First-Order Logical Validity and the Hilbert-Bernays Theorem

Gary Ebbs and Warren Goldfarb

Draft 8/1/2017

Forthcoming in Juhl and Schechter, eds., *Philosophy of Logic and Inferential Reasoning: Philosophical Issues*, Volume 18 (2018)

What we call the Hilbert-Bernays (HB) Theorem establishes that for any satisfiable first-order quantificational schema S, there are expressions of elementary arithmetic that yield a true sentence of arithmetic when they are substituted for the predicate letters in S. Our goals here are, first, to explain and defend W. V. Quine's claim that the HB theorem licenses us to define the first-order logical validity of a schema in terms of predicate substitution; second, to clarify the theorem by sketching an accessible and illuminating new proof of it; and, third, to explain how Quine's substitutional definition of logical notions can be modified and extended in ways that make it more attractive to contemporary logicians.

1. The standard set-theoretical definition of validity and consequence

We take for granted the widely-accepted Quine-Tarski schematic conception of validity and implication, and restrict our discussion to first-order logical schemata that can be constructed from quantifiers, variables, truth-functional connectives, and schematic predicate letters with any number of argument places.

The standard characterization of an interpretation of first-order logical schemata comprises

(1) a non-empty universe of discourse, U;

(2) for each n-place predicate letter P that occurs in any of the schemata, an assignment of a set of n-tuples of members of U to be the extension of P; and

(3) to each free variable that occurs in any of the schemata, an assignment of a member of U to be its value.

A schema is valid if and only if it is true under every interpretation; schema R implies schema S (i.e. S is a logical consequence of R) if and only if every interpretation of R and S that makes R true also makes S true (or, equivalently, the material conditional with R as antecedent and S as consequent is valid).

For sentences that have a first-order grammar and are either true or false, we may use the above definitions of validity and implication to define derivative notions of validity and implication, as follows (the subscript 'a' is short for "applied"): A sentence is $valid_a$ if and

only if it can be obtained by substitution of predicates for predicate letters in a valid schema; one sentence, r, *implies*$_a$ another, s (i.e. s is a *logical consequence*$_a$ of r) if and only if, there are schemata R and S such that r and s can be obtained by uniform substitutions of predicates for predicate letters in R and S, respectively, and R implies S (or, equivalently, the material conditional with r as antecedent and s as consequent is valid$_a$).

The central interest of these classifications lies in their consequences for our pursuit of truth. If we have determined that a sentence is valid$_a$, for instance, we are entitled to conclude that it is true. But why? The standard set-theoretical account of interpretation does not directly answer this question, because it does not explain the link between a true set-theoretical interpretation of a schema and sentences that one can construct by substituting predicates for predicate letters in the schema.

What we call a *bivalent language* is a language every sentence of which has a first-order grammar and is either true or false, and every open sentence of which determines a set. A sentence that one obtains from a schema S by substituting linguistic expressions of a bivalent language L for predicate letters in S provides what we call a *linguistic interpretation* of S. If sentence s is a linguistic interpretation of a schema S, then there is a corresponding set-theoretical interpretation of S. The universe of discourse U is either implicitly settled by our use of L or explicitly specified; if P is an n-place predicate of L that we can substitute for an n-place predicate letter in S, the corresponding extension of the predicate letter is the set of all and only n-tuples of objects from U that satisfy P. The extension that corresponds to the substitution of 'wise' for F, for instance, is $\{x \mid x \text{ is wise}\}$; the extension that corresponds to the substitution of 'loves' for G is $\{\langle x, y\rangle \mid x \text{ loves } y\}$. In the same way, for each n-place predicate P of L that is built up from basic predicates of L, variables, truth functional connectives, or quantifiers, a corresponding extension can be specified as the set of all and only those n-tuples of objects in U that satisfy P. These corresponding extensions, together with assignments of objects to the free variables in S, if any, amount to a set-theoretic interpretation of S. The key point is that a linguistic interpretation s of S is true if and only if S is true under the set-theoretic interpretation of S that corresponds to s in the way just explained.[1] This elementary reasoning establishes

Lemma. If L is a bivalent first-order language and schema S is valid (i.e. true for all set-theoretical interpretations) then every sentence of L that can be obtained by substituting predicates of L for predicate letters in S is true.

[1] This equivalence fails in any language, such as the language of ZF set theory, in which some open sentences do not determine a set. To explain the relationship between set-theoretical validity and sentences of such languages, the reasoning in this paragraph and the Lemma it supports would have to be modified.

It is natural to wonder whether the converse of this lemma is also true. Suppose every sentence of a bivalent language L that can be obtained by substituting predicates of L for predicate letters in S is true. Does it follow that S is valid, in the sense we defined above, using the set-theoretical understanding of interpretation? The answer obviously depends on the richness of the modes of expression of L. It may seem, however, that no language could be rich enough in modes of expression to guarantee that if every sentence of L that can be obtained by substituting predicates of L for predicate letters in S is true, then S is true for all set-theoretical interpretations. The apparent problem, as Quine points out, is that "it has been an accepted tenet of classical set theory from Cantor onward that the classes ... outrun the expressions [of any language]" (Quine 1982, p. 212) How then can we be sure that a substitutional characterization of validity is co-extensive with the standard set-theoretical of validity? The HB theorem decisively answers this question.

2. The HB Theorem and coextensivity

The Skolem-Löwenheim Theorem tells us that for any satisfiable schema S of quantification theory there is an interpretation that makes S true whose universe of discourse is the natural numbers. The Hilbert-Bernays Theorem sharpens this result by showing that the predicate letters in the schema may be interpreted as sets definable in first-order arithmetic, or, equivalently, there are open sentences of arithmetic that, when substituted for the predicate letters of S, yield a sentence of arithmetic that is true when the variables are construed as ranging over the natural numbers. (The HB Theorem also yields a restriction on the complexity of the open sentences, as we'll show in §4, but that is not needed for the purposes of this section.) We shall apply this theorem to show that the substitutional definition of validity is coextensive with the set-theoretical definition for all languages that are rich enough to contain arithmetic. Let us call a bivalent language *arithmetical* iff its universe of discourse includes the natural numbers and it contains predicates or function signs that express identity, addition, and multiplication on the natural numbers, as well as a predicate (primitive or defined) that holds just of the natural numbers.

Theorem. If S is a satisfiable first-order quantificational schema and L is an arithmetical language, then there are open sentences of L that, when substituted for the predicate letters of S, yield a truth of L.

Proof. If the universe of discourse of L is the natural numbers, this is just the HB theorem. If the universe of discourse of L is larger, we need to use the tactic outlined in Quine 1982 (p. 117), namely, any interpretation that verifies a schema of quantification theory (without identity) can be expanded to include new objects in the universe by making those new

objects indistinguishable from some object in the original universe of discourse. Let f be the function on the universe of L defined by $f(x) = x$ if x is a natural number and $f(x) = 0$ if x is not a natural number. Note that f is definable in L, since L contains a predicate true of just the natural numbers. The HB Theorem yields open sentences of arithmetic that are to replace the predicate letters of S; for each such open sentence $P(x_1, \ldots, x_n)$ let $P'(x_1, \ldots, x_n)$ be the open sentence $P(f(x_1), \ldots, f(x_n))$. It follows that when the primed open sentences replace the predicate letters of S, the resulting sentence is true when the universe of discourse is that of L.

Note that this tactic is dependent on the identity sign not occurring in S, that is, the theorem only holds for schemata S of quantification theory without identity.

It follows at once from the Theorem that if L is a bivalent arithmetical language and S is any schema of quantification theory (without identity) that comes out true whenever its predicate letters are replaced by open sentences of S, then S is valid (that is, true for all set-theoretical interpretations). For if S were not valid, then $\neg S$ would be satisfiable (true for at least one set-theoretical interpretation), and so, by the Theorem there would be open sentences of L that, when substituted for the predicate letters of S, yield a false sentence of L.

From this fact and the Lemma of §1, we have the desired result: if L is a bivalent arithmetical language, then a schema S of quantification theory without identity is valid (true for all set-theoretical interpretations) if and only if every sentence obtained from S by substituting open sentences of L for the predicate letters is true.

3. Quine's substitutional definition

Quine takes this coextensivity result to license defining validity for first-order logical schemata substitutionally: "a schema is valid if substitution in it yields none but true sentences."[2] (Quine 1986, p. 51) This is an explication, not a conceptual analysis, of validity. To explicate a linguistic expression e that one finds useful in some ways yet problematic in others is to decide to use, in place of e, a different linguistic expression that preserves and clarifies what one takes to be useful about e and avoids what one takes to be the problems with e (Quine 1960, §53). To evaluate Quine's substitutional definition of 'valid' one therefore needs to consider how it avoids what Quine takes to be the problems with the set-theoretical definition of 'valid', or, to put it positively, what advantages Quine thinks our adoption of the substitutional definition would bring. "The evident philosophical advantage

[2] Quine's substitutional definition of validity for first-order logical schemata should not be confused with a substitutional definition of quantification. Quine takes all quantifiers to be objectual.

of resting with th[e] substitutional definition [of validity], and not broaching model theory," Quine writes, "is that we save on ontology. Sentences suffice, sentences even of the object language, instead of a universe of sets specifiable and unspecifiable." (Quine 1986, p. 55) The substitutional definition "renders the notions of validity and logical truth independent of all but a modest bit of set theory; independent of the higher flights."[3] (Quine 1986, p. 56) Even if we are committed for other reasons to the "higher flights" of set theory, Quine argues, we do well to keep track of the ontological commitments of different parts of our overall theory, so that "when occasions arise for revising theories, we are in a position to favor theories whose demands are lighter." (Quine 1986, p. 55)

Despite these theoretical advantages of adopting Quine's substitutional definition of validity, three main objections to it have become entrenched in the literature.

The first objection is that the acceptability of the substitutional definition depends in a counter-intuitive way on the richness of the language relative to which the substitutions are defined. Carnap presses a version of this objection in a letter he wrote to Quine in 1943. We reconstruct his objection (changing it slightly to correct a technical mistake in Carnap's version[4]) as follows. Let L be a first-order language that contains only one non-

[3] The "modest bit of set theory" that Quine alludes to in this passage is the set theory that we need to define truth in Tarski's way for the language in which the substitutions are defined (Quine 1986, pp. 40–46) and to explicate sentences as sets of sets of uttered or written tokens (Quine 1986, pp. 55–56). We could do without even this modest bit of set theory if we take the notions of *truth* and *sentence* as primitive.

[4] Carnap asserts that in a language L with just one two-place predicate that "happens to be symmetric," the L-sentence that expresses the assumption that the two-place predicate is symmetric in L "is the only instance of its logical form in L," and so Quine's substitutional criterion of validity is fulfilled for the schema '$\forall x \forall y(Rxy \rightarrow Ryx)$' (Carnap 1943, pp. 304–305, with 'S' changed to 'L'). Suppose that 'admires' is the L-predicate in question. Then, according to Carnap, '$\forall x \forall y(x$ admires $y \rightarrow y$ admires $x)$', the L-sentence that expresses the assumption that 'admires' is symmetric in L, is the only L sentence that interprets the schema '$\forall x \forall y(Rxy \rightarrow Ryx)$'. Since '$\forall x \forall y(x$ admires $y \rightarrow y$ admires $x)$' is true by assumption, Carnap concludes that if L is the language we use for interpreting logical schemata, Quine's substitutional criterion of validity is fulfilled for the schema '$\forall x \forall y(Rxy \rightarrow Ryx)$'. In fact, however, the L-sentence '$\forall x \forall y((x$ admires $x \wedge x$ admires $y) \rightarrow (y$ admires $y \wedge y$ admires $x))$' also interprets '$\forall x \forall y(Rxy \rightarrow Ryx)$'. And since '$\forall x \forall y((x$ admires $x \wedge x$ admires $y) \rightarrow (y$ admires $y \wedge y$ admires $x))$' may be false even if 'admires' is symmetric in L, the schema '$\forall x \forall y(Rxy \rightarrow Ryx)$' is not guaranteed to be valid in Quine's substitutional sense by the assumption that '$\forall x \forall y(x$ admires $y \rightarrow y$ admires $x)$' is true. It does not help to add the constraint that 'admires' is reflexive in L, for in that case there is a different L-sentence that interprets '$\forall x \forall y(Rxy \rightarrow Ryx)$' but is not guaranteed to be true, namely, the sentence that results when we substitute the predicate 'x admires $y \wedge \exists z \exists z'[(x$ admires $z) \wedge (x$ admires $z') \wedge \neg(z$ admires $z')]$' for 'R' in '$\forall x \forall y(Rxy \rightarrow Ryx)$'.

logical predicate, namely, a two-place (fully interpreted) predicate 'admires', and suppose that, as it happens, 'admires' is true of all pairs of objects in L's domain. Then Quine's substitutional definition of validity classifies the schema '$\forall x \forall y(Rxy \rightarrow Ryx)$' as valid, since every L-sentence that interprets it is true in L. By the set-theoretical definition of validity, however, '$\forall x \forall y(Rxy \rightarrow Ryx)$' is not valid. Carnap concludes that Quine's substitutional definition should be rejected. Hinman, Kim, and Stich 1968 press the same kind of objection using a more complicated example.

From a Quinean point of view, the main problem with this objection is that it presupposes that a definition of validity is a sort of conceptual analysis of validity, and should be evaluated accordingly, relative to the assumption that the set-theoretical definition of validity is at least co-extensive with the concept of validity. It follows from this presupposition that if a proposed definition of validity does not yield the same results in all cases as the set-theoretical definition of validity, then it is unsatisfactory. Quine rejects this presupposition. He does not ask that a definition work in all cases, but only in cases of interest. If we know in advance that we are interested in the set-theoretical definition of validity because of the way it contributes to inquiries that we conduct in languages that are rich enough to express elementary arithmetic, and we are not aiming at a conceptual analysis of the notion of validity, then the HB theorem guarantees that a definition of validity in terms of substitution in such a language will suffice for our purposes. In this context it is irrelevant that for some toy languages the substitutional definition would yield results that differ from the results of the set-theoretical definition.[5]

The second of the entrenched objections to Quine's substitutional definition of validity is illustrated by a sentence such as '$\exists x \exists y(x \neq y)$' in a first-order language in which we treat '$=$' as a logical constant. While the set-theoretical account classifies '$\exists x \exists y(x \neq y)$' as not valid, because '$\exists x \exists y(x \neq y)$' comes out false under any set-theoretical interpretation with a universe of discourse with just one member, the substitutional account provides no way to vary the size of the universe, which remains fixed and infinite for all sentences (hence all substitutional interpretations) of any arithmetical theory. The substitutional account therefore appears to be unable to get the right result for '$\exists x \exists y(x \neq y)$'. (Read 1995, p. 41)

This second objection is not directly relevant to Quine's substitutional definition of validity, which is designed in the first instance only for first-order logical schemata without identity. The substitutional definition would be of limited value, however, if it could not be extended to yield satisfactory results for first-order logic with identity. Fortunately there is such an extension, if we accept a surrogate for the identity relation. For every language

[5] A parallel criticism of Hinman, Kim, and Stich 1968 is made in Berlinski and Gallin 1969, but they overlook what is highlighted here, namely that arithmetical languages suffice.

L that contains '=' and only a finite number of other simple predicates, there is a complex two-place predicate I_L that contains all the simple predicates of L except '=' and holds of objects x and y in L's domain if and only if x and y are indiscernible by any of the simple predicates of L other than '=' (Quine 1986, p. 63). We may without loss of expressive power restrict ourselves to languages L for which I_L and '=' are extensionally equivalent from the point of view of L.[6] For such languages, we may treat I_L as a surrogate for '=' for purposes of classifying sentences as logically true or not. For first-order logic without identity as a logical primitive, to say that sentence is logically true is just to say that it is valid$_a$, as defined above (i.e. can be obtained by substitution from a valid schema in which '=' does not occur). To define logical truth for first-order logic with identity as a logical primitive, we use schemata that contain '='. To interpret such schemata set-theoretically, we need, in addition the assignments described by clauses (1)–(3) of §1, an assignment to '=' of a set of ordered pairs of the form $\langle o, o\rangle$ for all objects o in the interpretation's universe of discourse. (With this understanding of an interpretation, we may view formulas such as '$\forall x(x = x)$', which contain no schematic predicate letters, as schemata.) In terms of these expanded notions of logical schema and interpretation, we can define a sentence as logically true if and only if it can be obtained by substitution of open sentences for predicate letters in a set-theoretically valid first-order schema, where the schema in question may (but need not) contain '='. The key point for present purposes is that a sentence s of L is logically true in this set-theoretical sense if and only if the sentence that results when one substitutes I_L for '=' in s is valid$_a$. In particular, all the sentences that result when one substitutes I_L for '=' in sentences that can be obtained by substitution of open sentences for predicate letters

[6] One might wish to stipulate that there exist objects x and y such that $x \neq y$ but x and y are indiscriminable by all the simple predicates of a language L other than '='. In such a language, of course, '=' are I_L are not coextensive. In place of this stipulation, however, one may simultaneously stipulate that there exist two distinct objects x and y and introduce by implicit definition two new simple monadic predicates, here represented by 'A' and 'B', that are uniquely true of x and y, respectively, as follows: $\exists x \exists y(Ax \wedge By \wedge \forall z(Az \rightarrow x = z) \wedge \forall z(Bz \rightarrow y = z) \wedge x \neq y)$. (Such a stipulation is not ad hoc, since if we can refer differently to objects x and y, then there is some predicate that we are using to distinguish them, even if the simple predicates of L do not.) Let L' be L with the addition of these implicitly defined simple monadic predicates 'A' and 'B'. By replacing the first stipulation with the second one, and expanding L to L', we ensure that the (now expanded) stipulation does not entail that I'_L and '=' are not extensionally equivalent from the point of view of L'. In this way, stipulations of the existence of objects that are not identical to each other but are indiscriminable by any of the predicates of a given language can be replaced by more complex stipulations that introduce by implicit definition new simple monadic predicates that discriminate between the stipulated objects.

in a set-theoretically valid first-order schema in which '=' occurs as a primitive, including such sentences as '$x = x$' and '$(x = y \wedge x$ is wise$) \rightarrow y$ is wise', are valid$_a$ (Quine 1986, p. 63). As required to address Read's challenge, however, the sentence that results when one substitutes I_L for '=' in '$\exists x \exists y(x \neq y)$' is not valid$_a$. If we treat I_L as a surrogate for '=', we may therefore mirror the standard set-theoretical classifications of sentences containing '=' as logically true or not by classifying their counterparts, with I_L for '=', as valid$_a$ or not, where valid$_a$ is defined substitutionally.

One might object that since the proposed surrogates for the identity relation (the I_L for variable L) will differ depending on the number and type of simple predicates in the language for which they are defined, they do not express the identity relation, which is univocal across languages. (There is an exactly parallel objection to a Tarski-style definition of truth for a given language: it does not provide a univocal definition of truth for *all* languages.) This would be another way of insisting that a definition of validity$_a$ should provide a conceptual analysis of validity. As we have seen, however, Quine rejects conceptual analysis and endorses the method of explication instead. If we judge that the theoretical advantages of the above surrogate-relative substitutional definition of validity for first-order logical with identity trump the goal of providing a univocal account of the meaning of '=' across languages, then it is not unreasonable for us to define validity substitutionally. (Similar reasoning explains why one may reasonably accept a Tarski-style definition of truth for a given language, even though it does not provide a univocal definition of truth for all languages.)

The third of the entrenched objections to Quine's substitutional definition of validity, due to George Boolos, is that the predicate-substitutional account does not generalize to yield a plausible account of the relation of logical consequence when applied to arbitrary infinite sets of schemata. A set Γ of schemata logically implies a schema S if and only if the set $\Gamma \cup \{\neg S\}$ is unsatisfiable. To provide the predicate-substitutional account to the relation of logical consequence, then, we need to characterize satisfiability of a set of schemata in terms of predicate-substitution. The problem is that for each arithmetical language L, there is at least one satisfiable infinite set Γ of schemata for which no simultaneous, uniform substitution of open sentences of L for the predicate letters in the schemata in Γ yields a set of true sentences of L. (This follows from Tarski's undefinability theorem. See Boolos 1975, pp. 526–527 for details.) Boolos concludes that we cannot plausibly define the set-theoretical notion of logical consequence, which encompasses cases in which Γ is an infinite set of schemata, in predicate-substitutional terms.

Boolos himself points out, however, that by the set-theoretical compactness theorem, "a set [of schemata] is satisfiable if and only if all its finite subsets are satisfiable." (Boolos

1975, p. 525) We may therefore define satisfiability in predicate-substitutional terms, as follows: "a set [of schemata] is satisfiable just in case every conjunction of its members has a true substitution instance". (Boolos 1975, p. 525) His response to this definition, which he of course knows to be extensionally correct, is that it "has no . . . plausibility as an account of satisfiability. It even sounds wrong." (Boolos 1975, p. 526)

If explication, not conceptual analysis, is our aim, however, Boolos's complaint that a substitutional definition of satisfiability "has no . . . plausibility" and "even sounds wrong" is irrelevant to our decision about whether to adopt such a definition. To evaluate a definition from a Quinean point of view, we must instead ask whether it would further our goals to adopt it. If we wish to minimize the set-theoretical commitments of our definition of satisfiability, then we may find it useful to rely on set-theoretical compactness and define a set of schemata as satisfiable just in case every conjunction of its members has a true substitution instance in an arithmetical language. We may infer from this definition that a set Γ of schemata implies a schema S if and only if not every conjunction of the members of the set $\Gamma \cup \{\neg S\}$ has a true substitution instance in an arithmetical language. By adopting these definitions, which are guaranteed to be extensionally correct, we render the notions of logical implication and satisfiability independent of the "higher flights" of set theory. We therefore have good reasons to adopt them as explications of the notions of satisfiability and logical implication.[7]

[7] Some writers respond to Boolos's objection by dropping the restriction that the substitutions in terms of which satisfiability is defined be drawn from a fixed language (McKeon 2004, p. 221, note 17; Dogramaci 2017, p. 85). This response faces three serious problems. First, if these writers propose an indefinite array of languages that could be used, then it is hard to understand how there could be a language in which to propose the definition, since it would have to contain truth-definitions for any of the languages which might need to be invoked, but one cannot define a truth predicate for a language whose vocabulary is not specified. Second, if the array of languages was meant to be completely extensive, then the language in which the definition is proposed would have to contain a truth-predicate for itself, but by Tarski's Undefinability Theorem no such language exists. Third, if, to avoid the first two problems, these writers propose a fixed stock of languages to be used, then their proposal rests on a conjecture that there is no infinite class of schemata that is satisfiable in the set-theoretical sense but not in their proposed substitutional sense. This conjecture looks doubtful in light of both Tarski's Undefinability Theorem and Boolos's technique of exploiting that theorem (Boolos 1975, pp. 526–527) to define arithmetical languages L and infinite satisfiable sets of schemata Γ for which for which no simultaneous, uniform substitution of open sentences of L for the predicate letters in the schemata in Γ yields a set of true sentences of L.

4. A new proof of the HB theorem

We shall prove the HB theorem in the following form (equivalent to Kleene 1952, Theorem 35, which makes precise the original result in Hilbert and Bernays 1939, §4.2): if S is a satisfiable schema of first-order quantification theorem then there are Δ_2-predicates of first-order arithmetic that, when substituted for the predicate letters of S, yield a truth when the quantifiers are taken to range over the natural numbers. (A Δ_2-predicate is one that can be expressed as both a Π_2-predicate $\forall x \exists y Q$ and a Σ_2-predicate $\exists x \forall y R$, where Q and R are recursive relations.) To show this, it suffices to show it just for schemata S that are closed and prenex, since a schema with free variables is satisfiable iff its existential closure is satisfiable, and every schema has a prenex equivalent, whether we define "equivalent" in terms of the set-theoretic or predicate-substitutional sense. (A schema is prenex iff all quantifiers stand in an initial row governing the rest of the schema. A schema is closed iff it contains no free variables.)

Let S be the schema, and let $u_0, u_1, u_2 \ \ldots$ be variables that do not occur in S. We shall call them u-variables. We generate schemata from S by an instantiation procedure with the following properties:

(1) if an existential schema $(\exists w)\Phi(w)$ is generated, then at some later point an instance $\Phi(u_i)$ is also generated, and the variable u_i is not used in any other instantiation of an existential schema;

(2) if a universal schema $(\forall w)\Phi(w)$ is generated, then for each i at some later point the instance $\Phi(u_i)$ is also generated.

We call the generated schemata *u-schemata*. In some cases — those in which no universal quantifier of S governs an existential quantifier — instead of (2) we could get by with only a finite number of u-variables.[8] We will ignore these cases in the rest of the proof. All the u-schemata that can be generated can be deduced from S in a sound natural deduction system for first-order logic: detailed ways of enacting this can be found in Quine 1982, pp. 205–206 and Goldfarb 2003, p. 221. Hence every conjunction of quantifier-free schemata that are generated is truth-functionally satisfiable, since otherwise S would be refutable in a sound deduction system, and hence unsatisfiable. Alternatively, purely model-theoretically, one can note (as Skolem did in 1923, 1928, and 1929) that any interpretation that makes S true can be extended to one that makes any conjunction of the generated schemata true

[8] In these cases there are recursive predicates that, when substituted for the predicate letters in S, yield a truth when the quantifiers are taken to range over the first n positive integers, where n is the number of existential quantifiers in S.

simply by appropriate choice of values for the u-variables, so that any such conjunction is also satisfiable.

Let n be the earliest stage at which the instantiation process generates a quantifier-free u-schema; and for each p let $U(p)$ be the conjunction of quantifier-free u-schemata generated by stage $n+p$. Let V be the set of atomic schemata formed from predicate letters of S and u-variables. Thus each $U(p)$ is a truth-function of members of V. We assume a standard gödel numbering of V, such that the gödel numbers of the members of V form a recursive set. An atomic schema A in V is *earlier* than an atomic schema A' in V iff the gödel number of A is less than that of A'. We identify "truth" with 0 and "falsity" with 1. A function $\varphi : V \to \{0,1\}$ is *p-acceptable* iff the truth-assignment it represents makes $U(p)$ true. Obviously, if φ is p-acceptable, it is also q-acceptable for all $q < p$. If φ and ψ are two distinct functions, we say φ *precedes* ψ iff at the earliest point A of difference between them, $\varphi(A) = 0$ and $\psi(A) = 1$.

Now define, for any integer p and any atomic schema A in V, $\Phi(p, A) = \varphi(A)$ where φ is the minimal p-acceptable function, that is, the p-acceptable function such that no other p-acceptable function precedes φ.

Note that Φ is recursive, since for any given p the question whether or not a function $\varphi: V \to \{0,1\}$ is p-acceptable can be answered by a truth-functional calculation in a finite number of steps, as can the question of whether φ is minimal (the minimal p-acceptable function is one of the finitely many functions ψ such that for every A in V that does not occur in $U(p)$, $\psi(A) = 0$).

Lemma 1. For each A in V the limit as p goes to infinity of $\Phi(p,\mathrm{A})$ exists: that is, for $\delta = 0$ or $\delta = 1$, $\exists q \forall r > q(\Phi(r, A) = \delta)$.

Proof. By induction on the ordering of V. Suppose for all A' earlier than A, the limit of $\Phi(p, A')$ exists. Let q be such that, for all A' earlier than A, $\Phi(r, A')$ is the limit value whenever $r > q$. (In the base case, where A is the earliest atomic schema in V, the supposition is vacuously satisfied, so we may take $q = 0$ in the reasoning that follows.) Either $\Phi(r, A) = 0$ for all $r > q$, in which case the limit of $\Phi(p, A)$ exists and is 0, or there exists an $r > q$ such that $\Phi(r, A) = 1$, so that the minimal r-acceptable function φ has $\varphi(A) = 1$. In the latter case, we claim, there can be no $s > r$ such that $\Phi(s, A) = 0$, so that the limit of $\Phi(p, A)$ exists and is 1. For suppose there were such an s. Then the minimal s-acceptable function ψ has $\psi(\mathrm{A}) = 0$. Since $s > r$, ψ is also r-acceptable. And since $r > q$, by our inductive hypothesis $\psi(A') = \varphi(A')$ for all A' earlier than A. Hence A is the earliest point at which ψ and φ differ. But $\psi(A) = 0$ and $\varphi(A) = 1$, so that ψ precedes φ. This

contradicts the assumption that φ is the minimal r-acceptable function. (The main idea behind this proof is in Skolem 1923, point 3).

For each $A \in V$, let $\Psi(A)$ be the limit of $\Phi(p, A)$ as p goes to infinity.

Lemma 2. Ψ is p acceptable for each p.

Proof. Let $q > p$ be large enough that $\Phi(q, A)$ is the limit value for each A in V that occurs in $U(p)$. Then $\Psi(A) = \Phi(q, A)$ for each A in V that occurs in U(p). $\Phi(q, \_)$ is q-acceptable by definition, so it is also p-acceptable, since $q > p$. Hence Ψ is p-acceptable.

Lemma 3. The relation $\Psi(A) = 0$ is Δ_2.

Proof. Recall that Φ is recursive. Also, $\Psi(A) = 0$ iff $\forall x \exists y > x(\Phi(y, A) = 0)$ iff $\exists x \forall y > x(\Phi(y, A) = 0)$. Hence Ψ is both Σ_2 and Π_2, and so is Δ_2.

To obtain an interpretation of S we proceed by exploiting the truth-assignment that Ψ provides. To each variable u_i we assign the value i. The universe of discourse is the set of integers so assigned, that is, all of the natural numbers, since we have required that every u-variable be used in the instantiation procedure. An n-adic predicate letter P that occurs in S is true of just the n-tuples of integers that Ψ declares that P is true of, when the variables are assigned numbers as just described. For example, if P is dyadic, we interpret P to be true just of those pairs $\langle m, n \rangle$ such that $\Psi(Pu_m u_n) = 0$. Since Ψ is Δ_2, so is this interpretation of P. Under these interpretations of the u-variables and predicates, all the unquantified u-schemata are true. Moreover, since the procedure instantiates each existential u-schema once and each universal u-schema by every u-variable, we may conclude that if we substitute these Δ_2-predicates for the predicate letters in S, the resulting arithmetical sentence is true when the universe of discourse is the natural numbers.

The argument of our Lemmas 1 - 3, framed more abstractly, amounts to a proof of the Weak König Lemma that is different from the standard one. The Weak König Lemma states that every infinite subtree of the full binary tree contains an infinite path. Here we'll consider only recursive subtrees. The standard proof stems from König 1927 and is the proof invariably presented in textbooks (e.g. Goldfarb 2003, pp. 218–220; Jeffrey 1991, p. 72; Kleene 1952, p. 391; Quine 1982, pp. 204–205). In that proof, an infinite path through the subtree is defined by requiring, at each stage, a non-recursive choice: in one common formulation, having arrived at a node, if the left successor node has infinitely many descendants in the subtree then pick it as the next node on the path, and if not pick the right successor node. The proof in Kleene 1952 (Theorem 35, pp. 394–395) uses an intricate double recursion to unpack this infinite succession of non-recursive choices into

one definition of the whole path and relies on a theorem by Post to establish that the path so defined is Δ_2. The proof presented here, based on Skolem 1923, proceeds differently. At each stage n a recursive path is defined down to level n, but at later stages the node at level n on the path may have to be altered. However (Lemma 2 above), such alterations can occur at most 2^n times. The final determination of the path is then the limit of a recursive process, and so easily shown to be Δ_2. Thus our proof provides a simple and direct argument that if the subtree is recursive then there exists a path through it that is Δ_2.[9]

5. Identity and a modified explication

As we have seen, Quine does not take identity to be a primitive in the schematic language, preferring instead, no doubt for reasons of economy, to define a surrogate using indiscernibility. In this section we present a modified Quinean position, in which "=" is taken as a logical primitive, as most logicians currently do. We first extend the argument of §4 to prove an extended version of the HB Theorem that holds for quantificational schemata with identity as a logical primitive; we then formulate a modified substitutional definition of validity that mirrors the set-theoretical account of validity for quantificational schemata with identity treated as a logical primitive.

Given a schema S containing "=", the *associated laws of identity* are "$\forall x(x = x)$" as well as the universal closures of all schemata

$$x = y \rightarrow (A \leftrightarrow B)$$

where A and B are atomic schemata containing either "=" or a predicate letter appearing in S, such that B differs from A by having "y" in some or all places that A has "x". If S is prenex, closed, and satisfiable, we now proceed to generate u-schemata from S and the associated laws of identity by instantiation, obeying the rules (1) and (2) given in §4. Treating "=" for the moment as if it were a predicate letter, as before, we can define a Δ_2 truth-assignment Ψ that makes all the u-schemata true. Call two u-variables u_i and u_j Ψ-*identified* iff Ψ assigns the value "true" to the schema $u_i = u_j$. The u-instances of the associated laws of identity insure that Ψ-*identified* is a congruence relation: if u_i is Ψ-identified with u_j, then Ψ assigns the same truth-value to an atomic schema containing u_i and any atomic schema that results from it by replacing some or all of the occurrences

[9] The existence of this alternative proof of the Weak König Lemma has not been recognized in the literature. In Skolem 1970, Hao Wang in his Introduction erroneously claims Skolem's proof is fallacious (p. 21); Wang's Introduction is reprinted without correction in Gabbay and Woods 2009.

of u_i with occurrences of u_j. Let μ be the function that takes each natural number i to the least k such that u_i and u_k are Ψ-identified, and let M be the range of μ. Note that for any distinct i and j in M, Ψ assigns falsity to the atomic schema $u_i = u_j$, and assigns truth to $u_i = u_i$. Thus, for members of M, Ψ treats "=" as identity. Moreover, since Ψ-*identified* is a congruence relation, Ψ makes true any schema that results from a u-schema by replacing every u_i with $u_{\mu(i)}$.

Note that M can be defined arithmetically from Ψ: a number i is in M iff $\forall j(j < i \rightarrow \Psi(u_i = u_j) = 1)$. In fact, because Ψ is Δ_2, so is M (bounded quantifiers do not affect complexity). Thus if we use the open sentence "$\forall j(j < i \rightarrow \Psi(u_i = u_j) = 1)$" to specify the domain of the quantifiers in S, paraphrasing each universal quantification $\forall x(\ldots x \ldots)$ that occurs in S by an expression of the form $\forall x(\forall j(j < x \rightarrow \Psi(u_x = u_j) = 1) \rightarrow (\ldots x \ldots))$ and each existential quantification $\exists x$ that occurs in S by an L expression of the form $\exists x(\forall j(j < x \rightarrow \Psi(u_x = u_j) = 1)\ \&\ (\ldots x \ldots))$, and use Ψ, just as in §4, to define open sentences of arithmetic to replace the predicate letters of S, the result will be a true arithmetical sentence that interprets S.

If M is infinite, then since it is Δ_2 there is a Δ_2 bijection between M and the set of natural numbers. The image of the interpretation just given under the bijection is then a Δ_2 interpretation whose universe of discourse is the set of natural numbers. We may conclude that if a quantificational schema with identity is satisfiable (in the set-theoretical sense), then either it is satisfiable over a finite universe or else there are Δ_2 predicates of arithmetic that, when substituted for the predicate letters, yield a truth when the quantifiers in the schema are construed as ranging over the natural numbers.

As noted in §2, the argument that gets us from the HB Theorem to the coextensiveness of the set-theoretical and substitutional definitions of validity works only for first-order quantificational schemata in which "=" does not occur as a logical primitive. In order to use the extended version of the HB Theorem to define validity for first-order quantificational schemata in which "=" occurs as a logical primitive, we need a somewhat different account of substitution, one that allows us to mirror, in substitutional terms, set-theoretical specifications of a domain of discourse for the quantifiers in a schema. No change is needed in our account of substitution for predicate letters in a schema S: for these we substitute open sentences of an arithmetical language L. Instead of substituting quantifiers of L for the quantifiers in S, as we did above, however, we now substitute complex expressions of L for the quantifiers in S, as follows. Where P is an open sentence of L that we think of as specifying the domain of the quantifiers in S, for each universal quantification $\forall x(\ldots x \ldots)$ that occurs in S we substitute an L expression of the form $\forall x(Px \rightarrow (\ldots x \ldots))$, and for each existential quantification $\exists x$ that occurs in S, we substitute an L expression of the form

$\exists x(Px \wedge (\ldots x \ldots))$. By making such substitutions for the predicate letters and quantifiers in S we obtain a sentence of L that is true if and only if S is true under the set theoretical interpretation that takes $\{x \mid Px\}$ as universe of discourse and that assigns to each predicate letter in S the set of objects from $\{x \mid Px\}$ that satisfy the open sentence we substitute for that predicate letter. By a straightforward extension of the reasoning we used to support the Lemma of §1, we may conclude that if S is true for all set-theoretical interpretations, then every substitutional interpretation of S is true. From the extended HB theorem that we proved in this section, we may also conclude that if every substitutional interpretation of S is true, then S is true for all set-theoretical interpretations. These conclusions together establish that a schema that contains "=" is valid (in the set-theoretical sense) if and only if every substitutional interpretation of the schema is true. We may therefore explicate validity for schemata that contain "=" substitutionally, while still regarding "=" as a logical primitive.

6. Infinite sets of schemata

In §3 we considered Boolos's objection that Quine's predicate-substitutional account of logical validity does not generalize to yield a plausible account of satisfiability (or of logical consequence, when it is defined in terms of satisfiability). Boolos points out that for each arithmetical language L, there is a satisfiable infinite set Γ of schemata for which no simultaneous substitution of open sentences of L for the predicate letters in the schemata in Γ yields a set of true sentences of L. He takes this to imply that we cannot in general define satisfiability for infinite sets Γ of schemata in terms of the substitution of open sentences of an arithmetical language for the predicate letters in the schemata in Γ. In reply we defined a set of schemata as satisfiable just in case every conjunction of its members has a true substitution instance, and relied on set-theoretical compactness to guarantee that this definition is co-extensive with the set-theoretical one. This reply is correct as far as it goes, but it may also leave one with the impression that a substitutional account of satisfiability is expressively much more limited than the set-theoretical one. This impression is quite misleading, as we can see if we consider how to apply the argument for the HB Theorem to infinite sets of schemata.

Let Γ be a satisfiable infinite set of schemata. We modify the instantiation procedure so that every schema in Γ, as well as each associated law of identity, is at some point put on the list of generated schemata, interspersed among the instantiations of previously generated schemata. There is no substantial change needed to the argument: the truth-assignment Ψ that renders all u-schemata true is defined as in §4; from Ψ we may obtain, as in §5, an interpretation that makes all the schemata in Γ true. However, Lemma 3 must be modified,

since the generation of the schemata will not be recursively describable if Γ is not itself recursive. In this case, $\Phi(p, A)$ is not recursive, but only recursive *in* Γ (more precisely, in the set of gödel numbers of schemata in Γ). The rest of the argument can proceed as before, yielding the fact that the open sentences obtained using Ψ are Δ_2 in Γ. (A function is recursive in a set Γ if it can be obtained using the usual recursive means from a primitive predicate true just of members of Γ. A predicate is Δ_2 in Γ iff it can be expressed in the forms $\forall x \exists y Q$ and $\exists x \forall y R$ where Q and R are relations that are recursive in Γ.)

We may conclude that for any satisfiable set Γ, there are predicates definable from Γ and arithmetical notions that yield, via simultaneous substitutions of the kind described in §5, an infinite set of true sentences that interpret the schemata in Γ. The result shows that there is no expressive gap between the specification of a satisfiable set of schemata and the provision of open sentences that yield truths when substituted for the predicate letters. For in any given language L, our substitutional horizon for satisfiability extends at least as far as the open sentences of L that are Δ_2 in any satisfiable set *that we can specify in* L.[10]

[10] We presented earlier drafts of this paper at the University of Pittsburgh (October 2014), Indiana University, Bloomington (December 2014), the Midwest PhilMath Workshop (October 2015), the Society for Exact Philosophy (May 2016), and the University of Texas, Austin (April 2017). For helpful comments on these and other occasions we thank Josh Dever, Sinan Dogramici, Anil Gupta, Ulf Hlobil, Hans Kamp, Jon Litland, Jeff Pelletier, Thomas Ricketts, Michael Thompson, Joan Weiner, Max Weiss, and Mark Wilson.

Works Cited

Berlinski, David and Gallin, Daniel

1969 "Quine's Definition of Logical Truth," *Noûs*, Vol. 3, No. 2, pp. 111–128.

Boolos, George

1975 "On Second Order Logic," *The Journal of Philosophy*, Vol. 72, pp. 509–527.

Carnap, Rudolf

1943 Letter to W.V. Quine, dated January 21, 1943, in Creath 1990, pp. 302–310.

Creath, Richard

1990 *Dear Carnap, Dear Van*. Berkeley: University of California Press.

Dogramici, Sinan

2017 "Why is a Valid Inference a Good Inference?" *Philosophy and Phenomenological Research*, Vol. 94, No. 1, pp. 61-96

Fenstad, Jens Erik and Hao Wang

2009 "Thoraf Albert Skolem," in Gabbay and Woods 2009, pp. 127–194

Gabbay, Dov and John Woods

2009 *Handbook of the History of Logic*, vol. 5, *Logic from Russell to Church*. Amsterdam: North-Holland

Goldfarb, Warren

2003 *Deductive Logic*. Indianapolis: Hackett Publishing Company, Inc.

Hilbert, David, and Paul Bernays

1939 *Grundlagen der Mathematik*, Vol. 2. Berlin: Springer.

Hinman, Peter, Kim, Jaegwon, and Stich, Stephen,

1968 "Logical Truth Revisited," *The Journal of Philosophy*, Vol. 65, pp. 495–500.

Jeffrey, Richard

1991 *Formal Logic: Its Scope and Limits*, 3rd Edition. New York: McGraw-Hill, Inc.

Kleene, S. C.

1952 *Introduction to Metamathematics*. Amsterdam and New York: Van Nostrand Reinhold.

König, Dénes

1927 "Über eine Schlussweise aus dem Endlichen ins Unendliche," in *Acta litterarum*

ac scientiarum Regiae Universitatis Hungaricae Francisco-Josephinae, sectio scientiarum mathematicarum 3, pp. 121–130.

McKeon, Matthew

2004 "On the Substitutional Characterization of First-Order Logical Truth," *History and Philosophy of Logic*, Vol. 25, pp. 205–224.

Quine, W. V.

1982 *Methods of Logic*, 4th ed. Cambridge, Mass.: Harvard University Press.

1986 *Philosophy of Logic*, 2nd ed. Cambridge, Mass.: Harvard University Press.

Read, Stephen

1995 *Thinking About Logic*. Oxford: Oxford University Press.

Skolem, Thoralf

1923 "Einige Bemerkungen zur axiomatischen Begründung der Mengenlehre," in *Matimatikerkongressen i Helsingfors den 4–7 Juli 1922, Den femte skandinaviska matematikerkongressen, Redogörelse* (Akademiska Bokhandeln, Helsinki 1923), 217–232. Translated as "Some remarks on axiomatized set theory," and incorrectly dated 1922, in van Heijenoort, ed., *From Frege to Gödel* (Cambridge, Mass.: Harvard University Press, 1967), 200–301.

1970 *Selected Works in Logic*, Jens E. Fenstad, editor. Oslo: Universitetsforlaget.