Boolos and the Metamathematics of Quine's Definitions of Logical Truth and Consequence

Günther Eder
University of Vienna
guenther.eder@univie.ac.at

Abstract. The paper is concerned with Quine's substitutional account of logical truth. The critique of Quine's definition tends to focus on miscellaneous odds and ends, such as problems with identity. However, in an appendix to his influential article *On Second Order Logic*, George Boolos offered an ingenious argument that seems to diminish Quine's account of logical truth on a deeper level. In the article he shows that Quine's substitutional account of logical truth cannot be generalized properly to the general concept of logical consequence. The purpose of this paper is threefold: first, to introduce the reader to the metamathematics of Quine's substitutional definition of logical truth; second, to make Boolos' result accessible to a broader audience by giving a detailed and self-contained presentation of his proof; and, finally, to discuss some of the possible implications and how a defender of the Quinean concepts might react to the challenge posed by Boolos' result.

Keywords: Boolos, Quine, Logical Truth, Logical Consequence

1 Introduction

Philosophers and logicians have been wary of Quine's definition of logical truth since he first introduced a version of it in his *Truth by Convention*. Counterexamples were quickly found and prompted Quine to further refine his definition. In what follows, I will focus on the definition he seemed to have settled on, the one he proposed in his Quine 1970.[1]

The paper will be organized as follows. In the first few sections, I will briefly review the standard Tarskian as well as Quine's more idiosyncratic approach to logical truth and consequence while also highlighting their commonalities. I will then discuss some qualifications concerning the Quinean approach, which had been at the centre of the criticism. The next sections are devoted to setting the stage for a presentation of Boolos' result (established in Boolos 1975), implying that Quine's notion of logical truth does not properly generalize to the concept of logical consequence, followed by a detailed and self-contained presentation of Boolos' proof. In the remainder of the paper, I will address possible strategies to cope with the situation ensuing from Boolos' result.

[1] Quine's *Truth by Convention* is reprinted in Quine 1966, pp. 70–99. A revised version of his definition of logical truth was presented in his *Carnap and Logical Truth*, reprinted in Quine 1966, pp. 100–125. Hinman et al. 1968 and Berlinski and Gallin 1969 provide some critique as well as clarification. For a philosophical discussion compare Saguillo 2001. The final version of his substitutional definition of logical truth, the one we shall be concerned with for the most part, is discussed in Quine 1970, p. 48. More recently McKeon 2004 defends a version of Quine's account, which seems to draw on another Quinean definition of logical truth, the one defined 'in terms of grammar' (Quine 1970, p. 58). In this paper, I will exclusively deal with *Quine's* substitutional definitions of logical truth. Hence, I have little to say about other philosophers who have proposed similar definitions. In doing so, I obviously do not want to diminish the accomplishments of these other philosophers, like, for instance, Bolzano. For an illuminating discussion on Bolzano's account of logical truth and its relation to Quine's, see Lapointe 2014.

2 Common Ground

In what follows, we will be concerned with interpreted first-order languages, where an interpreted language is specified by a grammar together with a semantics. More specifically, we assume that the grammar of such a language is fixed by first delineating a class of *logical* terms and a class of *non-logical* or *lexical* terms. For the purpose of this article, we will assume that the logical terms are exhausted by the truth-functional connectives and first-order quantifiers—i.e. quantifiers that bind individual variables only.[2] The non-logical terms, consisting of some set of n-ary predicate symbols, are supposed to express particular properties specific to a given discourse. Atomic formulas are formed by applying n-ary predicates to variables and, finally, the class of well-formed formulas (or simply *formulas*) is defined as the class that results from closing the atomic formulas under truth-functional connectives and quantification. Sentences are formulas that do not contain free variables.

In order to get a full-fledged language—i.e. an *interpreted* language—we then supply the grammar with a *semantics*. That is, we define, as Tarski taught us in his celebrated *The Concept of Truth in Formalized Languages*, a truth predicate that applies to sentences of such a language.[3] In order to do so, we have to assume that we are given a set D, which is supposed to provide the range of the quantifiers and an interpretation function I, which assigns subsets of D^n to the primitive n-ary predicates. (In ordinary discourse, mathematical or otherwise, the domain and the interpretation function are fixed by the context. This is commonly referred to as the 'intended interpretation' of the language in question.) For each language L of the kind just described, we can then define a truth predicate in the following way: We first define recursively the *satisfaction relation* 's satisfies φ', a dyadic relation holding between assignments s of objects of the domain to the individual variables and interpreted formulas of L. Finally, using Tarski's trick, we stipulate that a sentence (i.e. a formula not containing any free variables) is true if it is satisfied by *all* assignments of objects and false if there is *no* such assignment.

3 Tarskian vs. Quinean Logical Truth

Up until now, this is common ground. However, we shall soon see that the ways depart from here on out. But before we turn to the Tarskian and Quinean explications of the informal concept of logical truth, let us briefly examine some of its traits.

There are a number of proposals as to what the key features of logical truth are, but two of the most common informal conceptions are the following:

(1) A logical truth is a sentence that is true regardless of what the non-logical terms occurring in it *mean*

(2) A logical truth is a sentence that is true in virtue of its *logical form*

Informal as these descriptions may be, they provide some hints about how logical truth is to be explicated. But, of course, they leave some room for disagreement. For instance,

[2] Of course, determining what the logical constants are is regarded by many philosophers as *the* main problem when it comes to the task of defining metatheoretical notions, such as logical truth and consequence. See, for instance, Tarski 1986.

[3] See Tarski 1956, pp. 152–278.

some philosophers think the notion of a 'logical law' should be explicated in terms of *correct inferences*. Considerable effort has been spent in developing what has come to be called *proof-theoretical semantics*. The majority, however, believes that an explication of logical truth should be spelled out in terms of *semantic* features, such as denotation, truth or satisfaction. Both conceptions have their merits, but in the following we shall exclusively consider explications that are based on variations of a broadly semantic approach. As we shall see, even here, it is the details that matter.[4]

The standard approach to logical truth and consequence, again essentially due to Tarski, proceeds by giving priority to the first informal conception. Instead of considering only *one* particular interpretation—i.e. the intended interpretation—we also take into account *reinterpretations* of the language in question. Here, a reinterpretation (in the following, the 're-' will be suppressed) is the same kind of thing as the *intended* interpretation.

Definition. *A **Tarskian interpretation** (T-interpretation for short) $\boldsymbol{M}$ for some language L is a pair $\langle D, I\rangle$, consisting of a set D, providing the range of the quantifiers and a function I, assigning a set $P^I \subseteq D^n$ to each n-ary predicate symbol P.*

In complete analogy with the earlier definition of truth *simpliciter*, we may then define a relative concept of truth—*viz. truth in a T-interpretation*. The only difference from the definition of truth for some interpreted language L, as outlined earlier, is that we now allow the 'built-in interpretation' of L to vary. More specifically, we define a triadic satisfaction relation 's satisfies φ relative to the interpretation $\mathbf{M}$' in the following way:[5]

Definition. ***Truth in a T-interpretation**:*

(i) $\boldsymbol{M}, s \vDash_T \varphi$ *iff.* $\langle s(x_1), \ldots s(x_n)\rangle \in P^I$, *if* $\varphi = Px_1 \ldots x_n$ *is atomic, then*

(ii) $\boldsymbol{M}, s \vDash_T \neg\varphi$ *iff.* $\boldsymbol{M}, s \nvDash_T \varphi$

(iii) $\boldsymbol{M}, s \vDash_T (\varphi \to \psi)$ *iff.* $\boldsymbol{M} \nvDash_s \varphi$ *or* $\boldsymbol{M} \vDash_T \psi$

(iv) $\boldsymbol{M}, s \vDash_T \exists x\varphi$ *iff. there is an x-variant s' of s, such that:* $\boldsymbol{M}, s' \vDash_T \varphi$

Finally, we stipulate that an L-sentence φ is true in $\boldsymbol{M}$, $\boldsymbol{M} \vDash_T \varphi$, if φ is satisfied by any $\boldsymbol{M}$-assignment s.[6]

Again, it is evident that the only difference between the truth definition outlined earlier and the definition of truth in a T-interpretation is that we now regard the interpretation $\mathbf{M}$ as

[4]There is a further informal conception of logical truth, which is sometimes referred to as the 'modal conception'. According to this conception, a sentence is logically true if it is *impossibly false*. For a general critique of the Tarskian definition of logical truth (which equally applies to Quine's), particularly concerning the question of whether Tarski's definition accounts for this modal intuition, see Etchemendy 1990, Sher 1996 and Hanson 1997. For a general outline of proof-theoretical semantics, see Prawitz 1974. Further discussion on this project can be found in Schroeder-Heister 2006 and Kahle and Schroeder-Heister 2006.

[5]In keeping with the Quinean spirit of economy, we will regard the universal quantifier and other truth-functional connectives as being defined in the usual way. We also assume that our language does not contain function symbols or names. This, again, is not a serious limitation since names and function symbols can be eliminated in favour of predicates. The focus on languages lacking these devices has a further purpose, which will be discussed in Section 4.

[6]Here, an $\mathbf{M}$-assignment s is a function that assigns objects of the domain of the interpretation $\mathbf{M}$ to the variables of the language. An x-variant s' of a given $\mathbf{M}$-assignment s is an assignment that agrees with s, except (possibly) for the variable x.

an explicit, metatheoretic variable. In other words, Tarski's definition of truth *simpliciter* is simply a particular instance of the relative concept of truth—namely, the special case where **M** is the intended interpretation. Finally, the concept of logical truth based on Tarskian interpretations is defined as follows:

Definition. *An L-sentence φ is a* ***T-logical truth*** *or* ***T-valid*** *if φ is true in each T-interpretation.*

Thus, the idea of the Tarskian definitions is straightforward. If we think of the domain of the interpretation as the meaning of the quantifiers and the interpretation function I as specifying extensional meanings for the non-logical terms, a sentence being true in each T-interpretation is true regardless of what the non-logical terms occurring in it mean, in accordance with the informal characterization (1) mentioned earlier. Furthermore, the concept of logical truth thus defined can be generalized immediately to the more general notions of logical consequence and satisfiability in the following way:

Definition. *A sentence φ is a* ***T-logical consequence*** *of Γ, $\Gamma \vDash_T \varphi$ for short, if φ is true in each T-interpretation in which each sentence in Γ is true and a set of sentences Γ is* ***T-satisfiable*** *if there is some T-interpretation in which each sentence in Γ is true.*

Before we go on, two somewhat obvious observations may be in order. First, it should be clear that T-logical consequence and T-satisfiability are interdefinable since a sentence φ is a T-logical consequence of Γ if $\Gamma \cup \{\neg\varphi\}$ is not T-satisfiable. Conversely, a set of sentences Γ is T-satisfiable if there is at least one sentence that is not a T-logical consequence of Γ. Second, the importance of the concepts of T-logical consequence and T-satisfiability lies in their general range of applicability. If the set Γ in the definitions of T-logical consequence and T-satisfiability is assumed to be *finite*, both concepts can be reduced to T-logical truth since we can stipulate that φ is a T-logical consequence of the premises $\varphi_1, \ldots, \varphi_n$ if the sentence $(\varphi_1 \wedge \ldots \wedge \varphi_n) \rightarrow \varphi$ is a T-logical truth. Hence, the interest in the general concepts of logical consequence and satisfiability lies in the fact that Γ may be *any* set of sentences, finite or infinite, 'reasonably specified' or not.

Now what's important for our purposes is that, in the Tarskian definitions of logical truth and consequence, we are quantifying in the metatheory over *sets*, because quantification over T-interpretations *ipso facto* means quantification over all sets that might be the domain of some T-interpretation. Moreover, since predicates are T-interpreted by subsets of the given domain, quantification over T-interpretations *ipso facto* means quantification over each subset of such a domain. So the Tarskian, model-theoretic concepts of logical truth and consequence are intimately bound up with set theory, a fact that Quine finds spurious. According to Quine, in defining such a basic concept as logical truth, we should try to get along without sets as far as possible. This is not to say that Quine thinks there is a fundamental problem with set theory. Nor does he think that there is something 'intrinsically wrong' with the Tarskian, model-theoretic definitions of logical truth and consequence. But, from Quine's point of view, a concept as basic as that of validity should be as independent as possible from advanced branches of mathematics like set theory. I will return to this important point later on.

Quine's leading idea in order to avoid set theory is simple enough: As in the Tarskian conception, a sentence will be valid if it is true in every interpretation. The difference between both conceptions arises from Quine's modifying the notion of an *interpretation*.

Recall that a Tarskian interpretation was specified by providing a certain set as domain together with specifications of the extensions of the non-logical vocabulary. A Quinean interpretation, however, is given in a more 'syntactical' fashion. Giving priority to the second informal characterization of logical truth mentioned earlier—that a sentence is valid if each sentence with the same logical form is true (where the logical form of a sentence is determined by its truth-functional/quantificational structure)—Quine stipulates that a sentence φ is a logical truth if each *substitutional instance* of φ is true. In other words, φ is a logical truth if by simultaneous, uniform replacement of the non-logical terms in φ by (complex or simple) terms of the same grammatical category, we *only* get truths. In order to make this more precise, we first define the notion of a *Quinean interpretation* in the following way:

Definition. *A* ***Quinean interpretation*** *(Q-interpretation for short) for an interpreted language L is a function $\boldsymbol{J}$ that assigns formulas $\varphi_P(x_1, ...x_n)$ to each non-logical predicate P of L.*

Similarly to Tarski, we then define a relative concept of truth by stipulating that a sentence φ is true in a Q-interpretation $\mathbf{J}$ if the respective substitutional instance $\varphi^{\mathbf{J}}$ is *true simpliciter*. Bearing in mind that we are dealing with interpreted languages L (so that *true simpliciter* means 'true with respect to the intended interpretation $\mathbf{M}$ of L'), we can proceed as follows:

Definition. ***Truth in a Q-interpretation****: We first recursively define the notion of a substitution instance $\varphi^{\boldsymbol{J}}$ of a formula φ with respect to some given Q-interpretation $\boldsymbol{J}$:*

(i) $\text{If } \varphi = Px_1...x_n \text{ is atomic, then } \varphi^{\boldsymbol{J}} = \varphi_P(x_1, ...x_n)$

(ii) $(\neg\varphi)^{\boldsymbol{J}} = \neg\varphi^{\boldsymbol{J}}$

(iii) $(\varphi \rightarrow \psi)^{\boldsymbol{J}} = (\varphi^{\boldsymbol{J}} \rightarrow \psi^{\boldsymbol{J}})$

(iv) $(\exists x\varphi)^{\boldsymbol{J}} = \exists x\varphi^{\boldsymbol{J}}$

Finally, we stipulate that a sentence φ is true in a Q-interpretation $\boldsymbol{J}$, $\boldsymbol{J} \vDash_Q \varphi$, if $\boldsymbol{M} \vDash_T \varphi^{\boldsymbol{J}}$.

Again, this definition merely captures the idea that a sentence φ is true in some Q-interpretation $\mathbf{J}$ if its translation $\varphi^{\mathbf{J}}$ is true with respect to the intended interpretation $\mathbf{M}$. The Tarskian definition of logical truth can then be conferred almost verbatim:

Definition. *An L-sentence φ is a* ***Q-logical truth*** *or* ***Q-valid*** *if φ is true in each Q-interpretation.*

Here again, two observations may be in order. First, although it is an important difference that a domain D is explicitly mentioned in the Tarskian definition, but not in the Quinean, at least for the present purpose, this difference is inessential. In order to guarantee the correctness of the Quinean definition of logical truth, heavy constraints have to be imposed on the object language anyway. These constraints will guarantee that the lack of domain variation will do no harm.[7] Second, the salient feature that Quine finds so troublesome with

[7]It should be mentioned, however, that the difference is not as 'incidental' (Quine 1970 p. 52) as Quine wants his readers to think it is. Domain variation is considered by many philosophers to be *the* distinctive feature of the modern (model-theoretic) conception of semantics and, therefore, an essential prerequisite for

the Tarskian definition of validity—*viz.* that it invokes *set theory*—is no longer present. We are no longer quantifying over *sets* but only over *expressions* of the given object language.

Although Quine does not do so (and for a good reason, as we shall see later), the concepts of logical consequence and satisfiability may be defined in straightforward analogy to the Tarskian ones as follows:

Definition. *A sentence φ is a* ***Q-logical consequence*** *of Γ, $\Gamma \vDash_Q \varphi$ for short, if φ is true in each Q-interpretation in which each sentence in Γ is true and a set of sentences Γ is* ***Q-satisfiable*** *if there is some Q-interpretation in which each sentence in Γ is true.*

Note that, at least *prima facie*, the Quinean concept of logical truth (and, accordingly, the Quinean concepts of satisfiability and consequence) can claim to be just as good a candidate for explicating the pre-theoretic concept of logical truth as the Tarskian. Both accounts seem to capture important traits of the informal notion of logical truth. Tarski, somewhat privileging the semantic feature (1) mentioned above, explicates logical truth in terms of reinterpretations of the *denotations* of the primitive vocabulary. Quine, somewhat privileging the formal feature (2), explicates it in terms of substitutions of *expressions* of the primitive vocabulary. In his *On the Concept of Logical Consequence* Tarski notably considered a substitutional account of logical truth quite similar to that of Quine. Yet, Tarski rejects it for reasons relating to possible limitations of the resources available in a given object language.[8] We shall now examine this issue more closely.

4 Sufficiently Rich Languages

Before we can go on, a number of comments regarding the sort of languages for which Quine wants to define validity have to be made. First of all, it is common to include linguistic devices such as names and function symbols in first-order languages. The standard language of arithmetic, for instance (which will be important later on), includes a name 0, designating the number zero, as well as function symbols S, $+$ and $\times$, designating the successor, addition and multiplication functions, respectively. In a Quinean 'standard language', on the other hand, no such devices are supposed to occur. It is well-known that Quine eschews names and function symbols for principled reasons. From his point of view, robust ontological commitment should be expressed in terms of quantifiers and variables instead of names that *presuppose* the existence of the objects named. Similarly, according to Quine, predicates should be used instead of function symbols. As is well-known, this is not an actual restriction since names and function symbols can be eliminated in the usual way.[9]

the concepts of logical truth and consequence. To make T-interpretations and Q-interpretations look more similar, we could define a Q-interpretation to be a pair $\langle \psi_D(x), \mathbf{J} \rangle$, where $\mathbf{J}$ is as before and $\psi_D(x)$ is a 'domain formula' that 'reinterprets' the range of the quantifiers. A substitution instance for a quantified formula $\exists x \varphi$ may then be defined as a formula of the form $\exists x(\psi_D(x) \wedge \varphi^{\mathbf{J}})$. Quinean interpretations would then be much like *syntactical translations*. Yet, this creates other problems, both interpretational as well as technical. For instance, such an account cannot—at least in a straightforward way—claim to elucidate the idea of logical truth as *truth in virtue of logical form*, unless one is willing to accept that $\exists x \varphi$ has the same logical form as $\exists x(\psi_D(x) \wedge \varphi^{\mathbf{J}})$. I am grateful to Jason Turner for drawing my attention to this point. Jason also deserves credit for various suggestions concerning notation in Section 4.

[8] Tarski's important paper is reprinted in Tarski 1956, pp. 409–420. His definition (F) corresponds to Quine's substitutional definition.

[9] It should be mentioned that it is not strictly necessary stick to such 'quinized' languages in what follows. However, it will facilitate the discussion in some respects and adapt to Quine's views on what a properly

A more important constraint on the languages to which Quine's definitions are supposed to apply is that, contrary to most standards of logicality, the identity sign cannot be counted as a logical notion. It has been pointed out by many commentators that this restriction is necessary in order to avoid overgeneration.[10] The sentence $\exists x \exists y(x \neq y)$, for instance, is true in each interpreted language whose built-in domain contains more than one object. Hence, if identity were to count as a logical constant and, therefore, not open to substitution, $\exists x \exists y(x \neq y)$ would, on the Quinean account, not only be true, but *logically* true. But this just seems wrong, for it is not a matter of logic to tell us how many things there are. Considerable effort has been spent on explaining away problems of this kind. The major alternatives here are to either argue that Quine is, for independent reasons, not committed to the view that identity is a logical constant anyway, or by modifying Quine's account in order to admit at least the *possibility* of counting identity as purely logical. *Prima facie*, both seem to be live options, for Quine himself wavers regarding the question of whether or not identity should count as a purely logical notion.[11] However, I will not discuss this issue here any further since there are more serious problems with Quine's account. In the following I will simply assume that identity is not a purely logical notion.

Another significant feature of the languages to which Quine's definitions are to be applied is that they have to be heavily constrained in terms of their 'ontology' (how big is the built-in domain of the language?) as well as their 'ideology' (what can be expressed by means of the basic concepts of the language?). For instance, we do not want a sentence like

$$\forall x \forall y \forall z(Rxy \wedge Ryz \rightarrow Rxz) \wedge \forall x \exists y Rxy \rightarrow \exists x Rxx$$

to be logically true. Or, at any rate, *Quine* does not want this. But if the built-in domain of the language in question is finite, this sentence *will* be a logical truth. Similar 'counterexamples' to Quine's definition can be found if the language, though rich enough with respect to its ontology, is not sufficiently *expressive*.

The Quinean 'solution' to these problems is simply to exclude such languages as too impoverished to be worthwhile defining metatheoretical concepts such as logical truth for them. But this gives rise to the question: For which languages *does* Quine's definition yield an informally correct characterization of validity? Now, if we agree that the Tarskian concept of logical truth indeed correctly characterizes the pre-theoretic notion of validity, there is a large class of languages for which it can be *proved* that the Quinean concept is correct, namely languages extending the language of elementary number theory. In other words, it can be proved that the Tarskian and the Quinean concepts of logical truth agree for these languages, a result we shall call the *weak Tarski-Quine equivalence theorem* (**WTQE** for short). But before looking at the proof, let us fix some notation for later.

As I mentioned earlier, the language of elementary number theory usually contains a constant 0 and function symbols for successor, addition and multiplication functions, respectively. Using the identity sign (and, as usual, boldface letters **n** as metatheoretical abbreviations for expressions of the form '$S(S...(0))$', where 0 is preceded by n occurrences

regimented language should look like. Aside from that, in presenting Boolos' result in Section 8, I wanted to stay as close as possible to his original proof as presented in his Boolos 1975.

[10] E.g. Boolos 1975, Hinman et al. 1968 or Hanson 1997. See also Quine's discussion of identity in his 1970, pp. 61–64.

[11] McKeon 2004 provides an extensive discussion of Quine's view on identity. Also, in his modified version of Quine's account of logical truth (cf. McKeon 2004, p. 219), identity may indeed be taken to be logical.

of S), we can formally express (or, rather, metatheoretically abbreviate the expressions of) equations like '$2+3=5$' by '$\mathbf{2}+\mathbf{3}=\mathbf{5}$'. As we saw earlier though, none of these expressions is allowed to occur in a Quinean 'standard language'. In what follows, we will accordingly only consider languages that are adequate for arithmetic, where names and function symbols have been eliminated in the usual way. More specifically, we shall be concerned with languages that contain at least a one-place predicate Zx true of 0 (only), a two-place relation Sxy true of all and only pairs $\langle 0,1\rangle, \langle 1,2\rangle, ...$ of consecutive numbers, and three-place predicates $Axyz$ and $Mxyz$ true of all and only triples $\langle n,m,k\rangle$ for which $n+m=k$ and $n\times m=k$ respectively.[12] In particular, instead of using standard numerals, we have to use numerical *predicates* like $\mathbf{2}(x)=\exists x_0\exists x_1(Zx_0\wedge Sx_0x_1\wedge Sx_1x)$, expressing the property of *being the number* 2. Hence, in a language meeting Quine's standards, '$2+3=5$' can be expressed by

$$\forall x\forall y\forall z(\mathbf{2}(x)\wedge\mathbf{3}(y)\wedge\mathbf{5}(z)\rightarrow Axyz)$$

More generally, the standard numerical predicate $\mathbf{n}(x)$, applying to only the natural number n, is defined by the formula

$$\exists x_0...\exists x_{n-1}(Zx_0\wedge Sx_0x_1\wedge ...\wedge Sx_{n-1}x)$$

A sentence having the form $\varphi(\mathbf{n})$ in the usual language of arithmetic will accordingly be expressed by a sentence of the form

$$\forall x(\mathbf{n}(x)\rightarrow\varphi(x))$$

In the following, we shall refer to Quinean languages that are adequate for elementary arithmetic as *sufficiently* or *reasonably rich* languages.

5 Completeness, Löwenheim and a 'Remarkable Concurrence'

Quine's proof of **WTQE**, i.e. the theorem stating that T- and Q-logical truth agree for sufficiently rich languages, depends crucially on two famous theorems concerning first-order logic. The first is the *weak completeness theorem* and the second is a strengthening of a theorem first proven by Leopold Löwenheim, which Quine sometimes refers to as the *Hilbert-Bernays-Löwenheim theorem*. Let us look at weak completeness first.

Up until now, we have not mentioned syntactical proof procedures at all. Both T- and Q-logical truth (as well as T- and Q-logical consequence) were defined in essential reference to the concept of *truth*; hence, both are based on a semantic notion. Naturally, however, one is also interested in the concept of *derivability* with respect to some calculus, in which theorems can be proved by using some basic set of axioms and inference rules without having to consider interpretations or the like. Plenty of deductive calculi are available for this purpose, such as calculi of natural deduction, Hilbert-style calculi, sequent or resolution calculi. A sentence is defined to be a *theorem* if it is derivable, in finitely many steps, from the empty set of premises by means of the syntactically specified rules of the calculus in

[12] If required, identity could be introduced either as a primitive non-logical predicate or, if there are only finitely many predicates in the language, definitionally, as suggested by Quine in his 1970, p. 63.

question. (As usual, we will be silent about the particular calculus and simply write $\vdash \varphi$ if φ is derivable from the empty set of premises and $\Gamma \vdash \varphi$ is φ is derivable from the premise-set Γ.) Given some language L, we then have the following

Theorem. *Weak Completeness Theorem (**WC** for short): For every L-sentence φ: if $\vDash_T \varphi$, then $\vdash \varphi$.*

Thus, each T-logical truth is derivable in any of these calculi. Note that completeness in this sense is a relational property, holding between some *semantic* notion of logical truth and some *syntactic* notion of derivability. As semantic notions of logical truth and consequence are usually considered to be more basic than syntactic ones, what the completeness theorem (in combination with the soundness of the given proof procedure) establishes is that the complex notion of T-logical truth, defined in terms of *all* T-interpretations (which is a lot!) can be 'reduced' to the much more mundane concept of derivability.

The second theorem that will play a decisive role in Quine's argument is a strengthening of the Löwenheim theorem. Recall that the Löwenheim theorem states that each T-satisfiable sentence is already true in some *numerical* T-interpretation—i.e. some T-interpretation whose domain consists solely of the natural numbers $\mathbb{N}$. Since the T-logical truth of φ is just the non-T-satisfiability of $\neg\varphi$, this implies that a sentence is a T-logical truth if and only if it is true in all *numerical* T-interpretations.

What Hilbert and Bernays proved is that we can do even better than that. Not only is a sentence a T-logical truth if true in all numerical interpretations, it is already a T-logical truth if true in each numerical T-interpretation where all the primitive concepts are interpreted by *definable* sets of natural numbers. Here, a set of natural numbers A is said to be *definable* if there is some formula $\varphi(x)$ in the language of elementary arithmetic, such that for each natural number n: $n \in A$ if and only if $\varphi(x)$ *is true of* n. Call a T-interpretation $\langle \mathbb{N}, I \rangle$ *denumerical* if the interpretation function I assigns arithmetically definable subsets of $\mathbb{N}$ to the primitive concepts of the language L in question. We can then state the Hilbert-Bernays-Löwenheim theorem as follows:[13]

Theorem. *Hilbert-Bernays-Löwenheim Theorem (**HBL** for short): If an L-sentence φ is true in every denumerical T-interpretation, then φ is T-valid.*

Therefore, with regard to sufficiently rich languages, the difference between denumerical T-interpretations and Q-interpretations essentially vanishes since each denumerical T-interpretation gives rise to a substitutional interpretation and *vice versa*. Call this fact **OBV**.

With everything in place, Quine proves the following

Theorem. *Weak Tarski-Quine Equivalence Theorem (**WTQE** for short): For each sufficiently rich language L and each L-sentence φ: φ is a T-logical truth if and only if φ is a Q-logical truth.*

> *'Proof'.* This is an immediate consequence of what has been said so far. Using the abbreviations Val_T, Val_Q, Val_D and Der for *T-validity*, *Q-validity*, *validity in all denumerical T-interpretations* and *derivability* respectively, we know that $Val_T(\varphi)$ implies $Der(\varphi)$ by **WC**. Then, because of the soundness of the

[13] For a detailed proof, see Kleene 1952 p. 394.

usual proof procedures with respect to Quinean logical truth, $Der(\varphi)$ implies $Val_Q(\varphi)$.[14] Thus, Tarskian validity implies Quinean validity. On the other hand, $Val_Q(\varphi)$ implies $Val_D(\varphi)$ by **OBV**. But, by **HBL**, $Val_D(\varphi)$ implies $Val_T(\varphi)$ and so Quinean validity also implies Tarskian validity. ■

Hence, Quinean and Tarskian logical truth agree with respect to sufficiently rich languages. In particular, all of the usual deductive calculi are weakly complete with respect to Quinean logical truth. But what is the philosophical payoff of this result? In the next section, we will flesh out the sketchy remarks from Section 3.

6 Quine on the Significance of WTQE

As we have already hinted at in Section 3, the major advantage of the Quinean definition of logical truth—at least from Quine's point of view—is simply *less ontology*. It is well-known that, from the very beginning of his career, Quine had always been anxious to avoid 'ontological excess'. We all know his memorable confession at the beginning of *Steps towards a Constructive Nominalism*, written together with Goodman: 'We don't believe in abstract entities.' (cf. Goodman and Quine 1947) Although he eventually came to reject his initial nominalism in favour of some kind of pragmatic realism, the question of ontology remained a prominent issue in Quine's writings. In fact, an issue he has been quite obsessed with. Hence, the one huge advantage Quine sees in his approach to logical truth is that, whereas the Tarskian definition ontologically commits us to a universe of sets and makes the definition of logical truth dependent on some *theory* of sets, his own substitutional definition does not. As he writes in his 1970 on p. 55: 'The evident philosophical advantage of resting with this substitutional definition, and not broaching model theory, is that we save on ontology. Sentences suffice, sentences even of the object language, instead of a universe of sets specifiable and unspecifiable.'[15] In order not to get a wrong impression, it should be noted that, for Quine, the fact that the model-theoretic account of logical truth rests on set theory is not problematic *per se*. Nevertheless, according to Quine, it has certain shortcomings. For one thing, there are several mutually incompatible set theories, none of which can claim to be *the* theory of sets. From Quine's point of view, there is simply no principled reason to prefer one set theory over another. Various set theories may be equally adequate for the purpose of science in the sense that they yield the mathematics that is necessary for doing physics, chemistry or biology. Yet, different set theories may radically differ regarding which sets are postulated to exist in addition and, therefore, may

[14] The soundness of our unspecified proof procedure with respect to Quinean logical truth is guaranteed by the fact that a derivation remains correct if the primitive vocabulary is replaced by formulas according to some Q-interpretation. Some of the proof procedures are 'visibly sound', as Quine writes in his 1970, p. 54.

[15] Now, this is not the whole truth. Quine's definition of logical truth rests on a standard (Tarskian) theory of *truth* for the object language in question. This much is obvious from the definition of 'truth in a Q-interpretation' given earlier. Such a theory of truth will itself need, as Quine is aware, some 'heavy equipment from set theory' (see Quine 1970, p. 42). More specifically, turning the inductive clauses of Tarski's truth definition into an explicit definition will require quantification over sets. However, a truth definition for first-order arithmetic requires, at least by common set-theoretic standards, not *that* 'heavy equipment' after all (a truth definition for first-order arithmetic can be expressed by a Σ^1_1-formula). Still, as we shall see, Quine is right that the 'ontological costs' do not increase when we pass from truth to *logical* truth on the substitutional definition, whereas, they *do* increase when we pass from truth to logical truth on the Tarskian definition.

lead to different concepts of logical truth (at least in principle). Moreover, even if there were some single privileged set theory, we still could not be absolutely sure that such a theory might not turn out to be inconsistent after all. The stronger the set theory adopted in the background, the more likely it is for it to be inconsistent. This seems to be hinted at when Quine writes concerning the merits of his substitutional definition that '[in] this way, when occasions arise for revising theories, we are in a position to favor theories whose demands are lighter' (Quine 1970, p. 55).

In Quine's account, we are ontologically committed to just *expressions*—for it is expressions that are substituted for expressions in the Quinean approach. Thus, the only thing that is required is a theory of syntax for the object language and the concept of truth. As Quine famously puts it at the end of the chapter on logical truth in his 1970: 'Logic is, in the jargon of mechanics, the resultant of two components: grammar and truth.' Of course, a theory of syntax (a 'classical, infinite theory of finite strings of signs') will be equivalent to some weak set theory. But the emphasis is on 'weak'.[16] Quine, being aware of this, claims that the retreat to 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' (Quine 1970 p. 56). Hence, according to Quine, part of the significance of **WTQE** derives from its ensuring that nothing is lost when we define the concept of logical truth in the ontologically more parsimonious way suggested by him. **WTQE** assures us that we can stick to either definition without having to pay the ontological costs.

But a further motive is important to Quine, one that is closely tied to his conviction that logic is in some sense exhausted by first-order logic. This verdict is again partly justified on grounds of ontological commitment. First-order logic ontologically commits us to nothing but one object. And even this should not be misunderstood as some 'philosophical dogma about necessary existence' (Quine 1970, p. 52). Set theory, on the other hand (or higher-order logic, which, from Quine's point of view, is just 'set theory in sheep's clothing'), *does* have substantial ontological commitments. But besides these ontological considerations, Quine has another point to make in this regard when he writes that 'a remarkable concurrence of diverse definitions of logical truth [...] already suggested to us that the logic of quantification as classically bounded is a solid and significant unity' (Quine 1970, p. 91). So the 'remarkable concurrence of diverse definitions of logical truth', established by **WTQE** (between Q- and T-logical truth) and **WC** (between both T- and Q-logical truth and derivability), is simply more grist on Quine's mill. It seems to provide a further reason to draw the line between logic and mathematics the way Quine wants it to be drawn, for no such 'remarkable concurrence' can be shown for, say, higher-order logic. Hence, Quine's view seems to be that weak completeness is an essential ingredient for something called 'logic'.[17] But why, one might ask, should we take *weak completeness* to be the deciding factor? Why not, say, *decidability*? Decidability seems to be a property that is just as important as completeness. Moreover, why *weak* and not *strong* completeness?[18] Recall that *strong completeness*

[16]It is well-known that Robinson's arithmetic Q, for instance, suffices for a development of a theory of syntax. Q, in turn, can be interpreted in set theories like 'general set theory' (mentioned by Boolos in his 1987) and even weaker systems.

[17]With respect to the 'scope of logic', Quine, in a response to Hao Wang, mentions three criteria that contributed to his decision to demarcate logic from mathematics (which he identifies with *set theory*) the way he does: ontological neutrality, completability and multiple set theories. See Schilpp and Hahn 1986 p. 646.

[18]These questions have already been raised by Boolos in his 1975, p. 524, and, eventually, led him to the

is concerned with the notion of logical consequence (or, equivalently, with the general notion of satisfiability). It relates the semantic concept of logical consequence with the notion of derivability with respect to *arbitrary sets of premisses* Γ. It is well-known that, for Tarskian logical consequence, we can indeed show that the following is true:

Theorem. *Strong Completeness Theorem: For every set of L-sentences* Γ *and each L-sentence* φ*: if* $\Gamma \vDash_T \varphi$*, then* $\Gamma \vdash \varphi$*.*

So in addition to assuring us of the concurrence of Tarskian validity and derivability from the empty set of premises, the strong completeness theorem (together with soundness) assures us of the equivalence of various syntactical notions of derivability from *any* set of premisses and Tarskian logical consequence. It is easy to see that strong and weak completeness are in fact equivalent if we assume the *compactness* of the semantic consequence relation under consideration. Here, a consequence relation $\vDash_X$ is called compact if $\Gamma \vDash_X \varphi$ always implies that there is some *finite* subset $\Gamma_0 \subseteq \Gamma$ such that $\Gamma_0 \vDash_X \varphi$. (In particular, each deductive consequence relation is compact *by definition*, since a derivation is specified as a finite sequence of formulas.) Equivalently, compactness is often expressed thus: If every finite subset $\Gamma_0 \subseteq \Gamma$ is X-satisfiable, then Γ itself is X-satisfiable. Therefore, *modulo* some further requirements that are met in classical first-order logic, the only thing that distinguishes weak from strong completeness is compactness.[19]

Now, what Boolos proved in the appendix to his 1975 is that Quinean logical consequence as defined earlier is *not* compact.[20] There are T-satisfiable (hence consistent) sets of sentences Γ of some sufficiently rich language L that are not Q-satisfiable, although each finite subset of Γ *is*. So, in particular, none of the standard notions of derivability is strongly complete with respect to Quinean consequence. This, however, seems to be a major blow to Quine's account in at least two respects. First, it seems to undermine Quine's argument that first-order logic is somehow privileged on the ground that first-order logic is complete. One might argue more or less successfully that completeness is the (or, at any rate, *one*) deciding factor regarding what should be counted as belonging to something that is properly called 'logic'. But it is hard to see why it should be weak, as opposed to strong completeness. But second, and more importantly, it leaves us with the question of the *correctness* of Quine's account. That some sets of sentences are T-satisfiable but not Q-satisfiable shows that the Tarskian and Quinean concepts do not agree. At least one of them has to be incorrect and it seems hard to swallow that it is *Tarski's* that goes wrong. But let us not prejudge the issue before even looking at Boolos' proof. I shall return to this topic in due course.

proof that we shall examine in Section 8.

[19] Suppose $\vDash_X$ is compact, weakly completable and $\Gamma \vDash_X \varphi$. By compactness, we have $\{\varphi_1, ...\varphi_n\} \vDash_X \varphi$ for some finite subset $\{\varphi_1, ...\varphi_n\} \subseteq \Gamma$. But this is equivalent to $\vDash_X \varphi_1 \wedge ... \wedge \varphi_n \rightarrow \varphi$. By weak completeness, we then have $\vdash \varphi_1 \wedge ... \wedge \varphi_n \rightarrow \varphi$. So, $\{\varphi_1, ...\varphi_n\} \vdash \varphi$ and, therefore, $\Gamma \vdash \varphi$ by the monotonicity of the deducibility-relation $\vdash$. Moreover, due to the interdefinability of satisfiability and logical consequence, the strong completeness theorem can be stated in the form: If Γ is consistent, then Γ is T-satisfiable. In what follows, I will switch between these versions without much comment.

[20] I am unsure about what Shapiro means when he writes in his 2000 that 'Quine's notion of logical consequence is compact because first-order logic is complete'. McKeon in his 2004, p. 219, suspects that Shapiro is, without explicitly noting it, talking about a related (but distinct) Quinean definition of logical consequence, the one 'in terms of grammar' (cf. Quine 1970 pp. 58–60).

7 Preliminaries to Boolos' Appendix Argument

To begin with, let's examine some basic notions from computability theory. (Readers already familiar with this material may safely skip this section.) The central concept of computability theory is that of a computable number theoretic function—i.e. the notion of a function $f : \mathbb{N} \longrightarrow \mathbb{N}$, each value of which can be calculated in a finite number of steps. Various formal explications of this concept have been offered. Alan Turing was one of the first to define rigorously the notion of a computable function by introducing the concept of a 'computing machine'. Turing machines, as they are called nowadays, can be conceived of as (very) primitive computers that are able to perform certain (very) basic operations.[21] For the following, an informal grasp of a computable function as a function that can be algorithmically calculated and of a Turing machine as performing these calculations will do. We will derivatively call a *set* A of natural numbers computable (decidable, recursive) if its corresponding characteristic function

$$\chi_A(n) = \begin{cases} 1 & \text{if } n \in A \\ 0 & \text{if } n \notin A \end{cases}$$

is computable. Thus, a set of (n-tupels of) natural numbers is computable if membership in this set can be mechanically decided. The first thing we will need is the following

Fact 1. *If L is reasonably rich, then each computable function (each computable set) is definable in L.*

The proof of this is straightforward, but tedious. So I refer the interested reader to Boolos and Jeffrey 1974 to check the details.

We will also need the notion of a *Gödel numbering*. A Gödel numbering for an arithmetical language L is a function assigning 'code numbers' to finite sequences of symbols of L in some reasonable way. To be reasonable here means that the function itself as well as its inverse should be computable. (Otherwise, it would not serve the purpose of *coding* sequences. Clearly, we want to be able to restore the information coded by a number by following some mechanical procedure.) By means of Gödel numbers, we can then talk about expressions of some sufficiently rich language L *within* L itself. In particular, we can define formulas $Var(x)$, $Form(x)$, $Sent(x)$, etc., that define the set of Gödel numbers of variables, formulas, sentences etc. (for details, see again Boolos and Jeffrey 1974).

The last notion we will need in Boolos' proof is the concept of *relative computability*, which is based on the notion of an *oracle*. Recall that there are sets of natural numbers, even definable ones, that are not decidable.[22] But we can still make fair sense of the

[21] There are various other explications of the notion of computability (recursive functions, register machines, etc.). It can be shown that all of them are equivalent in that they determine the same class of functions. Therefore, if we further assume the Church-Turing thesis, stating that each informally computable function is Turing-computable, each explication is as good as the other in determining the class of intuitively computable functions. For details, consult the classical Boolos and Jeffrey 1974.

[22] A famous example is the set of Gödel numbers of sentences that are derivable from some recursive set of axioms Γ. Just as we are able to define (in some sufficiently rich language L) the sets of Gödel numbers of variables, sentences, etc., we can also define a predicate $Prov(x)$ being true of all and only the Gödel numbers of Γ-theorems. Yet, we know from Church's theorem that this set, call it P, is not *decidable*. That is, there is no effective algorithm that decides, of an arbitrary given sentence φ, whether its Gödel number is in P.

question of which sets of natural numbers *would* become decidable if we had a 'black box' for a given non-decidable set A. Such an oracle for A would immediately give us an answer to each question of the form 'Is n in A?'. 'Concatenating' such oracles with standard Turing machines, we arrive at the notion of *relative computability*. That is, we define a set B to be computable relative to some set A, if we can build a Turing machine that decides elementhood in B and which is permitted to ask a given oracle for A. More generally, B is computable relative to the sets $A_1, ...A_n$ if there is some Turing machine computing B, given oracles for the sets $A_1, ...A_n$. It can then be proved that, for sufficiently rich languages L, we have the following

Fact 2. *If $A_1, ...A_n$ are definable in L and B is computable relative to $A_1, ...A_n$, then B is definable in L.*[23]

The only further thing we need before we can look at Boolos' proof is Tarski's celebrated theorem on the undefinability of truth. Tarski's undefinability theorem, applied to arithmetical languages, simply states that arithmetical truth is not arithmetically definable. More generally, given some interpreted language L (with built-in interpretation $\mathbf{M}$) that extends the language of arithmetic, we have:

Theorem. *Tarski's Undefinability Theorem: There is no L-formula $Tr(x)$ such that for all sentences φ of L: $\boldsymbol{M} \vDash_T Tr(\ulcorner \varphi \urcorner)$ if and only if $\boldsymbol{M} \vDash_T \varphi$.*

Hence, there is no sufficiently rich language that contains a (simple or complex) predicate $Tr(x)$, which is true of all and only the Gödel-numbers of true sentences of this language.

Having everything in place then, let us look at Boolos' argument.

8 Boolos' Appendix Argument

In order to make it more vivid, we will split up Boolos' original proof into a series of lemmas. Our goal will be to establish the following

Theorem. *Boolos' Theorem: There are sufficiently rich languages L and sets of L-sentences $\mathcal{B}$ such that $\mathcal{B}$ is T-satisfiable, but not Q-satisfiable in L.*

First, we will describe the language L and the set $\mathcal{B}$, which we will show to be Q-unsatisfiable. In order to do so, we also need an auxiliary language L^*, which is syntactically identical to L, yet differs from L with regard to its built-in interpretation.

Being sufficiently rich interpreted languages, both L and L^* have $\mathbb{N}$ as their built-in domain and contain predicates Z, S, A and M standing for the zero, successor, addition and multiplication predicates respectively. Furthermore, both languages contain a predicate T. T, however, is interpreted differently in L and L^*. Whereas L specifies that T is to be true of all natural numbers, L^* specifies that T is true of all and only the Gödel numbers of the *true* sentences of L. Thus, T is a truth predicate for L in L^*. Let T^* be the extension of T in L^*. Now, consider the set

$$\mathcal{B} := \{\varphi \in L^* : \varphi \text{ is a true } L^*\text{-sentence}\}$$

[23] This follows from **Fact 1** together with two further facts: that (1) the 'concatenation' of Turing machines can be represented in a sufficiently rich language by the composition of computable functions representing these Turing machines and that (2) we can define the composition of two or more definable functions.

Evidently, $\mathcal{B}$ is T-satisfiable, because it is T-satisfied by the built-in interpretation of L^*. But from a syntactical point of view, $\mathcal{B}$ is also a set of L-sentences. Therefore, we may ask: Is $\mathcal{B}$, as a set of L-sentences, Q-satisfiable in L?

Suppose for *reductio* that it were. That is, assume that we are given some Q-interpretation $\mathbf{J}$ specified by L-formulas $\varphi_Z(x), \varphi_S(x,y), \varphi_A(x,y,z), \varphi_M(x,y,z)$ and $\varphi_T(x)$ such that each sentence in $\mathcal{B}$ comes out true in L when these formulas are substituted for the formulas Zx, Sxy, $Axyz$, $Mxyz$ and Tx. In what follows, we will use E_Z as shorthand for the set of natural numbers satisfying $\varphi_Z(x)$, E_S for the set of pairs satisfying $\varphi_S(x,y)$ and E_T for the set of numbers satisfying $\varphi_T(x)$.

In a first step, we will prove

Lemma 1. *(i) If $n \in T^*$, then '$\forall x(\boldsymbol{n}(x) \rightarrow Tx)$' $\in \mathcal{B}$*

(ii) If $n \notin T^*$, then '$\forall x(\boldsymbol{n}(x) \rightarrow \neg Tx)$' $\in \mathcal{B}$

> *Proof.* Since n is the only number satisfying $\mathbf{n}(x)$ and T is (in L^*) interpreted by the set T^*, '$\forall x(\mathbf{n}(x) \rightarrow Tx)$' just *says* that n is in T^*. So, '$\forall x(\mathbf{n}(x) \rightarrow Tx)$' will be a true L^*-sentence and, therefore, a member of $\mathcal{B}$. Similarly, '$\forall x(\mathbf{n}(x) \rightarrow \neg Tx)$' just *says* that n is *not* in T^*. Hence, if n is not in T^*, '$\forall x(\mathbf{n}(x) \rightarrow \neg Tx)$' will be a member of $\mathcal{B}$. ■

Next, we show

Lemma 2. *If k satisfies $[\boldsymbol{n}(x)]^{\mathbf{J}}$, then $k \in E_T$ iff. $n \in T^*$.*

> *Proof.* First, recall that $[\mathbf{n}(x)]^{\mathbf{J}}$ is the result of replacing each occurrence of Zx and Sxy in $\mathbf{n}(x)$ by the substitution formulas $\varphi_Z(x)$ and $\varphi_S(x,y)$ respectively. Now, suppose that k satisfies $[\mathbf{n}(x)]^{\mathbf{J}}$ and $n \in T^*$. Then, by **Lemma 1**, the sentence '$\forall x(\mathbf{n}(x) \rightarrow Tx)$' is in $\mathcal{B}$. Hence, by our assumption that $\mathbf{J}$ models the set $\mathcal{B}$, $[\forall x(\mathbf{n}(x) \rightarrow Tx)]^{\mathbf{J}}$ is true in L. Therefore, by the properties of Q-interpretations, $\forall x([\mathbf{n}(x)]^{\mathbf{J}} \rightarrow \varphi_T(x))$ is true in L. But, since k satisfies $[\mathbf{n}(x)]^{\mathbf{J}}$ and E_T is the extension of $\varphi_T(x)$, this sentence just says that $k \in E_T$. Hence, if k satisfies $[\mathbf{n}(x)]^{\mathbf{J}}$ and $n \in T^*$, then $k \in E_T$. By an analogous argument (using the second part of **Lemma 1**), one can easily show the converse as well—i.e. If k satisfies $[\mathbf{n}(x)]^{\mathbf{J}}$ and $n \notin T^*$, then $k \notin E_T$. ■

Next we show

Lemma 3. *T^* is computable relative to E_Z, E_S and E_T.*

> *Proof.* Suppose we are given oracles for E_Z, E_S and E_T and n is some arbitrary natural number. Then, the oracles for E_Z and E_S enable us to find some natural number k satisfying $[\mathbf{n}(x)]^{\mathbf{J}}$. (We ask the oracle for E_Z to find some number satisfying $\varphi_Z(x_0)$. If we have found such a number, say k_0, we ask the second oracle for a number k_1 satisfying $\varphi(k_0, x_1)$. Having found such a number k_1, we ask for a number k_2 satisfying $\varphi_S(k_1, x_2)$ and so on until we have found a number k satisfying $[\mathbf{n}(x)]^{\mathbf{J}}$.) We then ask the oracle for E_T whether $k \in E_T$. If the oracle answers positively, by **Lemma 2**, we may conclude that $n \in T^*$. If the

oracle answers negatively, we have $n \notin T^*$ by **Lemma 2**. Since the procedure outlined here is purely mechanical, T^* is computable relative to E_Z, E_S and E_T. ■

This finally yields

Lemma 4. *T^* is definable in L.*

Proof. This follows immediately from **Lemma 3** together with **Fact 2** and the fact that $\varphi_Z(x), \varphi_S(x, y)$ and $\varphi_T(x)$ define the sets E_Z, E_S and E_T in L. ■

By **Lemma 4** then, T^* would be defined in L by some L-formula $Tr(x)$. But recall that T^* is just the set of Gödel numbers of true L-sentences; hence, L-truth would be L-definable. But this flatly contradicts Tarski's theorem on the undefinability of truth, which states that no such formula exists. Therefore, we have to conclude that, contrary to our assumption, no Q-interpretation exists that makes each sentence in $\mathcal{B}$ true in L. Hence, $\mathcal{B}$ is not Q-satisfiable in L.

9 So, what now?

As we have seen, Tarskian and Quinean consequence and satisfiability do not coincide—even for sufficiently rich languages. The set of L-sentences $\mathcal{B}$ is a straightforward counterexample—for $\mathcal{B}$ is T-satisfiable, but not Q-satisfiable in L. So the Quinean concepts of satisfiability and logical consequence are not compact and, hence, none of the usual deductive calculi is strongly complete with respect to Quinean consequence. Boolos seems to take for granted that this in itself shows the Quinean concept of consequence to be inadequate. Boolos apparently believes that the result also sheds doubt on the adequacy of the Quinean concept of logical *truth*, for Quinean consequence as defined earlier is just its natural generalization to arguments with an arbitrary number of premises (see Boolos 1975, pp. 525–526). But it seems to me that this conclusion is premature. After all, what is established by Boolos' result is that two explications of a pre-theoretic concept of logical consequence (or two explications of two plausible pre-theoretic conceptions of logical consequence) do not concur. As far as I am aware, Quine never addressed the issue raised by Boolos' result in any of his published writings directly. In fact, Quine is rarely concerned with the general concept of logical consequence. Hence, the question that I will pursue in this section is this: How could we react to Boolos' result if we insist on preserving the Quinean concepts, or at least their spirit?

Two general strategies can be distinguished here. The first is to bite the bullet, take the result that the Quinean and Tarskian concepts diverge at face value, and still argue that this does not *a priori* go against the Quinean concepts. The second is to further refine (or redefine) the Quinean concepts in some principled way and try to bring them in line with the Tarskian concepts.[24]

[24] A third alternative is to deny that the Boolos set $\mathcal{B}$ is a 'genuine' counterexample. As is evident from Boolos' proof, the set $\mathcal{B}$ is defined in essential reference to the set of Gödel numbers of first-order truths of L (that is, the set T^*). As is well-known, this set is a fairly complex, non-recursive set of natural numbers. The idea then is that one can simply deny that Quine is committed to a theory that is strong enough to guarantee the *existence* of such a set. However, this reply does not seem to work since what is needed to

As to the first general alternative, we have seen that none of the standard deductive calculi is strongly complete with respect to Quinean consequence since Quinean consequence is not compact. The Boolos set $\mathcal{B}$ is a direct counterexample, for each finite subset of $\mathcal{B}$ is Q-satisfiable (because T-satisfiable, and, for finite sets, T-satisfiability and Q-satisfiability are equivalent), but the entire set $\mathcal{B}$ is not. But one might well ask: Why is compactness important anyway?

A defender of the Quinean concepts might argue along the following lines: Well, the fact that Quinean consequence is not compact (and, therefore, none of the standard calculi strongly complete with respect to Q-consequence) is interesting, but does not *per se* speak against Q-consequence. After all, both T- and Q-consequence have some antecedent plausibility as explications of the informal concept of logical consequence. Unless we have some further reason to prefer T- over Q-consequence, would it not be opportunistic to favour T-consequence simply on the ground that it yields a result we like? It is still remarkable that various concepts of logical truth agree for first-order logic, a fact that contributes—among other things—to regarding first-order logic as a 'significant and solid unity', thereby contributing to the decision to mark the difference between logic and mathematics here.[25] That various natural concepts of logical consequence (that is, natural generalizations of concurrent notions of logical truth) *diverge* is remarkable as well, but unless we provide some argument for why compactness is an indispensable feature of a 'correct' explication of the pre-theoretic notion of consequence, we have no reason to think that Tarski is 'right' and Quine 'wrong'. One might add that compactness is not only not indispensable for a correct explication of logical consequence, but not even *desirable*. Proponents of second-order logic, for instance, typically claim that compactness is a kind of defect of the Tarskian standard concept of first-order consequence.[26] So why would we want compactness anyway?

I think that this line of reasoning or something similar has at least *some* plausibility as far as it goes, but, as we shall see, under closer scrutiny, it will turn out to be incoherent—at

guarantee the existence of this set is (relatively) modest. We may, for instance, argue along the following lines: Since we are concerned with sets of natural numbers, a natural setting to measure the strength of the required theory is by considering subsystems of second-order arithmetic. (For details concerning subsystems of second-order arithmetic, consult Simpson 2009.) In this direction, one can prove, e.g. that the first-order consequences of the theory ACA_T = ACA_0 + 'The set of Gödel numbers of first-order arithmetical truths exists' are precisely those of ACA. (Both ACA and ACA_0 are subsystems of full second-order arithmetic, where the second-order comprehension scheme is restricted to first-order arithmetical formulas.) In terms of set-theoretical strength, it seems that ACA_T should be interpretable in, say, Kripke-Platek set theory (a subsystem of ZF set theory), a system that should not provoke any Quinean resentments about set-theoretical excess. Thus, at least in comparison to standard set theories like ZF, Boolos' counterexample does not seem to require *that much* set theory. In other respects, however, ACA_T is not that innocent after all. ACA_T (and even its subtheory ACA_0) *does* prove a good deal of ordinary analysis as is needed in, say, physics (for details, see again Simpson 2009). Therefore, whether the objection that Boolos' counterexample might not be 'genuine' is cogent cannot be judged lightly and much more work has to be done here (both technical and philosophical) than I am able to in this article. I am grateful to Chris Fermüller, Michael Rathjen and Michael Toppel for drawing my attention to and helping sharpen this (possible) objection.

[25] Even more can be said: Tarskian and Quinean logical consequence and satisfiability agree for infinite sets of sentences as well (if we assume a sufficiently rich object language), as long as they are *recursive*. See Kleene 1952, pp. 397–398. So the Tarskian and the Quinean concepts yield the same results if applied to any 'natural' set of premises (i.e. axiomatic theories like first-order number theory, set theory, etc.) as well. It seems that Quine must have been aware of this coincidence since, in his 1954, he explicitly refers to Kleene 1952 as one of his main sources. Maybe because of his awareness of this further concurrence, he was not too shocked by Boolos' result. As I will argue, however, this would be a mistake.

[26] See e.g. Van Dalen in his 1980, p. 107 or Stewart Shapiro in his 1991, p. 122 and p. 159.

least for someone close enough to Quine in spirit.

First of all and to state the obvious, I can hardly imagine that Quine would want to cite second-order logic to account for the informal non-satisfiability of certain sets of sentences. Second-order logical consequence and logical truth are typically rejected by Quine and his followers, essentially for two reasons: First, full second-order consequence is incompletable. Thus, there is no proof procedure that is complete with respect to second-order consequence (or even second-order logical truth). Second, as is well-known, second-order logic is, according to Quine, merely 'set theory in sheep's clothing' (Quine 1970, p. 66); hence, appealing to second-order consequence/satisfiability would reintroduce set theory through the back door once again. Besides, alluding to second-order satisfiability would not even help for the case at hand because the Boolos set $\mathcal{B}$ is a set of *first-order* sentences. Furthermore, $\mathcal{B}$ is not just satisfied by some outlandish T-interpretation, but by a 'standard interpretation', where quantifiers range over the natural numbers only and each arithmetical predicate has its usual meaning. $\mathcal{B}$ is even a *true* set of sentences and one might well wonder: What is an explication of satisfiability worth if not even *true* sets of sentences are declared to be satisfiable by such an account? It seems that no matter how odd a certain set of sentences might appear to be (finite or infinite, recursive or non-recursive), if we know it is *true* in its intended interpretation, it should not be unsatisfiable. So I take it that whatever argument one *might* come up with to rule out compactness as a genuine feature of the pre-theoretic notion of first-order satisfiability, it had better not rule out the informal satisfiability of sets like $\mathcal{B}$.

Furthermore, remember that none of the usual deductive calculi is strongly complete with respect to Q-logical consequence. This means that certain arguments with infinitely many premises that are valid in the Quinean sense are not deductively valid. Assuming that Q-logical consequence were indeed the 'correct' explication of the pre-theoretic notion of consequence, the usual deductive calculi would then have to be regarded as somehow *deficient*, since they would fail to capture certain 'valid' forms of inference. However, one is tempted to say that, in this particular case, it is not the deductive calculi that are to be blamed, but the Quinean consequence relation. (Unlike in the case of, say, a standard calculus of natural deduction for classical logic, which lacks a rule for negation introduction.) But, at the moment, the adequacy of the various deductive calculi with respect to *Tarskian* consequence cannot be used as an argument for this. Hence, we would have to provide independent reasons for the 'correctness' of the deductive calculi rather than the Quinean consequence relation. But there is a simpler argument to the effect that the problem is not with the deductive calculi. The point is simple: The notion of a deductively valid argument does not carry with it any restriction as to the number or nature of its premises. Derivability is defined for arguments with infinitely many premises in the same way as it is for arguments with finitely many (or none), namely, as the the existence of a finite sequence of formulas such that each formula is either a premise or the result of an application of some inference rule to formulas already established. The concept of derivability is blind, as it were, regarding the number or nature of the premises that are allowed to be used in a derivation. *Any* set of sentences can be used as a premise set. Therefore, *if* one acknowledges that the usual proof procedures are not deficient as applied to arguments with *finitely* many (or no) premises, one has to acknowledge that they are not deficient as applied to arguments with *infinitely* many premises as well. But this means that *if* the usual calculi are weakly complete with respect to a given notion of logical truth (and the usual calculi are indeed weakly complete

with respect to Q-logical truth), they *have to be* strongly complete as well.

The only other option I see is to admit that the notion of deductive consequence is, appearances notwithstanding, equivocal and has to be spelt out differently in case there are infinitely many premises. But at least from a broadly Quinean perspective, it is far from clear how this would lead to a notion of deductive consequence that can claim to be 'elementary' in any sense that would blend with Quine's general views on first-order logic.

Thus, it seems that we have reached an impasse as regards our first general line of reasoning. So perhaps we should stick to the second line—admit that the Quinean definitions are defective as they stand and try to refine them in some principled way. Three approaches may be distinguished:

(1) We further restrict the languages for which the Quinean concepts are applicable in order to rule out the possibility of not having available enough Q-interpretations

(2) We broaden the notion of a Q-interpretation to include possible extensions of a given object language

(3) We redefine Q-logical consequence and Q-satisfiability in a way that excludes counterexamples like $\mathcal{B}$ *by fiat* by building compactness into the definitions of these concepts

In the remainder of this section, I will discuss each of these alternatives and conclude that, at the end of the day, none of them is convincing.

9.1 Restricting Languages

Let us start with the first alternative. Recall that the substitutional approach to logical truth overgenerated when applied to languages containing identity as a logical constant. Certain sentences were wrongly declared to be logically true by the Quinean definitions. As a result, we ignored languages that contained identity as a logical constant. For similar reasons, we excluded languages that were too impoverished with respect to their ontology or their expressive richness. So one way out of our dilemma might be to further constrain the languages to which the substitutional concepts can be applied in order make them agree with their Tarskian counterparts after all. The task then would be to find a class of languages $\mathcal{L}$ such that for each language $L \in \mathcal{L}$ and each set of L-sentences Γ, Γ is Q-satisfiable in L if and only if Γ is T-satisfiable.

The trouble with this suggestion is that it just does not work. The Boolos argument is perfectly general and is not tied to a specific class of languages. Hence, we cannot fix in advance a certain range of languages and expect that, for each language L of this class (however rich those languages may be), Quinean and Tarskian satisfiability agree. We can set up the same conditions for L as in Boolos' proof: Simply add some dummy predicate T to L and consider another language L^* that contains T interpreted as a truth predicate for $L + T$. (Here and in the following, '$L + T$' stands for the interpreted language that results from adding a fresh, interpreted predicate T to the language L.) We will again obtain a 'Boolos set' for the language $L + T$—i.e. a set of sentences which is T-satisfiable, yet not Q-satisfiable in L.

9.2 Allowing Possible Extensions

The foregoing discussion indicates, however, how to overcome this problem by adopting alternative (2). The basic idea of this strategy is to not try to delineate a class of languages *in advance*, but to consider possible extensions of a language. Thus, on this account, substitutional formulas may be drawn from possible extensions L' of a given object-language L. This approach, already mentioned in Quine 1970 (and further refined by McKeon 2004), is supposed to be something in between the Tarskian and the original Quinean account.[27] Applied to the Boolos argument, we can see that under such a redefinition of the notion of a Q-interpretation, the Boolos set $\mathcal{B}$ will indeed become Q-satisfiable in L, because Q-satisfiability is no longer bound to the resources provided by the language L. We may simply consider an expanded language $L + T'$, where T' is interpreted as a truth predicate for L^* and replace T with T' in each sentence in $\mathcal{B}$ (leaving everything else as it stands). On this Quinean 'reinterpretation', each sentence in $\mathcal{B}$ will indeed be true in $L + T'$.[28]

The problem with this suggestion is that it is not a viable alternative for a Quinean, since, by way of ontology, it is not a real *alternative*. This can be seen if we look at the completely parallel problem of interpreting *quantifiers*. Here, again, we have the standard, objectual interpretation of quantifiers, where the truth of a quantified sentence is defined *via* the notion of *satisfaction* referring to the objects of the domain of quantification. On the other side of the spectrum we have the substitutional interpretation of quantifiers, according to which truth of a quantified sentence is defined in terms of the truth of its substitution instances. Evidently, the accounts diverge if the domain of the interpreted language contains objects that are not denoted by any individual term in the language. It has been pointed out, however, that there is a third account (often attributed to Frege), which is sometimes presented as an alternative to both accounts and which, in a sense, combines the virtues of each account without having its vices.[29] Here, the truth of a quantified sentence of some fixed language L is defined, neither *via* satisfaction nor *via* the truth of substitution instances in L, but by means of 'auxiliary names' a that do not belong to the language L itself, but to some extension $L + a$. According to this approach, the truth of $\forall x\varphi(x)$ is defined as the truth of each substitution instance $\varphi(a)$ *for each interpreted extension* $L + a$ *of* L. What is relevant here is that this approach is as 'objectual' as the standard Tarskian. The only difference is that instead of letting *assignments* and *satisfaction* do the work, on this account, *interpretations* are responsible for the (objectual) interpretation of the quantifiers.[30] Thus,

[27] Cf. Quine 1970, p. 58 and McKeon 2004, p. 219.

[28] Alternatively, instead of redefining Q-interpretations, we may redefine Q-satisfiability along the following lines: We now regard Q-satisfiability as defined earlier as a *relative* concept (relative to some language L, whether sufficiently rich or not) and further define an *absolute* notion of Q-satisfiability by requiring that a set of sentences Γ of some interpreted language L is (absolutely) Q-satisfiable if there is some sufficiently rich interpreted language L' such that Γ is (relatively) Q-satisfiable in L. The points I will make apply to this suggestion as well.

[29] See Dummett's 1973, ch. 15, Mates 1972, ch. 4 and Evans 1977, p. 474. More recently, Richard Heck rediscovered this account in the context of a discussion of Fregean semantics in his Heck 2007.

[30] This can be made more precise by first stipulating that the *extension* $L + a$ of an interpreted language L is the interpreted language whose formulas are precisely the formulas of L plus all formulas that can be formed by means of the fresh name a. Second, we require that the built-in interpretation $\mathbf{M}_a$ of each expansion $L + a$ agrees with the interpretation $\mathbf{M}$ of the original language L. In particular, the domains of $\mathbf{M}$ and $\mathbf{M}_a$ agree. The 'auxiliary name' a may be interpreted by *any* object of this domain (thereby securing the 'objectual' character of the quantifiers). Finally, the new clause for the quantifier is this:

$\mathbf{M} \vDash \forall x\varphi(x)$ iff. for each extension $L + a$: $\mathbf{M}_a \vDash \varphi(x/a)$

instead of taking a detour through satisfaction, the truth conditions for a quantified sentence are formulated directly in terms of *truth*.

The foregoing discussion directly confers to the problem of Quinean satisfiability and logical consequence.[31] Just as the 'auxiliary name semantics' for quantifiers is not a real alternative to the Tarskian standard account from an 'ontological' point of view, the revised Quinean notion of consequence, defined *via* possible extensions, is not a genuine alternative to the Tarskian notion of consequence. More precisely, it is not a genuine alternative for someone who wants to preserve the point of introducing a substitutional concept of logical truth in the first place—namely, *ontological parsimony*. In allowing Q-interpretations to vary over possible extensions of some interpreted language, we allow metatheoretical quantification over *any* set, no matter how it is defined or whether it is definable in this or that language, for each language may be extended by new predicates that are interpreted by *any set whatsoever*. Thus, we can see that ontological parsimony, which originally led Quine to introduce his substitutional concept of logical truth, is squandered on the possible extensions approach.[32]

9.3 Building Compactness into the Definition of Consequence

Since the first two alternatives do not seem promising from Quine's point of view, we might be lucky with the third. Here, the basic strategy is to redefine Q-logical consequence, so that counterexamples like the Boolos set are ruled out *by fiat*. A version of this strategy has already been considered (and dismissed) by Boolos himself (see again Boolos 1975, p. 525). The idea is to redefine Q-logical consequence in such a way that φ is a Q-logical consequence of Γ if there is a finite subset $\{\varphi_1, ...\varphi_n\}$ of Γ such that $\varphi_1 \wedge ... \wedge \varphi_n \rightarrow \varphi$ is Q-logically true (similarly for Q-satisfiability). Compactness is now built into the *definition* of Q-logical consequence and, hence, T-logical consequence and Q-logical consequence are concurrent after all due to the compactness theorem for Tarskian consequence. Yet, Boolos complains that such a manoeuver is not convincing at all. That some notion of logical consequence or satisfiability is equivalent to the usual, Tarskian one is not at all interesting if that notion is

Hence, the truth condition for a quantified sentence can be stated directly in terms of truth without having to consider (variants of) assignments or satisfaction.

[31] In his discussion of substitutional quantification, Quine himself mentions the analogy between objectual vs. substitutional quantification on the one hand and the model-theoretic vs. substitutional conception of logical truth on the other. See Quine 1970, p. 93.

[32] One has to distinguish here various versions of the possible extensions approach. McKeon 2004, for instance, adopts a variant that allows for the 'importation' of new predicates whose extension is not even a subset of the domain of the 'old' language. (See McKeon 2004, p. 219.) It is clear that, in this approach (and without further restrictions), we are committed to quantification over *any* set whatsoever, because an interpreted predicate is simply a pair consisting of an expression and *some* set. If, on the other hand, we restrict ourselves to sufficiently rich languages again and require that each 'fresh' predicate be interpreted by any subset of the *old* domain, say, the set of natural numbers, then we no longer have to quantify over any set whatsoever but 'only' over all subsets of the natural numbers. This is still a considerable leap however. So it remains that the 'ontological costs' dramatically rise if we adopt a possible extensions approach. This being the case, it seems incorrect to me to say, as McKeon does in his 2004, that 'the alteration [...] does not contradict any view espoused by Quine in print on the nature of logical truth'. In particular, it does not seem to me correct that 'the revised account does not require anything by way of ontology that is not already required in Quine's account: it is independent of all but a modest bit of set theory; independent of the higher flights', as McKeon claims (citing Quine) in his 2004, p. 220. Again, in quantifying in the metatheory over predicates of possible extensions of some interpreted language, we are quantifying—as Tarski does—over possible denotations of such predicates and which might be *any* set.

hastily cobbled together only to get what we want. Although Boolos is prepared to concede that Quine's substitutional definition of logical truth has some 'antecedent plausibility', he complains that the 'definition of satisfiability of a set as "truth of some instance of each conjunction of schemata in the set" has no such plausibility as an account of satisfiability. It even sounds wrong' (Boolos 1975 p. 526). Given Quine's aims, however, there might be a reasonable reply to Boolos' criticism. One might simply point to the fact that Quine should be allowed to *fix the extension* of the concept of logical consequence as he likes. In order to appreciate this idea, it is instructive to look at some of Quine's remarks on rival definitions of logical *truth* in his 1970. Most of what will be said in the following will directly confer to our discussion of the revised definition of Q-logical consequence.

This is the background: Quine, in his 1970, at one point considers defining logical truth by recourse to some complete proof procedure. That is, he considers stipulating that a sentence is logically true if derivable by means of the chosen proof procedure. With an eye to the other candidate definitions (substitutional and model-theoretic), he mentions two possible objections to this strategy:

> The theorems establishing equivalence among very unlike formulations of a notion—logical truth or whatever—are of course the important part. Which of the formulations we choose as the somehow official definition is less important. But even in such verbal matters there are better and worse choices. The more elementary of two definitions has the advantage of relevance to a wider range of collateral studies. It should be said, however, that part of the resistance to this elementary way of defining logical truth has a special reason: the arbitrariness of choice among proof procedures. One feels he has missed the essence of logical truth when his definition is arbitrary to that degree. (Quine 1970, p. 57)

So there are two issues that would make Quine hesitate to adopt a particular definition of a concept (like logical truth) rather than some other definition: lack of relevance to a wider range of collateral studies and arbitrariness that would result in not getting the 'essence' of the concept in question right. Let us look at the second point first.

I am not sure how seriously Quine's talk about the 'essence of logical truth' should be taken here. His general and well-known worries about essential properties and related concepts point to interpreting this statement with a grain of salt. 'Essences' are un-Quinean, and so is the task of trying to somehow 'capture the essence of a concept'. Moreover, the 'arbitrariness' to which Quine refers in this passage is quite specific. There is indeed a bathful of complete proof procedures available that would serve the purpose of fixing the extension of the concept of logical truth. Some of these proof procedures differ from each other by only a narrow margin and each of them is specified, in principle, in the same way—*viz.* by syntactic operations on signs. But the fact remains that, if extensional adequacy were the *only* adequacy criterion on a proposed definition, Quine could not complain about defining logical truth by appealing to some specific proof procedure or the other. The degree of arbitrariness of a proposed definition can be as high as it wants to be if the definition only gets the extension right. Quine, however, *does* feel—in line with the intuitions of most philosophers—that this cannot be the whole truth.

I suggest that a similar objection can be made to the redefinition of logical consequence mentioned earlier. Such a definition seems to be 'arbitrary' in a way that is similar to a definition of logical truth that appeals to some specific proof procedure. It is similar, I suggest, in the following respect: In defining logical consequence as the existence of a

finite subset $\{\varphi_1, ...\varphi_n\}$ of the premise set Γ such that $\varphi_1 \wedge ... \wedge \varphi_n \rightarrow \varphi$ is a logical truth, we are appealing to notions that are not essential to the concept of logical consequence. For instance, having the notion of *finitude* (and, thus, *infinitude*) at our disposal does not seem to be required in order to grasp the concept of logical consequence. Therefore, it is unreasonable to let this notion already figure in a *definition* of logical consequence. To provide a somewhat frivolous analogy, doing so would be like stipulating that n is a prime number if and only if, say, the set of truth functions $\mathbf{Ł}_{n+1}$, generated by the Lukasiewicz truth functions, is functionally precomplete in the set of all truth functions $\mathbf{P}_{n+1}$. Even though this 'definition' of prime number provably gets the extension right, one 'feels he has missed the essence of prime number if his definition is arbitrary to that degree', to paraphrase Quine.

The matter is complex though. Following this line of reasoning, one quickly gets sucked into problems concerning analysis, analytic definitions and standards for a 'correct' analysis/analytic definition. Problems, alas, which I am not able to solve here. So let me say a few things concerning Quine's second worry, which has a more pragmatic flavour.

To repeat, according to Quine, 'the more elementary of two definitions [of logical truth] has the advantage of relevance to a wider range of collateral studies'. Quine's subsequent discussion makes it clear that he thinks that a definition of logical truth in terms of some proof procedure is indeed 'more elementary':

> It describes rules of proofs and thus talks of strings of signs. On this score it is on a par with the definition that appeals to substitution of sentences; it operates, in effect, at the level of elementary number theory. But it keeps to that level, whereas the other definition invoked also the notion of truth. (Quine 1970, p. 58)

So, in terms of 'elementariness', Quine's ranking of the available definitions of logical truth is this: First, there is the Tarskian, model-theoretic definition, which is 'clearly' non-elementary, since it 'heavily' relies on set theory. Second, there is the substitutional definition, which is more elementary than the model-theoretic definition since it keeps set theory to a minimum and, hence, is independent of the 'higher flights'. And third, there are definitions that are based on some proof procedure. These definitions are even more elementary than the substitutional definition since they do not even appeal to the notion of *truth*. Furthermore, as Quine makes clear on another occasion, what makes the notion of truth non-elementary is that its definition requires 'heavy equipment' from set theory (cf. Quine 1970, p. 42).

What this once again shows is, first, that Quine is not concerned with an 'analysis' of the concept of logical truth and, second, that Quine's yardstick of elementariness is, as one might have expected, calibrated in terms of ontology and, indeed, set theory. So even though we do not get a criterion by which we could decide which of two definitions is more likely to capture the 'essence' of logical truth (or any other concept), we *do* get something like a 'scale of elementariness' of rival definitions. Moreover, according to this scale, the revised definition of Q-logical consequence (existence of a finite subset, such that...) is indeed more elementary than the usual, model-theoretic definition. So were we too quick in discarding (the redefinition of) Q-logical consequence? After all, the new definition is extensionally correct *and* 'more elementary' in the Quinean sense. But two questions arise immediately in this respect: First, does this imply, as suggested by Quine, that the redefinition has 'the advantage of relevance to a wider range of collateral studies'? And second, is the scale

advanced by Quine a *reasonable* scale of elementariness? Concerning the second question, it seems wrong to me to think that less ontology *alone* is already a warrant for a definition to be more elementary. 'Conceptual priority', as one might dub it, seems to be just as important, if not more important. The 'definition' of prime number mentioned earlier, for instance, is clearly not elementary in any reasonable sense. But this non-elementariness is not effected by some fact concerning ontology, but by the fact that complex notions like many-valued truth functions, precompleteness, etc., figure in the definition, notions whose mastery requires substantive knowledge about mathematical logic. As already suggested earlier, I think something similar is true of the revised definition of Q-logical consequence and, indeed, the original definition of Q-logical truth as well. A sure telltale sign for non-elementariness is the complexity of the kind of *theorems* that are required in order to ensure the extensional adequacy of a certain definition. In order to see that the mentioned definition of prime number is correct we need to be able to prove a bunch of things about many-valued truth functions, functional completeness, etc. Similarly, in order to be able to recognize the extensional adequacy of Quine's definition of logical truth and the redefinition of logical consequence, we need to know a bunch of things about definability, different magnitudes of infinity and to be able to prove theorems like the Hilbert-Bernays-Löwenheim theorem, the completeness theorem, etc., for the *standard* concepts of logical truth and consequence first.[33] So I think the revised definition of Q-logical consequence should not count as 'basic' or 'elementary'—at least, not more elementary than the standard, model-theoretic definition—in any reasonable sense, and neither should the Quinean definition of logical truth.

There still remains the question of whether the Quinean concepts of logical truth and consequence have 'relevance to a wider range of collateral studies' than the usual Tarskian ones. To be sure, it is far from clear what the relevant notion of 'relevance' is in this context and what (still) counts as a 'collateral study'. But I think that, once again, the Quinean concepts score worse in this respect than the standard ones, given some reasonable standard that is based on actual research in mathematical logic. On a very general level, for instance, the Tarskian definitions can arguably be said to have paved the way for (and were preceded by) a number of non-trivial results in mathematical logic. As already noted, Quine's definitions would not even be recognizable as extensionally correct definitions were it not for Löwenheim, Skolem, Bernays, Hilbert, Gödel and others, who were relying on—and proving things about—informal versions of the model-theoretic conceptions of logical truth and consequence. Results like various versions of the Löwenheim theorem, the completeness theorem, the existence of 'non-standard models' of first-order arithmetic, etc., would not have been possible if logical truth and consequence were understood according to Quine's definitions. In addition to these general considerations, there are more specific reasons to think that Quine's definitions are inferior to Tarski's with respect to their support of 'collateral studies'. For example, 'collateral studies' concerning mathematical fields like group theory, topology, etc., seem to be ruled by the Quinean definitions from the very beginning, since the applicability of Q-logical truth and consequence requires a sufficiently rich, interpreted object language. We want to be able to say, for instance, that the law of commutativity is not a logical consequence of the axioms of group theory. But we cannot (at

[33] That is why Boolos, at the end of his 1975, p. 526, writes that Quine's definition of logical truth is 'a definition of logical truth *only* in virtue of a remarkable theorem about first-order logic' [emphasis by the author].

least not in a straightforward way) if we adopt the Quinean concept of logical consequence, because the language of group theory simply is not rich enough and, indeed, not interpreted in the first place.

So, overall it appears that Quine's definitions come off worse than the usual definitions, both with respect to their 'elementariness' (given some reasonable standard of 'elementariness' that is admittedly not Quine's) and with respect to the 'collateral studies' that are supported by these definitions.

10 Conclusion

In conclusion, let us take stock and see what we have shown in the previous sections. We have seen that the usual, model-theoretic definition of logical consequence differs essentially from the definition of logical consequence we get when we generalize Quine's definition of logical truth in the most natural way. Furthermore, there seems to be no way to argue, on grounds acceptable to a Quinean, that this is unproblematic. Therefore, we had to conclude that the original Quinean definitions are indeed defective. Three alternatives have been considered regarding how one might react to this situation. The first two alternatives (further restricting the languages to which the Quinean concepts apply and redefine the notion of a Q-interpretation to allow for possible extensions of the original language) have been dismissed. The first alternative simply did not work and the second led to definitions of logical truth and consequence that were no more ontologically parsimonious than the usual ones. Hence, from a Quinean point of view, there is no decisive advantage in adopting these definitions rather than the standard ones. Now, even though there seems to be no knock-down argument to the third alternative (redefining Q-logical consequence by building compactness into the definition), we saw that, even here, a case can be made that this leads to undesirable consequences. The reasons for not taking this course are, in my opinion, good reasons for not adopting the (Quinean version of the) substitutional definition of logical truth in the first place. And even though there might be philosophical leeway in this direction, it seems to me that, all things considered, Boolos was right and certain reservations about the Quinean definitions are indeed justified.

References

Berlinski, D. and D. Gallin, 1969: Quine's Definition of Logical Truth. *Nous*, **3 (2)**, 111–128.

Boolos, G., 1975: On Second Order Logic. *The Journal of Philosophy*, **72(16)**, 509–527.

Boolos, G., 1987: The Consistency of Frege's Foundations of Arithmetic. *On Being and Saying: Essays in Honor of Richard Cartwright*, J. Thomson, Ed., Mit Press, 3–20.

Boolos, G. and R. Jeffrey, 1974: *Computability and Logic*. Third edition ed., Cambridge University Press.

Dummett, M., 1973: *Frege: Philosophy of Language*. Duckworth.

Etchemendy, J., 1990: *The Concept of Logical Consequence*. Cambridge, MA: Harvard University Press.

Evans, G., 1977: Pronouns, Quantifiers, and Relative Clauses (I). *Canadian Journal of Philosophy*, **7 (3)**, 467–536.

Goodman, N. and W. V. Quine, 1947: Steps Toward a Constructive Nominalism. *Journal of Symbolic Logic*, **12 (4)**, 105–122.

Hanson, W. H., 1997: The Concept of Logical Consequence. *Philosophical Review*, **106 (3)**, 365–409.

Heck, R., 2007: Frege and Semantics. *Grazer Philosophische Studien*, **75 (1)**, 27–63.

Hinman, P. G., J. Kim, and S. P. Stich, 1968: Logical Truth Revisited. *Journal of Philosophy*, **65 (17)**, 495–500.

Kahle, R. and P. Schroeder-Heister, 2006: Introduction: Proof-Theoretic Semantics. *Synthese*, **148 (3)**.

Kleene, S. C., 1952: *Introduction to Metamathematics*. North-Holland Publishing Co., Amsterdam - P. Noordhoff N.V. - Groningen.

Lapointe, S., 2014: Bolzano, Quine and Logical Truth. *A Companion to W.V.O. Quine*, G. Harman and E. Lepore, Eds., Wiley Blackwell, 296–312.

Mates, B., 1972: *Elementary Logic*. 2d ed., New York,Oxford University Press.

McKeon, M., 2004: On the Substitutional Characterization of First-Order Logical Truth. *History and Philosophy of Logic*, **25 (3)**, 205–224.

Prawitz, D., 1974: On the Idea of a General Proof Theory. *Synthese*, **27 (1?2)**, 63–77.

Quine, W., 1954: Interpretations of Sets of Conditions. *The Journal of Symbolic Logic*, **19 (2)**, 97–102.

Quine, W., 1966: *The Ways of Paradox and other Essays*. Random House New York.

Quine, W., 1970: *Philosophy of Logic*. 2d ed., Harvard University Press.

Saguillo, J. M., 2001: Quine on Logical Truth and Consequence. *AGORA - Papeles de Filosofia*, **20 (1)**, 139–156.

Schilpp, P. A. and L. E. Hahn, (Eds.) , 1986: *The Philosophy of W. V. Quine*. Second 1998 ed., Open Court Publishing.

Schroeder-Heister, P., 2006: Validity Concepts in Proof-Theoretic Semantics. *Synthese*, **148 (3)**, 525–571.

Shapiro, S., 1991: *Foundations Without Foundationalism: A Case for Second-Order Logic*. 174, Oxford University Press, 127 pp.

Shapiro, S., 2000: The Status of Logic. *New Essays on the A Priori*, P. Boghossian and C. Peacocke, Eds., Oxford Universtity Press, Clarendon, 333–366.

Sher, G., 1996: Did Tarski commit "Tarski's Fallacy"? *The Journal of Symbolic Logic*, **61 (2)**, 653–686.

Simpson, S. G., 2009: *Subsystems of Second Order Arithmetic*. Cambridge University Press.

Tarski, A., 1956: *Logic, Semantics, Metamathematics*. Oxford, Clarendon Press.

Tarski, A., 1986: What Are Logical Notions? *History and Philosophy of Logic*, **7 (2)**, 143–154.

van Dalen, D., 1980: *Logic and Structure*. Springer-Verlag.