Notre Dame Journal of Formal Logic
Volume 39, Number 2, Spring 1998

On a Consistent Subsystem of Frege's *Grundgesetze*

JOHN P. BURGESS

Abstract Parsons has given a (nonconstructive) proof that the first-order fragment of the system of Frege's *Grundgesetze* is consistent. Here a constructive proof of the same result is presented.

1 System Russell showed that the system of Frege's *Grundgesetze* is inconsistent. But the theme of much recent work on Frege (as represented, for instance, in Demopolous [2]) has been that the inconsistent system has consistent subsystems in which a significant amount of mathematics can be developed. In particular, Parsons [4] (see also the discussion by Boolos [1]) has proved the consistency of the first-order fragment of Frege's system. Heck [3] has extended the proof to cover predicative second-order fragments, while moreover showing that a well-known system **Q** of formal arithmetic can be interpreted in such a fragment.

These results leave a gap between the strongest system that has been interpreted within a predicative fragment of Frege's system and the weakest system in which the consistency of such a fragment has been proved. For on the one hand, whereas **Q** is by no means a trivial system, it is also by no means as strong a system as first-order Peano Arithmetic **PA**. And on the other hand, the original consistency proof for the first-order fragment and its extension to predicative second-order fragments are model-theoretic and nonconstructive and cannot be formalized even in **PA**. The first step toward narrowing this gap would be to produce a proof-theoretic and constructive proof of the consistency of the first-order fragment of Frege's system. This first step is taken in the present note. Further steps toward pinning down just how weak a system suffices to prove the consistency of predicative fragments of Frege's system, and just how strong a system can be interpreted in such fragments, must await the publication of Heck's results.

The first-order fragment of Frege's system may be presented as a first-order theory in the following way. Let $\mathbf{L}_0$ be the language of the first-order theory of identity.

Received February 4, 1997; revised March 17, 1998

Let $\mathbf{L}_1$ add to $\mathbf{L}_0$ a function symbol $\mathbf{e}_\varphi(\mathbf{y})$ for every formula $\varphi(x, \mathbf{y})$ of $\mathbf{L}_0$. (Here $\mathbf{y}$ represents a "vector" of any length $m \geq 0$ of additional free variables $y_1, \ldots, y_m$.) Let $\mathbf{L}_{n+2}$ add to $\mathbf{L}_{n+1}$ a function symbol $\mathbf{e}_\varphi(\mathbf{y})$ for every formula $\varphi(x, \mathbf{y})$ of $\mathbf{L}_{n+1}$ that is not already a formula of $\mathbf{L}_n$. Let $\mathbf{L}_\omega$ be the union of the $\mathbf{L}_n$ and let $\mathbf{T}_\omega$ be the theory in $\mathbf{L}_\omega$ having as axioms the following for all pairs of formulas of $\mathbf{L}_\omega$:

(V) $\quad \forall\mathbf{y}\forall\mathbf{z}(\mathbf{e}_\varphi(\mathbf{y}) = \mathbf{e}_\psi(\mathbf{z}) \longleftrightarrow \forall x(\varphi(x, \mathbf{y}) \longleftrightarrow \psi(x, \mathbf{z}))).$

Then $\mathbf{T}_\omega$ is a notational variant of the first-order fragment of Frege's system. A more traditional notation would be $\{x|\varphi(x, \mathbf{y})\}$ for $\mathbf{e}_\varphi(\mathbf{y})$. Let $\mathbf{T}_n$ be the subtheory of $\mathbf{T}_\omega$ in the language $\mathbf{L}_n$ whose axioms are all axioms of $\mathbf{T}_\omega$ that are formulas of $\mathbf{L}_n$. Then $\mathbf{T}_\omega$ is the union of the $\mathbf{T}_n$, and to prove the consistency of the first-order fragment of Frege's system it suffices to prove the consistency of each $\mathbf{T}_n$. This will be proved in Section 3 on the basis of three lemmas established in Section 2. The lemmas may be of some independent interest, but also seem individually so elementary that it is hard to believe they have not already been noted by others in some context, though I know of no reference in the literature.

2 Lemmas

Lemma 2.1 *Let* $\mathbf{T}$ *be a first-order theory implying the existence of infinitely many objects. Then the extension of* $\mathbf{T}$ *obtained by adding the axioms*

(A1) $\quad o \neq \pi(x, y)$ *and*
(A2) $\quad \pi(x, y) = \pi(u, v) \rightarrow (x = u \wedge y = v)$

is consistent.

Proof: It is to be understood that what are taken as axioms are the universal closures of what is displayed in (A1) and (A2). It is also to be understood that the constant omicron (o) and the two-place function symbol pi (π) do not already occur in the language of $\mathbf{T}$. Similar remarks apply in the other lemmas below. What is meant by saying that $\mathbf{T}$ implies the existence of infinitely many objects is that for each k, the formula $\mathbf{I}_k$ of the first-order language of identity saying that there exist more than k objects is a theorem of $\mathbf{T}$.

Toward proving the lemma, let $\mathbf{S}$ be the theory with o and π as its only nonlogical vocabulary and with (A1) and (A2) as its only nonlogical axioms. There will in addition be the logical axioms of identity, namely, reflexivity and indiscernibility of identicals for atomic formulas:

(A0a) $\quad x = x,$
(A0b) $\quad (x = u \wedge y = v) \rightarrow \pi(x, y) = \pi(u, v).$

First note that $\mathbf{S}$ is consistent. For by Herbrand's theorem, if it were inconsistent there would be some finite set of instances of (A0), (A1), and (A2), obtained by substituting terms of the language of $\mathbf{S}$ for the variables, that was *truth-functionally* unsatisfiable. But this is impossible, since any finite set of such instances is truth-functionally satisfiable by assigning the value 'true' to all and only those identities $t = s$ in which the terms t, s on the two sides are literally the same sequence of symbols. Further note that $\mathbf{S}$ itself implies each $\mathbf{I}_k$. Indeed, if we let

$$\mathbf{0} = o, \mathbf{1} = \pi(o, \mathbf{0}), 2 = \pi(o, \mathbf{1}), \ldots$$

then it is a theorem of **S** that $\mathbf{0}, \mathbf{1}, \ldots, k$ are all distinct.

What is to be proved is that $\mathbf{T} \cup \mathbf{S}$ is consistent. If not, then there would be a finite conjunction τ of axioms of **T** and a finite conjunction σ of axioms of **S** such that $\tau \to \neg\sigma$ is a theorem of first-order logic. But then by the Craig interpolation theorem there would be a formula φ such that $\tau \to \varphi$ and $\varphi \to \neg\sigma$ are theorems of first-order logic, and φ contains no nonlogical vocabulary except what is common to the languages of **T** and **S**, which is to say, contains no nonlogical vocabulary at all, but only the identity predicate. But as is well known, the first-order theory of identity is decidable by elimination of quantifiers, and the quantifier elimination shows that any closed formula of the language is equivalent to a truth-functional compound of the $\mathbf{I}_k$ for various k. Since **T** is consistent and implies each $\mathbf{I}_k$, and since $\mathbf{I}_k$ implies $\mathbf{I}_h$ for $h < k$, it follows that φ is implied by some $\mathbf{I}_k$ for k sufficiently large. But then since **S** is consistent and implies $\mathbf{I}_k$, $\varphi \to \neg\sigma$ cannot be a theorem of first-order logic. □

Lemma 2.2 *Let* **T** *be a consistent first-order theory whose axioms include (A1) and (A2) above. Then the extension of* **T**, *obtained by adding for every formula* φ *in the language of* **T** *the axiom,*

(B1) $\quad \forall \mathbf{y} \exists u \forall x (\varphi(x, \mathbf{y}) \longleftrightarrow \Delta(u, x)),$

is consistent.

Proof: It suffices to show that for any finite number of formulas $\varphi_1, \ldots, \varphi_n$, there is a formula $\delta(u, x)$ of the language of **T** such that instances of Axiom B1 for these φ_i become theorems of **T** when δ is substituted for Δ. And indeed if the total number of free variables additional to x occurring in these φ_i is m, then writing $\pi^2 = \pi$ and

$$\pi^k(y_1, y_2, \ldots, y_k) = \pi(y_1, \pi^{k-1}(y_2, \ldots, y_k)),$$

it suffices to let δ be the disjunction for $i = 1, \ldots, n$ of

$$\varphi_i(x, \mathbf{y}) \wedge u = \pi(\mathbf{i}, \pi^m(\mathbf{y})).$$

□

For technical purposes connected with the next lemma, note that (B1) implies that $\forall x(\Delta(u, x) \longleftrightarrow \Delta(v, x))$ is an equivalence relation, and that it has infinitely many equivalence classes (since the u corresponding to the formulas $x = \mathbf{0}$, $x = \mathbf{1}$, $x = \mathbf{2}$, ... must all be distinct). Moreover, it may be assumed that each equivalence class is infinite and that the following axiom holds.

(B2) $\quad \forall u \forall x(\Delta(u, x) \to \exists t(u = \pi(\mathbf{1}, t)).$

For if not, simply replace the original Δ by Δ' defined as follows.

$$\Delta'(u, x) \longleftrightarrow \exists v \exists w(u = \pi^3(\mathbf{1}, v, w) \wedge \Delta(w, x)).$$

Lemma 2.3 *Let* **T** *be a first-order theory and* $\varepsilon(u, v)$ *a formula with two free variables in the language of* **T** *such that* **T** *implies that* ε *is an equivalence relation, that it has infinitely many equivalence classes, and that each equivalence class is infinite. Then the extension of* **T***, obtained by adding the axiom,*

(C1) $\quad \forall u\, \varepsilon(u, \upsilon(u)) \wedge \forall u \forall v(\varepsilon(u, v) \rightarrow \upsilon(u) = \upsilon(v)),$

is consistent.

Proof: The hypothesis that **T** implies that there are infinitely many equivalence classes, and that each equivalence class is infinite, is not actually needed but does simplify the proof. What is meant by this hypothesis is that for each m and n the formula $\mathbf{E}_{>m,n}$ of the first-order theory of one equivalence relation saying that there exist at least m equivalence classes each having more than n elements is a theorem of **T**, and so is the negation of the formula $\mathbf{E}_{=m,n}$ saying that there exist at least m equivalence classes each having exactly n elements. It may be assumed that ε is a primitive two-place predicate, since such a predicate could always be added to the language with an axiom defining it to be equivalent to any desired formula with two free variables.

If the lemma failed, there would be a finite conjunction τ of axioms of **T** such that $\tau \rightarrow \neg\gamma$ is a theorem of first-order logic, where γ is the formula displayed in (C1). But then by the Craig interpolation theorem there would be a formula φ such that $\tau \rightarrow \varphi$ and $\varphi \rightarrow \neg\gamma$ are theorems of first-order logic, and φ contains no nonlogical vocabulary except what is common to the languages of **T** and γ, which is to say, contains no nonlogical vocabulary except the two-place predicate ε. But as is well known, the first-order theory of an equivalence relation is decidable by elimination of quantifiers, and the quantifier elimination shows that any closed formula of the language is equivalent to a truth-functional compound of the $\mathbf{E}_{>m,n}$ and $\mathbf{E}_{=m,n}$ for various m and n. And indeed, letting $\mathbf{F}_k$ be the conjunction of $\mathbf{E}_{>k,k}$ and the negations of the $\mathbf{E}_{=m,n}$ for all $m, n \leq k$, since each $\mathbf{F}_k$ is a theorem of **T** and since $\mathbf{F}_k$ implies $\mathbf{F}_h$ for $h < k$, it follows that φ is implied by some $\mathbf{F}_k$ for k sufficiently large. But each $\mathbf{F}_k$ has a finite model with just $k \cdot (k+1)$ elements, and any such finite model can be expanded to a finite model of (C1). So $\varphi \rightarrow \neg\gamma$ cannot be a theorem of first-order logic. □

3 Proof Let $\mathbf{T}^0$ be the theory in the first-order language of identity whose axioms are just the $\mathbf{I}_k$ for $k = 2, 3, 4, \ldots$, and apply Lemmas 2.1 and 2.2 and 2.3 to $\mathbf{T}^0$ to add the pairing apparatus o and π and predicate Δ_1 and function symbol υ_1 for which (A1), (A2), (B1), (B2), and (C1) all hold. Write $\Delta_1^*(u, x)$ for

$$\Delta_1(u, x) \wedge u = \upsilon(u).$$

Then for any formula φ of the language of $\mathbf{T}^0$, the following is a theorem.

(D1) $\quad \forall \mathbf{y} \exists! u\{[\neg\exists x \varphi(x, \mathbf{y}) \wedge u = \pi(\mathbf{0}, \mathbf{0})] \vee$
$[\exists x \varphi(x, \mathbf{y}) \wedge \forall x(\varphi(x, \mathbf{y}) \longleftrightarrow \Delta_1^*(u, x))]\}.$

Add function symbols $\mathbf{e}_\varphi$ and axioms defining $\mathbf{e}_\varphi(\mathbf{y})$ to be the unique u as in (D1) and call the resulting extension $\mathbf{T}^1$. Then $\mathbf{T}^1$ is consistent and moreover it has as theorems all pertinent instances of (V) in §1.

Now apply Lemmas 2.2 and 2.3 to $\mathbf{T}^1$ to add a predicate Δ_2 and function symbol υ_2 for which (B1), (C1), and the following variant of (B2) hold:

$$\forall u \forall x(\Delta_2(u, x) \to \exists t(u = \pi(\mathbf{2}, t)).$$

Write $\Delta_2^*(u, x)$ for

$$\neg\exists v \forall z(\Delta_2(u, z) \longleftrightarrow \Delta_1(v, z)) \wedge \Delta_2(u, x) \wedge u = \upsilon_2(u).$$

Then for any formula φ of the language of $\mathbf{T}^1$, the following is a theorem.

(D2) $\quad \forall \mathbf{y} \exists ! u \{ [\neg \exists x \varphi(x, \mathbf{y}) \wedge u = \pi(\mathbf{0}, \mathbf{0})] \vee$
$[\exists x \varphi(x, \mathbf{y}) \wedge \forall x(\varphi(x, \mathbf{y}) \longleftrightarrow \Delta_1^*(u, x)) \vee$
$\forall x(\varphi(x, \mathbf{y}) \longleftrightarrow \Delta_2^*(u, x))] \}.$

Add function symbols $\mathbf{e}_\varphi$ and axioms defining $\mathbf{e}_\varphi(\mathbf{y})$ to be the unique u as in (D2) and call the resulting extension $\mathbf{T}^2$. Then $\mathbf{T}^2$ is consistent and, moreover, it has as theorems all pertinent instances of (V) of Section 1. Iterating, obtain $\mathbf{T}^3$, $\mathbf{T}^4$, $\mathbf{T}^5$, ... in the same way and consider their union $\mathbf{T}^\omega$. This $\mathbf{T}^\omega$ is consistent and has all axioms of $\mathbf{T}_\omega$ of §1 as theorems, completing the proof that $\mathbf{T}_\omega$ is consistent. □

Acknowledgments I am grateful to Richard Heck first for calling my attention to the problem treated in this note, and second for teaching me that it was a nontrivial one by shooting down several naïve and fallacious proposals for a quick and easy solution that I floated.

REFERENCES

[1] Boolos, G., "Whence the contradiction?," *Aristotelian Society Supplementary*, vol. 67 (1993), pp. 213–33. Zbl 0961.03529 1

[2] Demopoulos, W., editor, *Frege's Philosophy of Mathematics*, Harvard University Press, Cambridge, 1995. Zbl 0915.03004 MR 96h:03015 1

[3] Heck, R., "*Grundgesetze der Arithmetik I* §§29–32," *Notre Dame Journal of Formal Logic*, vol. 38 (1997), pp. 437–74. Zbl 0915.03005 MR 2000a:03001 1

[4] Parsons, T., "On the consistency of the first-order portion of Frege's logical system," *Notre Dame Journal of Formal Logic*, vol. 28 (1987), pp. 161–88. Zbl 0637.03005 MR 88h:03002 1

Department of Philosophy
Princeton University
Princeton, NJ 08544-1006
email: jburgess@princeton.edu