

Fregean abstraction in Zermelo-Fraenkel set theory: a deflationary account

JOEL DAVID HAMKINS

Abstract. The standard treatment of sets and definable classes in first-order Zermelo-Fraenkel set theory accords in many respects with the Fregean foundational framework, such as the distinction between objects and concepts. Nevertheless, in set theory we may define an explicit association of definable classes with sets $F \mapsto \varepsilon F$ in such a way, I shall prove, to realize Frege's Basic Law V as a ZF theorem scheme, Russell notwithstanding. A similar analysis applies to the Cantor-Hume principle and to Fregean abstraction generally. Furthermore, because these extension and abstraction operators are definable, they provide a deflationary account of Fregean abstraction, one expressible in and reducible to set theory — every assertion in the language of set theory allowing the extension and abstraction operators εF , $\#G$, αH is equivalent to an assertion not using them. The analysis thus sidesteps Russell's argument, which is revealed not as a refutation of Basic Law V as such, but rather as a version of Tarski's theorem on the nondefinability of truth, showing that the proto-truth-predicate " x falls under the concept of which y is the extension" is not expressible.

I am grateful to Paddy Blanchette for insightful discussion and for pointing me to several relevant passages in Frege and grateful as well for insightful audience commentary at the talks I gave on this material in Oxford, Pavia, New York, and Paris. Commentary can be made about this article on my blog at <http://jdh.hamkins.org/fregean-abstraction-deflationary-account>.

§ 1. — Introduction.

The treatment of sets and classes in Zermelo-Fraenkel set theory instantiates in many respects the principal features of the Fregean foundational framework, in which Frege distinguishes sharply between *objects* and *concepts*. The realm of objects, a universe of individuals, constitutes the domain of discourse, a domain over which the quantifiers may range, and this domain is considered under the guise of diverse Fregean concepts, each serving in effect as a predicate on that domain, picking out a collection of those objects, for a given object either falls under a concept or does not. Frege analyzes concepts through their course of values, viewing them in effect as functions mapping every object either to an object standing for truth or one for falsity.

The usual practice and development of sets and classes in Zermelo-Fraenkel set theory accords very well with this Fregean framework, for set theorists consider the cumulative set-theoretic universe V , the collection of all sets, as constituting a domain of individual objects, the sets, about which we make mathematical assertions, with our quantifiers $\forall x$ and $\exists x$ ranging over all the sets in this universe, and we consider these sets in the light of the various set-theoretic properties $\varphi(x)$ expressible in the language of set theory, forming the corresponding definable classes $\{x \mid \varphi(x)\}$, thereby undertaking predication by those properties. Thus, we have a universe of first-order objects, the sets, in the context of diverse Fregean concepts, predicates on that domain, the classes. We refer to or name such a concept or class $\{x \mid \varphi(x)\}$ by providing the definition φ along with any parameters, if any, used in that definition. Different intensional descriptions φ or ψ may end up describing the same class extensionally, if $\varphi(x) \leftrightarrow \psi(x)$ holds for every individual x . In this way, we may come to view the usual understanding of sets and classes in ZF set theory in a Fregean light.

Andreas Blass expresses this view, that proper classes are not part of the set-theoretic ontology, but rather are merely a convenience for predication.

Proper classes are not objects. They do not exist. Talking about them is a convenient abbreviation for certain statements about sets. (For example, $V=L$ abbreviates “all sets are constructible.”) If proper classes were objects, they should be included among the sets, and the cumulative hierarchy should...continue much

farther, but in fact, it already continues arbitrarily far.
 ...Considerations like these are what prompt me to view
 ZFC as a reasonable foundational system, in contrast to
 Morse-Kelley set theory. [Bla11]

Thus, Blass appears to work in precisely the set theory I have described, a set-theoretic realm having only sets as objects, and considering classes only as a shorthand manner of undertaking definable predication.

In his well-known treatise on large cardinals, Kanamori [Kan04] similarly works in this theory. Notably, he regards the famous Kunen inconsistency, asserting that there is no nontrivial elementary embedding $j : V \rightarrow V$, as “a schema of theorems, one for each j .” That is, he considers the theorem as a scheme of statements making separate claims for each possible definition of such a class j .⁽¹⁾

Thus, it appears that mainstream set theory is often undertaken in the Fregean style I have described, where the universe of sets constitutes a first-order realm of objects, and the various definable classes occupy the second-order realm of concepts. The classes “do not exist” in the sense that they are not themselves objects — they are not sets — but rather the classes serve as predications of the objects. Set theorists nevertheless use and refer to classes, of course, in that same Fregean manner, just as Frege uses and refers to concepts. In short, in set theory we have sets as objects and definable classes as concepts.

In this article, I aim to investigate set theory in this Fregean light, to see how well we can implement Frege’s ideas in Zermelo-Fraenkel set theory. Let me emphasize, however, that my project here is not to elucidate Frege’s exact system in set theory, but rather to be inspired by Fregean ideas as a way of understanding what is going on in set theory. I aim to investigate Fregean-inspired set theory, not to present a set-theoretical Frege.

⁽¹⁾I find it curious for Kanamori to emphasize this point with this particular theorem, however, in light of the fact that one may easily refute the existence of nontrivial *definable* elementary embeddings $j : V \rightarrow V$ with a soft minimality argument showing that no particular definition (allowing parameters) can admit a smallest critical point of such an embedding; see [Suz98; Suz99]. This soft argument furthermore makes no use of the axiom of choice, while for the class version this is a major open question leading to the so-called choiceless large cardinals.

A quick review of terminology. Before proceeding to the main argument, let me quickly review some of the terminology I shall be using, which is all standard. By ZFC I refer to the usual Zermelo-Fraenkel axioms of set theory, including the axiom of choice, expressed in the first-order language of set theory, whose only nonlogical symbol is the binary membership relation $x \in y$. This is a first-order theory, whose models have the form $\langle M, \in^M \rangle$, where M is a domain of individuals, the sets of the model, and $x \in^M y$ is the membership relation on those sets. The first-order quantifiers $\forall x$ range over the elements x in that domain M . The theory without the axiom of choice is denoted ZF.

Gödel-Bernays set theory GB (sometimes denoted NGB), in contrast, is expressed in the second-order language, allowing for both sets and classes, with models of the form $\langle M, \in^M, \mathcal{X} \rangle$, where $\langle M, \in^M \rangle$ is a model of Zermelo-Fraenkel set theory, and the second-order part $\mathcal{X}$ is a family of classes $X \subseteq M$. The first-order quantifiers $\forall x$ range over the elements of M , while the second-order quantifiers $\forall X$ range over the classes $X \in \mathcal{X}$. The theory GB posits the first-order class comprehension axiom, asserting that $X = \{x \mid \varphi(x, z, Z)\}$ is a class for any set parameter z and class parameter Z , where φ is any assertion involving only first-order quantifiers. This is expressible by a single second-order axiom asserting that the classes defined by atomic assertions exist and the classes are closed under intersection, complements, and projections; it follows that GB is altogether finitely axiomatizable. The theory GB expresses the replacement axiom in the form: the image of a set under a class function is a set, and we use GBc to denote the theory with the axiom of choice for sets, while GBC is the theory with the (strictly stronger) global choice principle, asserting the existence of a global well order of the universe.

In the main theorem, I shall be using in effect the models of GBc that arise from a given model $\langle M, \in^M \rangle$ of ZFC, by equipping the model with all and only its first-order definable classes (allowing parameters). That is, we consider $\langle M, \in^M, \mathcal{X} \rangle$, where $\mathcal{X}$ consists of the classes $X \subseteq M$ that are definable $X = \{x \in M \mid M \models \varphi(x, z)\}$ in $\langle M, \in^M \rangle$ by some first-order formula φ with parameter $z \in M$. This is easily seen to be a model of GBc, and furthermore since every model of ZFC can be expanded in this way to a model of GBc, it shows that GBc is conservative over ZFC for assertions about sets. A corresponding statement is also

true about GBC by adding also a generic class well-ordering of the universe, which can be done by forcing without adding any sets.

The theory GBC is altogether weaker than Kelley-Morse set theory KM, which has the second-order class comprehension axiom, asserting that $\{x \mid \varphi(x, z, Z)\}$ is a class, where z and Z are set/-class parameters, and where now φ is allowed to have second-order quantifiers $\forall X$, ranging over the classes $X \in \mathcal{X}$ of the model. This axiom is expressed as a scheme of statements in the second-order language.

§ 2. — Basic Law V.

One of Frege's central ideas was that we might associate every concept F with an object εF , the *extension* of F , an object that names or represents this concept. In this way, Frege is able to treat concepts in effect as objects by means of their extensions — he can quantify over them and refer to them using the objects that represent them.

The *extension* terminology carries today the connotation of a collection — the extension of a concept is nowadays commonly understood in set theory as the set or class of fulfilling instances of that concept. Nevertheless, this collection connotation was not necessarily part of Frege's conception. The extension object εF of a concept F was simply an object of some kind that served to name or represent the concept and needn't itself be understood as a set, class, or collection.

Importantly, Frege insisted that the assignment of extension objects to concepts should be a conceptual invariant — the extensions of equivalent concepts should be the same, and the extensions of inequivalent concepts should be different. In other words, εF should represent the concept F extensionally, rather than intensionally, and this is expressed by his Basic Law V.

Basic Law V. *The extension of a concept F is the same as the extension of concept G if and only if $Fx \leftrightarrow Gx$ for every x .*

$$\varepsilon F = \varepsilon G \iff \forall x (Fx \leftrightarrow Gx).$$

Let me refer to the problem of finding a system of extension assignments $F \mapsto \varepsilon F$ fulfilling Basic Law V as the *extension-assignment problem*. Are we able in general to define extension objects for all our concepts in a way that fulfills this conceptual invariance requirement?

Basic Law V is an instance of the more general process of *Fregean abstraction*, by which we move from a notion of equivalence to the abstraction of that equivalence, a classification invariant for that equivalence. With Basic Law V, concepts F and G are logically equivalent in the sense that they have the same course of values $\forall x (Fx \leftrightarrow Gx)$ if and only if they have the same corresponding abstraction, that is, the same extensions of those concepts $\varepsilon F = \varepsilon G$. Similarly, according to the Cantor-Hume principle, considered in section 8, concepts F and G are *equinumerous* in that there is a one-to-one correspondence of the instances of F with instances of G if and only if they have the same corresponding abstractions, that is, the same number of elements $\#F = \#G$. And more generally, for any notion of equivalence $\sim$, the corresponding Fregean abstractions αF would be classification invariants in the sense that $F \sim G$ if and only if $\alpha F = \alpha G$.

In this article, I shall provide deflationary accounts for all of these various Fregean abstraction principles, providing ZF definitions of suitable extension objects εF , number objects $\#F$, and the various abstraction objects αF for any concept of equivalence, and prove as ZF theorem schemes that they fulfill the desired classification invariant requirements expressed by Basic Law V, the Cantor-Hume principle, and Fregean abstraction generally. I describe this account as deflationary because the abstractions I provide are expressible and defined in set theory, and as a consequence they are entirely eliminable — every assertion in the language with the abstraction operators, I shall prove, is equivalent to an assertion in the plain language of set theory without them.

§ 3. — Basic Law V holds in ZF set theory.

Let me prove the main result, namely, that Basic Law V holds in the set-theoretic Fregean framework I have described. I shall define an explicit association of definable classes F with sets εF representing them in such a way that the association $F \mapsto \varepsilon F$ is definable in set theory in various senses and such that Basic Law V is fulfilled. The theorem thus realizes Basic Law V as a provable theorem scheme in Zermelo-Fraenkel set theory.

Theorem 1. *In Zermelo-Fraenkel ZF set theory, there is an explicit association of definable classes F with sets εF , such that:*

1. *The association $F \mapsto \varepsilon F$ is definable in the following senses:*

- (a) *For any explicitly given class $F = \{x \mid \varphi(x, z)\}$, the extension object is first-order definable in set theory, using the same parameter z as provided in the description of F :*

$$\forall y (y = \varepsilon F \text{ if and only if } \theta(y, z)).$$

- (b) *For any class I and class function $i \mapsto F_i$ for $i \in I$ and classes F_i , presented uniformly as a definable class $F = \{(i, x) \mid i \in I, x \in F_i\} \subseteq I \times V$, the function $i \mapsto \varepsilon F_i$ is first-order definable in set theory, again with the same parameters used to define F .*

- (c) *The full mapping $F \mapsto \varepsilon F$ is second-order definable in any model of ZF equipped with its definable classes:*

$$\forall F \forall y (y = \varepsilon F \text{ if and only if } \Theta(F, y))$$

2. *The association $F \mapsto \varepsilon F$ fulfills Basic Law V:*

$$\varepsilon F = \varepsilon G \text{ if and only if } \forall x (Fx \leftrightarrow Gx).$$

Proof. To prove the theorem, I shall begin by explaining how to define the extension objects εF . Suppose that we are presented with a particular definable class $F = \{x \mid \varphi(x, z)\}$, the class of all sets x fulfilling the first-order formula $\varphi(x, z)$, using a fixed set parameter z . To have provided a class concept means to have provided such a formula φ and parameter z explicitly. I shall associate this class F with a certain set εF in such a manner so as to fulfill all the claims of the theorem. Technically, I shall be providing for each formula $\varphi(x, z)$ a definable mapping $z \mapsto y$, where $y = \varepsilon F$ will be the extension of the class $F = \{x \mid \varphi(x, z)\}$ defined by φ with that parameter z . We cannot in general take the class $\{x \mid \varphi(x, z)\}$ itself as the extension, since this might be a proper class, not a set, but we need εF to be a set. Instead, what I propose is to use the data of the presentation itself to define εF .

Let us fix in the metatheory a particular enumeration $\psi_0, \psi_1, \psi_2, \dots$ of the formulas of the first-order language of set theory. The formula φ appears somewhere in this enumeration. Let ψ_n be the earliest formula in that enumeration that is capable, with some choice of parameter, of defining the same class that φ

defines with parameter z . That is, ψ_n is the earliest formula on the list for which there is some set p for which $\forall x(\varphi(x, z) \leftrightarrow \psi_n(x, p))$. We may find such a parameter p of minimal possible rank in the set-theoretic hierarchy. Indeed, let u be the set of all such parameters p of that minimal rank that define with that formula ψ_n the same class as φ does with parameter z . We now define εF be the pair $(\ulcorner \psi_n \urcorner, u)$, where $\ulcorner \psi_n \urcorner$ is the set coding the formula ψ_n . In short: the extension of a concept is the earliest formula on the list that is capable of defining it with some parameter, together with the set of minimal-rank parameters that do so.

Let us now prove that εF is first-order definable in the senses mentioned in the theorem. Suppose that we are presented with a class $F = \{x \mid \varphi(x, z)\}$, presented intensionally — that is, we have been provided explicitly with the defining formula φ and the parameter z . The formula φ itself appears as some ψ_n in the list, and there are only finitely many preceding formulas $\psi_0, \dots, \psi_n = \varphi$ in the enumeration. To express the property $y = \varepsilon F$, therefore, we may take a disjunction over $i \leq n$ of the assertions that say,

“ y is a pair $(\ulcorner \psi_i \urcorner, u)$, whose first coordinate is the code of the formula ψ_i and whose second coordinate u is a nonempty set of parameters $p \in u$ for which

$$\forall x (\varphi(x, z) \leftrightarrow \psi_i(x, p)), \text{ and}$$

and furthermore u is the set of all such minimal-rank p for which this is true, and finally, none of the formulas ψ_j for $j < i$ admit such a parameter.”

This property exactly describes what it means for y to be the extension object εF as we defined it above, and since we have the formula φ as a metatheoretic formula in hand and there are only finitely many preceding formulas ψ_i , all of this is expressible by a formula θ in the first-order language of set theory

$$y = \varepsilon F \iff \theta(y, z).$$

Thus I have established claim (1a) as a ZF theorem scheme.

Next, suppose that we are provided with a definable class functional $i \mapsto F_i$, associating every object i in a class I with a class F_i . We assume that this class functional has been provided intensionally, by means of a uniform defining formula φ and parameter z , such that

$$x \in F_i \iff \varphi(i, x, z).$$

The uniformity of this definition means that it simply reduces to case (1a), for we can simply treat i as another parameter, thereby defining the extensions

$$y = \varepsilon F_i \iff \theta(y, i, z)$$

using the same analysis as above. This provides a uniform definition of the class function

$$i \mapsto \varepsilon F_i.$$

Thus, whenever we have a concept for a function from objects to concepts, we also have a concept for the extensions of those concepts, establishing (1b) as a ZF theorem scheme.

For statement (1c), let us assume we work in a model of ZF equipped for the Henkin semantics with all and only its definable classes, a model of Gödel-Bernays set theory GBc (not necessarily with the global choice principle). I claim that the map $F \mapsto \varepsilon F$ is second-order definable in this context. Here we assume that we are given the class F as a class only, without knowing the formula and parameter by which it was defined. Nevertheless, with a second-order assertion ranging over the definable classes, we shall be able to recover the defining formula and thereby recognize and provide the extension object εF . We may assume that the meta-theoretic enumeration of formulas $\psi_0, \psi_1, \psi_2, \dots$ that we used in the definition is also available internally to the object theory, arising as the standard part of a definable enumeration. Note that for standard k , there is a definable Σ_k -truth predicate. That is, we have first-order definable classes providing these partial truth predicates. Using these classes, we can express the relation $y = \varepsilon F$ as follows:

y is a pair $\langle \ulcorner \psi_n \urcorner, u \rangle$, where ψ_n is a Σ_k formula for some k , such that there is a class T that is a Σ_k truth predicate, u is the nonempty set of minimal-rank parameters p such that $\forall x (Fx \leftrightarrow T(\ulcorner \psi_n \urcorner, \langle x, p \rangle))$, and no preceding formula ψ_i has this property.

The key point is that since F is definable, it is defined by some formula ψ_n having some complexity Σ_k , and because the least such k will be standard there will be a definable Σ_k truth predicate class T available, and with this class we can express truth for $\psi_n(x, p)$ and the fact that no earlier ψ_i succeeds in defining F . So we have proved (1c).

Finally, let me prove that this definition fulfills Basic law V. We had defined εF to be $\langle \ulcorner \psi \urcorner, u \rangle$ where ψ is earliest formula to define

F and u is set of minimal-rank parameters that do so. Thus, the extension object εF depends only on the class F considered as a predicate, and not on the original presentation of F as defined by a formula and parameter $\varphi(x, z)$. If we had used a different formula and parameter, but thereby defined the same class, then the extension object εF would still be picking out the least formula ψ_n in the metatheoretic enumeration that is able to define that class, and u would again be the set of minimal-rank parameters that work with this formula to define F . Therefore, different equivalent presentations of the definable class will get assigned the same object:

$$\varepsilon F = \varepsilon G \quad \text{if and only if} \quad \forall x (Fx \leftrightarrow Gx).$$

And this is precisely what it means for Basic Law V to be fulfilled. $\square$

Notice how the method exhibits Frege's distinction between identity of concepts and extensional equivalence of concepts. Namely, we can often describe a class in two distinct manners $\{x \mid \varphi(x)\} = \{x \mid \psi(x)\}$, that is, using different formulas φ and ψ or perhaps using the same formula with different parameters. So we have given different names to this class, by picking it out by means of different descriptions, and in this sense we have different concepts, but those concepts have the same extension in the case that $\forall x (\varphi(x) \leftrightarrow \psi(x))$, and in this case we shall have $\varepsilon\{x \mid \varphi(x)\} = \varepsilon\{x \mid \psi(x)\}$, or perhaps it is better to write it simply as $\varepsilon\varphi = \varepsilon\psi$.

I described in theorem 1 an interpretation of extension objects for unary classes $\{x \mid \varphi(x, z)\}$ defined by a formula $\varphi(x, z)$ with one free variable x ranging over the defined class and the other variable z held constant as a parameter. In the most general case, however, one would naturally want to consider higher-arity formulas $\varphi(x_1, \dots, x_k, z_1, \dots, z_r)$ defining a k -ary relation in the first k variables, using the other variables z_i as constants or parameters in the definition. For this one could introduce a formalism $\varepsilon_{x_1, \dots, x_n} \varphi$ to indicate exactly which variables are taken as part of the defined relation, with the others held as parameters in the definition (and Frege has an analogous notation for exactly this kind of case). Without saying much more about the details of such a system, it is clear that the argument of theorem 1 works essentially just the same with this more general treatment.

What I claim about theorem 1, furthermore, is that it is deflationary. Every use of the extension operator ε in the semantics that

are provided is entirely eliminable — every assertion in the language of set theory expanded to allow the extension operator $\varepsilon\varphi$ and more generally the higher-arity versions $\varepsilon_{x_1, \dots, x_n}\varphi$ is equivalent to an assertion in the base language of set theory alone.

Corollary 2. *In the semantics for extensions of definable classes $\varepsilon\varphi$ provided by theorem 1, including the higher-arity versions $\varepsilon_{x_1, \dots, x_n}\varphi$, every use of the extension operator ε is eliminable. That is, every assertion in the expanded language of set theory allowing the extension operator ε is equivalent to an assertion in the language of set theory. Furthermore, one can find the ε -free translation by a primitive recursive translation function.*

Proof. One can prove this by induction on formulas, using statements (1a) and (1b) of theorem 1. One can systematically eliminate $\varepsilon\varphi$ from formulas φ not involving ε . The point is statements (1a) and (1b) provide ε -free identity criterion for $y = \varepsilon\varphi$, and so any assertion made about this extension object can be made equivalently without need for any ε expression. The same idea works in the higher-arity case $\varepsilon_{x_1, \dots, x_n}\varphi$. The translation process is clearly primitive recursive, being defined by a syntactic recursion on formulas. $\square$

It follows immediately that Basic Law V, expressed in the language expanding the language of set theory with the extension operator $\varepsilon\varphi$, is a conservative extension of Zermelo-Fraenkel set theory in the base language of set theory. No new facts about sets will be proved by adopting the $\varepsilon\varphi$ formalism using the semantics I have defined.

This conservativity observation is essentially related to the conservativity of the Hilbert-Bernays epsilon operator (see [AZ20]), but there is an important difference: that operator is not definable or eliminable in the way I have just proved for $\varepsilon\varphi$ in corollary 2. My operator is eliminable, deflationary, but the Hilbert-Bernays epsilon operator is not.

One might want to consider the quality of the extension assignments in any proposed solution to the Basic Law V abstraction problem. The extension objects I provided in the proof above fulfill the biconditional requirement of Basic Law V, but there is no claim here of any canonicity. To be sure, perhaps one regards the abstraction requirement itself as expressing all that matters concerning a solution to the abstraction problem — every solution might be taken as equally good or valid as another.

Nevertheless, some may find a given solution dissatisfying or unnatural. My standard response when someone objects to a given solution as being unnatural is to inquire precisely which further features are desired? It is the duty of those who object on naturality grounds to articulate a precise account of what counts as natural. In the case of the extension-assignment problem, what further properties should we seek that are not expressed by Basic Law V?

Let me defend my extension assignment solution as natural in several further respects. First, my extension objects carry the natural information that one would expect when naming or presenting classes. Namely, whenever we present a particular class concept, we do so by presenting the class $\{x \mid \varphi(x, z)\}$ with a particular defining formula φ and possibly a parameter z . And the extension object I define for this case similarly consists of a pair $\langle \psi, u \rangle$, where ψ is such formula (possibly simpler than φ) and u is a set of such parameters (of minimal rank), defining the very same class. Indeed, the formula ψ might be the very same formula φ used in the presentation, if that one had already been optimal. And conversely, every extension object $\langle \psi, u \rangle$ for a class naturally provides the means for us to present the class in that natural way, namely, presenting it as $\{x \mid \psi(x, p)\}$ for any particular $p \in u$. These are senses in which my solution is natural.

But furthermore, more importantly, my solution is also natural in fulfilling the definability features achieved in theorem 1 statement (1). This is precisely the deflationary aspect of my solution. I shall argue in section 5 that we should naturally count such deflationary definability requirements as amongst the further desirable features we would want to see in any attractive solution of the extension-assignment problem — the extension assignments become almost useless or incoherent without this. This definability feature of my solution is missing in the prior art discussed in section 4.

In section 8 we shall see further instances where one seeks to place further naturality requirements on an abstraction problem. Namely, in connection with the Cantor-Hume abstraction principle for equinumerosity, we shall discuss that not every solution of the cardinal-assignment problem is also a solution of the cardinal-selection problem, whereas the latter solutions seem more desirable.

§ 4. — Prior art.

Let me mention some relevant earlier work in the same direction as the project of this article. Terence Parsons [Par87] showed the consistency of the first-order fragment of Frege's system by providing a model in which Basic Law V is fulfilled for first-order definable classes. He begins with the observation that every countable model in a countable language admits, of course, only countably many definable classes. Given a countably infinite model, he divides the domain (externally to the model) into infinitely many infinite classes, reserving these objects to be assigned as extensions εF for the definable classes F in a series of stages stratified by the ε depth of the definitions. That is, at the first stage, using the first class of objects, he assigns extensions for the definable classes that are defined by formulas not involving extensions at all; at the next stage, using the next class of objects, he handles classes defined by formulas that make reference to the extensions, but only the extensions of extension-free formulas; and so on. The main point is that at each stage, all the relevant extension objects for a given formula to be successfully interpreted will have been already assigned at the earlier stages, and in this way he is able to assign all the extension objects εF in turn, realizing Basic Law V.

John L. Bell [Bel94] strengthens the Parsons result with a very clear argument, expanding the language Henkin-style with constants c_A for each formula A , in such a way that $c_A = c_B \leftrightarrow \forall x(Ax \leftrightarrow Bx)$, which is to say that the objects c_A serve as conceptually-invariant extensions for the classes A . See also his later work [Bel00], detailing a theory for the systematic labeling of classes.

John Burgess [Bur98] carries out the Parsons argument a little more explicitly, using compactness to achieve consistency in expanded language. Burgess [Bur05, p. 89] mounts a similar cardinality argument using Löwenheim-Skolem to reduce from full second-order to Henkin semantics with countably many classes.

I greatly admire this earlier work, but what I should also like to mention about it is that none of these prior arguments provide a deflationary solution of the extension-assignment problem in the sense I have claimed for my solution in theorem 1. In the Parsons construction, for example, the labeling of concepts with objects is performed from outside the model, as though by the hand of God (or at least the hand of Parsons), and similarly in effect for the Bell construction and the compactness arguments. These methods consequently do not provide a uniform definability of the identity

criterion, as in (1a), (1b), and (1c) of theorem 1. In the case of Bell's construction, of course one gets the expressibility of $y = \varepsilon A$ in the expanded language simply by using the formula $y = c_A$, but in general there is no guarantee that the individual named by c_A is definable in the original language (and indeed the original model may have no definable elements), and the method does not achieve a uniform identity criterion with respect to parameters. That is, if one defines classes A_i uniformly by a formula $A_i(x) \leftrightarrow \varphi(x, i)$, then there will generally be no uniform way to express $y = c_{A_i}$ in the manner of (1b) of theorem 1, since each parameter i leads to a syntactically different constant c_{A_i} , and these constants cannot all appear in one formula.

§ 5. — The identity criterion and definability.

I find it frankly strange to neglect the identity criteria and definability questions arising with the extension-assignment problem, for it seems to me that having an expressible identity criterion should be taken as a core part of what it means to solve the extension-assignment problem. Not only must there be an assignment of extension objects εF to concepts F , but one should also be able to recognize which objects are extensions, which objects go with which concepts, which concepts go with which objects and so on.

In order actually to use the extensions εF , after all, to form the extensions for our concepts at hand and to use them in our analysis, we would want to know not just that there are objects out there, somewhere, that have been assigned to concepts in a conceptually invariant manner, somehow, with nothing more to be said about which objects are used or how. Rather, we would want to know much more about it. Which objects get assigned as extensions for which concepts? How do we recognize for a given concept F which object is its extension?

$$y = \varepsilon F$$

How do we recognize for a given object y for which concept F it is the extension?

$$y = \varepsilon F$$

How do we recognize whether a given object y is the extension of any concept?

$$\exists F y = \varepsilon F$$

What use will an extension assignment $F \mapsto \varepsilon F$ be, after all, if performed externally to the domain of discourse, if it is invisible within it? If the assignments are made Parsons-style from outside the model, as it were, then we cannot seem to work with the resulting solution in the original object theory. But what good is that? The only way to find out which object is the extension εF of an expressible concept F , in those systems, is to already know what it is.

In my view, the identity criterion problem for a given solution to the extension-assignment problem is akin to the famous Julius Ceasar problem. Namely, when Frege despairs at the impossibility of determining whether Julius Ceasar is a number, at bottom he is pointing out his lack of an identity criterion, his lack of a decision procedure for determining whether a given individual is a number. Similarly, a solution of the extension-assignment problem without a corresponding identity criterion will lack a decision procedure for determining whether a given object or individual is the extension of a concept.

The definability of the extension assignments amounts to having what I have referred to as a deflationary account of extensions. When we have an extension-assignment $F \mapsto \varepsilon F$ that admits an expressible identity criterion, one for which we achieve the definability properties mentioned in theorem 1, then we are in effect able to refer to the extensions of concepts in the original language and theory. Precisely because the features of the extension assignment are expressible in the base language, we have no need to expand the language to include direct reference to extensions. In short, to have a solution of extension-assignment problem that admits of definable identity criterion is exactly to be able to reduce the treatment of extensions to the original language and theory, to deflate the extensions to the original theory, as described in corollary 2.

For this reason, I find the solution of the extension-assignments provided by my theorem 1 to be an advance on the earlier work mentioned in section 4, at least for the context of set theory as the base theory. Not only can we have a solution of the extension-assignment problem realizing Basic Law V in set theory, but we

can do so in a deflationist manner, so that we can recognize in the object language which objects get assigned to which concepts.

Frege builds the expressivity of the extension assignment into his syntax for the expanded language with the extension operator ε , because he allows himself to form expressions such as $y = \varepsilon\varphi(x, p)$ for any formula φ , including any formula in the expanded language. This formal assertion expresses the relation of y being the extension of the concept defined by $\varphi(x, p)$. But ultimately one must recognize that Frege provided no semantics for these expressions and did not seem to make any claims about whether his use of ε should be eliminable.

Robert May had emphasized to me in discussion at the 2022 AILA conference in Pavia that Frege had sought a “logically perfect” language, one in which every concept has a name and every well-formed name denotes a concept (see [Fre03, §28]). To my way of thinking, however, merely having a syntactic expression $\varepsilon\varphi(x, p)$ serving as a name to refer to the extension object of a given concept is quite different from knowing which object it is or whether it is possible to have a semantics for these expressions having all the properties claimed of them. Ultimately, I shall argue, Frege’s syntactic approach led him astray, because it was built into his language that he could express the concept “object x falls under the concept of which y is the extension,” and it is this, rather than Basic Law V, that is the true downfall of his system, as I shall explain in theorem 3.

§ 6. — Reconciling Russell.

Russell is commonly described as having refuted Frege’s system and he is often described specifically as having shown Basic Law V to be inconsistent. How are we to reconcile this with my claim in theorem 1 that Basic Law V is a provable theorem scheme of Zermelo-Fraenkel set theory? Wouldn’t this show ZF also to be inconsistent?

To be sure, Russell’s argument is most naturally construed as a refutation of the general comprehension principle.

General Comprehension Principle. *For any property φ , one may form the set of all x with property $\varphi(x)$. That is,*

$$\{ x \mid \varphi(x) \}$$

is a set.

Russell's argument refutes even very simple instances of this, in the case of the formula $x \notin x$. Namely, by the general comprehension principle, we may form the set R consisting of all sets x that are not elements of themselves.

$$R = \{ x \mid x \notin x \}.$$

If this is indeed a set, then $R \in R$ if and only if $R \notin R$, which is a contradiction. Thus, the general comprehension principle has contradictory instances.

Although this is often how the Russell paradox is presented, the historical difficulty is that Frege does not state the general comprehension principle explicitly as part of his system. Rather, that principle is essentially hidden away into his manner of forming concepts and denoting them. In his letter replying to Russell, Frege blames Basic Law V, yet also hints at a systemic problem:

It seems, then, that transforming the generalization of an equality into an equality of course-of-values (§9 of my *Grundgesetze*) is not always permitted, that my rule V (§20, p. 36) is false, and that my explanations in §31 are not sufficient to ensure that my combinations of signs have a meaning in all cases. I must reflect further on the matter. It is all the more serious since, with the loss of my rule V, not only the foundations of my arithmetic, but also the sole possible foundations of arithmetic, seem to vanish.

–Frege, 22 June 1902 letter to Russell [Hei67, p.127]

responding to Russell's fateful letter

Let me explain how Russell's argument can be used to refute Basic Law V in combination with a specific comprehension-like principle.

Theorem 3 (Russell). *Basic Law V is inconsistent with the existence of the concept “ x does not fall under any concept of which x is the extension.”*

Proof. Assume Basic Law V and let R be the concept “ x does not fall under any concept of which x is the extension.” If this is indeed a concept, then let $r = \varepsilon R$ be the extension of it, then by Basic Law V we see that any concept of which r is the extension will agree extensionally with R , and so r falls under R if and only if it does not, which is a contradiction. $\square$

The main point here is that Russell's argument is not at its core a refutation of Basic Law V as such, but rather a refutation of a certain concept-formation principle, which is an instance of class comprehension. It would suffice for Russell's argument if we had the *falling-under* concept: object x falls under the concepts of which y is the extension. In this case, we could negate and project onto the diagonal to form the forbidden concept mentioned in theorem 3.

In fact, the argument does not require the full biconditional of Basic Law V, but only the forward implication, namely, that if r is the extension of a concept, then that concept agrees with R . For example, it would suffice that every concept F had a distinct extension εF . This situation falls far short of Basic Law V, and it is similar to the situation of every formula φ having a distinct Gödel code $\ulcorner \varphi \urcorner$, but we don't generally insist that equivalent formulas get the same code. The point is that Russell's argument shows that even this weak form of the law is inconsistent with the existence of the concept " x falls under the concept of which y is the extension." In this sense, Russell's argument has little to do with Basic Law V as such.

I should like to emphasize that the falling-under concept amounts to a proto-truth predicate. That is, to say that object x falls under the concept of which y is the extension is exactly to say that that concept is true of x . In short, the extension-assignments εF are a kind of Gödel-coding for predicates F , and the falling-under concept is the corresponding truth predicate.

The true mistake of Frege's system, in my view, is not Basic Law V as such, but rather the fact that Frege has built this proto truth predicate into his syntax. Namely, one can easily express in Frege's system that x falls under the concept of which y is the extension by the assertion:

$$\exists F (y = \varepsilon F \text{ and } Fx).$$

Thus, Frege has made the falling-under concept available simply as a matter of notation. And he makes pervasive use of it throughout his work. He uses the falling-under concept, for example, when defining what it means to be a natural number: a natural number is any object that falls under every property that holds of 0 and is passed from every object to its successor.

Perhaps Frege looked upon the idea of "falling under" as innocent, it being merely a special case of functional application, which Frege took as fundamental. That is, if one understands concepts as functionals mapping objects to *true* or *false*, then for an object x to

fall under a concept F is simply to compute the value of $F(x)$. If this is *true*, then x falls under F , and if it is *false*, then it does not.

To be sure, the difficulty of the falling-under relation is not when it is used like this in a particular case with a particular concept F , for one can express that x falls under the concept F simply by Fx . Rather, the source of all the trouble is the falling-under relation as a relation of objects to objects, taking the latter objects to represent concepts, as in the relation “object x falls under the concept of which y is the extension.” This uniform treatment of the falling-under relation enables Russell’s refutation as described in theorem 3 and is what amounts to a truth predicate.

§ 7. — Proving Tarski’s theorem directly from Russell.

Let me explain next how this kind of Russellian reasoning can be used to provide a direct proof of Tarski’s theorem on the non-definability of truth. Since Gödel’s incompleteness theorem can be seen as a consequence of Tarski’s theorem, this argument therefore also provides a proof of Gödel’s theorem directly from Russell. In short, I would like to explain how truly close Russell was to proving Gödel’s theorem. I find this remarkable, because contemporary accounts of Tarski’s theorem usually invert this presentation, commonly proving Tarski’s theorem only after Gödel’s, and indeed often treating Tarski’s theorem essentially as an afterthought to Gödel.

Tarski’s theorem is commonly described in arithmetic or set theory, given a particular encoding of formulas φ with objects $\ulcorner \varphi \urcorner$, as the assertion that there can be no formula $T(y, x)$ in the same language expressing the truth of those formulas, that is, fulfilling every instance of the truth conditions:

$$\forall x [T(\ulcorner \varphi \urcorner, x) \leftrightarrow \varphi(x)].$$

A typical proof of Tarski’s theorem proceeds by arguing that if there were such a truth predicate, then by the Gödel fixed-point lemma there would be a sentence asserting its own nontruth according to the predicate, and then by the Liar-paradox reasoning this sentence would be true if and only if the predicate declared it not to be true, contradicting the truth-predicate requirement.

That is fine, but allow me here to state a rarified version of Tarski’s theorem and prove it directly by Russell’s method, bypassing Gödel, with no explicit use of self-reference or the fixed-point

lemma. In particular, this formulation of the theorem avoids the need to consider any encoding of syntax whatsoever. We could state the theorem for arithmetic or for set theory, but in fact the argument applies to any first-order theory at all. It is a totally elementary argument.

Theorem 4 (Tarski). *In any first-order theory, truth is not definable. That is, there is no formula $T(y, x)$ in the same language, such that for every formula $\varphi(x)$ it is consistent with the theory that there is some y for which*

$$\forall x [T(y, x) \leftrightarrow \varphi(x)].$$

Proof. Given any formula $T(y, x)$ with two free variables, let $R(x) = \neg T(x, x)$. This is a perfectly good formula with one free variable. There can be no r for which $\forall x [T(r, x) \leftrightarrow R(x)]$, however, since by its definition $R(r)$ asserts $\neg T(r, r)$, which contradicts $T(r, r)$. So there is no formula T such that every formula can arise consistently with the theory as a section of T . $\square$

Let me elaborate in the case where we have fixed a representation of formulas $\ulcorner \varphi \urcorner$, which as I had mentioned is a common way of talking about Tarski's theorem. Given any proposed truth predicate T , let $R(x) = \neg T(x, x)$, which on the truth-predicate understanding of T could be seen as expressing, "it is not the case that x represents a formula that is true of x ." Let $r = \ulcorner R \urcorner$ and form the sentence $R(r)$, which asserts $\neg T(r, r)$. But this contradicts $T(r, r)$, which is the same as $T(\ulcorner R \urcorner, r)$, and so $T(\ulcorner R \urcorner, r)$ is not equivalent to $R(r)$, which violates the desired equivalence. So there can be no such definable truth predicate T .

I find this argument to be a direct parallel of Russell's original argument against Frege. The proof here is essentially identical to Russell's proof of theorem 3, simply using $\ulcorner R \urcorner$ in place of ϵR . The Russell formula $R(x)$ asserting that x is not a formula that is true of itself is directly analogous to the Russellian concept of all concepts that do not apply to themselves, or the set of all sets x that are not members of themselves.

A very similar manner of presenting Russell's argument as a proof of Tarski's theorem is also provided in [Fit17], and accords with arguments of Smullyan, who had emphasized that much of the fascination surrounding Gödel's theorem may be more properly directed at Tarski's theorem. The inevitable road from Russell to Gödel is the related similar theme of Kripke's [Kri14].

Of course, one can find key elements of Gödel's argument here. The construction of the sentence $R(r)$, for example, where $R(x) = \neg T(x, x)$ and $r = \ulcorner R(x) \urcorner$, is exactly what ultimately appears in the usual proof of Gödel's fixed-point lemma, and so the resulting sentence in this Russellian argument is the same as what one finds in the Gödelian fixed-point proof of Tarski's theorem. In that sense, the two approaches to proving Tarski's theorem are not very different. Nevertheless, the straightforward Russellian treatment above is simpler than the Gödelian proof and entirely avoids confusing issues of self-reference. In any case, I find it far more natural to view Gödel's fixed-point lemma as generalizing Russell's method here, rather than viewing Russell's argument as an application of Gödel's lemma, which is how Tarski's theorem is often presented. Surely Gödel found his proof of the fixed-point lemma by contemplating and generalizing Russell's argument.

I have stated Tarski's theorem in a two-dimensional version, ruling out a definable satisfaction relation rather than sentential truth. But one also commonly finds a sentential version of the theorem, stating that there can be no arithmetically expressible predicate T such that

$$T(\ulcorner \sigma \urcorner) \leftrightarrow \sigma$$

for every sentence σ , using the standard Gödel coding of sentences $\ulcorner \sigma \urcorner$. Of course arithmetic satisfaction and sentential truth are closely related in arithmetic, since one expects that a formula φ is true at a number x if and only if the sentence $\varphi(1 + 1 + \cdots + 1)$ is true, using the term $1 + 1 + \cdots + 1$ with x summands. The point is that PA proves that every number x is the value of such a term, possibly nonstandard, in a nonstandard model.

Apart from that, however, the sentential formulation of Tarski's theorem can also be proved along Russellian lines, but the proof will require that we have an encoding of syntax in the object theory, rather than merely in the metatheory as above, and this admittedly makes the argument closer to Gödel. Namely, suppose that T is a sentential truth predicate. Let us form the Russell formula $R(x)$ asserting " x is the code of a formula φ that is false at the point x ." This can be expressed as

$$R(x) \quad =_{\text{def}} \quad \neg T(\ulcorner \overbrace{\varphi(1 + 1 \cdots + 1)}^x \urcorner \urcorner).$$

The term $1 + 1 + \cdots + 1$ with x summands is formed and then substituted into φ all in the object theory — this is where we use that

syntactic operations can be undertaken in the object theory, in order to know that R is expressible. Now let $r = \ulcorner R \urcorner$, and consider the sentence $R(r)$, meaning $R(1 + 1 + \cdots + 1)$ with r summands. By the definition of R the sentence $R(r)$ asserts $\neg T(\ulcorner R(r) \urcorner)$, but this contradicts the truth predicate property of T . So there can be no such sentential truth predicate.

Perhaps when thinking of this sentential version, some people argue that Tarski's theorem relies fundamentally on the use of Gödel codes $\ulcorner \varphi \urcorner$ and the arithmetization of syntax, whereas arithmetization is rightly considered a core contribution of Gödel. I am in complete agreement on the profound nature of Gödel's arithmetization ideas; see my remarks at [Ham21, p. 233]. A rebuttal to the specific point here, however, is that the core of Tarski's theorem simply requires far less coding than Gödel's. First of all, the set-theoretic version of Tarski's theorem has no need for coding specifically into arithmetic at all, since one can use set theory as a foundation of mathematics to represent essentially arbitrary mathematical constructions — including sequences and syntax and parse trees — a usage widely recognized by Hilbert and others before Gödel. And there is no need in Tarski's theorem to undertake the coding of any formal proof system. But furthermore, in the two-dimensional satisfaction form of Tarski's theorem, as I have stated in theorem 4, there is no explicit need for the coding of syntax at all. In particular, this is why the proof I gave above for theorem 4 is much simpler than the corresponding treatment in [Fit17], which aims instead at the sentential formulation. The simple fact is that Tarski's theorem in the two-dimensional satisfaction form applies in a sweeping manner to any possible representation of formulas in the object theory, even a very rudimentary representation, for which syntactic operations may not be expressible.

To my way of thinking, the idea of having an internal representation of formulas in the object theory can be traced to Frege himself. Namely, I see a very strong analogy between the extension $\varepsilon\varphi$ of a concept φ and the Gödel code $\ulcorner \varphi \urcorner$ of a formula, for in both cases we have an internal object-theory representation of a formula or concept from the metatheory. And as I have explained, Tarski's theorem already applies with such a representation, even without any additional involvement of Basic Law V. Namely, when proving Tarski's theorem we did not require that the representations respect logical equivalence, that equivalent formulas are represented by the same code.

Let me offer a slogan form of Tarski's theorem in a Fregean-style language:

"Truth is not a concept."

Just as Russell's argument showed in theorem 3 a sense in which the falling-under relation is not a concept, similarly the Russellian proof of theorem 4 shows that truth is not expressible. The core of the argument in both cases is about our lack of uniformity in expressing these notions. That is, in the context of a given instance, we can easily express that a given object x falls under a given concept F by asserting Fx . What we cannot do is express in a uniform manner that "object x falls under the concept of which y is the extension." Similarly, in the context of Tarski's theorem, there is no difficulty for us to express that a particular formula φ holds at a point x — we simply assert $\varphi(x)$ itself — what we cannot do is express in a uniform manner that " y represents a formula that is true at x ."

Now that we have proved Tarski's theorem directly via Russell, let me deduce Gödel's incompleteness theorem as a consequence. Namely, since truth is not definable, but provability is definable (and for this one must implement the proof system in the object theory), they cannot be the same thing. QED.

In this way, one realizes the power of Tarski's theorem over Gödel's. Whereas Gödel is showing that truth is not the same as provability, a Σ_1 notion, Tarski shows the stronger result that truth is not Σ_n expressible for any n .

§ 8. — The Cantor-Hume principle.

Frege had used his extension objects εF and Basic Law V in order to prove other abstraction principles, including especially the Cantor-Hume principle, upon which he founded his arithmetic, which in many respects was the focal effort of his project. What I would like to do is explain how the ideas of theorem 1 can be used also to fulfill these other Fregean abstraction principles in Zermelo-Fraenkel set theory.

Let me begin with the central instance, the Cantor-Hume principle, also known simply as Hume's principle, which is the Fregean abstraction principle concerning numbers as a classification invariant of the equinumerosity relation. Specifically, one class X is

equinumerous with another Y , written $X \simeq Y$, if there is a one-to-one correspondence between them. The *cardinal-assignment problem*, which we might also call the *number-assignment problem*, is the problem of assigning a cardinal or number object $\#X$ to every class X , in such a way that it is an equinumerosity invariant, so that equinumerous classes get assigned the same number and nonequinumerous classes get assigned different numbers. The Cantor-Hume principle is the assertion that indeed there is a solution to this problem.

Cantor-Hume principle. *There is a number-assignment scheme, assigning to every class X a number object $\#X$, for which two classes get assigned the same number if and only if they are equinumerous:*

$$\#X = \#Y \quad \text{if and only if} \quad X \simeq Y.$$

Let me begin by mentioning several easy solutions to the cardinal-assignment problem that are available in Zermelo-Fraenkel set theory for special instances of the principle.

Consider, for example, the set-only version of the cardinal-assignment problem. This is completely solved in ZFC using the axiom of choice by the standard treatment of cardinalities in ZFC. Namely, in ZFC every set x can be well-ordered and is therefore equinumerous with some smallest ordinal, and so we may simply define the cardinality $|x|$ of the set x to be the smallest ordinal with which it is equinumerous. This not only solves the cardinal-assignment problem, but does so in a way that is simultaneously a solution of the *cardinal-selection problem*, the problem of choosing a unique representative from each equinumerosity class. That is, for any set x , the cardinality of x defined as the smallest ordinal equinumerous with x is not only an equinumerosity invariant, but it is one of the sets that is itself equinumerous with x — this ZFC definition provides a definable way to pick a canonical member in every equinumerosity class of sets.

This same solution works to an extent in ZF, without the axiom of choice, at least for the well orderable sets — for every such set x we can define the cardinality $|x|$ to be the smallest ordinal equinumerous with it.

But in fact, we do not need the axiom of choice at all to solve the cardinal-assignment problem for sets.

Theorem 5. *In Zermelo-Fraenkel set theory ZF without the axiom of choice, there is a definable solution of the cardinal-assignment problem for*

sets, thereby verifying the Cantor-Hume principle for sets. That is, there is a definable assignment

$$x \mapsto \#x$$

assigning to each set x a set $\#x$, in such a way that provides a classification invariant for equinumerosity:

$$\#x = \#y \quad \text{if and only if} \quad x \simeq y.$$

The theorem has become a folklore result in set theory, but it seems that Gödel had it by 1951, see [Göd48, footnote to Dfn 6.2, added 1951]. The main idea has become known as Scott's trick, in light of Dana Scott's discovery of it at around that same time, while an undergraduate student of Tarski's in Berkeley, but it would seem reasonable to me for it to be named the Gödel-Scott trick.⁽²⁾ Scott's trick resolves an important issue in the definition of ultrapowers of the set-theoretic universe, and thus continually arises and rearises in large cardinal set theory, which is preoccupied with such ultrapowers as a central method, and it also arises similarly when defining the Boolean ultrapower of the universe by a forcing notion, thus becoming a key technique in the diverse mutual interpretability results provided by forcing.

The essence of Scott's trick is to handle a proper-class sized equivalence class for an equivalence relation defined on a proper class not by picking representatives of each class, which might not be possible without the global choice principle, but rather by considering the set of rank-minimal elements in each class, which will form a set and which will canonically represent that equivalence class.

Proof. The key idea is to drop the cardinal-selection aspect of the minimal-ordinal solution in ZFC, and strive instead merely for an equinumerosity invariant as required by the Cantor-Hume principle. For this, we can for any set x define $\#x$ to be the set of rank-minimal sets in the equinumerosity class of x . This is simply to use Scott's trick to select a subset of the equinumerosity class of the set. The set of rank-minimal members of the equinumerosity class determines the equinumerosity class and is invariant with respect to equinumerosity, and so it fulfills the Cantor-Hume principle for sets. $\square$

⁽²⁾I corresponded with Dana Scott about his trick, and he explained that Tarski put to the class the problem of finding a set-sized invariant for a definable equivalence relation on the set-theoretic universe. Scott's minimal-rank idea solved it, and Tarski was taken with Scott's solution, dubbing it "Scott's trick," but Scott himself was sheepish about the honor, emphasizing to me his desire to be remembered for other mathematical achievements, several of which he itemized to me.

That seems to polish off the set version of the cardinal-assignment problem. But the Cantor-Hume principle, of course, is not just about assigning cardinals to sets, but requires us also to assign numbers to classes, and so we have not yet achieved a proof of the full Cantor-Hume principle in ZF.

To find a solution for the class case, let's observe first that if the global choice principle happens to hold, for example, if $V = \text{HOD}$, then we can define a well-ordering of the universe in order type Ord , and from this it follows that all proper classes are equinumerous. In this event, therefore, there will be only one additional equinumerosity class to be handled, the proper class equinumerosity class, and so we can simply assign $\#X$ either to be the smallest ordinal equinumerous with X , if X is a set, or some default value ∞ that is not an ordinal, if X is a proper class. This will solve the cardinal-assignment problem under global choice, fulfilling the Cantor-Hume principle.

If the global choice principle fails, however, then not all proper classes are equinumerous, even in ZFC. For example, the whole universe V will not be equinumerous with the class of ordinals, since such a bijection would imply global choice.

So let me now explain a completely general solution to the number-assignment problem, for both sets and classes, one which fulfills the Cantor-Hume principle in ZF without using any choice principle at all. The main difficulty here is that we may have many different proper classes, not equinumerous, but we need to associate to each of them a cardinal number in such a way that fulfills the Cantor-Hume principle.

Theorem 6. *In Zermelo-Fraenkel set theory ZF with definable classes only, there is a number assignment scheme $F \mapsto \#F$, assigning to every definable class F a set $\#F$, called the number of the class F , with the following properties:*

1. *For any particular definable class $F = \{x \mid \varphi(x, z)\}$, the identity criterion for being the number $\#F$ is expressible by a formula Θ in the second-order language of set theory, using the same set parameter z , if any, appearing in the presentation of F :*

$$y = \#F \quad \text{if and only if} \quad \Theta(y, z).$$

2. *The identity criterion of whether a given object x is a number is expressible by a formula $\Lambda(x)$ in the second-order language of set theory.*

3. The numbers $\#F$ are well-defined with respect to the equinumerosity equivalence of the classes, in the sense that

$$\#F = \#G \quad \text{if and only if} \quad F \text{ is equinumerous with } G.$$

In other words, the Cantor-Hume principle holds for these number assignments.

Proof. The main difficulty, of course, is to describe how we are to assign numbers to each class. We shall provide a general procedure that works for any definable class F . Namely, suppose we are faced with a particular definable class $F = \{x \mid \varphi(x, z)\}$, named intentionally by the formula φ with parameter z . Let us define $\#F$ to be the pair $\langle \ulcorner \psi \urcorner, u \rangle$, where ψ is the first formula on the enumeration of all formulas for which there is some z such that $\{x \mid \psi(x, z)\}$ is equinumerous with F , and u is the set of all minimal-rank such sets z for which this is true.

Let's verify statement (1) of the theorem. Given any particular definable class $F = \{x \mid \varphi(x, z)\}$, let $\Theta(y, z)$ assert that y is a pair $\langle \ulcorner \psi \urcorner, u \rangle$ whose first coordinate is the code of a formula ψ and whose second coordinate is the set of minimal-rank parameters p for which $\{x \mid \psi(x, p)\}$ is a class that is equinumerous with F by a definable bijection, and such that no earlier formula ψ_i preceding ψ on the list can define with some parameter such a class equinumerous with F . This assertion is expressible in second-order set theory, in the context where the classes are all the first-order definable classes, because we can quantify over the partial truth predicates and thereby refer to the existence of a definable bijection between the classes.

For statement (2), the question of whether a given object x is the number $\#F$ of some definable class F is simply the question of whether x is a pair $\langle \ulcorner \psi \urcorner, u \rangle$ whose first coordinate is the code of a formula for which there is a truth predicate available and whose second coordinate is the set of minimal-rank parameters all defining via ψ classes that are definably equinumerous with one another, such that no earlier formula works in this way.

For statement (3), the number assignment $\#F$ clearly determines the equinumerosity class of F and depends only on that class and not on the particular presentation of F . So it will fulfill the full Cantor-Hume principle as desired. $\square$

One can understand statement (2) as an answer of sorts to the Julius Caesar problem, for it provides a criterion deciding whether

a given object x is a number on this account of number. Of course, Frege was concerned with the Julius Caesar problem in so far as that the Cantor-Hume principle, by itself, did not seem automatically to provide an identity criterion for what is a number. Here, in contrast, statement (2) is providing such an identity criterion for the specific number-assignment defined in the proof of this theorem; there is no claim, of course, that this is the only solution to the number-assignment problem and other assignments will naturally have a different identity criterion.

Perhaps one objects to this theorem in comparison with the realization of Basic Law V in theorem 1 because here we have only second-order definitions for the number identity criterion. But to my way of thinking, this objection must retreat in the face of the fact that the equinumerosity relation itself is a second-order relation. That is, the invariance property of Basic Law V is the first-order property of logical equivalence of two classes, $\forall x [Fx \leftrightarrow Gx]$, whereas with the Cantor-Hume principle the equivalence relation of equinumerosity $F \simeq G$ is itself asking for a definable bijection between F and G , a second-order quantification. So it is hardly surprising that the corresponding identity criterion is also second-order expressible.

§ 9. — Fregean abstraction in ZF.

Frege emphasized the process by which we abstract from a notion of equivalence to the abstraction of that equivalence. We have a notion of parallel lines, for example, which all have the same direction, but what is the *direction* of a line? According to Frege, it is an abstraction of parallelism. The direction of a line is an invariant of the parallelism relation — parallel lines have the same direction and nonparallel lines have different directions. Similarly, we have the notion of concepts being logically equivalent, and the *extension* of a concept is the corresponding abstraction — according to Basic Law V, logically equivalent concepts have the same extension and logically inequivalent concepts have different extensions. We have the notion of classes being equinumerous, and the *number* of a class is the abstraction — according to the Cantor-Hume principle, equinumerous classes have the same number of elements and nonequinumerous classes have a different number of elements.

Thus Fregean abstraction generalizes both Basic Law V and the Cantor-Hume principle. In the general case, we have an equivalence relation $x \sim y$, and we seek a system of abstractions $x \mapsto \alpha x$ with the property that equivalent instances get assigned the same abstraction and inequivalent instances get assigned different abstractions.

Fregean abstraction principle. *For any given equivalence relation $x \sim y$, the Fregean abstraction for this relation is a system of abstractions*

$$x \mapsto \alpha x$$

which respect the equivalence relation:

$$\alpha x = \alpha y \quad \text{if and only if} \quad x \sim y.$$

In short, Fregean abstraction for an equivalence relation $\sim$ provides a classification invariant for that relation, an assignment $x \mapsto \alpha x$ that is constant on each equivalence class and with different values on different equivalence classes.

In this section, I should like to investigate how well we might achieve the Fregean abstraction principle in Zermelo-Fraenkel set theory. Let me start as I had earlier with easier special instances of the principle. Let us begin with the easy case of a first-order definable equivalence relation $x \sim y$ on the sets.

Theorem 7. *In Zermelo-Fraenkel set theory ZF, suppose that $x \sim y$ is a first-order definable equivalence relation defined on a class of sets. Then there is definable system of abstractions, $x \mapsto \alpha x$, associating to each set x in the domain an object αx called the abstraction of x , which form a classification invariant for $\sim$ and thereby fulfill the Fregean abstraction principle:*

$$\alpha x = \alpha y \quad \text{if and only if} \quad x \sim y$$

Proof. In this easy case, the equivalence relation lives on sets only, and so we can simply employ Scott's trick. Namely, we define the abstraction αx of any object x in the relevant class to be the set of rank-minimal members of the $\sim$ -equivalence class of x . This is an invariant for the equivalence relation, because if $x \sim y$, then they have the same equivalence class and consequently the same set of rank-minimal members of that class, and so $\alpha x = \alpha y$. And conversely, the set of rank-minimal members of an equivalence class determines that class, and so if $\alpha x = \alpha y$, then indeed $x \sim y$. So this is a full solution of Fregean abstraction in ZF for definable class equivalence relations on the sets. $\square$

Model theorists will recognize this theorem as the elimination of imaginaries in ZF. This theorem covers the case of parallel lines, if one regards the lines as interpreted in set theory, and many further cases of that nature, including the case of equinumerosity for sets. In particular, this simple theorem covers the case of equinumerosity for finite classes, since these must be sets, and as a consequence this theorem provides a sufficient level of Fregean abstraction to undertake much of Frege's development of natural number arithmetic, which requires only sets rather than classes, since every finite class is a set.

But of course, several of Frege's other principal applications of Fregean abstraction arise not from equivalence relations on sets, but on classes. This is a more difficult realm, in which Scott's trick is not necessarily applicable. So let me make the step up to classes, treating first the case of a first-order expressible equivalence relation on definable classes:

$$F \approx G.$$

The instances F and G here will be definable classes in ZF set theory, and the notion of equivalence $F \approx G$ will be expressible by a formula $\varphi(F, G)$ involving only first-order quantification, in which those classes F and G appear as predicates. This is the kind of equivalence arising in Basic Law V, for example, because to express the course-of-values equivalence for concepts F and G , which is to say, that they are logically equivalent, is to say $\forall x (Fx \leftrightarrow Gx)$, which is therefore first-order expressible in terms of F and G .

Theorem 8. *In Zermelo-Fraenkel set theory, suppose that $\approx$ is a first-order definable equivalence relation on definable classes. Then there is a system of abstractions $F \mapsto \alpha F$, associating to each definable class F an abstraction object αF , which fulfill the Fregean abstraction principle:*

$$\alpha F = \alpha G \quad \text{if and only if} \quad F \approx G.$$

Furthermore, for any particular definable class $F = \{x \mid \varphi(x, z)\}$, the identity criterion $y = \alpha F$ is first-order expressible in set theory, using only the same parameter z , if any, appearing in the presentation of F :

$$\forall y (y = \alpha F \leftrightarrow \theta(y, z)).$$

If I is any class and $i \mapsto F_i$ is a definable sequence of classes, then the identity criterion $y = \alpha F_i$ is first-order expressible in set theory, using the same parameter used to define the sequence:

$$\forall y \forall i [y = \alpha F_i \leftrightarrow \theta(y, i, z)]$$

Proof. We use the same idea as in theorem 1. Namely, given any class $F = \{x \mid \varphi(x, z)\}$, define the abstraction αF to be the pair $\langle \ulcorner \psi_n \urcorner, u \rangle$, whose first coordinate is the code of a formula ψ_n that with some parameter p defines a class that is equivalent to F , that is, for which $\{x \mid \psi_n(x, p)\} \approx \{x \mid \varphi(x, z)\}$, and where ψ_n is the first formula to achieve this with some parameter and u is the set of rank-minimal parameters that do so with ψ_n . We don't insist that all parameters in u define the same class, but rather just that they define an $\approx$ -equivalent class. This assignment will fulfill the Fregean abstraction principle, because αF is determined by and determines the $\approx$ -class of F , and so we will achieve $\alpha F = \alpha G \iff F \approx G$.

The identity criterion $y = \alpha F$ will be first-order expressible, if we are provided with the explicit definition of $F = \{x \mid \varphi(x, z)\}$, since there will be only finitely many earlier competitor formulas in the metatheoretic list of formulas, and to express that a given definable class is $\approx$ -equivalent to F is first-order expressible.

That case immediately implies that we can also express the identity criterion $y = \alpha F_i$ for a uniformly definable sequence of classes $i \mapsto F_i$, defined by a first-order formula $F_i(x) \leftrightarrow \varphi(x, i, z)$. The point is that this amounts simply to regarding i as another parameter. $\square$

As I mentioned, this theorem has Basic Law V as a special case, since the equivalence relation of logical equivalence for definable classes is first-order expressible. In this sense, theorem 1 is a special case of theorem 8.

Let me finally treat the most difficult case, where we have an equivalence $F \equiv G$ on definable classes that is second-order expressible in Zermelo-Fraenkel set theory equipped with definable classes only.

Theorem 9. *In Zermelo-Fraenkel set theory, suppose that $\equiv$ is a second-order definable equivalence relation on definable classes, in the context where all classes are first-order definable. Then there is a system of abstractions $F \mapsto \alpha F$, associating to each definable class F an abstraction object αF , which fulfill the Fregean abstraction principle:*

$$\alpha F = \alpha G \quad \text{if and only if} \quad F \equiv G.$$

Furthermore, the identity criterion $y = \alpha F$ is second-order expressible

$$y = \alpha F \quad \text{if and only if} \quad \Theta(y, F).$$

Proof. We may proceed just as in theorem 6. Namely, the abstraction object αF is a pair $\langle \ulcorner \psi_n \urcorner, u \rangle$ where ψ_n is the first formula on

the metatheoretic list that is able with some parameter p to define a class $\{x \mid \psi_n(x, p)\}$ that is $\equiv$ -equivalent to F . This is easily seen to fulfill the Fregean abstraction principle requirement, since αF determines and is determined by the $\equiv$ -equivalence class of F . The identity criterion $y = \alpha F$ will be expressible in second-order set theory in our context, because first of all, we have available the partial truth predicates sufficient to determine the truth of any particular formula $\psi_n(x, p)$, and the $\equiv$ -equivalence relation was assumed to be second-order expressible. $\square$

One must disparage the fact that the identity criterion $y = \alpha F$ is merely second-order expressible here, which is surely unsatisfying, but it should not be surprising at all in light of the fact that this is the case where the equivalence relation $\equiv$ itself is merely second-order definable. Theorem 9 includes the case of equinumerosity of classes, and thus has theorem 6 as a special case.

§ 10. — Gödel-Bernays and Kelley-Morse set theories.

Let us now investigate the extent to which we may achieve the Fregean abstraction principles in the various standard second-order set theories, such as Gödel-Bernays set theory and Kelley-Morse set theory. In the earlier main theorems, we had considered Zermelo-Fraenkel set theory in the context of all its definable classes, which is a model of Gödel-Bernays GB set theory (without the axiom of choice or global choice). All those arguments generalize, however, to the context of a *principal* model of Gödel-Bernays set theory, which is a model in which there is a class Z from which all other classes are first-order definable. An important example of these kinds of models arising when performing class forcing over a model of ZFC, for in such a case one gets a principal model of Gödel-Bernays set theory when taking as classes everything that is definable from the generic filter class G . Indeed, the principal models are preserved generally by tame class forcing, since one can amalgamate the new generic filter with the old class parameter.

Observation 10. *All the results of theorems 1, 6, and 9 hold not just in all models of Zermelo-Fraenkel with definable classes, but also in all principal models of Gödel-Bernays set theory, provided one allows a fixed class parameter in the definitions.*

The earlier proofs simply relativize, accommodating the fixed class parameter Z into the definitions of the abstraction objects. The extension αF of a class $F = \{x \mid \varphi(x, z, Z)\}$, for example, will be the pair $\langle \ulcorner \psi_n \urcorner, u \rangle$ where ψ_n is the first formula on the list able to define F with some parameter p and class Z , and u is the set of rank-minimal such parameters p that do so. And similarly with the other Fregean abstraction invariants $\#F$ and αF for the other instances of Fregean abstraction.

The main result of [HLR13] shows that every countable model of Gödel-Bernays GBC set theory (with countably many sets and classes) has an extension to a pointwise definable model, a model of GBC in which every set and class is first-order definable without parameters. Such a model falls under the scope of the original theorem, since the first-order ZFC content of the model suffices to recover definably all the original classes. In this sense, any countable set-theoretic universe can be extended to one where there are definable solutions for Fregean abstraction.

What I should like to observe next, however, is that one cannot achieve this in a model of Kelley-Morse set theory, for Russell's argument can be implemented in KM with force.

Theorem 11. *Kelley-Morse set theory is inconsistent with a definable solution of the extension-assignment problem (even second-order definable), which is to say, KM is inconsistent with a definable solution to Basic Law V.*

Proof. Assume in Kelley-Morse set theory that we have a definable assignment $F \mapsto \varepsilon F$, in the sense that

$$\forall x \forall F [x = \varepsilon F \leftrightarrow \theta(x, F)]$$

for some formula $\theta(x, F)$ in the second-order language of set theory. It would be fine for the argument if this definition involved second-order quantifiers. It follows that the Julius-Caesar-like property of x being an extension $\exists F x = \varepsilon F$, is second-order expressible in Kelley-Morse set theory.

Let us also assume that this system of extension-assignment fulfills Basic Law V:

$$\forall F, G [\varepsilon F = \varepsilon G \leftrightarrow \forall x (Fx \leftrightarrow Gx)].$$

In the Russellian style, let R be the class of all sets x that do not fall under any concept of which it is the extension. That is,

$$R(x) \iff \forall F [x = \varepsilon F \rightarrow \neg Fx].$$

If $r = \varepsilon R$, then $R(r)$ asserts that $\neg F(r)$ for any F for which $r = \varepsilon F$, and by the Basic Law V requirement all such F agree on whether $F(r)$ or not. But R is such an F , and so $R(r)$ if and only if $\neg R(r)$, which is a contradiction. $\square$

Theorem 1 shows that the corresponding theorem is not provable in mere Gödel-Bernays set theory, because that theorem shows that every model of Zermelo-Fraenkel ZF set theory, when equipped with its definable classes, does admit a definable system of extension assignments $F \mapsto \varepsilon F$.

§ 11. — Fregean ontology without infinity.

Although I have provided the analysis of this paper for Zermelo-Fraenkel ZF set theory, which includes the axiom of infinity, nevertheless the entire analysis can be undertaken in finite set theory $\text{ZF}^{\neg\infty}$, which replaces the infinity axiom with its negation and which includes the foundation axiom in the form of the $\in$ -induction schema. I find finite set theory to be quite a nice realm for Fregean thinking.

Observation 12. *All the results of theorems 1, 6, and 9 hold in models of finite set theory $\text{ZF}^{\neg\infty}$, equipped with their definable classes.*

The proofs go through essentially unchanged. One thing that is a bit nicer in finite set theory is that there is a definable global well-ordering — this theory proves that the whole universe is in bijective correspondence with the finite von Neumann ordinals (just as V_ω can define a bijection of itself with ω), and so we are in the easy case for definably picking representatives in every definable class.

§ — References.

- [AZ20] Jeremy Avigad and Richard Zach. “The Epsilon Calculus”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Fall 2020. Metaphysics Research Lab, Stanford University, 2020. <https://plato.stanford.edu/archives/fall2020/entries/epsilon-calculus/>.
- [Bel00] John L. Bell. “Sets and Classes as Many”. *Journal of Philosophical Logic* 29.6 (2000), pp. 585–601. doi: 10.1023/a:1026564222011.
- [Bel94] John L. Bell. “Fregean extensions of first-order theories”. *Math. Logic Quart.* 40.1 (1994), pp. 27–30. issn: 0942-5616. doi: 10.1002/malq.19940400104.
- [Bla11] Andreas Blass. *Are proper classes objects?* MathOverflow answer. 2011. <https://mathoverflow.net/q/71773> (version 01 August 2011).
- [Bur05] John P. Burgess. *Fixing Frege*. Princeton Monographs in Philosophy. Princeton University Press, Princeton, NJ, 2005, pp. x+257. isbn: 978-0-691-12231-1; 0-691-12231-8.
- [Bur98] John P. Burgess. “On a consistent subsystem of Frege’s *Grundgesetze*”. *Notre Dame J. Formal Logic* 39.2 (1998), pp. 274–278. issn: 0029-4527. doi: 10.1305/ndjfl/1039293068.
- [Fit17] Melvin Fitting. “Russell’s paradox, Gödel’s theorem”. In: *Raymond Smullyan on self reference*. Vol. 14. Outst. Contrib. Log. Springer, Cham, 2017, pp. 47–66.
- [Fre03] Gottlob Frege. *Grundgesetze der Arithmetik, Band I/II*. Verlag Herman Pohle, 1893/1903.
- [Göd48] K. Gödel. “The consistency of the axiom of choice and of the generalized continuum hypothesis with the axioms of set theory”. *Uspehi Matem. Nauk (N.S.)* 3.1(23) (1948), pp. 96–149. issn: 0042-1316.
- [Ham21] Joel David Hamkins. *Lectures on the Philosophy of Mathematics*. MIT Press, 2021. isbn: 9780262542234. <https://mitpress.mit.edu/books/lectures-philosophy-mathematics>.
- [Hei67] Jean van Heijenoort. *From Frege to Gödel. A Source Book in Mathematical Logic, 1879–1931*. Harvard University Press, 1967, xi+660 pp. (1 plate).

- [HLR13] Joel David Hamkins, David Linetsky, and Jonas Reitz. “Pointwise definable models of set theory”. *Journal of Symbolic Logic* 78.1 (2013), pp. 139–156. issn: 0022-4812. doi: 10.2178/jsl.7801090. arXiv:1105.4597[math.LO]. [http : / / jdh . hamkins . org / pointwisedefinablemodelssettheory/](http://jdh.hamkins.org/pointwisedefinablemodelssettheory/).
- [Kan04] Akihiro Kanamori. *The Higher Infinite*. Corrected 2nd edition. Springer-Verlag, 2004.
- [Kri14] Saul Kripke. “The Road to Gödel”. In: *Naming, Necessity and More: Explorations in the Philosophical Work of Saul Kripke*. Ed. by Jonathan Berg. Palgrave-Macmillan, 2014, pp. 223–241.
- [Par87] Terence Parsons. “On the consistency of the first-order portion of Frege’s logical system”. *Notre Dame J. Formal Logic* 28.1 (1987), pp. 161–168. issn: 0029-4527. doi: 10.1305/ndjfl/1093636853.
- [Suz98] Akira Suzuki. “Non-existence of generic elementary embeddings into the ground model”. *Tsukuba J. Math.* 22.2 (1998), pp. 343–347. issn: 0387-4982.
- [Suz99] Akira Suzuki. “No elementary embedding from V into V is definable from parameters”. *J. Symbolic Logic* 64.4 (1999), pp. 1591–1594. issn: 0022-4812. doi: 10.2307/2586799.

Joel David Hamkins,
 O’Hara Professor of Logic, University of Notre
 Dame, 100 Malloy Hall, Notre Dame, IN 46556 USA
 & Associate Faculty Member, Professor of Logic,
 Faculty of Philosophy, University of Oxford, UK

★
 ★ ★