x \neq x)$, which in turn holds just if $-(\exists x)Fx$.[18]

Induction Hypothesis: Suppose $Nx : Fx = n_f \longleftrightarrow (\exists_n x)Fx$. We need to show that it follows that

$$Nx : Fx = (n+1)_f \longleftrightarrow (\exists_{n+1} x)Fx.$$

(*Left to right*) Consider any $F$ such that $Nx : Fx = (n+1)_f$. By Lemma 51 and the reflection that $\mathrm{Nat}(n_f)$, $n_f = Nx : [\mathrm{Nat}(x)\ \&\ P^*xn_f]$. So by the Hypothesis, $(\exists_n x)(\mathrm{Nat}(x)\ \&\ P^*xn_f)$. But by Lemma 52, $-P^*n_f, n_f$. So $(\exists_n x)(\mathrm{Nat}(x)\ \&\ (P^*xn_f \vee x = n_f)\ \&\ x \neq n_f)$. So $(\exists y)(\mathrm{Nat}(y)\ \&\ (P^*yn_f \vee y = n_f)\ \&\ (\exists_n x)(\mathrm{Nat}(x)\ \&\ (P^*xn_f \vee x = n_f)\ \&\ x \neq y))$. So by the recursion for the quantifiers, $(\exists_{n+1} x)(\mathrm{Nat}(x)\ \&\ (P^*xn_f \vee x = n_f)$. But by Lemma 5121 and since $P(n_f, n+1_f)$, we have that $(\forall x)(\mathrm{Nat}(x)\ \&\ (P^*xn_f \vee x = n_f) \longleftrightarrow \mathrm{Nat}(x)\ \&\ P^*(x, n+1_f))$. So $(\exists_{n+1} x)(\mathrm{Nat}(x)\ \&\ P^*(x, n+1_f))$.

That establishes the desired result for one concept of which $(n+1)_f$ is the number. But by HP, any $G$ such that $(n+1)_f = Nx : Gx$ will admit a one-one correspondence with that concept. So a lemma to the following effect will now suffice:

$$(\forall F)(\forall G)((\exists R)(F\ \text{1-1}_R\ G) \rightarrow ((\exists_{n+1} x)Fx \longleftrightarrow (\exists_{n+1} x)Gx).$$

A proof by induction—strictly, at third order—suggests itself:

Base: It suffices to show $(\forall F)(\forall G)((\exists R)(F\ \text{1-1}_R\ G) \rightarrow ((\forall x) - Fx \longleftrightarrow (\forall x) - Gx)$.

Hypothesis: Suppose $(\forall F)(\forall G)((\exists R)(F\ 1\text{-}1_R\ G) \rightarrow ((\exists_n x)Fx \longleftrightarrow (\exists_n x)Gx)$.

Consider any $H$ such that $(\exists_{n+1}x)Hx$. Then $(\exists x)(Hx\ \&\ (\exists_n y)(Hy\ \&\ y \neq x))$. Let $a$ be such that $Ha\ \&\ (\exists_n y)(Hy\ \&\ y \neq a)$. Let $J$ be one-one correlated with $H$ by $R$. Let $b$ be such that $Jb\ \&\ Rab$. Then $R$ one-one correlates $Hx\ \&\ x \neq a$ with $Jx\ \&\ x \neq b$. So, by the Hypothesis, $(\exists_n x)(Jx\ \&\ x \neq b)$. So $(\exists x)(Jx\ \&\ (\exists_n x)(Jx\ \&\ x \neq b))$. So $(\exists_{n+1}x)Jx$.

(*Right to left*) Consider any $F$ such that $(\exists_{n+1}x)(Fx)$. Then there is some $a$, such that, $Fa\ \&\ (\exists_n y)(Fy\ \&\ y \neq a)$. So by the Hypothesis $Ny(Fy\ \&\ y \neq a) = n_f$. So, by HP, there is an $R$ such that $(Fy\ \&\ y \neq a)(1\text{-}1_R)(\mathrm{Nat}(x)\ \&\ P^*xn_f)$. Let $R^{\#}$ correlate $(Fy\ \&\ y \neq a)$ with $(\mathrm{Nat}(x)\ \&\ P^*xn_f)$ in just the fashion of $R$, and let it also correlate $a$ with $n_f$. Then $(Fy)(1\text{-}1_{R^{\#}})(\mathrm{Nat}(x)\ \&\ (P^*xn_f \vee x = n_f))$. But, as established above, $(\forall x)(\mathrm{Nat}(x)\ \&\ (P^*xn_f \vee x = n_f) \longleftrightarrow \mathrm{Nat}(x)\ \&\ P^*(x, n+1_f))$. So $Nx : Fx = (n+1)_f$.

## NOTES

1. At pp. 158–69. An outline of a proof of the Peano Axioms from Hume's Principle is also given in the Appendix to [3]. The derivability of Frege's Theorem is first explicitly asserted in Parsons [18]; see remark at p. 194. My own "rediscovery" of the theorem was independent. I do not know what form of proof Parsons had in mind but the reconstruction of the theorem is trickier than Frege's own somewhat telegraphic sketch suggests. For an excellent recent overview of the ins and outs of the matter—early on, they remark that

   > §§82–83 offer severe interpretative difficulties. Reluctantly and hesitantly, we have come to the conclusion that Frege was at least somewhat confused in these two sections and that he cannot be said to have outlined, or even to have intended, any correct proof there. (p. 407)

   —see [5].

2. [4], p. 251. See more generally pp. 248–54, ibid.; also [2] at p. 231 and [3] at pp. 246–48. (The latter references are to the pagination in [7].)

3. [2]; for a detailed proof, see the first appendix to [5].

4. A plausible general principle (suggested to me by Hale) of which this exception would be a special case would be this: that a nonsortal concept, $F$, may nevertheless have a determinate cardinal number if every sortal restriction of it has the same cardinal number. This would not, of course, legitimate anti-zero, since the cardinality of sortal restrictions of the form, $Gx\ \&\ x = x$, will vary with that of $G$. But it would save the standard Fregean definition of zero. (Would there be any instances of this principle other than those mere predicables which are necessarily uninstantiated?)

   More generally, we might—indeed, ought to—allow that a nonsortal $F$ may determine a number if we know that all and only $F$-things are $G$, where $G$ is sortal and nonindefinitely extensible. (But again, are there any such cases?)

5. Dummett first introduced this notion—which, of course, ultimately derives from one strand in Russell's Vicious Circle Principle—in [8] (reprinted in [11], pp. 186–201). It is central to the argument of the concluding chapter of Dummett [9]. See also his "What is Mathematics About?" [10] at pp. 429–45.

6. For details see [22] at pp. 221–25.

7. A tidied version of the characterization offered in [22] (at note 49, p. 232) would be as follows. Let

   $$(\Sigma) \qquad (\forall\alpha_i)(\forall\alpha_j)(\Sigma(\alpha_i) = \Sigma(\alpha_j) \longleftrightarrow \alpha_i \approx \alpha_j),$$

   be any abstraction. Introduce a predicate, $Sx$, to be true of exactly the referents of the $\Sigma$-terms and no other objects. Define the $\Sigma$-*restriction* of a sentence $T$ to be the result of restricting the range of each objectual quantifier in $T$ to non-$S$ items, —thus each subformula of $T$ of the form $(\forall x)Ax$ is replaced by one of the form $(\forall x)(-Sx \rightarrow Ax)$ and each subformula of the form $(\exists x)Ax$ is replaced by one of the form $(\exists x)(-Sx \,\&\, Ax)$. The $\Sigma$-*restriction* of a theory $\theta$ is correspondingly the theory containing just the $\Sigma$-*restrictions* of the theses of $\theta$. Let $\theta$ be any theory with which $\Sigma$ is consistent. Then $\Sigma$ is conservative with respect to $\theta$ just in case, for any $T$ expressible in the language of $\theta$, the theory consisting of the union of $(\Sigma)$ with the $\Sigma$-restriction of $\theta$ entails the $\Sigma$-restriction of $T$ only if $\theta$ entails $T$. The requirement on acceptable abstractions is, then, that they be conservative with respect to any theory with which they are consistent. (The tidying referred to, for which I am indebted to Alan Weir, consists in having the reference to the $\Sigma$-restriction of $\theta$, rather than as originally one simply to $\theta$, in the clause for 'conservative with respect to $\theta$'.)

8. I here rely again on formulations given in Wright [24].

9. I reproduce in Appendix B the proof of this claim given at pp. 366–68 of Wright [23].

10. See Section V of [23] and—for a supplementary consideration in response to an objection by Dummett—Section VI of [24].

11. This is a point that Boolos enthusiastically accepted:

    > I want to endorse Wright's . . . suggestion that the problems and possibilities of a Fregean foundation for mathematics remain [wide?] open and [his] remark . . . that "the more extensive epistemological programme which Frege hoped to accomplish in *Grundgesetze* is still a going concern." ([4], p. 246)

    For interesting preliminary steps toward the extension of the neo-Fregean program to the classical theory of the reals, see Hale [15].

12. This is schema (D) discussed in some detail in [22]; see pages 216 and following.

13. Alan Weir's puckish term.

14. Cf. [22], pp. 220–21.

15. See Shapiro and Weir [20]. In rough outline: we can derive the Burali-Forti paradox on the assumption that the concept, Ordinal, is not Big; but if Ordinal is Big, then there is a 1-1 correlation between Ordinal and $x = x$. So $x = x$ may be well ordered by that correlation.

16. Assuming that there no fewer singleton concepts than there are objects.

17. On standard cardinal-arithmetical assumptions.

18. As Boolos remarked to me, Frege himself observes, at *Grundlagen* §75 and §78, that he is in a position to obtain proofs of $N^q$ for $0_f$ and $1_f$, respectively.

## REFERENCES

[1] Boolos, G., "Saving Frege from contradiction," *Proceedings of the Aristotelian Society*, vol. 87 (1986), pp. 137–51; reprinted on pp. 438–52 in *Frege's Philosophy of Mathematics*, edited by W. Demopoulos, Harvard University Press, Cambridge, 1995.

[2] Boolos, G., "The consistency of Frege's *Foundations of Arithmetic*," pp. 3–20, in *On Being and Saying: Essays in Honor of Richard Cartwright*, edited by J. J. Thompson, The MIT Press, Cambridge, 1987; reprinted on pp. 211–33 in *Frege's Philosophy of Mathematics*, edited by W. Demopoulos, Harvard University Press, Cambridge, 1995.

[3] Boolos, G.. "The standard of equality of numbers," pp. 261–77 in *Meaning and Method: Essays in Honor of Hilary Putnam*, edited by G. Boolos, Cambridge University Press, Cambridge, 1990; reprinted on pp. 234–54, in *Frege's Philosophy of Mathematics*, edited by W. Demopoulos, Harvard University Press, Cambridge, 1995.

[4] Boolos, G., "Is Hume's Principle analytic?," pp. 245–61 in *Language, Thought and Logic*, edited by R. G. Heck, Jr., The Clarendon Press, Oxford, 1997.

[5] Boolos, G., and R. G. Heck, Jr., "Die Grundlagen der Arithmetik §§82-83," pp. 407–28 in *Philosophy of Mathematics Today*, edited by M. Schirn, The Clarendon Press, Oxford, 1998.

[6] Clark, P., "Dummett's argument for the indefinite extensibility of set and real number," pp. 51–63 in *Grazer Philosophische Studien 55, New Essays on the Philosophy of Michael Dummett*, edited by J. Brandl and P. Sullivan, Rodopi, Vienna, 1998.

[7] Demopoulos, W., *Frege's Philosophy of Mathematics*, Harvard University Press, Cambridge, 1995.

[8] Dummett, M., "The philosophical significance of Gödel's Theorem," *Ratio*, vol. 5 (1963), pp. 140–55.

[9] Dummett, M., *Frege: Philosophy of Mathematics*, Duckworth, London, 1991.

[10] Dummett, M., *The Seas of Language*, The Clarendon Press, Oxford, 1993.

[11] Dummett, M., *Truth and Other Enigmas*, Duckworth, London, 1978.

[12] Field, H., "Critical notice of Crispin Wright *Frege's Conception of Numbers as Objects*," *Canadian Journal of Philosophy*, vol. 14 (1984), pp. 637–62.

[13] Field, H., "Platonism for cheap? Crispin Wright on Frege's context principle," pp. 147–70 in *Realism, Mathematics and Modality*, Basil Blackwell, Oxford, 1989.

[14] Hale, B., "*Grundlagen* §64," *Proceedings of the Aristotelian Society*, vol. 97 (1997), pp. 243–61.

[15] Hale, B., "Reals by Abstraction," *Philosophia Mathematica*, vol. 8 (2000), pp. 100–23.

[16] Heck, R. G., Jr., "Finitude and Hume's Principle," *Journal of Philosophical Logic*, vol. 26 (1997), pp. 589–617.

[17] Oliver, A., "Hazy totalities and indefinitely extensible concepts: an exercise in the interpretation of Dummett's *Philosophy of Mathematics*," pp. 25–50 in *Grazer Philosophische Studien 55, New Essays on the Philosophy of Michael Dummett*, edited by J. Brandl and P. Sullivan, Rodopi, Vienna, 1998.

[18] Parsons, C., "Frege's theory of number," pp. 180–203, in *Philosophy in America*, edited by M. Black, Allen and Unwin, London; reprinted on pp. 182–210 in *Frege's Philosophy of Mathematics*, edited by W. Demopoulos, Harvard University Press, Cambridge, 1995.

[19] Shapiro, S., "Induction and indefinite extensibility: the Gödel sentence is true but did someone change the subject," *Mind*, vol. 107 (1998), pp. 597–624.

[20] Shapiro S., and A. Weir, "New V, ZF and abstraction," *Philosophia Mathematica*, vol. 7 (1999), pp. 293–321.

[21] Wright, C., *Frege's Conception of Numbers as Objects*, Aberdeen University Press, Aberdeen, 1983.

[22] Wright, C., "On the philosophical significance of Frege's Theorem," pp. 201–44 in *Language, Thought and Logic*, edited by R. G. Heck, Jr., The Clarendon Press, Oxford, 1997.

[23] Wright, C., "On the harmless impredicativity of $N^{=}$ ('Hume's Principle'), " pp. 339–68 in *Philosophy of Mathematics Today*, edited by M. Schirn, The Clarendon Press, Oxford, 1998.

[24] Wright, C., "Response to Dummett," pp. 389–405 in *Philosophy of Mathematics Today*, edited by M. Schirn, The Clarendon Press, Oxford, 1998.

*Department of Logic and Metaphysics*
*University of St. Andrews*
*St. Andrews, Fife, KY169AJ*
*Scotland*
*UNITED KINGDOM*
*email: cjgw@st-and.ac.uk*

*Department of Philosophy*
*Columbia University*
*New York NY 10027*