Notre Dame Journal of Formal Logic
Volume 28, Number 1, January 1987

A Model-Theoretic Reconstruction of Frege's Permutation Argument

PETER SCHROEDER-HEISTER*

1 Introduction In Section 10 of [3] (p. 17[1]) Frege claims:

> . . . without contradicting our setting '$\overset{,}{\epsilon}\Phi(\epsilon) = \overset{,}{\epsilon}\Psi(\epsilon)$' equal to '$-\overset{\mathfrak{a}}{\smile}-\Phi(\mathfrak{a}) = \Psi(\mathfrak{a})$' it is always possible to stipulate that an arbitrary course-of-values is to be the True and another the False.

In what follows this assertion will be called the *identifiability thesis* since it states that two arbitrary but different courses-of-values can be identified with the truth-values. Frege considers the identifiability thesis a consequence of his previous argumentation ([3], p. 17, lines 23-36)[2] which, following Dummett ([1], p. 408), will be called the *permutation argument*, because the concept of a one-one mapping from the considered domain of objects onto itself, i.e. a permutation, is essential for it. More precisely, Frege gives a specific permutation which interchanges the True and the False with two objects denoted by names of the form '$\overset{,}{\eta}\Phi(\eta)$'. This paper attempts to show that the permutation argument is correct, but that it is no argument *for* the identifiability thesis, and that the same holds for related arguments using arbitrary transformations of the domain of objects into itself instead of permutations. This contradicts every interpretation of Section 10 of the "Basic Laws" with which I am familiar, even the most careful and detailed presentation by Thiel [6].[3] The validity of the identifiability thesis itself and the conclusions which can be drawn from it[4] are of course quite independent of this result. However, at the end a counterexample will be given which

*This work was partly stimulated by discussions with the participants of a seminar on Frege's "Basic Laws of Arithmetic" held in winter 1981/82 at the University of Konstanz, especially with Franz-Viktor Kuhlmann and Pirmin Stekeler-Weithofer. An earlier version of the manuscript was written during a stay at the Institute for Advanced Studies in the Humanities of the University of Edinburgh which was financially supported by the British Council. The results have been presented in part at the Second Frege Conference, Schwerin, September 10-14, 1984. I should like to thank an anonymous referee for helpful suggestions and Stella Lewis for checking the English.

Received March 19, 1984; revised July 2, 1984

shows that the identifiability thesis does not hold in the generality claimed by Frege.

Frege's argument will be reconstructed within a model-theoretic framework. However, due to the particular features of Frege's system, the following modifications of the usual model-theoretic procedure are necessary:

(1) Truth-values are objects in the sense that they belong to the range of quantifiers and that they can be arguments of the identity function like any other object. Syntactically this means that first-level function names can be combined even with names of truth-values as argument signs; e.g., the sentence '$\overset{\mathfrak{a}}{\smile}(\mathfrak{a} = \mathfrak{a})$' may occur to the right or left of the identity sign '='. In particular, '=' can function as a sentence connective.[5]

(2) Closed abstracts (i.e., expressions of the form 'ἐΦ(ε)' without free variables) are the smallest semantic units; an interpretation will be an assignment of objects to closed abstracts. Thus closed abstracts function as the only nonlogical constants of the language being considered here.[6] This view is chosen because nothing is assumed about courses-of-values: If in the definition of the value of an expression under an interpretation one wanted to take into account the internal structure of abstracts, one would already have to presuppose that closed abstracts are used to denote sets and in particular that the abstraction principle (Frege's Basic Law V) holds. But this is just what is *not* intended. We want to investigate which interpretations are compatible with the abstraction principle and which are not.

2 The language When constructing a model-theoretic semantics, we must first specify the language to be interpreted semantically. We use the means of expression provided by Frege in the first nine sections of the "Basic Laws". That is, we use a language with the horizontal ('—'), the negation-sign ('$\top$') and the identity-sign ('=') as function symbols, the first-order universal quantifier ('$\smile$'), and the abstraction-sign for courses-of-values ('᾿'). In doing so, we restrict ourselves to the first-order subsystem of the system developed in the "Basic Laws" containing course-of-values abstraction but not second-order quantification. This is justified because in Section 10 Frege argues for his identifiability thesis solely on the basis of the system he has thus far developed.[7]

We follow Frege's original notation fairly closely, but in some places we use a terminology that is more modern. Furthermore we distinguish between metalinguistic (i.e., syntactical) variables and signs of the object-language, and use, following Quine, quasi-quotation marks '⌜' and '⌝' to refer to the result of a concatenation of signs which are themselves partly referred to by syntactical variables.

Symbols of the language under consideration are:

(1) *Parameters* 'a', 'b', 'c' (with and without indices)
(Syntactical variables: 'u', with and without indices)

(2) *Bound variables of two sorts*:
Quantifier-variables '$\mathfrak{a}$', '$\mathfrak{b}$', '$\mathfrak{c}$' (with and without indices)
(Syntactical variables: 'x', with and without indices)
Abstractor-variables 'α', 'ϵ' (with and without indices)
(Syntactical variables: 'κ', 'μ')

(3) *Function expressions* ‘—’, ‘┬’, ‘=’
(4) *Universal quantifier* (briefly: quantifier) ‘$\smile$’
(5) *Course-of-values abstractor* (briefly: abstractor) ‘ ’ ’
(6) *Round brackets* ‘(’, ‘)’.

An *expression* is the result of an arbitrary concatenation of symbols. Syntactical variables for expressions are ‘X’, ‘Y’, ‘Z’. A three-place *substitution operation* [·:] is defined as follows: For expressions X, Y, Z, where Y is a symbol, $[X^{Y}_{Z}]$ is the result of replacing all occurrences of Y in X by Z. (If Y does not occur in X, $[X^{Y}_{Z}]$ is identical with X.)

Terms are defined as follows:

(1) Every parameter is a term.
(2) If X and Y are terms, then so are $\ulcorner - X\urcorner$, $\ulcorner \text{┬} X\urcorner$ and $\ulcorner (X = Y)\urcorner$.
(3) If X is a term which contains no occurrence of a parameter u within a part $\ulcorner \overset{x}{\smile} Y\urcorner$ of X, then $\overset{x}{\smile}[X^{u}_{x}]$ is a term.
(4) If X is a term which contains no occurrence of u within a part $\ulcorner \overset{'}{\kappa} Y\urcorner$ of X, then $\ulcorner \overset{'}{\kappa}[X^{u}_{\kappa}]\urcorner$ is a term.

Syntactical variables for terms are ‘A’, ‘B’, ‘C’, ‘D’, ‘E’, with and without indices. When considering expressions $[A^{u}_{x}]$ or $[A^{u}_{\kappa}]$ we shall always assume that A contains no occurrence of u within a subexpression of the form $\ulcorner \overset{x}{\smile} X\urcorner$ or $\ulcorner \overset{'}{\kappa} X\urcorner$, respectively. Outer brackets may be omitted. Note that our way of bracketing is somewhat different from Frege’s.

A *formula* is a term of the form $\ulcorner - A\urcorner$, $\ulcorner \text{┬} A\urcorner$, $\ulcorner A = B\urcorner$ or $\ulcorner \overset{x}{\smile}[A^{u}_{x}]\urcorner$. An *abstract* is a term of the form $\ulcorner \overset{'}{\kappa}[A^{u}_{\kappa}]\urcorner$. Terms are called *closed* if they do not contain any parameter; otherwise they are called *open*. Closed formulas are called *sentences*. The set of all closed abstracts will be denoted by ‘$\mathfrak{C}$’.

The *universal closure* A^{c} of a term A is defined to be $\ulcorner \overset{x_n}{\smile}[\ldots \overset{x_2}{\smile}[\overset{x_1}{\smile}[A^{u_1}_{x_1}]^{u_2}_{x_2}] \ldots {}^{u_n}_{x_n}]$, where $u_1, \ldots, u_n$ are the parameters occurring in A and $x_1, \ldots, x_n$ are quantifier-variables that do not occur in A. Here a certain standard order of the parameters and of the quantifier-variables is assumed in order to guarantee uniqueness of A^{c}.

‘$\equiv$’ will be used as the metalinguistic identity-sign. ‘$\top$’ and ‘$\bot$’ are (metalinguistic) abbrevations of ‘$\overset{\mathfrak{a}}{\smile}(\mathfrak{a} = \mathfrak{a})$’ and ‘$\text{┬}\overset{\mathfrak{a}}{\smile}(\mathfrak{a} = \mathfrak{a})$’ (i.e. $\ulcorner \text{┬}\top\urcorner$), respectively.

The *rank* of a term is defined as follows:

(1) Parameters, abstracts, $\top$ and $\bot$ are of rank 0.
(2) If A and B are of rank n and m, respectively, then $\ulcorner - A\urcorner$ is of rank $n + 1$, $\ulcorner \text{┬} A\urcorner$ is of rank $n + 1$ if different from $\bot$, $\ulcorner A = B\urcorner$ is of rank $max(n,m) + 1$, and $\ulcorner \overset{x}{\smile}[A^{u}_{x}]\urcorner$ is of rank $n + 1$ if different from $\top$.

This definition of “rank” underlies the inductive definition of the following semantics. Abstracts are of rank 0, because, if closed, they are counted as smallest semantic units. $\top$ and $\bot$ are given rank 0 in order to have, in every case, standard names of rank 0 at our disposal not only for courses-of-values, but also for truth-values.

3 Semantics A *structure* $\mathfrak{A}$ is a triple $\langle U, t, f \rangle$, where U is a set, called the *domain* of $\mathfrak{A}$ (also denoted by '$|\mathfrak{A}|$'), and where t and f are two distinct elements of U, called the truth-values of $\mathfrak{A}$ (also denoted by '$t_{\mathfrak{A}}$' and '$f_{\mathfrak{A}}$', respectively). An *interpretation* $\mathfrak{I}$ in $\mathfrak{A}$ is a mapping from the set of all closed abstracts to U, such that $\mathfrak{I}(\mathfrak{C}) \cup \{t,f\} \equiv U$. (I.e., at most t and f are not values of closed abstracts under $\mathfrak{I}$.) The elements of $\mathfrak{J}(\mathfrak{C})$ are called the courses-of-values under $\mathfrak{I}$.

Let an interpretation $\mathfrak{I}$ in $\mathfrak{A}$ be given. A mapping $\mathfrak{I}^*$ from the set of all closed terms onto U, called the *extension* of $\mathfrak{I}$, is inductively defined as follows:

(1) $\mathfrak{I}^*(A) \equiv \mathfrak{I}(A)$ for closed abstracts A

(2) $\mathfrak{I}^*(\top) \equiv t$, $\mathfrak{I}^*(\bot) \equiv f$

(3) $\mathfrak{I}^*(\ulcorner -A \urcorner) \equiv \begin{cases} t & \text{if } \mathfrak{I}^*(A) \equiv t \\ f & \text{otherwise} \end{cases}$

(4) $\mathfrak{I}^*(\ulcorner \top\!\!A \urcorner) \equiv \begin{cases} f & \text{if } \mathfrak{I}^*(A) \equiv t \\ t & \text{otherwise} \end{cases}$

(5) $\mathfrak{I}^*(\ulcorner A = B \urcorner) \equiv \begin{cases} t & \text{if } \mathfrak{I}^*(A) \equiv \mathfrak{I}^*(B) \\ f & \text{otherwise} \end{cases}$

(6) $\mathfrak{I}^*(\ulcorner \overset{x}{\smile}[A^u_x] \urcorner) \equiv \begin{cases} t & \text{if for each } B \text{ which is either } \top, \bot \text{ or a closed} \\ & \text{abstract: } \mathfrak{I}^*([A^u_B]) \equiv t \\ f & \text{otherwise.} \end{cases}$

A term A is *valid* under $\mathfrak{I}$ in $\mathfrak{A}$ ($\mathfrak{I}$ is a *model* of A, $\mathfrak{I} \models A$), if $\mathfrak{I}^*(A^c) \equiv t$ (A need not be a formula!). $\mathfrak{I}$ is called a model of the *abstraction principle*, if for all u, κ, μ, x, A, B: $\mathfrak{I} \models \ulcorner (\grave{\kappa}[A^u_\kappa] = \grave{\mu}[B^u_\mu]) = \overset{x}{\smile}([A^u_x] = [B^u_x]) \urcorner$.

One could object to this semantics that the universal quantifier is interpreted *substitutionally*, i.e. with reference to the instances of the quantified formula, and not (as in Frege's introduction of the quantifiers in Section 8 of the "Basic Laws") *referentially*, i.e. with reference to the objects of the domain. This is not a problem because we assume that, for all elements of the domain, names are at our disposal in the formal language: It holds that $\mathfrak{I}^*(\mathfrak{C} \cup \{\top, \bot\}) \equiv |\mathfrak{A}|$. The assumption that each object can be denoted is certainly not un-Fregean. A referential interpretation of quantification is not possible within our framework, because closed abstracts are smallest semantic units and open abstracts cannot be interpreted satisfactorily, not even with respect to a valuation of parameters. Terms cannot be semantically decomposed down to the level of parameters.

4 Frege's metalinguistic usage of abstracts—abstracts of the form '⁖. . . .' and '⁘. . . .' (In this section we use Frege's original notation '$\grave{\epsilon}\Phi(\epsilon)$' for abstracts.) Frege did not yet have at his disposal the model-theoretic way of speaking about signs and their denotations, in particular not the strict distinction between object-language and metalanguage. In model-theoretic argumentations one talks about signs as well as about objects in the metalanguage without using any expression of the object-language. We can, for example, say that the object ω_1 is assigned to '$\grave{\epsilon}\Phi(\epsilon)$' by the interpretation $\mathfrak{I}_1$, and that ω_2 is assigned to the same sign '$\grave{\epsilon}\Phi(\epsilon)$' by the interpretation $\mathfrak{I}_2$. In doing so we use

‘ω_1’ and ‘ω_2’ as metalinguistic names for objects which have nothing to do with the signs of the object-language. Frege, however, also uses, when speaking about courses-of-values, signs of the object-language like ‘$\acute{\epsilon}\Phi(\epsilon)$’ as names. According to Frege, ‘$\acute{\epsilon}\Phi(\epsilon)$’ is always a name for the object $\acute{\epsilon}\Phi(\epsilon)$, which means that for Frege closed abstracts are always names with courses-of-values as fixed denotations.

Thus Frege is, strictly speaking, not in a position to discuss the non-uniqueness of the denotations of abstracts as he does in Section 10 of the “Basic Laws”. If ‘$\acute{\epsilon}\Phi(\epsilon)$’ denotes the course-of-values $\acute{\epsilon}\Phi(\epsilon)$, then one cannot even formulate that another denotation of ‘$\acute{\epsilon}\Phi(\epsilon)$’ is consistent with the abstraction principle. In not only mentioning but also using abstracts throughout his investigations, Frege presupposes their denotative relationship to fixed objects. (This is not the case with sentences and truth-values: Here Frege never uses signs of the object-language (such as ‘$-\!\overset{\mathfrak{a}}{\smile}\!-(\mathfrak{a} = \mathfrak{a})$’ or ‘$\top\!-\!\overset{\mathfrak{a}}{\smile}\!-(\mathfrak{a} = \mathfrak{a})$’) in order to refer to the truth-values, but always the metalinguistic names ‘the True’ and ‘the False’!)

In order to treat this problem nevertheless, Frege uses *another sort of term*, which he writes as ‘$\tilde{\epsilon}\Phi(\epsilon)$’.[8] He expresses the view that the abstraction principle determines no unique model but admits different interpretations $\mathfrak{I}_1$ and $\mathfrak{I}_2$ where, e.g., $\mathfrak{I}_1$(‘$\acute{\epsilon}\Phi(\epsilon)$’) $\equiv \omega_1$ and $\mathfrak{I}_2$(‘$\acute{\epsilon}\Phi(\epsilon)$’) $\equiv \omega_2$ about as follows: In addition to terms ‘$\acute{\epsilon}\Phi(\epsilon)$’ we may introduce terms ‘$\tilde{\eta}\Phi(\eta)$’ by the abstraction principle “without the identity of $\acute{\epsilon}\Phi(\epsilon)$ and $\tilde{\eta}\Phi(\eta)$ being derivable from this” ([3], p. 17).[9] That means, instead of directly speaking metalinguistically about different objects, he introduces new signs into the object-language and uses them within his metalinguistic reflections as names for objects.

Since in our model-theoretic investigations a sign of the object-language can never be *used*, there is no reason for us to introduce a new kind of abstract. Rather, from this point of view, two languages, whose abstracts are formed with ‘ ’ ’ and ‘˜’, respectively, appear to be identifiable: The form of the symbols of the object-language is quite unimportant.

5 The permutation argument

According to the previous section we need not take into account Frege's distinction between abstracts of different forms. So we propose the following reconstruction of his permutation argument: Let an interpretation $\mathfrak{I}_1$ in the structure $\mathfrak{A} \equiv \langle U, t, f \rangle$ be given. Let ω_1 and ω_2 be different courses-of-values under $\mathfrak{I}_1$, i.e., $\omega_1 \not\equiv \omega_2$ and $\omega_1 \equiv \mathfrak{I}_1(A)$, $\omega_2 \equiv \mathfrak{I}_1(B)$ for closed abstracts A, B (thus it holds: $\mathfrak{I}_1 \models \ulcorner \top(A = B) \urcorner$). Let p_F be the permutation of U which is defined as follows:

$$p_F(\omega) \equiv \begin{cases} \omega_1 \text{ if } \omega \equiv t \\ \omega_2 \text{ if } \omega \equiv f \\ t \text{ if } \omega \equiv \omega_1 \\ f \text{ if } \omega \equiv \omega_2 \\ \omega \text{ otherwise.} \end{cases}$$

That is, p_F interchanges t and f with ω_1 and ω_2, respectively, and leaves other arguments unchanged. Let $\mathfrak{B}$ be the structure $\langle U, \omega_1, \omega_2 \rangle$ and $\mathfrak{I}_2$ the interpretation in $\mathfrak{B}$ which is defined as follows: $\mathfrak{I}_2(C) \equiv p_F(\mathfrak{I}_1(C))$ for all closed abstracts C. Then for each term D: $\mathfrak{I}_1 \models D$ iff $\mathfrak{I}_2 \models D$. In particular, $\mathfrak{I}_2$ is a model of the abstraction principle iff $\mathfrak{I}_1$ is a model of the abstraction principle.

This can be proved for *arbitrary* permutations; so the permutation argument is a special case of the following theorem:

Theorem 1 *Let* $\mathfrak{I}_1$ *be an interpretation in* $\mathfrak{A} \equiv \langle U, t, f\rangle$. *Let* p *be a permutation of* U. *Let* $\mathfrak{B}$ be $\langle U, p(t), p(f)\rangle$. *Let* $\mathfrak{I}_2$ *be the interpretation* $p \circ \mathfrak{I}_1$ (*composition of the mappings* $\mathfrak{I}_1$ *and* p) *in* $\mathfrak{B}$. *Then for each term* D, $\mathfrak{I}_1 \models D$ *iff* $\mathfrak{I}_2 \models D$.

Proof: We show

(*) $p(\mathfrak{I}_1{}^*(C)) \equiv \mathfrak{I}_2{}^*(C)$ for any closed term C.

Then the assertion can be inferred as follows:

$$\begin{array}{lll}
\mathfrak{I}_1 \models D & \text{iff } \mathfrak{I}_1{}^*(D^c) \equiv t & (\text{definition of } \mathfrak{I}_1 \models D) \\
& \text{iff } p(\mathfrak{I}_1{}^*(D^c)) \equiv p(t) & (p \text{ one-one}) \\
& \text{iff } \mathfrak{I}_2{}^*(D^c) \equiv p(t) & ((*)) \\
& \text{iff } \mathfrak{I}_2 \models D & (\text{definition of } \mathfrak{I}_2 \models D).
\end{array}$$

We prove (*) by induction on the rank of C.

$$\begin{array}{lll}
(1) & \multicolumn{2}{l}{\text{If } C \text{ is a closed abstract, we have}} \\
& \multicolumn{2}{l}{p(\mathfrak{I}_1{}^*(C)) \equiv p(\mathfrak{I}_1(C)) \equiv \mathfrak{I}_2(C) \equiv \mathfrak{I}_2{}^*(C).} \\
(2) & \multicolumn{2}{l}{p(\mathfrak{I}_1{}^*(\top)) \equiv p(t) \equiv \mathfrak{I}_2{}^*(\top).} \\
& \multicolumn{2}{l}{p(\mathfrak{I}_1{}^*(\bot)) \equiv p(f) \equiv \mathfrak{I}_2{}^*(\bot).} \\
(3) & p(\mathfrak{I}_1{}^*(\ulcorner - C\urcorner)) \equiv p(t) & \text{iff } \mathfrak{I}_1{}^*(\ulcorner - C\urcorner) \equiv t \\
& & \text{iff } \mathfrak{I}_1{}^*(C) \equiv t \\
& & \text{iff } p(\mathfrak{I}_1{}^*(C)) \equiv p(t) \\
& & \text{iff } \mathfrak{I}_2{}^*(C) \equiv p(t) \qquad (\text{induction hyp.}) \\
& & \text{iff } \mathfrak{I}_2{}^*(\ulcorner - C\urcorner) \equiv p(t). \\
& p(\mathfrak{I}_1{}^*(\ulcorner - C\urcorner)) \equiv p(f) & \text{iff } \mathfrak{I}_2{}^*(\ulcorner - C\urcorner) \equiv p(f) \text{ analogously.} \\
(4) & p(\mathfrak{I}_1{}^*(\ulcorner -\!\!\top\!\! C\urcorner)) \equiv p(t) & \text{iff } \mathfrak{I}_2{}^*(\ulcorner -\!\!\top\!\! C\urcorner) \equiv p(t) \text{ analogously.} \\
& p(\mathfrak{I}_1{}^*(\ulcorner -\!\!\top\!\! C\urcorner)) \equiv p(f) & \text{iff } \mathfrak{I}_2{}^*(\ulcorner -\!\!\top\!\! C\urcorner) \equiv p(f) \text{ analogously.} \\
(5) & p(\mathfrak{I}_1{}^*(\ulcorner C_1 = C_2\urcorner)) \equiv p(t) & \text{iff } \mathfrak{I}_1{}^*(\ulcorner C_1 = C_2\urcorner) \equiv t \\
& & \text{iff } \mathfrak{I}_1{}^*(C_1) \equiv \mathfrak{I}_1{}^*(C_2) \\
& & \text{iff } p(\mathfrak{I}_1{}^*(C_1)) \equiv p(\mathfrak{I}_1{}^*(C_2)) \\
& & \text{iff } \mathfrak{I}_2{}^*(C_1) \equiv \mathfrak{I}_2{}^*(C_2) \qquad (\text{ind. hyp.}) \\
& & \text{iff } \mathfrak{I}_2{}^*(\ulcorner C_1 = C_2\urcorner) \equiv p(t). \\
& p(\mathfrak{I}_1{}^*(\ulcorner C_1 = C_2\urcorner)) \equiv p(f) & \text{iff } \mathfrak{I}_2{}^*(\ulcorner C_1 = C_2\urcorner) \equiv p(f) \text{ analogously.} \\
(6) & p(\mathfrak{I}_1{}^*(\ulcorner \overset{x}{-\!\smile\!-} [C^u_x]\urcorner)) \equiv p(t) & \text{iff } \mathfrak{I}_1{}^*(\ulcorner \overset{x}{-\!\smile\!-} [C^u_x]\urcorner) \equiv t \\
& & \text{iff } \mathfrak{I}_1{}^*([C^u_E]) \equiv t \text{ for each } E \text{ which is} \\
& & \quad \top,\ \bot \text{ or a closed abstract} \\
& & \text{iff } p(\mathfrak{I}_1{}^*([C^u_E])) \equiv p(t) \text{ for such } E\text{'s} \\
& & \text{iff } \mathfrak{I}_2{}^*([C^u_E]) \equiv p(t) \text{ for such } E\text{'s} \\
& & \qquad\qquad\qquad (\text{ind. hyp.}) \\
& & \text{iff } \mathfrak{I}_2{}^*(\ulcorner \overset{x}{-\!\smile\!-} [C^u_x]\urcorner) \equiv p(t). \\
& p(\mathfrak{I}_1{}^*(\ulcorner \overset{x}{-\!\smile\!-} [C^u_x]\urcorner)) \equiv p(f) & \text{iff } \mathfrak{I}_2{}^*(\ulcorner \overset{x}{-\!\smile\!-} [C^u_x]\urcorner) \equiv p(f) \text{ analogously.}^{10}
\end{array}$$

If p is the Fregean permutation p_F and $\mathfrak{I}_1$ is a model of the abstraction principle, then $\mathfrak{I}_2$ is a model of the abstraction principle for which $\mathfrak{I}_2(A) \equiv t$

and $\mathfrak{I}_2(B) \equiv f$ holds. This result may suggest that we have obtained a model of the abstraction principle in which A denotes the True and B denotes the False, just as Frege claims in his identifiability thesis. This, however, is wrong: Which objects are truth-values depends on the chosen structure. Now t and f are the truth-values of $\mathfrak{A}$ whereas $p_F(t)$ and $p_F(f)$ are the truth-values of $\mathfrak{B}$. This means that, under $\mathfrak{I}_2$, A and B denote the truth-values of $\mathfrak{A}$ but not the truth-values of $\mathfrak{B}$ (provided $p_F(\{t,f\}) \not\equiv \{t,f\}$)—and, similarly, under $\mathfrak{I}_1$, A and B denote the truth-values of $\mathfrak{B}$ but not the truth-values of $\mathfrak{A}$. Therefore, if A and B do not already denote the truth-values of $\mathfrak{A}$ under $\mathfrak{I}_1$ they cannot denote the truth-values of $\mathfrak{B}$ under $\mathfrak{I}_2$. If the truth-values of $\mathfrak{A}$ are not already courses-of-values under $\mathfrak{I}_1$, the truth-values of $\mathfrak{B}$ are not courses-of-values under $\mathfrak{I}_2$.

When Frege writes, "that an arbitrary course-of-values is to be the True and another the False", he obviously means (in our terminology) that this must be the case in *one and the same model.* This can easily be seen if one reformulates, following Thiel, Frege's identifiability thesis in such a way that Frege's aim is "the explicit stipulation that two certain expressions having the given form of abstracts can serve for denoting the True and the False, respectively" ([6], p. 288; my translation). According to this interpretation, for which Thiel gives strong reasons, the stipulation should have the effect that $\ulcorner \top = A \urcorner$ and $\ulcorner \bot = B \urcorner$ become valid. That is, the denotations of A and B under $\mathfrak{I}_2$ must be determined in such a way that $\mathfrak{I}_2^*(\top) \equiv \mathfrak{I}_2(A)$ and $\mathfrak{I}_2^*(\bot) \equiv \mathfrak{I}_2(B)$, which means that A and B must under $\mathfrak{I}_2$ denote the truth-values of $\mathfrak{B}$ (which are, according to our semantics, assigned to $\top$ and $\bot$).

So the transition from the interpretation $\mathfrak{I}_1$ in $\langle U, t, f\rangle$ to the interpretation $\mathfrak{I}_2$ in $\langle U, p_F(t), p_F(f)\rangle$ by means of p_F does not contribute anything to the goal of determining certain abstracts to be names of the truth-values, because the truth-values are transformed by p_F as well. This argument is obviously independent of the specific choice of p_F and holds for *all* permutations p of $|\mathfrak{A}|$.

One could object to our formulation of Theorem 1 that Frege's permutation argument only concerns the *domain* U of $\mathfrak{A}$ but not the *structure* $\mathfrak{A}$ itself, i.e., the interpretation of sentences by truth-values. This means that although $\mathfrak{I}_2$ is defined as $p \circ \mathfrak{I}_1$ for a permutation p, $\mathfrak{I}_2$ should remain an interpretation in the same structure $\mathfrak{A}$ (i.e., in $\langle U, t, f\rangle$ instead of $\langle U, p(t), p(f)\rangle$). In other words, even under $\mathfrak{I}_2^*$, sentences should be interpreted by t and f, as under $\mathfrak{I}_1^*$, and not by $p(t)$ and $p(f)$.

In fact, the stated objections would become superfluous if Theorem 1, or at least its special case for p_F as p, held with $\mathfrak{I}_2$ considered an interpretation in $\mathfrak{A}$. Then $\mathfrak{I}_2 \models \ulcorner \top = A \urcorner$ and $\mathfrak{I}_2 \models \ulcorner \bot = B \urcorner$ would hold since $\mathfrak{I}_2^*(\top) \equiv t \equiv p(\mathfrak{I}_1(A)) \equiv \mathfrak{I}_2(A) \equiv \mathfrak{I}_2^*(A)$ and $\mathfrak{I}_2^*(\bot) \equiv f \equiv p(\mathfrak{I}_1(B)) \equiv \mathfrak{I}_2(B) \equiv \mathfrak{I}_2^*(B)$. So A and B would really be names of the True and the False under $\mathfrak{I}_2$ because t and f are also the truth-values of $\mathfrak{I}_2$.

However, for this alternative one would have to abandon the restriction $U \equiv \mathfrak{I}_2(\mathfrak{C}) \cup \{t,f\}$. For if t and f are not already courses-of-values under $\mathfrak{I}_1$ (i.e., if $\{t,f\} \not\subseteq \mathfrak{I}_1(\mathfrak{C})$), then $p(t)$ and $p(f)$ are not courses-of-values under $\mathfrak{I}_2$ (i.e., $\{p(t), p(f)\} \not\subseteq \mathfrak{I}_2(\mathfrak{C})$). Another possibility would be to consider $\mathfrak{I}_2$ an interpretation in $\langle U', t, f\rangle$ for $U' \subseteq U$. But not even this method would be viable. Not only no *permutation* but even no *mapping* exists which has the desired property, as the following theorem demonstrates.

Theorem 2 *Let $\mathfrak{I}_1$ be an interpretation in $\mathfrak{A} \equiv \langle U, t, f\rangle$ which is a model of the abstraction principle. Let h be a mapping from U to U, such that $h(\omega_1) \equiv t$ for an $\omega_1 \in \mathfrak{I}_1(\mathfrak{C})$ which is different from t. Let $\mathfrak{I}_2$ be the interpretation $h \circ \mathfrak{I}_1$ in $\langle U', t, f\rangle$ for $U' \equiv h(\mathfrak{I}_1(\mathfrak{C})) \cup \{t,f\} \subseteq U$. Then $\mathfrak{I}_2$ is not a model of the abstraction principle.*

Proof: Let A be a closed abstract such that $\mathfrak{I}_1(A) \equiv \omega_1$. Then $\mathfrak{I}_1 \models \ulcorner \grave{\epsilon}(—A = \epsilon) = \grave{\epsilon}(\bot = \epsilon)\urcorner$, since, because of $\omega_1 \not\equiv t$ (which implies $\mathfrak{I}_1{}^*(\ulcorner —A\urcorner) \equiv f$), it holds that $\mathfrak{I}_1 \models \ulcorner -\!\overset{\mathfrak{a}}{\smile}\!-((—A = \mathfrak{a}) = (\bot = \mathfrak{a}))\urcorner$. However, it does not hold that $\mathfrak{I}_2 \models \ulcorner -\!\overset{\mathfrak{a}}{\smile}\!-((—A = \mathfrak{a}) = (\bot = \mathfrak{a}))\urcorner$, since $\mathfrak{I}_2(A) \equiv h(\mathfrak{I}_1(A)) \equiv h(\omega_1) \equiv t$ (thus $\mathfrak{I}_2{}^*(\ulcorner —A\urcorner) \equiv t$). On the other hand we have $\mathfrak{I}_2 \models \ulcorner \grave{\epsilon}(—A = \epsilon) = \grave{\epsilon}(\bot = \epsilon)\urcorner$, because from $\mathfrak{I}_1(\ulcorner \grave{\epsilon}(—A = \epsilon)\urcorner) \equiv \mathfrak{I}_1(\ulcorner \grave{\epsilon}(\bot = \epsilon)\urcorner)$ it follows that $h(\mathfrak{I}_1(\ulcorner \grave{\epsilon}(—A = \epsilon)\urcorner)) \equiv h(\mathfrak{I}_1(\ulcorner \grave{\epsilon}(\bot = \epsilon)\urcorner))$, i.e., $\mathfrak{I}_2(\ulcorner \grave{\epsilon}(—A = \epsilon)\urcorner) \equiv \mathfrak{I}_2(\ulcorner \grave{\epsilon}(\bot = \epsilon)\urcorner)$.

A corresponding theorem can be proved for $h(\omega_2) \equiv f$ for $\omega_2 \in \mathfrak{I}_1(\mathfrak{C})$, $\omega_2 \not\equiv f$.

Theorem 2 shows that it is not possible to create a 'new' interpretation in $\mathfrak{A}$, which is a model of the abstraction principle and under which $\ulcorner A = \top\urcorner$ is valid for a closed abstract A, by simply adding a transformation of the domain of $\mathfrak{A}$ to a given interpretation in $\mathfrak{A}$ which is a model of the abstraction principle and under which A does not denote the True. Roughly speaking: Transformation of the domain is not sufficient for stipulating a closed abstract to be a name for the True.

Frege's aim can, however, be understood in yet another way. Whereas Theorem 2 dealt with the problem of reinterpreting closed abstracts A and B as names of the truth-values t and f of a given model, one could try to interpret A and B as names for the True and the False, independent of how truth-values were specified in a previously given model. More precisely, let an interpretation $\mathfrak{I}_1$ in $\mathfrak{A} \equiv \langle U, t, f\rangle$, which is a model of the abstraction principle, and let closed abstracts A, B be given such that $\mathfrak{I}_1(A) \not\equiv \mathfrak{I}_1(B)$. Then an interpretation $\mathfrak{I}_2$ in $\mathfrak{B} \equiv \langle U', \mathfrak{I}_1(A), \mathfrak{I}_1(B)\rangle$, for $U' \subseteq \mathfrak{I}_1(\mathfrak{C})$, is to be construed in such a way that $\mathfrak{I}_2(A) \equiv \mathfrak{I}_1(A)$ and $\mathfrak{I}_2(B) \equiv \mathfrak{I}_1(B)$. In this case we would obtain as desired $\mathfrak{I}_2 \models \ulcorner A = \top\urcorner$ and $\mathfrak{I}_2 \models \ulcorner B = \bot\urcorner$. However, this purpose too cannot (in nontrivial cases) be reached by simply adding to $\mathfrak{I}_1$ a transformation h from U to $\mathfrak{I}_1(\mathfrak{C})$, as can be shown analogously to Theorem 2:

Theorem 3 *Let $\mathfrak{I}_1$ be an interpretation in $\mathfrak{A} \equiv \langle U, t, f\rangle$, which is a model of the abstraction principle. Let A, B be closed abstracts such that $t \not\equiv \mathfrak{I}_1(A)$ and $\mathfrak{I}_1(A) \not\equiv \mathfrak{I}_1(B)$. Let h be a mapping from U to $\mathfrak{I}_1(\mathfrak{C})$ such that $h(\mathfrak{I}_1(A)) \equiv \mathfrak{I}_1(A)$. Let $\mathfrak{I}_2$ be the interpretation $h \circ \mathfrak{I}_1$ in $\mathfrak{B} \equiv \langle U', \mathfrak{I}_1(A), \mathfrak{I}_1(B)\rangle$ for $U' \equiv h(\mathfrak{I}_1(\mathfrak{C})) \cup \{\mathfrak{I}_1(A),\mathfrak{I}_1(B)\}$. Then $\mathfrak{I}_2$ is not a model of the abstraction principle.*

Proof: As in the proof of Theorem 2 we consider the closed abstracts $\ulcorner \grave{\epsilon}(—A = \epsilon)\urcorner$ and $\ulcorner \grave{\epsilon}(\bot = \epsilon)\urcorner$. By the abstraction principle and $t \not\equiv \mathfrak{I}_1(A)$ (which implies $\mathfrak{I}_1{}^*(\ulcorner —A\urcorner) \equiv f$), $\mathfrak{I}_1(\ulcorner \grave{\epsilon}(—A = \epsilon)\urcorner) \equiv \mathfrak{I}_1(\ulcorner \grave{\epsilon}(\bot = \epsilon)\urcorner)$ holds, and therefore $h(\mathfrak{I}_1(\ulcorner \grave{\epsilon}(—A = \epsilon)\urcorner)) \equiv h(\mathfrak{I}_1(\ulcorner \grave{\epsilon}(\bot = \epsilon)\urcorner))$, i.e. $\mathfrak{I}_2(\ulcorner \grave{\epsilon}(—A = \epsilon)\urcorner) \equiv \mathfrak{I}_2(\ulcorner \grave{\epsilon}(\bot = \epsilon)\urcorner)$. However, it does not hold that $\mathfrak{I}_2 \models \ulcorner -\!\overset{\mathfrak{a}}{\smile}\!-((—A = \mathfrak{a}) = (\bot = \mathfrak{a}))\urcorner$, because $\mathfrak{I}_2^*(A) \equiv \mathfrak{I}_2(A) \equiv h(\mathfrak{I}_1(A)) \equiv \mathfrak{I}_1(A)$, therefore $\mathfrak{I}_2^*(\ulcorner —A\urcorner) \equiv \mathfrak{I}_1(A)$, but $\mathfrak{I}_2^*(\bot) \equiv \mathfrak{I}_1(B) \not\equiv \mathfrak{I}_1(A)$.[11]

A corresponding result can be proved if one assumes $f \not\equiv \mathfrak{I}_1(B)$ instead of $t \not\equiv \mathfrak{I}_1(A)$ and $h(\mathfrak{I}_1(B)) \equiv \mathfrak{I}_1(B)$ instead of $h(\mathfrak{I}_1(A)) \equiv \mathfrak{I}_1(A)$.

From a given model of the abstraction principle no model of the abstraction principle can be construed just by transforming its domain, such that two given courses-of-values of the 'old' model are truth-values of the 'new' model and are at the same time denoted by the 'old' abstracts.

The proofs of Theorems 2 and 3 show that the stipulation of closed abstracts as names for the True and the False, together with the abstraction principle, has the effect that closed abstracts which have the same denotation, can lose their synonymity: $\ulcorner \acute{\epsilon}(-A = \epsilon) \urcorner$ and $\ulcorner \acute{\epsilon}(\bot = \epsilon) \urcorner$ have the same denotation as long as A does not denote the True, but obtain different denotations if A is defined to be a name of the True. If one wants to transform a model of the abstraction principle into a new model of the abstraction principle where a closed abstract which has not denoted the True in the 'old' model now denotes the True, then one has to proceed in such a way that abstracts which are synonymous under the old interpretation can receive different denotations under the new interpretation. This is not possible by simply mapping the domain of a structure. Such mappings can transform several objects into one (and so closed abstracts with different denotations into synonymous ones), but not one object into several (and so synonymous closed abstracts into closed abstracts with different denotations).

This does not mean that Frege's identifiability thesis is wrong but only that his permutation argument cannot serve as a foundation for it. It can, however, be shown in our framework that the identifiability thesis does not hold in full generality, namely that not any two *arbitrary* closed abstracts A, B which denote different courses-of-values can be defined to be names of the truth-values. Define A_1, B_1, A_2, B_2 as follows: $A_1 \equiv \ulcorner \acute{\epsilon}(\top = \epsilon) \urcorner$, $B_1 \equiv \ulcorner \acute{\epsilon}(A_1 = \epsilon) \urcorner$, $A_2 \equiv \ulcorner \acute{\epsilon}(\bot = \epsilon) \urcorner$, $B_2 \equiv \ulcorner \acute{\epsilon}(-\!\!\top\!\!-A_2 = \epsilon) \urcorner$) Let $\mathfrak{I}_1$ be an interpretation in $\mathfrak{A} \equiv \langle U, t, f \rangle$ which is a model of the abstraction principle. If $\mathfrak{I}_1(A_1) \not\equiv t$, then $\mathfrak{I}_1^*(A_1) \not\equiv \mathfrak{I}_1^*(B_1)$. But every model $\mathfrak{I}_2$ of the abstraction principle which is a model of $\ulcorner A_1 = \top \urcorner$ is a model of $\ulcorner A_1 = B_1 \urcorner$ and thus *not* a model of $\ulcorner B_1 = \bot \urcorner$. So A_1 and B_1, though interpreted differently by $\mathfrak{I}_1$, cannot denote the truth-values under an $\mathfrak{I}_2$ which is a model of the abstraction principle. If $\mathfrak{I}_1(A_1) \equiv t$, then $\mathfrak{I}_1(A_2) \not\equiv t$, and one can argue analogously for A_2 and B_2.

This example shows that there are closed abstracts A and B that have different denotations in a model $\mathfrak{I}_1$ of the abstraction principle but that receive the same denotation in any model $\mathfrak{I}_2$ of the abstraction principle, in which A is used as a name for the True (i.e., $\mathfrak{I}_2(A) \equiv \mathfrak{I}_2^*(\top)$). Thus they are not suitable as candidates for names of truth-values, since the True must be different from the False.

This leads to the question of which necessary and sufficient conditions closed abstracts must satisfy, so that for each given model $\mathfrak{I}_1$ of the abstraction principle there is a model $\mathfrak{I}_2$ of the abstraction principle in which A and B are names of truth-values, i.e., for which $\mathfrak{I}_2 \models \ulcorner A = \top \urcorner$ and $\mathfrak{I}_2 \models \ulcorner B = \bot \urcorner$ hold. As has been shown, it is not sufficient that A and B have different denotations under $\mathfrak{I}_1$ (i.e., $\mathfrak{I}_1(A) \not\equiv \mathfrak{I}_1(B)$). It is not even obvious that this is a necessary condition, for it seems that it cannot be precluded that $\mathfrak{I}_2(A) \not\equiv \mathfrak{I}_2(B)$ holds in spite of $\mathfrak{I}_1(A) \equiv \mathfrak{I}_1(B)$. If one could give *sufficient* conditions, one could check whether Frege's own choice of '$\overset{,}{\epsilon}(-\epsilon)$' and '$\overset{,}{\epsilon}(\epsilon = \top\!\!-\overset{\mathfrak{a}}{\smile}\!-(\mathfrak{a} = \mathfrak{a}))$' for *A* and *B* satisfy them. Concerning *necessary* conditions it would be interesting to see how far they restrict the choice of *A* and *B*, and whether Frege would have had essentially different possibilities in his choice of *A* and *B*. This could perhaps remove the impression of conventionalism (which is rather far from Frege's other thoughts) from the stipulation of the truth-values and names for truth-values in Section 10 of the "Basic Laws".

An even more basic question is whether there is a model of the abstraction principle *at all* and whether in such a model closed abstracts *A* and *B* can denote the truth-values. This problem does not concern the conditions *A* and *B* have to fulfill, with respect to a given model $\mathfrak{J}_1$ of the abstraction principle, in order to become names for the truth-values in *another* model $\mathfrak{J}_2$ of the abstraction principle, but the *initial* construction of such a model. The inconsistency of Frege's second-order system is no argument against the existence of a model because we have confined ourselves to its first-order part including the abstraction principle. Since in this restricted system the elementhood-relation (which is crucial for the derivation of the antinomy) is not definable, we have assumed its consistency throughout this paper.[12]

NOTES

[1] For the English translation, cf. [2], p. 48.

[2] In English, cf. [2], p. 47, line 28–p. 48, line 3).

[3] Also Dummett, whose sketch of the permutation argument ([1], p. 403 seq.) comes close to the model-theoretic interpretation proposed below, seems to consider this argument a correct foundation for the identifiability thesis.

[4] For example, the incompleteness of the first-order part of Frege's formalism; cf. Thiel [5].

[5] This is overlooked by Resnik ([4], p. 209) who in his short presentation of the permutation argument considers only courses-of-values but not truth-values to belong to the domain of the model he assumes to be given.

[6] In particular, our language contains no descriptive function or predicate symbols. Thus no conflict arises between Frege's central claim that denotations of function symbols are *functions* rather than objects and the model-theoretic way of assigning *objects* to descriptive symbols.

[7] It seems difficult to carry over our approach to the full second-order system of the "Basic Laws" because then, in order to deal with second-order quantifiers, we would need denotations of predicate or function symbols, something which is avoided here (see note 6). Besides that, results would probably become trivial since, due to the inconsistency of this second-order system, no model of the abstraction principle would exist.

[8] To be exact, Frege would have to write something like '$\tilde{\epsilon}\tilde{\Phi}(\epsilon)$', where '$\tilde{\Phi}(\epsilon)$' results from '$\Phi(\epsilon)$' by throughout replacing ',' by '$\tilde{}$'.

[9] In English, see [2], p. 47).

10. At the beginning of "Basic Laws", Section 10, Frege gives an argument which is in some respect similar to the later one and whose assertion can be reformulated as follows: Let $\mathfrak{A}$ and $\mathfrak{B}$ be structures with the same truth-values (i.e., $t_{\mathfrak{A}} \equiv t_{\mathfrak{B}}, f_{\mathfrak{A}} \equiv f_{\mathfrak{B}}$) and g be a one-one mapping from $|\mathfrak{A}|$ onto $|\mathfrak{B}|$ *leaving the truth-values fixed* (*i.e.*, $g(t_{\mathfrak{A}}) \equiv t_{\mathfrak{B}}$, $g(f_{\mathfrak{A}}) \equiv f_{\mathfrak{B}}$). Let $\mathfrak{J}_1$ be an interpretation in $\mathfrak{A}$ and $\mathfrak{J}_2$ be the interpretation $g \circ \mathfrak{J}_1$ in $\mathfrak{B}$. Then for each term D: $\mathfrak{J}_1 \models D$ iff $\mathfrak{J}_2 \models D$. This argument, which can be proved similarly to Theorem 1 and which shows that the abstraction principle does not completely determine a range of objects, differs from the permutation argument in that $|\mathfrak{A}|$ and $|\mathfrak{B}|$ may be different but that the truth-values must remain fixed.

11. It is not even necessary to require in Theorem 3 that $h(\mathfrak{J}_1(A)) \equiv \mathfrak{J}_1(A)$. It suffices to assume the existence of a closed abstract C such that $h(\mathfrak{J}_1(C)) \equiv \mathfrak{J}_1(A)$ and $\mathfrak{J}_1(C) \not\equiv t$, and to consider $\ulcorner \overset{,}{\epsilon}(-C = \epsilon)\urcorner$ instead of $\ulcorner \overset{,}{\epsilon}(-A = \epsilon)\urcorner$ in the proof.

12. T. Parsons has now established this consistency (see his contribution to this issue).

REFERENCES

[1] Dummett, M., *The Interpretation of Frege's Philosophy*, Duckworth, London, 1981.

[2] Frege, G., *The Basic Laws of Arithmetic: Exposition of the System*, trans., M. Furth, University of California Press, Berkeley and Los Angeles, 1964.

[3] Frege, G., *Grundgesetze der Arithmetik: Begriffsschriftlich abgeleitet*, vol. I, H. Pohle, Jena, 1893. Translated in part in [2].

[4] Resnik, M. D., *Frege and the Philosophy of Mathematics*, Cornell University Press, Ithaca and London, 1980.

[5] Thiel, C., "Die Unvollständigkeit der Fregeschen 'Grundgesetze der Arithmetik'," pp. 104-106 in *Vernünftiges Denken: Studien zur praktischen Philosophie und Wissenschaftstheorie*, ed., J. Mittelstrass and M. Riedel, de Gruyter, Berlin and New York, 1978.

[6] Thiel, C., "Wahrheitswert und Wertverlauf: Zu Freges Argumentation im §10 der 'Grundgesetze der Arithmetik'," pp. 287-299 in *Studien zu Frege I: Logik und Philosophie der Mathematik*, ed., M. Schirn, Frommann-Holzboog, Stuttgart-Bad Cannstatt, 1976.

Universität Konstanz
Fachgruppe Philosophie
7750 Konstanz
West Germany