KAI FREDERICK WEHMEIER and PETER SCHROEDER-HEISTER

FREGE'S PERMUTATION ARGUMENT REVISITED*

1. INTRODUCTION

In Section 10 of *Grundgesetze*, Volume I, Frege advances a mathematical argument (known as the *permutation argument*), by means of which he intends to show that an arbitrary value-range may be identified with the True, and any other one with the False, without contradicting any stipulations previously introduced (we shall call this claim the *identifiability thesis*, following Schroeder-Heister (1987)). As far as we are aware, there is no consensus in the literature as to (i) the proper interpretation of the permutation argument and the identifiability thesis, (ii) the validity of the permutation argument, and (iii) the truth of the identifiability thesis.[1] In this paper, we undertake a detailed technical study of the two main lines of interpretation, and gather some evidence for favoring one interpretation over the other.

2. *GRUNDGESETZE* I, SECTIONS 1 THROUGH 9

To give the reader some background to the problem discussed in Section 10, we briefly review what is going on in the preceding sections.

In Sections 1, 2, and 4 we find some general and condensed explanations of the notions of (unary and binary) *function*, *function-name*, *object*, and *proper name*. Section 3 is a very short introduction of the notions of *value-range* and *concept-extension*. The explanation Frege (1893, 7) offers for value-ranges is the following:

> I use the words 'the function $\Phi(\xi)$ has the same *value-range* as the function $\Psi(\xi)$' generally as meaning the same [gleichbedeutend] as the words 'the functions $\Phi(\xi)$ and $\Psi(\xi)$ always assume the same value for the same argument'.[2]

In Section 5 Frege introduces, besides the assertion sign that will not concern us here, a function $-\xi$ mapping the True to the True and everything

* Dedicated to Christian Thiel on the occasion of his 65th birthday.

Synthese (2005) 147: 43–61
DOI 10.1007/s11229-004-6206-6

else to the False, i.e., the characteristic function of the True, as it were. Section 6 is devoted to the negation function $\neg\xi$, mapping the True to the False and everything else to the True.[3] Identity is introduced in Section 7 as a binary function $\xi = \zeta$, i.e., as the characteristic function of the identity relation (as we would say today). The first-order universal (and, derivatively, existential) quantifier makes its first appearance in Section 8. Frege's official explanation is the following (1893, 12):

(...) let '$\forall x\Phi(x)$' denote the True, if the value of the function $\Phi(\xi)$ is the True for every argument, and otherwise, the False.[4]

The value-ranges, already briefly introduced in Section 3, are the subject of Section 9. Frege opens his discussion of them by reiterating (1893, 14):

If $\forall x(\Phi(x) = \Psi(x))$ is the True, then we may also say, according to our earlier stipulation (§3), that the function $\Phi(\xi)$ has the same value-range as the function $\Psi(\xi)$; that is: we may transform the generality of an equality into an equality of value-ranges, and vice versa.[5]

Following Heck (1999), we shall call this Frege's *Initial Stipulation*. There is, however, also another explanation that resembles that given for the quantifier (Frege 1893, 15):

(...) let '$\hat{x}\Phi(x)$' denote the value-range of the function $\Phi(\xi)$[6]

and as in the quantifier case, this is immediately supplemented with conventions regarding the scope of the value-range operator $\hat{x}$.

3. LATER DEVELOPMENTS

It is interesting to note which notions Frege has *not* yet introduced when he comes to Section 10. These are, most importantly, the definite description operator (Section 11), the conditional function (Section 12), and second-order quantification (Sections 20 and 24). Neither has Frege mentioned any of his *Basic Laws* yet; in particular, Basic Law V has not been enunciated:

$$(\hat{x}\mathsf{f}(x) = \hat{y}\mathsf{g}(y)) = (\forall z)(\mathsf{f}(z) = \mathsf{g}(z)),$$

where the free first-order function variables are understood to be bound by initial second-order universal quantifiers. This law occurs for the first time in Section 20, although, of course, the Initial Stipulation is just an informal version of it. For future use, we note that Frege, in Section 52, splits his Basic Law V up into

(Va) $\quad (\forall z)(\mathsf{f}(z) = \mathsf{g}(z)) \rightarrow (\mathsf{F}(\hat{x}\mathsf{f}(x)) = \mathsf{F}(\hat{y}\mathsf{g}(y)))$[7]

and

$$\text{(Vb)} \qquad (\hat{x}\mathsf{f}(x) = \hat{y}\mathsf{g}(y)) \rightarrow (\mathsf{f}(z) = \mathsf{g}(z)),$$

which, by Frege's rules for 'free' variables, is the same as

$$(\hat{x}\mathsf{f}(x) = \hat{y}\mathsf{g}(y)) \rightarrow (\forall z)(\mathsf{f}(z) = \mathsf{g}(z)).$$

4. THE CRUCIAL PASSAGES IN SECTION 10

For the reader's convenience, we provide translations of the two passages in Section 10 containing the permutation argument and the identifiability thesis.[8] Note that the first passage, designated 'A' here, contains a general version of the argument, which is then applied in the second passage ('B') to arrive at the identifiability thesis.

[A] Let us suppose that $X(\xi)$ is a function that never receives the same value for distinct arguments; then the same criterion of recognition holds for objects whose names have the form '$X(\hat{x}\Phi(x))$' as for those objects whose signs have the form '$\hat{x}\Phi(x)$'. For in that case, '$X(\hat{x}\Phi(x)) = X(\hat{y}\Psi(y))$' then also means the same as '$(\forall z)(\Phi(z) = \Psi(z))$' [footnote: This is not to say that the sense is the same.]. Hence, the denotation of a name like '$\hat{x}\Phi(x)$' is by no means completely determined by equating the denotation of '$\hat{x}\Phi(x) = \hat{y}\Psi(y)$' with that of '$(\forall z)(\Phi(z) = \Psi(z))$', at least if there is such a function $X(\xi)$ whose value for a value-range as argument is not always equal to that value-range. (Frege 1893, 16)

[B] It is possible to stipulate generally that '$\tilde{a}\Phi(a) = \tilde{b}\Psi(b)$' should mean the same as '$(\forall z)(\Phi(z) = \Psi(z))$', without it thereby being possible to infer the identity of $\hat{x}\Phi(x)$ and $\tilde{a}\Phi(a)$. We would then have, say, a class of objects which would have names of the form '$\tilde{a}\Phi(a)$' and for whose differentiation and recognition the same criterion would hold, as for the value-ranges. We could now determine the function $X(\xi)$ by saying that its value is to be the True for $\tilde{a}\Lambda(a)$ as argument and to be $\tilde{a}\Lambda(a)$ for the True as argument; further, that the value of the function $X(\xi)$ is to be the False for the argument $\tilde{b}M(b)$ and to be $\tilde{b}M(b)$ for the False as argument; that for every other argument the value of the function $\Phi(\xi)$ [sic; read: $X(\xi)$] is to coincide with the argument itself. If now the functions $\Lambda(\xi)$ and $M(\xi)$ do not always have the same value for the same argument, then our function $X(\xi)$ never has the same value for distinct arguments, and therefore, '$X(\tilde{a}\Phi(a)) = X(\tilde{b}\Psi(b))$' always means the same as '$(\forall z)(\Phi(z) = \Psi(z))$', too. The objects whose names would have the form '$X(\tilde{a}\Phi(a))$' would therefore then be recognized by the same means as the value-ranges, and $X(\tilde{a}\Lambda(a))$ would be the True and $X(\tilde{b}M(b))$ the False. Thus it is always possible, without contradicting the identification of '$\hat{x}\Phi(x) = \hat{x}\Psi(x)$' with '$(\forall z)(\Phi(z) = \Psi(z))$', to stipulate that an arbitrary value-range be the True and an arbitrary other one the False! (Frege 1893, 17)

5. THE TWO INTERPRETATIONS

There are, it seems to us, at least two initially reasonable ways of interpreting Frege's procedure in Section 10. The first reading, which we shall call the *metalogical* one, seems to be implicit in Thiel (1976, p. 288), Moore and Rein (1986, p. 375), Dummett (1991, p. 213), and Heck (1999, pp. 267–270), among others, but was first explicitly formulated and systematically investigated by Schroeder-Heister (1984, 1987). The following is a rough summary of that interpretation.

METALOGICAL READING. Prior to Section 10, Frege has introduced a fragment of a formal language, the *begriffsschrift*, whose primitive symbols combine to complex terms in various ways. Frege has in mind, if only implicitly, a certain notion of interpretation for this fragment of *begriffsschrift*. Such an interpretation $\mathfrak{I}$, with respect to a domain of objects U and two designated elements $\top$ (the True) and $\bot$ (the False) of U,[9] is, more or less, just a function assigning an object of U to each value-range name '$\hat{x}\Phi(x)$' formed from a *begriffsschrift* term $\Phi(x)$, where this assignment is subject to the Initial Stipulation, that is, the condition that '$\hat{x}\Phi(x)$' receive the same value as '$\hat{y}\Psi(y)$' if and only if the functions determined by $\Phi(x)$ and $\Psi(x)$ are extensionally the same. The interpretations of the function-names '$-\xi$', '$\neg\xi$', '$\xi = \zeta$', and of the quantifier are then determined by Frege's explanations for these symbols. The point of passage A is to show that, if we have one such interpretation $\mathfrak{I}$ (with respect to U, $\top$, and $\bot$), then any 1–1 function X from U into U gives rise to another such interpretation $X \circ \mathfrak{I}$ (again over U, $\top$, $\bot$) that also respects the Initial Stipulation. In passage B, Frege then applies this general argument to a specific permutation X: Given an interpretation $\mathfrak{I}$ and two value-range names '$\hat{x}\Lambda(x)$' and '$\hat{y}M(y)$' that are assigned distinct objects by $\mathfrak{I}$, we may define X by requiring that it interchange the True with the value under $\mathfrak{I}$ of '$\hat{x}\Lambda(x)$' and the False with the value under $\mathfrak{I}$ of '$\hat{y}M(y)$', and map every other object in U to itself. The interpretation $X \circ \mathfrak{I}$ will then not only, as argued in passage A, satisfy the Initial Stipulation, but will also assign the True to the value-range name '$\hat{x}\Lambda(x)$' and the False to the value-range name '$\hat{y}M(y)$'. Whence the identifiability thesis, as the metalogical interpretation reads it: Without contradicting the Initial stipulation, we may define any two value-range names, that are assigned distinct objects under some interpretation, to be names of the True and the False, respectively (in another interpretation).[10]

There is another reading of passages A and B, one that does not invoke notions like *interpretation* or *assignment of values to value-range names*. We shall call it the *mathematical* interpretation, as the permutation argument and the identifiability thesis may, according to it, be stated entirely in informal mathematical terms. As far as we are aware, this reading was first formulated explicitly and coherently by Ricketts (1997). In a way, it is implicit already in Moore's and Rein's work (1986) and especially (1987); however, the fact that they do not discriminate between their description of the permutation argument in (1986, p. 375), which is precisely the one used by Schroeder-Heister (1984, 1987), and the one in (1986, p. 382), and (1987), makes it hard to understand exactly where they see the differences. In any case, here is a rough description of the mathematical interpretation of Section 10.

MATHEMATICAL READING. Prior to Section 10, Frege has introduced a number of first- and second-order functions, such as the first-order function mapping the True to the True, and everything else to the False, which he writes as $-\xi$, and the second-order value-range function, written as $\hat{x}\varphi(x)$. This latter function has the property, according to Frege, that it maps functions $\Phi(\xi)$ and $\Psi(\xi)$ to the same value if and only if the functions $\Phi(\xi)$ and $\Psi(\xi)$ always return the same values for the same arguments. That is, expressed in modern terms: The value-range function maps the first-order functions, extensionally construed, 1–1 into the objects. Let us call this higher-order condition the *injectivity constraint*; so the Initial Stipulation, under the mathematical reading, is just the injectivity constraint.[11] The point of passage A then is to argue that, if $X(\xi)$ is a function mapping the domain of objects 1–1 into itself, then the composition of $X(\xi)$ with the value-range function also maps the first-order functions 1–1 into the objects, that is, if $\varphi(\xi) \mapsto \hat{x}\varphi(x)$ satisfies the injectivity constraint, then so does $\varphi(\xi) \mapsto X(\hat{x}\varphi(x))$. In passage B, Frege applies this general argument to a specific permutation $X(\xi)$, namely the one interchanging the value-range of some function $\Lambda(\xi)$ with the True, and the value-range of some function $M(\xi)$, assumed to be extensionally distinct from $\Lambda(\xi)$, with the False, and mapping every other object to itself. As $X(\xi)$ is clearly 1–1, by passage A the function $\varphi(\xi) \mapsto X(\hat{x}\varphi(x))$ satisfies the injectivity constraint, and it maps $\Lambda(\xi)$ to the True, and $M(\xi)$ to the False. Whence the identifiability thesis: Assuming the existence of a second-order function abiding by the injectivity constraint, it follows that there are (other) second-order functions satisfying the injectivity constraint and mapping some arbitrary first-order function to the True and some other, extensionally distinct function, to the False.

It may seem, at first sight, that the metalogical and the mathematical interpretation of Section 10 amount to the same thing, that they differ only in the way they are presented. Nothing could be further from the truth. While the argument as read by the mathematical interpretation is obviously correct, and in fact trivial, as it is just a straightforward application of the elementary fact that the composition of 1–1 functions is again 1–1, the metalogical version of the argument is actually invalid, and the metalogical version of the identifiability thesis false, as was demonstrated by Schroeder-Heister (1984, 1987).

6. SCHROEDER-HEISTER'S MODEL-THEORETIC RECONSTRUCTION

In this section, we shall make the metalogical interpretation more precise by providing a detailed account of the syntax and semantics that it sees at work in *Grundgesetze* I, Section 10. We then review how Schroeder-Heister (1984, 1987) arrives at his conclusions about permutation argument and identifiability thesis. Finally, we shall put forward some, as we feel, fatal objections to the metalogical interpretation.

SYNTAX. The syntactic primitives of the Fregean language L_1 are: an infinite stock of individual variables x, y, z, ..., the horizontal stroke $-$, the negation symbol $\neg$, the equality symbol $=$, the conditional symbol $\rightarrow$, the universal quantifier $\forall$, and the value range operator $\hat{}$. The symbols $-$ and $\neg$ behave as unary function symbols, $=$ and $\rightarrow$ are binary function symbols (for which we use infix notation). The class of L_1-terms is built up inductively by letting all individual variables count as terms, allowing application of the function symbols $-$, $\neg$, $=$ and $\rightarrow$ to form new terms from old, and allowing, for any individual variable x and any term t already constructed, the formation of new terms $\forall xt$ and $\hat{x}t$ (terms of this latter form shall be called *value-range terms* or *VR terms* for short). If C is any set, then the terms of $L_1(C)$ are constructed according to the same inductive definition as those of L_1, but allowing, besides the individual variables, also members of C as primitive terms (that is, as individual constants). The operators $\forall$ and $\hat{}$ bind individual variables as usual. We write $t_x[s]$ for the result of substituting the term s for all free occurrences of x in t, where it is understood that s is substitutible for x in t. A closed term is a term with no free occurrences of any variable.

REMARK. The reader will have noticed that we are restricting our attention (for the time being) to the first-order fragment of Frege's language. This is (a) reasonable, as Frege has not introduced second-order

quantification prior to Section 10, and (b) a good idea anyway, as the first-order fragment of the *Grundgesetze* theory is known to be consistent (cf. Parsons 1987).

SEMANTICS. A *structure* for L_1 is a triple

$$\mathcal{U} = (U, \top, \bot),$$

consisting of a domain of objects U (possibly the domain of all objects), and two distinguished elements $\top$ (the True) and $\bot$ (the False) of U. A *pseudo-interpretation* of L_1 over such an L_1-structure $\mathcal{U}$ is a function $\mathcal{I}$ assigning an element of U to each closed VR term of $L_1(U)$. With respect to a pseudo-interpretation $\mathcal{I}$ over $\mathcal{U}$, the closed terms t of $L_1(U)$ are assigned elements $[\![t]\!]_{\mathcal{I}}$ of U as values recursively as follows: For elements c of U, $[\![c]\!]_{\mathcal{I}}$ is just c itself. $[\![-t]\!]_{\mathcal{I}}$ is $\top$ if $[\![t]\!]_{\mathcal{I}}$ is; otherwise $[\![-t]\!]_{\mathcal{I}}$ is $\bot$; in other words, where $\chi_\top$ is the characteristic function of the singleton $\{\top\}$ over U, $[\![-t]\!]_{\mathcal{I}}$ is $\chi_\top([\![t]\!]_{\mathcal{I}})$. $[\![\neg t]\!]_{\mathcal{I}}$ is $\bot$ if $[\![t]\!]_{\mathcal{I}}$ is $\top$; otherwise $[\![\neg t]\!]_{\mathcal{I}}$ is $\top$. $[\![s = t]\!]_{\mathcal{I}}$ is $\top$ if $[\![s]\!]_{\mathcal{I}}$ is the same object as $[\![t]\!]_{\mathcal{I}}$; otherwise $[\![s = t]\!]_{\mathcal{I}}$ is $\bot$. In other words, where $\chi_=$ is the characteristic function of the identity relation over U, $[\![s = t]\!]_{\mathcal{I}}$ is $\chi_=([\![s]\!]_{\mathcal{I}}, [\![t]\!]_{\mathcal{I}})$. $[\![s \to t]\!]_{\mathcal{I}}$ is $\bot$ if $[\![s]\!]_{\mathcal{I}}$ is $\top$ and $[\![t]\!]_{\mathcal{I}}$ is not $\top$; otherwise $[\![s \to t]\!]_{\mathcal{I}}$ is $\top$. $[\![\forall x t]\!]_{\mathcal{I}}$ is $\top$ if for all a in U, $[\![t_x[a]]\!]_{\mathcal{I}}$ is $\top$; otherwise $[\![\forall x t]\!]_{\mathcal{I}}$ is $\bot$. Finally, $[\![\hat{x} t]\!]_{\mathcal{I}}$ is just $\mathcal{I}(\hat{x} t)$. For closed terms t of $L_1(U)$ we say that $\mathcal{I}$ satisfies t, or $\mathcal{I} \models t$, if $[\![t]\!]_{\mathcal{I}} = \top$. For arbitrary terms t of $L_1(U)$, $\mathcal{I} \models t$ is defined to mean that $\mathcal{I}$ satisfies some (and hence every) universal closure of t. It should be obvious that pseudo-interpretations for L_1 exist.

REMARK 1. This is essentially the emendation of Schroeder-Heister's (1984, 1987) semantics proposed by Parsons (1987), except that, where Parsons uses variable assignments, we use parameters from the domain of individuals.

REMARK 2. In pseudo-interpretations, logic generally goes awry. This is because value-range terms are treated as unstructured constants, thereby destroying compositionality. There is, for instance, no guarantee that in a pseudo-interpretation, $\hat{x}(x = x)$ and $\hat{y}(y = y)$ receive the same denotation. In the same vein, $\hat{x}(x = x)$ and $\hat{x}(x \to x)$ may well receive distinct values under a pseudo-interpretation, even though every pseudo-interpretation must validate $\forall x((x = x) = (x \to x))$. This leads to the failure of certain logical principles such as $(\forall x)s \to s_x[t]$. To see this, note that any pseudo-interpretation $\mathcal{I}$ assigns $\top$ to both $\forall x(x = x)$ and $\forall y(y \to y)$. As the VR terms $\hat{z}(z = \forall x(x = x))$ and $\hat{z}(z = \forall y(y \to y))$

are syntactically distinct, there are pseudo-interpretations $\mathcal{I}$ assigning them distinct values. Such pseudo-interpretations invalidate the term

$$(\forall x(x = x) = \forall y(y \to y)) \to \left(\hat{z}(z = \forall x(x = x)) = \hat{z}(z = \forall y(y \to y))\right).$$

Nevertheless, the term $\forall x \forall y \left(x = y \to \hat{z}(z = x) = \hat{z}(z = y)\right)$ denotes $\top$ in all pseudo-interpretations, and so the principle $(\forall x)s \to s_x[t]$ cannot generally hold in pseudo-interpretations.

To eliminate such maverick pseudo-interpretations, we invoke one direction of the Initial Stipulation: Let us say that an *interpretation* over the structure $\mathcal{U} = (U, \top, \bot)$ is a pseudo-interpretation $\mathcal{I}$ over $\mathcal{U}$ that assigns the same value to VR terms $\hat{x}s$ and $\hat{y}t$ if, for all a in U, $[\![s_x[a]]\!]_{\mathcal{I}}$ and $[\![t_y[a]]\!]_{\mathcal{I}}$ coincide. In other words, an interpretation is a pseudo-interpretation $\mathcal{I}$ that satisfies the L_1-schema

(sVa) $\quad \forall z(s = t) \to (\hat{x}s_z[x] = \hat{y}t_z[y]),$

which is just a first-order schematic version of Frege's Basic Law (Va).

Interpretations are logically well-behaved. Indeed, we can give a nice compositional description of the semantics of interpretations: Any $L_1(U)$-term t in at most one free variable defines a function on U, namely, the function mapping each object $a \in U$ to the object $[\![t_x[a]]\!]_{\mathcal{I}}$. Using lambda-notation, we may write this function as $\lambda a.[\![t_x[a]]\!]_{\mathcal{I}}$. In this way, $\mathcal{I}$ gives rise to a space $\mathcal{F}$ of unary functions over U, namely, the space of all functions defined by some term t as above. On $\mathcal{F}$, we may define a function $\mathsf{VR}_{\mathcal{I}}$ by letting $\mathsf{VR}_{\mathcal{I}}(f)$ be $\mathcal{I}(\hat{x}t)$ for some term t in at most the free variable x that defines the function f. As $\mathcal{I}$ is an interpretation, the function $\mathsf{VR}_{\mathcal{I}}$ is well-defined, because by (sVa) the value of $\mathsf{VR}_{\mathcal{I}}$ for the argument f is independent of the choice of a defining term for f. Now $\mathsf{VR}_{\mathcal{I}}$ serves nicely as the semantic value of the VR operator $\hat{}$. Let us give two examples: First, consider the VR term $\hat{x} - x$. The term $-x$ defines the function $\lambda a.[\![-a]\!]_{\mathcal{I}}$, which we also write as $\chi_\top$. So $-x$ signifies $\chi_\top$, and now $\hat{x} - x$ signifies $\mathsf{VR}_{\mathcal{I}}(\chi_\top)$, that is, the syntactic operation of applying the VR operator $\hat{x}$ to the term $-x$ corresponds to the semantic operation of applying the function $\mathsf{VR}_{\mathcal{I}}$ to the semantic value of the term $-x$. Second, consider the VR term $\hat{y}(y = \hat{x} - x)$. By definition of $\mathsf{VR}_{\mathcal{I}}$, we know that $\mathcal{I}(\hat{y}(y = \hat{x} - x))$ is $\mathsf{VR}_{\mathcal{I}}(\lambda b.[\![b = \hat{x} - x]\!]_{\mathcal{I}})$. For any b in U, $[\![b = \hat{x} - x]\!]_{\mathcal{I}}$ is $\chi_=(b, \mathcal{I}(\hat{x}(-x)))$. $\mathcal{I}(\hat{x}(-x))$, we already know, is $\mathsf{VR}_{\mathcal{I}}(\chi_\top)$. Putting things together, we see that $\mathcal{I}(\hat{y}(y = \hat{x} - x))$ is equal to $\mathsf{VR}_{\mathcal{I}}(\lambda b.\chi_=(b, \mathsf{VR}_{\mathcal{I}}(\chi_\top)))$. Again, this shows nicely how the syntactic VR operator corresponds, in an entirely compositional way, to the function $\mathsf{VR}_{\mathcal{I}}$. It must be noted, however, that we have not defined the notion of interpretation independently of the logically

odd notion of pseudo-interpretation. In order to decide whether a pseudo-interpretation is in fact an interpretation, we have to apply the peculiar semantics of pseudo-interpretations first.

We still have not captured the right notion of interpretation for Frege's purposes, as there is no guarantee that the *other* half of the Initial Stipulation holds in arbitrary interpretations. Thus, let us call an interpretation $\mathfrak{I}$ *good* if it assigns distinct objects to VR terms $\hat{x}s$ and $\hat{y}t$ whenever for some a in the domain of objects U, $[\![s_x[a]]\!]_{\mathfrak{I}}$ is distinct from $[\![t_y[a]]\!]_{\mathfrak{I}}$. In other words, an interpretation is good if it satisfies the L_1-schema

(sVb) $(\hat{x}s = \hat{y}t) \rightarrow \forall z(s_x[z] = t_y[z])$,

mimicking, in our first-order setting, Frege's second-order Basic Law (Vb). Putting things together, we see that a good interpretation is a pseudo-interpretation satisfying the schematic version of Basic Law (V), viz.

(sV) $\forall z(s = t) = (\hat{x}s_z[x] = \hat{y}t_z[y])$.

Recall that by Parsons' (1987) result, good interpretations for L_1 exist.

We are now ready to rehearse Schroeder-Heister's refutations of the permutation argument and the identifiability thesis, metalogically construed.

First, for the identifiability thesis. Suppose given a good interpretation $\mathfrak{I}$ over the structure $(U, \top, \bot)$ such that $\mathfrak{I}(\hat{x} - x)$ is not $\top$ (such good interpretations exist, by Parsons (1987)). Then $\mathfrak{I} \models \neg\forall y\,\big((y = \hat{x} - x) = -y\big)$. By (sVb), $\mathfrak{I}(\hat{x} - x)$ is distinct from $\mathfrak{I}(\hat{y}(y = \hat{x} - x))$. If the identifiability thesis were true, there would have to be a good interpretation $\mathfrak{I}'$ over some structure $(U', \top', \bot')$ such that $\mathfrak{I}'(\hat{x} - x)$ would be $\top'$ and $\mathfrak{I}'(\hat{y}(y = \hat{x} - x))$ would be $\bot'$. But if $\mathfrak{I}'(\hat{x} - x) = \top'$, then $\mathfrak{I}' \models \forall y\,\big((y = \hat{x} - x) = -y\big)$. By (sVa) in $\mathfrak{I}'$, we would then have $\mathfrak{I}'(\hat{x} - x) = \mathfrak{I}'(\hat{y}(y = \hat{x} - x))$, and so $\mathfrak{I}'(\hat{y}(y = \hat{x} - x))$ cannot be $\bot'$ (the argument shows, in effect, that there is not even an interpretation, and hence *a fortiori* no good interpretation, with the desired property).

The identifiability thesis (metalogically read) being false, it is clear that the permutation argument (metalogically read) must be invalid. Schroeder-Heister demonstrates this directly as follows: Let $\mathfrak{I}$ be a good interpretation over $\mathfrak{U} = (U, \top, \bot)$, let t be any closed VR term of L_1 and suppose that $\mathfrak{I}(t)$ is distinct from $\top$. $\mathfrak{I}$ then satisfies the closed term $\forall x((-t = x) = (\bot = x))$.[12] As $\mathfrak{I}$ is an interpretation, $\mathfrak{I}(\hat{x}(-t = x))$ and $\mathfrak{I}(\hat{x}(\bot = x))$ are the same. Now let X be any function from U into U (we need not even assume that X is 1–1) mapping $\mathfrak{I}(t)$ to $\top$. Then the pseudo-interpretation $X \circ \mathfrak{I}$ over $\mathfrak{U}$ is *not* a good interpretation: As X is a function, we still have

$X \circ \mathfrak{I}(\hat{x}(-t = x)) = X \circ \mathfrak{I}(\hat{x}(\bot = x))$. So if (sVb) were valid in $X \circ \mathfrak{I}$, we should have $X \circ \mathfrak{I} \models \forall x((-t = x) = (\bot = x))$. But this cannot be the case, as $X \circ \mathfrak{I}(t) = \top$. Hence $X \circ \mathfrak{I}$ is not a *good* interpretation.

This is what Schroeder-Heister (1987) proves about the operation $\mathfrak{I} \mapsto X \circ \mathfrak{I}$. But in fact, things are even worse: In general, $X \circ \mathfrak{I}$ is not even an interpretation! Assume, in addition to all of the above, that X is actually 1–1. As X maps an object that is not $\top$ to $\top$, we know that $X(\top) \neq \top$. Suppose that for some closed VR term s, $\mathfrak{I}(s)$ is $\top$. Then $\mathfrak{I} \models \neg\forall x\,((x = -s) = (x = \bot))$, so by (sVb) in $\mathfrak{I}$, $\mathfrak{I}(\hat{x}(x = -s)) \neq \mathfrak{I}(\hat{x}(x = \bot))$. As X is 1–1, we still have $X \circ \mathfrak{I}(\hat{x}(x = -s)) \neq X \circ \mathfrak{I}(\hat{x}(x = \bot))$. However, as $[\![-s]\!]_{X\circ\mathfrak{I}}$ is not $\top$, $X \circ \mathfrak{I}$ validates $\forall x\,((x = -s) = (x = \bot))$, contradicting (sVa). Thus, not even the class of interpretations – much less the class of *good* interpretations – is closed under the operation $\mathfrak{I} \mapsto X \circ \mathfrak{I}$.

Why is it that things go so terribly wrong? Suppose $\mathfrak{I}$ is an interpretation in $\mathcal{U}$, and let s be any closed VR term of $L_1(U)$. Suppose $\mathfrak{I}(s)$ is the element a of U, and that $b := X(a) \neq a$. Then clearly, the value of s under $X \circ \mathfrak{I}$ is b. So in the context of, say, a term of the form $y = s$, the pseudo-interpretation $X \circ \mathfrak{I}$ interprets s as b. The trouble is that, when s occurs embedded under a VR operator, $X \circ \mathfrak{I}$ interprets s not as b, as it should, but rather assigns it its *old* value under $\mathfrak{I}$: $X \circ \mathfrak{I}(\hat{y}(y = s))$ is of course $X(\mathfrak{I}(\hat{y}(y = s)))$; now since $\mathfrak{I}$ is an interpretation and $\mathfrak{I}(s) = a$, we have $\mathfrak{I}(\hat{y}(y = s)) = \mathfrak{I}(\hat{y}(y = a))$, and hence $X \circ \mathfrak{I}(\hat{y}(y = s))$ becomes equal to $X \circ \mathfrak{I}(\hat{y}(y = a))$, and not to $X \circ \mathfrak{I}(\hat{y}(y = b))$. We thus have $X \circ \mathfrak{I} \models \neg\big(b = s \rightarrow (\hat{y}(y = b) = \hat{y}(y = s))\big)$. It turns out that in all but trivial cases (where X is the identity function on the $\mathfrak{I}$-values of VR terms), $X \circ \mathfrak{I}$ is not an interpretation – the permutation argument, metalogically construed, *never* works! This is because the definition of $X \circ \mathfrak{I}$ is such that it can achieve a re-interpretation of *outermost* occurrences of the VR operator only, so VR terms receive different values according as they occur within the scope of another VR operator or not.

Reflecting about this situation a bit, it seems that we could have known there to be a problem right from the beginning. Consider again passage A. Frege notes that, due to the injectivity of $X(\xi)$, $X(\hat{x}\Phi(x))$ equals $X(\hat{x}\Psi(x))$ if and only if $\Phi(\xi)$ and $\Psi(\xi)$ always return the same value for the same argument. This much is true. But supposing that Φ and Ψ are schematic variables ranging over *arbitrary* terms of *begriffsschrift*, what does that consideration show? Certainly not what the metalogical interpretation takes Frege to be claiming, and quite obviously so. Recall our description of the semantics of (good) interpretations $\mathfrak{I}$ by means of the function $\mathsf{VR}_{\mathfrak{I}}$. What one would want to show is that, if we move from a good interpretation based on $\mathsf{VR}_{\mathfrak{I}}$ to a pseudo-interpretation based on

$X \circ \mathsf{VR}_{\mathit{1}}$, then that pseudo-interpretation is again a good interpretation. Now if the terms $\Phi(\xi)$ and $\Psi(\xi)$ contain subterms of the form $\hat{y}t(y)$, then, under the pseudo-interpretation based on $X \circ \mathsf{VR}_{\mathit{1}}$, these inner occurrences of the VR operator should also be interpreted by $X \circ \mathsf{VR}_{\mathit{1}}$. In short, the correct claim to be argued would be that $X(\hat{x}\Phi^X(x))$ equals $X(\hat{x}\Psi^X(x))$ if and only if $\Phi^X(\xi)$ and $\Psi^X(\xi)$ always return the same value for the same argument, where $\Phi^X(\xi)$ results from $\Phi(\xi)$ by replacing every occurrence of a term $\hat{y}t(y)$ within Φ by an occurrence of $X(\hat{y}t(y))$, and similarly for Ψ. It seems to us that Frege, who repeatedly calls the VR operator a second-level function, and so almost certainly thought of the semantics of that operator in terms of our VR_{1} function, would not have made the kind of mistake that the metalogical interpretation must here attribute to him. This, then, is our first objection to the metalogical interpretation: It makes Frege look like a rather careless mathematician.

But the story is not over at this point. For unless Frege was extremely confused while writing Section 10, he must have been aware of the phenomena Schroeder-Heister (1987) uses to refute the identifiability thesis: Consider the extended footnote in Section 10 of *Grundgesetze*. Frege there discusses the proposal that every object be identified with the value-range of a concept under which it alone falls, i.e., that $\hat{x}(\Delta = x) = \Delta$ hold generally (and not just, as Frege has stipulated in Section 10, for truth values Δ). In arguing against this proposal, Frege asks the reader to consider the case where Δ is a value-range $\hat{x}\Phi(x)$, and points out that the proposed stipulation in this case amounts to requiring $\hat{y}(\hat{x}\Phi(x) = y) = \hat{x}\Phi(x)$, which, by Basic Law V, is equivalent to $\forall y\left((\hat{x}\Phi(x) = y) = \Phi(y)\right)$, expressing the circumstance that $\hat{x}\Phi(x)$ is the unique object falling under $\Phi(\xi)$. It seems hard to believe that Frege could, at this point, have overlooked the fact that the simplest possible instance for $\Phi(\xi)$, viz., $-\xi$, yields the equivalence of $\hat{y}((\hat{x} - x) = y) = \hat{x} - x$ with $-\hat{x} - x$. That is, given that Frege obviously thought about the matter in some detail, he could hardly have failed to notice that '$\hat{x} - x$' cannot be made to denote the True without thereby also making '$\hat{y}((\hat{x} - x) = y)$' a name of the True – and these are precisely Schroeder-Heister's counterexamples to the metalogical identifiability thesis.

7. A MODEL-THEORETIC RECONSTRUCTION OF THE MATHEMATICAL READING

We have seen that the assumption that Frege takes the letters 'Φ' and 'Ψ' in passage A to be *schematic* variables ranging over arbitrary terms of *begriffsschrift* is implausible. The mathematical reading of Section 10 sug-

gests that we should read them rather as function parameters, and both the permutation argument and the identifiability thesis turn out true when seen in this light. We mentioned earlier that there is no *need* for the adherent of the mathematical interpretation to present Frege's system as consisting of an uninterpreted formal language and a general notion of interpretation for this language. Indeed, one may simply say that '$-\xi$', for instance, denotes the function $-\xi$, which is determined by the stipulation that it map the True (which is assumed to be a fixed object) to the True and everything else to the False (which is likewise assumed to be a specific fixed object), or that the sign '$\hat{x}\varphi(x)$' denotes the function $\hat{x}\varphi(x)$, mapping any first-level function $\mathsf{F}(\xi)$ to its value-range $\hat{x}\mathsf{F}(x)$ (which is more or less what Frege says in Section 9). To facilitate comparison with the metalogical reconstruction discussed in the preceding section, we may nevertheless recast Frege's discussion in the formal framework of a term logic, without unduly distorting the mathematical interpretation. As we are assuming Frege to be using function quantification, such a term logic clearly has to be (at least) second-order; for the purposes of comparison, we shall therefore extend, in the obvious way, the language L_1 of the preceding section to the second-order language L_2.

L_2 thus has, as additional syntactic primitives, unary function variables $\mathsf{f}, \mathsf{g}, \ldots$, as additional term-forming operations the application of a function variable to a term, yielding $\mathsf{f}(t)$ from term t, and the application of the second-order quantifier to a term, yielding $\forall \mathsf{f} t$ from t. A *structure* for L_2 is a quadruple $\mathcal{U} = (U, \mathcal{S}, \top, \bot)$, where $(U, \top, \bot)$ is an L_1-structure, and $\mathcal{S} \subseteq U^U$ is the range of the second-order quantifier. Again, we may extend L_2 by a set $\mathcal{F}$ of function constants and a set $\mathcal{C}$ of individual constants to obtain the language $L_2(\mathcal{F}, \mathcal{C})$. In practice, $\mathcal{F}$ will always be $\mathcal{S}$ and $\mathcal{C}$ will be U, for some L_2-structure $\mathcal{U} = (U, \mathcal{S}, \top, \bot)$. A *pseudo-interpretation* $\mathcal{I}$ in $\mathcal{U}$ is now a mapping from the set of closed VR terms of $L_2(\mathcal{S}, U)$ to U; it should be obvious how terms are to be evaluated with respect to such a pseudo-interpretation $\mathcal{I}$, that is, how the map $t \mapsto [\![t]\!]_{\mathcal{I}}$ is defined. A pseudo-interpretation is an *interpretation* if it satisfies (sVa), now viewed as an L_2-schema, and it is a *good interpretation* if it moreover satisfies (sVb), again taken as an L_2-schema.

The language at work in the mathematical interpretation will be called L_{w}, as it was first investigated (in a predicate logic version) by Wehmeier (1999).[13] It has the same syntactic primitives as L_2; however, the VR operator is construed as a second-order function constant. The terms of L_{w} are hence built up from individual variables by applying the function constants $-$, $\neg$, $=$, and $\rightarrow$, and first-order and second-order quantification to terms already constructed, by applying any function variable f to any term t,

and, finally, by applying the function constant $\hat{}$ to any function variable f, yielding terms of the form $\hat{\mathsf{f}}$. So the VR operator is, in L_{w}, only applied to syntactically primitive items, viz., the function variables, whereas in L_2, it can be applied to arbitrary terms.

The notion of structure is the same for L_2 and L_{w}. Given such a structure $\mathcal{U} = (U, \mathcal{S}, \top, \bot)$, we again extend L_{w} to $L_{\mathrm{w}}(\mathcal{S}, U)$ by using the elements F of $\mathcal{S}$ as function constants to be interpreted by themselves, and the elements a of U as individual constants, also to be interpreted by themselves. An interpretation now simply consists in assigning a function VR, mapping the elements of $\mathcal{S}$ into U, to the VR operator $\hat{}$, and so we will speak of VR as *being* an interpretation. We assume that it is obvious to the reader how an $L_{\mathrm{w}}(\mathcal{S}, U)$-term t receives its value $[\![t]\!]_{\mathsf{VR}}$, when VR interprets the VR operator.

It will not have gone unnoticed that, in the case of L_{w}, we have not defined a notion of pseudo-interpretation prior to the introduction of interpretations. This is obviously unnecessary, as the semantics of L_{w} is entirely compositional anyway. In other words, Frege's Basic Law (Va), in its second-order form, is a *logical* truth of L_{w}.

Prima facie, one might object to L_{w} that it cannot be an adequate language for reconstructing Frege's system because, unlike L_1 and L_2, it does not allow for the formation of complex VR terms, of which the *Grundgesetze* abound. Such an objection has little force, however, as we may, on the basis of certain other Fregean principles, introduce complex VR terms into L_{w} by means of definitional extension. In fact, we will show below that any L_{w} structure satisfying full second-order (function) comprehension can be expanded uniquely to an L_2-interpretation.

Digression on Comprehension. Frege's Basic Law (IIb) (cf. Section 25), together with his instantiation rule for free second-order variables (rule 9 in Section 48), and some propositional logic, immediately yields the following comprehension principle:

$$\text{(Term-CA)} \quad \exists \mathsf{f} \forall x\, (\mathsf{f}(x) = t),$$

where t is any term of *begriffsschrift* not containing a free occurrence of the function variable f. That is, there is a function in the second-order domain that is defined, with respect to the free variable x, by the term t. But more can be shown to hold. Due to the presence of a description operator, there even is a function corresponding to any term which describes a function graph:

$$\text{(CA)} \quad (\forall x \exists! z t) \rightarrow \exists \mathsf{f} \forall x \forall z\, ((\mathsf{f}(x) = z) = t)\,,$$

where t is again an arbitrary term, f does not occur free in t, and $\exists!zt$ abbreviates $\exists u\forall w(t_z[w] = (u = w))$. These instances of (CA) follow from (Term-CA), some elementary logic, and Basic Law (VI), which governs Frege's description operator. It should be noted that neither (Term-CA) nor (CA) depend on Basic Law (Vb) – the only assumption needed about the VR operator occurring in Basic Law (VI) is that it abides by Basic Law (Va) (although, of course, the validity of Basic Law (VI) does further constrain the value range function, in that extensionally distinct singleton concepts must receive distinct value ranges). It should further be noted that (Term-CA) follows from (CA) by elementary logical means if, in (CA), we take t to be $t = z$ (assuming that z is not free in t). Given that Frege was committed to these comprehension principles, we will, at least for the time being, appeal to them as well. Indeed, (CA) is the standard formulation of function comprehension today (see e.g., Enderton 2001, p. 284). □

Here, then, is the promised expansion theorem for L_w:

THEOREM 1. Let VR be an interpretation of L_w in $\mathcal{U} = (U, \mathcal{S}, \top, \bot)$, i.e., let VR be a function with domain $\mathcal{S} \subseteq U^U$ and values in U. Suppose that the interpretation VR satisfies (CA) as an L_w-schema. Then there is a unique L_2-interpretation $\mathcal{I}_\mathsf{VR}$ in $\mathcal{U}$ such that

1. for each element F of $\mathcal{S}$ and each individual variable x, $\mathcal{I}_\mathsf{VR}(\hat{x}\mathsf{F}(x)) = \mathsf{VR}(\mathsf{F})$, and
2. $\mathcal{I}_\mathsf{VR}$ satisfies (CA) as an L_2-schema.

Proof. Define the rank $\mathsf{rk}(t)$ of an $L_2(\mathcal{S}, U)$-term t recursively as follows. Individual variables and elements of U are terms of rank 0. The terms $-t$, $\neg t$, $\forall xt$, $\forall\mathsf{f}t$, and $\phi(t)$, where ϕ is either a function variable or an element of $\mathcal{S}$, all have the same rank as t itself. The rank of $s = t$ and of $s \to t$ is the maximum of the ranks of s and t, and finally, the rank of a VR term $\hat{x}t$ is $1 + \mathsf{rk}(t)$. Thus, the rank of a term t is the maximal depth of nestings of the VR operator in t.

We first prove that there is *at most one* L_2-interpretation $\mathcal{I}_\mathsf{VR}$ satisfying conditions 1 and 2 of the theorem. By condition 2 there must be, for any term t, a function F in $\mathcal{S}$ such that $\mathcal{I}_\mathsf{VR} \models \forall x(\mathsf{F}(x) = t)$; so $\lambda a.[\![t_x[a]]\!]_{\mathcal{I}_\mathsf{VR}} = \mathsf{F} \in \mathcal{S}$. As $\mathcal{I}_\mathsf{VR}$, being an interpretation, satisfies (sVa), we have $\mathcal{I}_\mathsf{VR}(\hat{x}\mathsf{F}(x)) = \mathcal{I}_\mathsf{VR}(\hat{x}t)$. By condition 1, we further know that $\mathcal{I}_\mathsf{VR}(\hat{x}\mathsf{F}(x)) = \mathsf{VR}(\mathsf{F})$. Putting things together, we see that for any t, $\mathcal{I}_\mathsf{VR}(\hat{x}t)$ must be equal to $\mathsf{VR}(\lambda a.[\![t_x[a]]\!]_{\mathcal{I}_\mathsf{VR}})$. Now if t is of rank 0, then $[\![t_x[a]]\!]_{\mathcal{I}_\mathsf{VR}}$ is determined by the structure $\mathcal{U}$ alone, and so VR uniquely determines $\mathcal{I}_\mathsf{VR}$ on VR terms of rank 1. Now suppose that $\mathcal{I}_\mathsf{VR}$ is uniquely determined on VR terms of rank at most n. If t is any term of rank n,

then $[\![t_x[a]]\!]_{\mathcal{I}_{\mathsf{VR}}}$ is thereby always uniquely determined; but $\mathcal{I}_{\mathsf{VR}}(\hat{x}t)$ must be $\mathsf{VR}(\lambda a.[\![t_x[a]]\!]_{\mathcal{I}_{\mathsf{VR}}})$. So $\mathcal{I}_{\mathsf{VR}}$ is uniquely determined on VR terms of rank $n+1$; by induction, it follows that there is at most one interpretation $\mathcal{I}_{\mathsf{VR}}$ as required in the theorem.

Now for *existence*. We define a sequence of mappings $\mathcal{I}_n$ ($n \geq 1$) such that $\mathcal{I}_n$ assigns values to all closed VR terms of $L_2(\mathcal{S}, U)$ of rank at most n, and $\mathcal{I}_n$ extends $\mathcal{I}_m$ if $n > m$. $\mathcal{I}_{\mathsf{VR}}$ will be the union of all the $\mathcal{I}_n$. $\mathcal{I}_1$ is defined in such a way that $\mathcal{I}_{\mathsf{VR}}$ satisfies condition 1. Every $\mathcal{I}_n$ is defined so as to make true all $L_2(\mathcal{S}, U)$-instances of (sVa) and (CA) of ranks at most n, so that $\mathcal{I}_{\mathsf{VR}}$ is indeed an interpretation satisfying condition 2. Before we start, let us note that the L_2-terms of rank 0 are precisely the terms of L_{w} in which the VR operator does not occur. The values of such terms are determined solely by the structure $\mathcal{U}$. Therefore, all $L_2(\mathcal{S}, U)$-instances of (CA) of rank 0 hold under any pseudo-interpretation in $\mathcal{U}$, because they are also $L_{\mathrm{w}}(\mathcal{S}, U)$-instances of (CA).

INDUCTION BASIS. Let t be an $L_2(\mathcal{S}, U)$-term of rank 0. By the comprehension axioms of rank 0, we know that $\lambda a.[\![t_x[a]]\!]_{\mathsf{VR}}$ is an element of $\mathcal{S}$, and so we may put $\mathcal{I}_1(\hat{x}t)$ equal to $\mathsf{VR}(\lambda a.[\![t_x[a]]\!]_{\mathsf{VR}})$. This obviously takes care of condition 1 and of all $L_2(\mathcal{S}, U)$-instances of (sVa) of rank 1. Comprehension is slightly more tricky. We will need the following translation procedure, both here and in the induction step. For any $L_2(\mathcal{S}, U)$-term t and every individual variable y not occurring in t, we define an $L_2(\mathcal{S}, U)$-term $I(t, y)$ recursively as follows. When σ is an individual variable or individual constant, we let $I(\sigma, y)$ be $\sigma = y$. $I(\phi(t), y)$ is $\exists z\,(I(t, z) \wedge \phi(z) = y)$ for function variables or function constants ϕ (where $\wedge$ is defined in the obvious way and z is a fresh individual variable). $I(-t, y)$ is $\exists z\,((z = \top) \wedge (I(t, z) = y))$, and $I(\neg t, y)$ is $\exists z\,((z = \top) \wedge ((\neg I(t, z)) = y))$. For $\circ$ being either equality or implication, $I(s \circ t, y)$ is $\exists z \exists w\,(I(s, z) \wedge I(t, w) \wedge ((z \circ w) = y))$. We set $I(\forall \rho t, y)$ equal to $\exists z\,((z = \top) \wedge ((\forall \rho I(t, z)) = y))$, where ρ is either an individual variable or a function variable. Most importantly, we define $I(\hat{x}t, y)$ to be the term $\exists \mathsf{f}\left(\forall x(\mathsf{f}(x) = t) \wedge (\hat{x}\mathsf{f}(x) = y)\right)$. For the remainder of the induction basis, let t be of rank 1. Since $\mathcal{I}_1$ validates the comprehension axioms of rank 0 and all instances of (sVa) of rank 1, it follows from our definition that $(t = y)$ and $I(t, y)$ receive the same value under $\mathcal{I}_1$. We also note that $I(t, y)$ contains VR terms of the form $\hat{x}\mathsf{f}(x)$ only. Now suppose that (a pseudo-interpretation extending) $\mathcal{I}_1$ satisfies $\forall x \exists! z t$, i.e., $\forall x \exists u \forall w\,(t_z[w] = (u = w))$. It follows, first, that $\forall x \forall z(t = -t)$; more-

over, we have $\forall x \exists! z \exists v((v = \top) \wedge (t = v))$. By our considerations above, this is $\mathfrak{I}_1$-equivalent to

$$(*) \qquad \forall x \exists! z \exists v((v = \top) \wedge I(t, v)).$$

Except for the occurrence of VR terms $\hat{x}\mathsf{f}(x)$ in $I(t, v)$, $(*)$ is the antecedent of some comprehension axiom of L_{w}. But of course $\mathfrak{I}_1$ evaluates $\hat{x}\mathsf{f}(x)$ just as the interpretation VR evaluates $\hat{\mathsf{f}}$, so modulo the identification of $\hat{x}\mathsf{f}(x)$ and $\hat{\mathsf{f}}$, $(*)$ holds in the L_{w}-interpretation VR. Therefore, the consequent of the pertinent comprehension axiom also holds under VR, from which it follows again that $\exists \mathsf{f} \forall x \forall z\, ((\mathsf{f}(x) = z) = \exists v\, ((v = \top) \wedge I(t, v)))$ holds under $\mathfrak{I}_1$. Now $I(t, v)$ is $\mathfrak{I}_1$-equivalent to $t = v$, so $\exists v\, ((v = \top) \wedge I(t, v))$ is $\mathfrak{I}_1$-equivalent to $-t$, which, as we already know, is the same as t. We conclude that $\mathfrak{I}_1$ guarantees the truth of all rank 1 L_2-instances of (CA).

INDUCTION STEP. Suppose $\mathfrak{I}_n$ has been defined so as to make true all $L_2(\mathfrak{S}, U)$-instances of (sVa) and (CA) of ranks at most n. Let t be a term of rank n. By the appropriate rank n instance of (CA), we know that $\lambda a.[\![t_x[a]]\!]_{\mathfrak{I}_n}$ is in $\mathfrak{S}$, and so we may define $\mathfrak{I}_{n+1}(\hat{x}t)$ as $\mathsf{VR}(\lambda a.[\![t_x[a]]\!]_{\mathfrak{I}_n})$. Again, this immediately takes care of the instances of (sVa) of rank $n + 1$. It remains to verify the comprehension axioms of rank $n + 1$. We note that, in $\mathfrak{I}_{n+1}$, $I(t, y)$ is equivalent to $t = y$ for terms t of rank at most $n + 1$, due to the instances of (CA) and (sVa) of rank n. Furthermore, $I(t, y)$ is of rank n if $\mathsf{rk}(t) = n + 1$. As in the induction basis, $\forall x \exists! z t$ implies $t = -t$ and $\forall x \exists! z \exists v(v = \top \wedge I(t, v))$. Also, $(\mathsf{f}(x) = z) = t$ is equivalent to $(\mathsf{f}(x) = z) = \exists v(v = \top \wedge I(t, v))$, because $t = -t$. Hence the comprehension axioms of rank $n + 1$ are $\mathfrak{I}_{n+1}$-equivalent to comprehension axioms of rank n, and so their truth under $\mathfrak{I}_{n+1}$ follows from the inductive hypothesis. This concludes the proof of our theorem.

We observe that an L_2-interpretation $\mathfrak{I}$ over some structure $\mathcal{U}$ satisfying (CA) induces in a natural way an L_{w}-interpretation $\mathsf{VR}_{\mathfrak{I}}$ over $\mathcal{U}$ satisfying (CA) (this time as an L_{w}-schema), whose unique expansion $\mathfrak{I}_{\mathsf{VR}_{\mathfrak{I}}}$ to L_2 is just $\mathfrak{I}$ itself, by setting $\mathsf{VR}_{\mathfrak{I}}(\mathsf{F}) := \mathfrak{I}(\hat{x}\mathsf{F}(x))$. We may therefore identify the class of L_2-interpretations satisfying (CA) with the class of L_{w}-interpretations satisfying (CA). So in the presence of (CA), we have a natural way of describing the L_2-interpretations without having to invoke the deviant semantics of pseudo-interpretations. It might be remarked that the class of L_2-interpretations $\mathfrak{I}_{\mathsf{VR}}$ satisfying (CA) is closed under the operation $\mathfrak{I}_{\mathsf{VR}} \mapsto \mathfrak{I}_{X \circ \mathsf{VR}}$, for any function $X \in \mathfrak{S}$, but, as our earlier results imply, not under Schroeder-Heister's operation $\mathfrak{I}_{\mathsf{VR}} \mapsto X \circ \mathfrak{I}_{\mathsf{VR}}$.

The situation is of course less clear-cut when it comes to good interpretations: Due to Russell's antinomy, there are no good interpretations, whether in terms of L_2 or L_w, that satisfy the full schema (CA). By a straightforward adaptation of Heck's consistency proof in (1996) to Frege's term-logical setting, it can be shown, however, that there are good L_2-interpretations satisfying *predicative* comprehension, that is, the schema

$$\text{(Pred-CA)} \qquad (\forall x \exists ! z t) \rightarrow \exists \mathsf{f} \forall x \forall z \left((\mathsf{f}(x) = z) = t \right),$$

where t *does not contain function quantifiers*. Such a good L_2-interpretation naturally determines a good L_w-interpretation VR satisfying (Pred-CA) as an L_w-schema, and if X is a 1–1 function present in the function universe of the underlying structure, then $X \circ \mathsf{VR}$ is again a good L_w-interpretation satisfying (Pred-CA). This, we suggest, is the most felicitous way of reconstructing Frege's permutation argument in a non-trivial model-theoretic setting.

ACKNOWLEDGEMENTS

Thanks to John MacFarlane, Robert May and Ed Zalta for discussions on the subject of this paper.

NOTES

[1] Given the inconsistency of the *Grundgesetze* theory, one might object the following to our formulation: 'The identifiability thesis is of course, strictly speaking, false, as the stipulations previously introduced already contradict each other. Similarly, as one premise of the permutation argument is the existence of a model (to speak anachronistically) for the previous stipulations, the argument is trivially valid'. It is not entirely clear whether the situation is so clear-cut (cf. footnote 5). One way to phrase (ii) and (iii) more cautiously would thus be the following: It has remained unclear whether, in the framework of a reasonable consistent reconstruction of the original Fregean setting, the permutation argument should turn out valid and the identifiability thesis true. This unclarity is of course due to an unclarity regarding the notion of a reasonable consistent reconstruction, of which more below.

[2] All English translations from *Grundgesetze* are ours. The reader may wish to also consult Furth (1964, p. 36).

[3] In order not to inconvenience our typesetters (that is, in these days of LaTeX, ourselves), we use present-day symbolism instead of Frege's original notation throughout this paper. It should be kept in mind, however, that the modern symbols are to be understood in Frege's sense, so that, e.g., the negation sign is not a sentence connective, but rather a function symbol.

[4] Cf. Furth (1964 p. 42).

[5] Cf. Furth (1964, pp. 43–44).

[6] Cf. Furth (1964, p. 44).

[7] In the presence of the other logical rules, one may, and we will, simplify this to $(\forall z)(\mathsf{f}(z) = \mathsf{g}(z)) \to (\hat{x}\mathsf{f}(x) = \hat{y}\mathsf{g}(y))$.

[8] Again, the reader may wish to compare Furth (1964, pp. 46–48).

[9] The metalogical interpreter may leave it open whether Frege envisaged this notion of interpretation to involve the possibility of varying the domain of objects U, or whether it should always be assumed to be the domain of all objects whatsoever. Similarly, she need not decide whether the two truth values are fixed once and for all, or whether they are allowed to vary with the interpretation. For all purposes of the present paper, it simply does not matter.

[10] Cf. Christian Thiel's (1976, p. 288) reading of Section 10: "The 'identification' of the truth values with value-ranges (...) is nothing more than the explicit stipulation that two specific expressions having the given form of value-range names serve to denote the True and the False, respectively" (our translation). See also Heck (1999, pp. 267–268).

[11] Note that Frege has not, prior to Section 10, introduced any axioms or rules that imply any kind of function comprehension. In fact, function quantification itself has not even made its appearance. Therefore, the assumption Frege makes (or seems to make) on the mathematical interpretation, that there is some injection from the first-order functions into the objects, does not, strictly speaking, contradict anything Frege has stipulated so far. In other words, nothing forces the function space to be too large to be injected into the individual domain at this point.

[12] Instead of the parameter $\bot$, we could equally well use the closed L_1-term $\neg\forall x(x = x)$.

[13] For a discussion of the relation between L_2 and L_{w} in the setting of second-order predicate logic, see the final section of Ferreira and Wehmeier (2002).

REFERENCES

Dummett, Michael: 1991, *Frege – Philosophy of Mathematics*, Harvard University Press, Cambridge, MA.

Enderton, Herbert B.: 2001, *A Mathematical Introduction to Logic*, 2nd edn., Harcourt/Academic Press, Burlington.

Ferreira, Fernando, and Kai F. Wehmeier: 2002, 'On the Consistency of the Δ^1_1-CA Fragment of Frege's *Grundgesetze*', *Journal of Philosophical Logic* **31**, 301–311.

Frege, Gottlob: 1893, *Grundgesetze der Arithmetik: Begriffsschriftlich abgeleitet, I. Band*, Hermann Pohle, Jena (reprinted 1962 and 1998: Georg Olms Verlagsbuchhandlung, Hildesheim).

Frege, Gottlob: 1964, *The Basic Laws of Arithmetic*, edited and translated by Montgomery Furth, University of California Press, Berkeley and Los Angeles.

Heck, Richard G.: 1996, 'The Consistency of Predicative Fragments of Frege's *Grundgesetze der Arithmetik*', *History and Philosophy of Logic* **17**, 209–220.

Heck, Richard G.: 1999, '*Grundgesetze der Arithmetik* I §10', *Philosophia Mathematica* **7**, 258–292.

Moore, Adrian W. and Andrew Rein: 1986, '*Grundgesetze*, Section 10', in L. Haaparanta and J. Hintikka (eds.), *Frege Synthesized*, D. Reidel, Dordrecht, pp. 375–384.

Moore, Adrian W. and Andrew Rein: 1987, 'Frege's Permutation Argument', *Notre Dame Journal of Formal Logic* **28**, 51–54.
Parsons, Terence: 1987, 'On the Consistency of the First-Order Portion of Frege's Logical System', *Notre Dame Journal of Formal Logic* **28**, 161–168.
Ricketts, Thomas: 1997, 'Truth-Values and Courses-of-Value in Frege's *Grundgesetze*', in W. W. Tait (ed.), *Early Analytic Philosophy: Frege, Russell, Wittgenstein*, Open Court, Chicago, pp. 187–211.
Schroeder-Heister, Peter: 1984, 'Frege's Permutationsargument. Zu §10 der *Grundgesetze der Arithmetik*', in G. Wechsung (ed.), *Frege Conference 1984*, Akademie-Verlag, Berlin, pp. 182–188.
Schroeder-Heister, Peter: 1987, 'A Model-Theoretic Reconstruction of Frege's Permutation Argument', *Notre Dame Journal of Formal Logic* **28**, 69–79.
Thiel, Christian: 1976, 'Wahrheitswert und Wertverlauf: Zu Freges Argumentation im §10 der *Grundgesetze der Arithmetik*', in M. Schirn (ed.), *Studien zu Frege I: Logik und Philosophie der Mathematik*, Frommann-Holzboog, Stuttgart-Bad Cannstatt, pp. 287–299.
Wehmeier, Kai F.: 1999, 'Consistent Fragments of *Grundgesetze* and the Existence of Non-Logical Objects', *Synthese* **121**, 309–328.

Kai F. Wehmeier
Department of Logic and Philosophy of Science
University of California, Irvine
3151 Social Science Plaza
Irvine, CA 92697
U.S.A.
E-mail: wehmeier@uci.edu

Peter Schroeder-Heister
Wilhelm-Schickard-Institut
Universität Tübingen
Sand 13
72076 Tübingen
Germany
E-mail: psh@informatik.uni-tuebingen.de