Hilbert's '*Verunglückter Beweis*', the first epsilon theorem, and consistency proofs[*]

Richard Zach
Department of Philosophy
University of Calgary
2500 University Drive N.W.
Calgary, Alberta T2N 1N4, Canada
rzach@ucalgary.ca

October 31, 2003

Abstract

In the 1920s, Ackermann and von Neumann, in pursuit of Hilbert's Programme, were working on consistency proofs for arithmetical systems. One proposed method of giving such proofs is Hilbert's epsilon-substitution method. There was, however, a second approach which was not reflected in the publications of the Hilbert school in the 1920s, and which is a direct precursor of Hilbert's first epsilon theorem and a certain 'general consistency result' due to Bernays. An analysis of the form of this so-called 'failed proof' sheds further light on an interpretation of Hilbert's Programme as an instrumentalist enterprise with the aim of showing that whenever a 'real' proposition can be proved by 'ideal' means, it can also be proved by 'real', finitary means.

1 Introduction

The aim of Hilbert's Programme for consistency proofs in the 1920s is well known: to formalize mathematics, and to give finitary consistency proofs of these systems and thus to put mathematics on a 'secure foundation'. What is perhaps less well known is exactly how Hilbert thought this should be carried out. Over ten years before Gentzen developed sequent calculus formalizations of arithmetic and used an elaboration of his cut-elimination procedure to give a consistency proof of Peano Arithmetic, Hilbert proposed a different approach. He believed that the principles criticized by intuitionists, the principle of the excluded middle in its application to infinite totalities and the use of unbounded existential quantifiers are, at root, the same. This root is the axiom of choice. In a course on the foundations of mathematics, he remarked that whereas the use of unbounded quantification results in significant problems for giving a consistency proof,

[*] To appear in *History and Philosophy of Logic*.

> the core of the difficulty lies at a different point, to which one usually only pays attention later: it lies with Zermelo's *axiom of choice*... We want to extend the axiom of choice. To each proposition with a variable $A(a)$ we assign an object for which the proposition holds only if is holds in general. So, a counterexample, if one exists.[1]

This counterexample is given by the τ-operator: $\tau_x A(x)$ is an object for which $A(x)$ is false, if there is one. The dual operator $\varepsilon_x A(x)$, is a witness, i.e., an object for which $A(x)$ is true, if $A(x)$ is true for anything.[2] The ε-operator is governed by the *transfinite axiom,*

$$A(a) \to A(\varepsilon_x A(x)).$$

A finitary consistency proof of mathematical theorems which allows the elimination of applications of the choice principle (in the form given to it by the transfinite axiom) would then show that such application is justified after all. It would also show that unbounded quantification is admissible in mathematics, since with the help of the transfinite axioms one can define quantifiers by

$$(\exists x)A(x) \equiv A(\varepsilon_x A(x)) \quad \text{and} \quad (\forall x)A(x) \equiv A(\varepsilon_x \neg A(x)).$$

ε-terms may be seen as ideal elements whose addition to the theory of finite propositions reintroduces the powerful methods of infinite mathematics, and 'round out the theory'. To show that their addition is permissible requires a proof that ε-terms can be eliminated from proofs of 'real', finitary, propositions. This elimination of ε-terms from formal proofs in arithmetical theories was to proceed according to the epsilon-substitution method. Hilbert's approach here was to define a finitary procedure which would produce, given a proof involving ε-terms, a substitution of these terms by actual numbers.[3] Applying this substitution to the proof would then result in a purely elementary proof about numbers which would contain no trace of the transfinite elements of the original proof. In addition, it is seen finitarily that all initial formulas, and hence also the end formula, of the resulting proof are true. Since such a proof cannot possibly have a contradiction as its last line, the consistency of arithmetic would be established. Hilbert presented his 'Ansatz' for finding such substitutions in Hilbert (1922c); it was extended by Ackermann (1924) and von Neumann (1927).

The epsilon-substitution method and its role in Hilbert's Programme are now relatively well understood. There was, however, a *second* proposal for proving consistency, also based on the epsilon calculus, which has escaped historical attention, and which was never presented in the publications of the Hilbert school before 1939. In the second volume of *Grundlagen der Mathematik* (Hilbert and Bernays, 1939), Bernays first developed in print a well worked-out theory of the epsilon calculus as an alternative formulation and extension of predicate logic, and proved the so-called first and second epsilon theorems. In Section 1.4, Bernays presented a 'general consistency theorem' based on the first epsilon theorem, which applies, e.g., to elementary geometry and to arithmetic with an open induction rule. This second approach to consistency proofs via the first epsilon theorem, however, dates back to the beginning of Hilbert's Programme. In a letter from Bernays to Ackermann of October 1929, Bernays refers to this second approach as Hilbert's '*verunglückter Beweis*' (the 'failed proof'). This failed proof is not mentioned in Hilbert's publications of the early 1920s nor into his lectures on the subject of 1922 and 1923. A record of the basic idea, a second '*Ansatz*', is, however, available in the form of a six-page note in Bernays's hand.

The aim of this paper is to present and analyze this second approach to proving consistency, and to show how Hilbert's 'verunglückter Beweis' precipitated the later proof of the first epsilon theorem by Bernays and Ackermann. Given the role envisaged by Hilbert for the ε-calculus and the consistency proofs based on it, such an analysis will help illuminate not just the genesis of an important proof-theoretic result (the epsilon theorem), but also Hilbert's aim and strategy for providing consistency proofs. In the following section, we will revisit the first epsilon theorem, and show how it can be used to establish consistency results. Following this discussion, I present the suggestion contained in Hilbert's second *Ansatz*, and outline why this approach was not pursued by Hilbert and his students in the 1920s. A concluding section discusses the relevance of the result in the wider context of Hilbert's consistency project.

2 The first epsilon theorem and the general consistency result

The epsilon calculus consists in the elementary calculus of free variables plus the 'transfinite axiom', $A(a) \to A(\varepsilon_x A(x))$. The elementary calculus of free variables is the quantifier-free fragment of the predicate calculus, i.e., axioms for propositional logic and identity, with substitution rules for free individual (a, b, ...) and formula (A, B, ...) variables and modus ponens.

One of the most basic and fruitful results concerning Hilbert's ε-calculus is the so-called epsilon theorem. It states that if a formula $\mathfrak{E}$ containing no ε-terms is derivable in the ε-calculus from a set of axioms which also do not contain ε-terms, then $\mathfrak{E}$ is already derivable from these axioms in the elementary calculus of free variables (i.e., essentially using propositional logic alone). A relatively easy consequence of this theorem (or rather, of its proof) is Herbrand's theorem, and, in fact, one of the first published correct proofs of Herbrand's theorem is that given by Bernays in *Grundlagen der Mathematik II* (Hilbert and Bernays, 1939) based on the first ε-theorem.[4] Leisenring (1969) even formulates the ε-theorem in such a way that the connection to Herbrand's theorem is obvious:

> If E is a prenex formula derivable from a set of prenex formulas Γ in the predicate calculus, then a disjunction $B_1 \lor \ldots \lor B_n$ of substitution instances of the matrix of E is derivable in the elementary calculus of free variables from a set Γ' of substitution instances of the matrices of the formulas in Γ.

Even without this important consequence, which was of course not discovered until after Herbrand's (1930) thesis, the first ε-theorem constitutes an important contribution to mathematical logic. Without the semantical methods provided by the completeness theorem for predicate logic, it is not at all clear that the addition of quantifiers in the guise of ε-terms and the axioms governing them is a conservative extension of the elementary calculus. Keeping in mind the role of epsilon-terms as ‘ideal elements’ in a proof, the eliminability of which is the main aim of a consistency proof of any mathematical system formulated with the aid of the epsilon calculus, the first epsilon theorem is also the main prerequisite for such a consistency proof.

Bernays stated the first and second epsilon theorem as follows:

> These theorems both concern a formalism F, which results from the predicate calculus by adding to its symbols the ε-symbol and also certain individual [constant], predicate, and function symbols, and to its axioms the ε-formula [the transfinite axiom] and furthermore certain *proper axioms* $\mathfrak{P}_1, \ldots, \mathfrak{P}_\mathfrak{k}$ *which do not contain the* ε*-symbol.* For such a formalism F, the two theorems state the following:
>
> 1. If $\mathfrak{E}$ is a formula derivable in F which does not contain any bound variables, and the axioms $\mathfrak{P}_1, \ldots, \mathfrak{P}_\mathfrak{k}$ also contain no bound variables, then the formula $\mathfrak{E}$ can be derived from the axioms $\mathfrak{P}_1, \ldots, \mathfrak{P}_\mathfrak{k}$ without the use of bound variables at all, i.e., with the elementary calculus of free variables alone (‘first epsilon theorem’).
> 2. If $\mathfrak{E}$ is a formula derivable in F which does not contain the ε-symbol, then it can be derived from the axioms $\mathfrak{P}_1, \ldots, \mathfrak{P}_\mathfrak{k}$ without the use of the ε-symbol, i.e., with the predicate calculus alone (‘second epsilon theorem’). (Hilbert and Bernays, 1939, 18)

The predicate calculus is formulated with a substitution rule for free individual and formula variables; the elementary calculus of free variables is the quantifier-free fragment of the predicate calculus (i.e., without quantifier axioms or rules) or equivalently, the epsilon calculus without transfinite axioms and without defining axioms for the quantifiers.

A proof that the ε-calculus is conservative over the elementary calculus of free variables in the way specified by the first epsilon theorem constitutes a proof of consistency of the ε-calculus and of mathematical theories which can be formulated in such a way that the first ε-theorem applies (i.e., the axioms are quantifier- and ε-free). Indeed Bernays used the first ε-theorem to prove a ‘general consistency result’ for axiomatic theories to which the first ε-theorem applies. Let us first outline the consistency proof for a very basic arithmetical theory. This theory results from the elementary calculus of free variables by adding the constant 0 and successor (+1) and predecessor (δ) functions. The additional axioms are:

$$\begin{aligned} 0 &\neq x+1 \\ x &= \delta(x+1) \end{aligned}$$

To prove that the resulting axiom system is consistent, assume there were a proof of $0 \neq 0$. First, by copying parts of the derivation as necessary, we can assume that every formula in the proof is used as the premise of an inference at most once. Hilbert and Bernays call this ‘resolution into proof threads’. The resulting proof is in tree form; a branch of this tree (beginning with an axiom and ending in the end-formula) is a *proof thread*. Next, we can substitute numbers for the free variables in the proof (‘elimination of free variables’). Bernays describes this as follows:

> We follow each proof thread, starting at the end formula, until we reach two successive formulas $\mathfrak{A}$, $\mathfrak{B}$ where the first results from the second by substitution. We record the substitution also in the formula $\mathfrak{B}$, so that we get instead of $\mathfrak{B}$ a repetition of the formula $\mathfrak{A}$.
>
> If $\mathfrak{B}$ is an initial formula [axiom], then the substitution has been transferred to the initial formula. Otherwise, $\mathfrak{B}$ was obtained by substitution into a formula $\mathfrak{C}$ or by repetition, or as conclusion of an inference
>
> $$\begin{array}{ccc} \mathfrak{C} & & \mathfrak{C} \to \mathfrak{B} \\ & \diagdown \quad \diagup & \\ & \mathfrak{B}. & \end{array}$$
>
> In the first case, we in turn replace $\mathfrak{C}$ by $\mathfrak{A}$, so that the substitutions leading from $\mathfrak{C}$ to $\mathfrak{B}$ and from $\mathfrak{B}$ to $\mathfrak{A}$ are recorded simultaneously. (In the case of repetition, only one substitution is recorded.)
>
> In the case of the inference schema [modus ponens], we record the substitution leading from $\mathfrak{B}$ to $\mathfrak{A}$ in the formulas $\mathfrak{C}$ and $\mathfrak{C} \to \mathfrak{B}$; this changes the formula $\mathfrak{C}$ if and only if it contains the variables being substituted for in the transition from $\mathfrak{B}$ to $\mathfrak{A}$. In any case, the original inference schema with conclusion $\mathfrak{B}$ is replaced by an inference schema
>
> $$\begin{array}{ccc} \mathfrak{C}^* & & \mathfrak{C}^* \to \mathfrak{A} \\ & \diagdown \quad \diagup & \\ & \mathfrak{A}. & \end{array}$$
>
> We can proceed in this way until we reach an initial formula in each thread. When the procedure comes to its end, each substitution has been replaced by a repetition, each inference schema by another inference schema, and certain substitutions have been applied to the initial formulas. (Hilbert and Bernays, 1934, 225)

Remaining free variables can now be replaced by 0 (for individual variables) and $0 = 0$ (for formula variables). We would thus obtain a proof of $0 \neq 0$ without free variables.

If we now reduce the variable-free terms in the resulting proofs to standard numerals by successively replacing $\delta(0)$ by 0 and $\delta(\mathfrak{t} + 1)$ by $\mathfrak{t}$, we get a proof where each initial formula is either an instance of a tautology, of an identity axiom, or, if the original formula was one of the axioms for $+1$ and δ, a formula of the form of either

$$\begin{array}{ccc} 0 & \neq & \mathfrak{n} + 1 \\ \mathfrak{n} & = & \mathfrak{n} \end{array}$$

(where $\mathfrak{n}$ is either 0 or of the form $0 + \cdots + 1$).

Call an equation of the form $\mathfrak{n} = \mathfrak{n}$ 'true' and one of the form $\mathfrak{n} = \mathfrak{m}$, where $\mathfrak{n}$ and $\mathfrak{m}$ are not identical, 'false'. This can be extended to propositional combinations of equations in the obvious way. We observe that the resulting proof has all true initial formulas, and since modus ponens obviously preserves truth, all other formulas are also true. Since $0 \neq 0$ is false, there can be no proof of $0 \neq 0$.

Hilbert presented this proof in his course on the foundations of mathematics in 1921/22 (1922a; 1922b) and outlined the basic approach in his 1922 talk at the meeting of the Deutsche Naturforscher-Gesellschaft in Leipzig (1923a). In a course given the following year (1923b; 1923a), Bernays and he extended it to axioms for primitive recursive functions; Ackermann (1924) further elaborated it to include second-order primitive recursive functions (see Zach 2003). The challenge was to extend it to the case where ε-terms and the transfinite axiom are also present, leading to Hilbert's ε-substitution method. There, the aim was to find substitutions not just for the free variables, but also for the ε-terms, ultimately also resulting in a proof without free or bound variables and with true initial formulas. An alternative method is this: Instead of treating ε-terms together with other terms of the system, eliminate ε-terms *first*. We introduce a step at the beginning of the proof which reduces a proof in the ε-calculus to one in the elementary calculus of free variables as in the first ε-theorem. Thus, with the first ε-theorem in hand, Bernays could later formulate the following 'general consistency theorem':

> Let F be a formalism which results from the predicate calculus by adding certain individual, predicate, and function symbols. Suppose there is a method for determining the truth value of variable-free atomic formulas uniquely. Suppose furthermore that the axioms do not contain bound variables [i.e., no quantifiers and no ε-terms] and are verifiable [i.e., every substitution instance is true]. Then the formalism is consistent in the strong sense that every derivable variable-free formula is a true formula. (Hilbert and Bernays, 1939, 36)

Suppose $\mathfrak{A}$ is variable-free and derivable in F. If a formalism F satisfies the conditions, then the first ε-theorem yields a proof of $\mathfrak{A}$ already in the elementary calculus of free variables. The procedures above (resolution into proof threads, elimination of free variables) yields a proof of $\mathfrak{A}$ from substitution instances of the axioms of F. Since the axioms of F are verifiable, these substitution instances are true, and again, modus ponens preserves truth. So $\mathfrak{A}$ is true. The requirement that the truth-value of variable-free atomic formulas is decidable ensures that this is a finitary proof: It can be finitarily verified that any *given* proof in fact has true initial formulas (and hence, a true end formula).

3 Hilbert's *Verunglückter Beweis*

The first ε-theorem is first formulated in print in Hilbert and Bernays (1939), but Hilbert had something like it in mind already in the early/mid 1920s. When working on *Grundlagen der Mathematik* in the late 1920s, Bernays revisited the idea, which had been abandoned in favour of the ε-substitution method. In correspondence with Ackermann in 1929 (discussed below), Bernays refers to 'Hilbert's second consistency proof for the ε-axiom, the so-called 'failed proof', and suggests ways in which the difficulties originally encountered could be fixed. Surprisingly, this 'failed proof', a precursor of the first ε-theorem, is not to be found in the otherwise highly interesting elaborations of lecture courses on logic and proof theory given by Hilbert (and Bernays) between

1917 and 1923. The only evidence that the ε-theorem predates Bernays's proof of it in Hilbert and Bernays (1939) are the letter from Bernays to Ackermann from 1929, and a sketch of the simplest case of the theorem.

The sketch in question is a six-page manuscript in Bernays's hand which can be found bound with the lecture notes to Hilbert's course *Elemente und Prinzipienfragen der Mathematik*, taught in the Summer Semester 1910 in Göttingen (Hilbert, 1910). Although it bears a note by Hilbert 'Insertion in WS [Winter Semester] 1920 [sic]', the notation used in the sketch suggests that it was written after sometime after 1922, when the epsilon notation was first introduced. Unfortunately, the only substantial discussion of the proof is found in the letter from Bernays to Ackermann from 1929 quoted below. The fact that it uses the ε-axioms in their final form, suggests that it was written after Hilbert and Bernays's 1922/23 course, in which εs were still used in their dual forms. One would also expect Hilbert to have presented the proof in the course, if it had had been available then. The proof is briefly alluded to in a letter from Ackermann to Bernays of June 1925 in a way that suggests that it was not a recent proof.[5] Thus, the proof likely dates from 1923 or 1924.

The sketch bears the title 'Consistency proof for the logical axiom of choice $Ab \to A\varepsilon_a Aa$, in the simplest case'. In it, we find a proof of the first ε-theorem for the case where the substitution instances of the transfinite axiom used in the proof, i.e., the so-called *critical formulas*

$$\mathfrak{A}(\mathfrak{t}) \to \mathfrak{A}(\varepsilon_x \mathfrak{A}(x))$$

are such that $\mathfrak{A}(x)$ contains no ε's, and furthermore the identity axioms are not used at all. The proof goes as follows. Suppose

$$\begin{array}{ccc}
\mathfrak{A}(\mathfrak{t}_1) & \to & \mathfrak{A}(\varepsilon_x \mathfrak{A}(x)) \\
 & \vdots & \\
\mathfrak{A}(\mathfrak{t}_{\mathfrak{n}}) & \to & \mathfrak{A}(\varepsilon_x \mathfrak{A}(x))
\end{array}$$

are all the critical formulas involving $\mathfrak{A}$ in a proof of $0 \neq 0$. First, replace every formula $\mathfrak{F}$ occurring in the proof by the conditional $\overline{\mathfrak{A}}(\mathfrak{t}_1) \to \mathfrak{F}$, and every application of modus ponens by the (derivable) inference

$$\frac{\overline{\mathfrak{A}}(\mathfrak{t}_1) \to \mathfrak{S} \qquad \overline{\mathfrak{A}}(\mathfrak{t}_1) \to (\mathfrak{S} \to \mathfrak{T})}{\overline{\mathfrak{A}}(\mathfrak{t}_1) \to \mathfrak{T}}$$

Every formula resulting thus from a substitution instance $\mathfrak{F}$ of an axiom (other than the critical formula for $\mathfrak{t}_1$) is then derivable by

$$\frac{\mathfrak{F} \qquad \mathfrak{F} \to (\overline{\mathfrak{A}}(\mathfrak{t}_1) \to \mathfrak{F})}{\overline{\mathfrak{A}}(\mathfrak{t}_1) \to \mathfrak{F}}$$

The formula corresponding to the ε-axiom involving $\mathfrak{t}_1$ is derived using

$$\frac{\begin{array}{c}\mathfrak{A}(\mathfrak{t}_1) \to (\overline{\mathfrak{A}}(\mathfrak{t}_1) \to \mathfrak{A}(\varepsilon_x \mathfrak{A}(x)) \\ (\mathfrak{A}(\mathfrak{t}_1) \to (\overline{\mathfrak{A}}(\mathfrak{t}_1) \to \mathfrak{A}(\varepsilon_x \mathfrak{A}(x))) \to (\overline{\mathfrak{A}}(\mathfrak{t}_1) \to (\mathfrak{A}(\mathfrak{t}_1) \to \mathfrak{A}(\varepsilon_x \mathfrak{A}(x)))\end{array}}{\overline{\mathfrak{A}}(\mathfrak{t}_1) \to (\mathfrak{A}(\mathfrak{t}_1) \to \mathfrak{A}(\varepsilon_x \mathfrak{A}(x))}$$

The premises of this inference are propositional axioms. Thus we obtain a proof of $\overline{\mathfrak{A}}(\mathfrak{t}_1) \to 0 \neq 0$ with only the critical formulas for $\mathfrak{t}_2, \ldots, \mathfrak{t}_n$.

Next, replace every formula in the original proof by the conditional $\mathfrak{A}(\mathfrak{t}_1) \to \mathfrak{F}$, and also replace $\varepsilon_a \mathfrak{A}(a)$ everywhere by $\mathfrak{t}_1$. The initial formulas of the resulting derivation (except those resulting from critical formulas) are again derivable as before. The formulas corresponding to the critical formulas are all of the form

$$\mathfrak{A}(\mathfrak{t}_1) \to (\mathfrak{A}(\mathfrak{t}_i') \to \mathfrak{A}(\mathfrak{t}_1))$$

which are propositional axioms. We therefore now have a proof of $\mathfrak{A}(\mathfrak{t}_1) \to 0 \neq 0$ without critical formulas. Putting the two proofs together and applying the law of excluded middle, we have found a proof of $0 \neq 0$ using only critical formulas for $\mathfrak{t}_2$, $\ldots$, $\mathfrak{t}_n$. By induction on n, there is a proof of $0 \neq 0$ using no critical formulas at all. In the resulting proof, we can replace $\varepsilon_x \mathfrak{A}(x)$ everywhere by 0.[6]

In a letter to Ackermann dated October 16, 1929, Bernays discusses this proof and suggests ways of extending the result to overcome problems that apparently had led Hilbert to abandon the idea in favour of consistency proofs using the ε-substitution method. The letter begins with a review of the problems the original idea suffered from:

> While working on the *Grundlagenbuch*, I found myself motivated to re-think Hilbert's second consistency proof for the ε-axiom, the so-called 'failed' proof, and it now seems to me that it can be fixed after all.
>
> Since I know that it is very easy to overlook something in the area of proofs like this, I would like to submit my considerations to you for verification.
>
> The stumbling blocks for the completion of the proof were threefold:
>
> 1. It could happen that due to the replacements needed for the treatment of one critical formula, a different critical formula lost its characteristic form without, however, thus resulting in a derivable formula.
>
> 2. Incorporating the second identity axiom, which can be replaced by the axiom
>
> $$(G) \qquad a = b \to (\varepsilon_x A(x,a) = \varepsilon_x A(x,b))$$
>
>   in its application to the ε-function [footnote: except in the harmless application consisting in the substitution of an ε-functional for an individual variable in the identity axiom]—only $\varepsilon_x \mathfrak{A}(x)$ are involved here, where x is an *individual* variable—caused problems.
>
> 3. Sometimes a new ε-functional appeared after successful elimination of an ε-functional, so that overall no reduction was achieved.[7]

The difficulties listed by Bernays arise already for the ε-theorem in the general case; dealing with number theory, i.e., the induction axiom, in the way outlined requires even further extensions of the method. Bernays acknowledges this in the letter, writing, 'With the addition of complete induction the method is no longer, i.e., at least not immediately, applicable. For that, your [Ackermann's] method of total substitution [i.e., a solving ε-substitution] would be the simplest way'. However, even if an extension to arithmetic is not immediately available, it seems that Bernays considered the 'second proof' valuable and interesting enough to fix. To summarize, there are two difficulties: The first is that the possibilities in which ε-terms can be nested in one another and in which cross-binding of variables can occur give rise to difficulties in their elimination. On the one hand, we replace $\varepsilon_x\mathfrak{A}(x)$ by $\mathfrak{t}_1$ in the second step. If ε-terms other than $\varepsilon_x\mathfrak{A}(x)$, but which contain $\varepsilon_x\mathfrak{A}(x)$, say, $\varepsilon_y\mathfrak{B}(y,\varepsilon_x\mathfrak{A}(x))$ are also present, we would obtain from a critical formula

$$\mathfrak{B}(\mathfrak{s},\varepsilon_x\mathfrak{A}(x)) \rightarrow \mathfrak{B}(\varepsilon_y\mathfrak{B}(y,\varepsilon_x\mathfrak{A}(x)),\varepsilon_x\mathfrak{A}(x))$$

a formula

$$\mathfrak{B}(\mathfrak{s},\mathfrak{t}_1) \rightarrow \mathfrak{B}(\varepsilon_y\mathfrak{B}(y,\mathfrak{t}_1),\mathfrak{t}_1)$$

which is a critical formula for a new ε-term (this is Bernays's point (3)). On the other hand, the formula $\mathfrak{A}(x)$ might contain another ε-expression, e.g., $\varepsilon_y\mathfrak{B}(x,y)$, in which case the corresponding ε-term would be of the form $\mathfrak{e} \equiv \varepsilon_x\mathfrak{A}(x,\varepsilon_y\mathfrak{B}(x,y))$. A critical formula corresponding to such a term is:

$$\begin{aligned}&\mathfrak{A}(\mathfrak{s},\varepsilon_y\mathfrak{B}(\mathfrak{s},y)) \rightarrow \mathfrak{A}(\varepsilon_x\mathfrak{A}(x,\varepsilon_y\mathfrak{B}(x,y)),\varepsilon_y\mathfrak{B}(\varepsilon_x\mathfrak{A}(x,\varepsilon_y\mathfrak{B}(x,y)),y)), \text{ i.e.,}\\ &\mathfrak{A}(\mathfrak{s},\varepsilon_y\mathfrak{B}(\mathfrak{s},y)) \rightarrow \mathfrak{A}(\mathfrak{e},\varepsilon_y\mathfrak{B}(\mathfrak{e},y))\end{aligned}$$

If in this formula the ε-term $\varepsilon_y\mathfrak{B}(\mathfrak{s},y)$ is replaced by some other term $\mathfrak{t}$, we get

$$\begin{aligned}&\mathfrak{A}(\mathfrak{s},\mathfrak{t}) \rightarrow \mathfrak{A}(\varepsilon_x\mathfrak{A}(x,\varepsilon_y\mathfrak{B}(x,y)),\varepsilon_y\mathfrak{B}(\varepsilon_x\mathfrak{A}(x,\varepsilon_y\mathfrak{B}(x,y)),y)), \text{ i.e.,}\\ &\mathfrak{A}(\mathfrak{s},\mathfrak{t}) \rightarrow \mathfrak{A}(\mathfrak{e},\varepsilon_y\mathfrak{B}(\mathfrak{e},y))\end{aligned}$$

which is no longer an instance of the ε-axiom. This is Bernays's point (1).

The second main difficulty is dealing with equality axioms, for again, the replacement of an ε-term $\varepsilon_x\mathfrak{A}(x,a)$ by $\mathfrak{t}$ might transform an instance of an quality axiom into

$$a = b \rightarrow \mathfrak{t} = \varepsilon_x\mathfrak{A}(x,b)$$

which no longer is an instance of an axiom. (This is Bernays's point (2)).

Bernays's proposed solution is rather involved and not carried out in general, but it seems to have prompted Ackermann to apply some methods from his own (1924) and von Neumann's (1927) ε-substitution proofs. Specifically, the final version of the first ε-theorem presented by Hilbert and Bernays (1939), where the solution of the difficulties is credited to Ackermann, uses double induction on the rank and degree of ε-expressions to deal with the first difficulty, and von Neumann's notion of ε-types to deal with the equality axiom.[8]

4 Hilbert's 'Conservativity Programme' and the practice of consistency proofs

A complete understanding of Hilbert's philosophy of mathematics requires an analysis of what may be called 'the practice of finitism'. Hilbert was, unfortunately, not always as clear as one would like in the exposition of his ideas about the finitary standpoint and of his project of consistency proofs. Only by analyzing the approaches by which he and his students attempted to carry out the consistency programme can we hope to get a complete picture of the principles and reasoning patterns he accepted as finitary, and about his views on the nature of logic and proof theory. The ε-substitution method, of course, was considered the most promising avenue in the quest for a consistency proof. The 'failed proof' discussed above shows that an alternative approach was, to a certain degree, pursued in parallel to the more well-known substitution method, and adds to the understanding we have of Hilbert's approach to proof theory and consistency proofs. The 'general consistency result' provides another example of how a consistency proof should be carried out according to Hilbert. Its particular interest lies in its general nature. Bernays's schematic formulation of the result underlines and makes explicit the conditions an axiomatic system should meet in order to be amenable to a consistency proof of the required form; it stresses once again the requirement of verifiability and decidability of atomic formulas.

Although Hilbert (and Bernays) have consistently presented the goal of the proof theoretic programme as giving a finitary consistency proof of classical, infinitary mathematics, it has become common among commentators to present the aim of Hilbert's Programme not as one of finding a proof of consistency, but of finding a proof of *conservativity of the ideal over the real*. Here, an appeal is made to Hilbert's distinction in (1926; 1928) between the ideal and real propositions in formalized mathematics: the real propositions are those that have finitary meaning, whereas the ideal propositions are those formulas which are added to the real part to 'round out' the theory, to make the uniform application of logical inferences possible, and which do not have a direct interpretation on finitist grounds (in particular, they may contain unbounded quantifiers). Hilbert likened the real propositions to those propositions of physical theories which can be verified by experiment (Hilbert, 1928, 475), and hence it is natural to interpret Hilbert's Programme as an instrumentalist enterprise where proof theory was supposed to show that whenever a real proposition can be proved by ideal methods, it can be proved by real methods alone. Following Detlefsen (1986), I shall call this *real-soundness* of ideal, formalized mathematics. The first one to present Hilbert's Programme as aiming for a finitary proof of real-soundness was von Neumann (1931). It was probably Kreisel who most consistently and influentially emphasized it in, e.g., (1951; 1958; 1968).

It must first of all be said that neither Hilbert nor Bernays presented the aim of the programme as that of finding a proof of real-soundness; they almost always talk of consistency, and never explicitly of conservativity or of justifying ideal mathematics by showing that whenever a real statement is provable by ideal means, it is also finitarily provable. In fact, there are only two places I could find where Hilbert formulates something close to a conservativity claim. In (1923), Hilbert writes: '[A] finite theorem can presumably [*vermutlich*] always be proved without the transfinite mode of inference [. . .] but this contention is of the same sort as the contention that every mathematical proposition can either be proved or refuted'. The use of 'presumably' and the qualification at the end suggest that Hilbert was not convinced that the transfinite is conservative over the finite. Moreover, this quote appears in the context of a general discussion of mathematical, not *meta*mathematical proof, and so should not be taken as a gloss on what the metamathematical consistency proof is supposed to establish. The second quote is from Hilbert (1926), where he writes

> For there is a condition, a single but absolutely necessary one, to which the use of the method of ideal elements is subject, and that is the *proof of consistency*; for, extension by the addition of ideals is legitimate only if no contradiction is thereby brought about in the old, narrower domain, that is, if the relations that result for the old objects whenever the ideal objects are eliminated are valid in the old domain. (p. 383)

This is likewise not as explicit and clear as one would like. Hilbert is here talking about the method of ideal elements in general, where this is readily understandable: when we extend real analysis to complex analysis by extending the domain to include imaginary numbers, the new theory should not prove any theorems about real numbers which aren't already true in the real numbers. The conservativity requirement formulated here is an explanation of consistency of the new ideal elements, theorems about them, and their consequences in the 'old, narrower domain', and it is the only place where it is so formulated.

The question then is: what is the 'old, narrower' domain in the case of proof theory; what are the real sentences for Hilbert? Hilbert reserved the label 'real formulas' quite clearly for variable-free formulas. In (1926), Hilbert did not use the term 'real' but he did say that the domain being extended by ideal objects are 'formulas to which contentual communications of finitary propositions [hence, in the main, numerical equations and inequalities] correspond' (380). Although general propositions of the form 'for all numerals $\mathfrak{a}$, $\mathfrak{b}$, $\mathfrak{a}+\mathfrak{b}=\mathfrak{b}+\mathfrak{a}$' are finitary propositions, the corresponding free-variable formula $a+b=b+a$ is 'no longer an immediate communication of something contentual at all, but a certain formal object . . . ' which does not mean anything in itself (380). Hilbert's (1926) was based on a lecture course given in the Winter semester 1924–25; and in the lecture notes to that course, Hilbert elaborates on the discussion from which the preceding quote is taken:

> The resulting formulas such as $a+b=b+a$ do not mean anything in themselves, any more than the numerals do, they are only images of our thoughts; but from these we can derive propositions, such as $2+3=3+2$ and we are thus led to consider these elementary propositions also as formulas, and to signify them as such; they then are formulas which mean those elementary unproblematic propositions. These formulas, which mean something, are the old objects, only in a new conception: all the added formulas, which do not mean anything in themselves, are the ideal objects of our theory.[9]

When Hilbert introduces the term 'real proposition', he likewise characterizes them as the 'formulas to which correspond contentual communication of finitary propositions (mainly numerical equations or inequalities, or more complex communications composed of these)' (Hilbert, 1928, 470).[10] Indeed, if the requirement of real-soundness is understood, as Hilbert and von Neumann do, by analogy with physical theories and observation statements, then real propositions must be *decidable*. So Hilbert (1928,

475) says that ‘only the real propositions are directly capable of verification’, and this only makes sense if ‘real proposition’ is understood as a variable-free decidable proposition about numerals.[11] Since, as Hilbert points out in a slightly different context (1928, 470), ‘one cannot, after all, try out all numbers’ a verification of a general (free-variable) proposition must consist of a general proof, and as such can hardly be called ‘direct’.[12]

Nevertheless, it has become common among commentators on Hilbert to take as real all quantifier-free formulas. Since in the formalisms considered, free-variable formulas are interderivable with their universal closures, one often takes the real formulas to be all Π_1 formulas. Smoryński (1977), for instance, claims that Hilbert's primary aim was that of establishing conservativity for Π_1-formulas, and that he pursued a consistency proof (only?) because it seemed more tractable and because consistency is equivalent to conservativity for Π_1-sentences. Prawitz (1981) similarly states the aim of the programme as aiming for ‘a demonstration in the real part of mathematics of the fact (if a fact) that every provable real sentence is true, i.e., that every sentence belonging to the real part which is proved by possible use of the ideal part is nevertheless true (according to the standards of the real part)’ (254) where ‘the real sentences comprise the decidable ones and the ones of the form $\forall x A(x)$ where each instance $A(t)$ is decidable’ (256). He stated furthermore that this was to be done by proving the claim that ‘[f]or each proof p in [ideal mathematics] T and for each real sentence A in T: if p is a proof of A in T, then A is true’ (257). Kitcher (1976) holds a similar view, and Giaquinto (1983, 124) adopts Smoryński's formulation of Hilbert's Programme. The question of whether Π_1-conservativity was a requirement on or an aim of Hilbert's Programme is also of interest because it underlies an argument against Hilbert's Programme based on the first incompleteness theorem (Detlefsen, 1990).

As pointed out above, such a formulation of the *aim* of Hilbert's Programme is not to be found in Hilbert. Hilbert's sparse remarks, as well as those contemporary formulations which do focus on conservativity such as von Neumann's (1931), call for a proof of conservativity for variable-free, decidable sentences only. Moreover, and here is where one can bring the historical discussion in the first part of this paper to bear, the *practice* of consistency proof only directly establishes conservativity for quantifier-free sentences. Consistency proofs based on the epsilon theorem (e.g., the general consistency result) as well as those based on ε-substitution showed that one can transform a proof of a closed, variable-free formula (perhaps using ideal methods, e.g., the transfinite axioms) into a purely real, variable-free proof which essentially amounts to a calculational verification of the numerical claim expressed by the end-formula.

It was only Bernays in the *Grundlagen der Mathematik* who drew the conclusion that the consistency proofs themselves actually established not only the truth of variable-free formulas provable by ideal methods, but also of free-variable theorems. In this context, Bernays used the term ‘verifiable’ (*verifizierbar*): a free-variable formula $\mathfrak{A}(a)$ is verifiable if each numerical instance $\mathfrak{A}(\mathfrak{z})$ is true. He then stated the following consequence of consistency proofs: every derivable free-variable formula is verifiable (This is a consequence of a consistency proof for quantifier-free formulations of systems of arithmetic in Hilbert and Bernays 1934, 248,298; Bernays also pointed it out as a consequence of the ‘general consistency result’ in Hilbert and Bernays 1939, 36). The idea is simple: If $\mathfrak{A}(a)$ (equivalently, $\forall x \mathfrak{A}(x)$ is derivable, then the following method constitutes a finitary proof that, for any $\mathfrak{z}$, $\mathfrak{A}(\mathfrak{z})$ is true. From the derivation of $\mathfrak{A}(a)$ we obtain a derivation of $\mathfrak{A}(\mathfrak{z})$ by substitution. The procedure given in the consistency proof transforms this derivation into a variable-free derivation of $\mathfrak{A}(\mathfrak{z})$, which codifies a finitary calculation that $\mathfrak{A}(\mathfrak{z})$ is true.[13]

So why has Hilbert been held to Π_1-conservativity of ideal mathematics? Kreisel (1951) cites Bernays's results; but in Kreisel (1960) and later, he instead points to an argument in (Hilbert, 1928, 474). This argument, like Bernays's, shows how a finitary consistency proof for a system T yields a finitary proof of every free-variable formula provable in T. Unlike Bernays's remark, it does not rely on a particular form of the consistency proof, but on the mere assumption that a finitary consistency proof is available. Assume there is a derivation of $\mathfrak{A}(a)$ (equivalently, of $\forall x \mathfrak{A}(x)$). The task is to show that, for any given $\mathfrak{z}$, $\mathfrak{A}(\mathfrak{z})$ is true. Suppose it weren't. Then $\neg\mathfrak{A}(\mathfrak{z})$ would be true, and, because T proves all true variable-free formulas, there would be a derivation of $\neg\mathfrak{A}(\mathfrak{z})$. But from the derivation of $\mathfrak{A}(a)$ we obtain, by substitution, a derivation of $\mathfrak{A}(\mathfrak{z})$, and hence T is inconsistent. But we have a finitary consistency proof of T, so this cannot be the case. Hence, $\mathfrak{A}(\mathfrak{z})$ must be true. (Note that this proof uses tertium non datur, but this is a finitarily admissible application to a variable-free numerical statement $\mathfrak{A}(\mathfrak{z})$.) This latter result is presented as a surprising application of proof theory. It came several years after the two models of consistency proofs—ε-substitution and the 'failed proof'—had already been worked out. Hence Hilbert here articulates a conservation property which, *as it turned out, follows* from consistency—and not, as it were, a property which the consistency proofs were all along supposed to establish.

To the extent Hilbert saw the original aim of his project as one of proving conservativity, this aim was to prove conservativity for decidable, variable-free 'real' propositions, but not for free-variable general propositions. There is only one passage in which conservativity for general statements is discussed prior to 1934 (viz., in 1928), and there it is presented as an *application* and not a statement of the project. Both the 'failed proof' and the ε-substitution method in the writings of Hilbert, Ackermann, and von Neumann at the time were given the formulation 'if there were an ideal proof of $0 = 1$, then there would be a variable-free (real) proof of $0 = 1$'. If conservativity had played a significant role in the minds of those involved, it would have been obvious to formulate the proofs so that they established conservativity. It is interesting to note that whereas the ε-substitution method only directly establishes conservativity for variable-free statements (i.e. by applying it to proofs of variable-free formulas instead of the specific $0 = 1$), the 'failed proof' could have been formulated for proofs of any ε-free formula (not just $0 \neq 0$), as it eventually was in the first ε-theorem. By itself, this would only have yielded a consistency proof for logic; for arithmetic, something like the strategy in the 'general consistency result' is needed. It is nevertheless conceivable that this possible generalization of the 'failed proof' suggested a strategy of how to remove in general ideal elements in the form of ε-terms from proofs of formulas not involving such ideal terms. However, the failed proof was never mentioned in print, and likewise it was never noted that it would be possible to generalize the then-existing strategies to give conservativity proofs (in particular, ε-substitution) until well into the 1930s.

In addition to the light they shed on the development of the conceptual framework of Hilbert's Programme, the results about the ε-calculus discussed above are, I think, of independent and genuine importance. Interest in the historical development of Hilbert's Programme has seen a marked increase in the last decade or so, and naturally the ε-calculus takes center stage in the work on logic in Hilbert's school. Independently of Hilbert studies, renewed interest in the theory and applications of the ε-calculus (Avigad and Zach, 2002) warrant a closer look at the foundations and origins of the epsilon calculus—the 'failed proof' is a rather important part of that story.

Acknowledgments

Many thanks to Jeremy Avigad and an anonymous referee for several helpful suggestions, and to Hans Richard Ackermann for permission to quote from the Bernays–Ackermann correspondence.

Notes

[1] Hilbert and Bernays (1923b, 30–31). Following Hilbert and Bernays (1934), we will use the following notation: $a, b, \ldots$ stand for free variables, whereas $x, y, \ldots$ are bound variables. $A, B, \ldots$ are formula variables. $\mathfrak{A}, \mathfrak{B}, \ldots$ are metavariables for formulas, and $\mathfrak{n}$, $\mathfrak{z}$ stand for numerical terms. For uniformity, the notation in some quotations has been adjusted.

[2] The τ-operator was mentioned in Hilbert (1922c) and formally introduced, together with the transfinite axiom, in Hilbert (1923). In a course given in Winter 1922/23 (Hilbert and Bernays, 1923b,a), Hilbert and Bernays introduce the ε-operator for the first time, although initially with the same interpretation (as a counterexample) as the τ-operator. In notes appended to the typescript (1923b), the final form of the ε-operator and ε-axioms appears for the first time. Kneser's notes to the course (Hilbert and Bernays, 1923a) do not contain this final version, suggesting that this change was made after the conclusion of the 1922/23 course.

[3] The basic idea was presented in Hilbert (1923) and in the course mentioned (Hilbert and Bernays, 1923b,a), for discussion, see Zach (2003). Roughly, the idea is this: first replace every ε-term by 0. The instances of the transfinite axiom for an ε-term $\varepsilon_x \mathfrak{A}(x)$ in the proof then become formulas of the form $\mathfrak{A}(\mathfrak{n}) \to \mathfrak{A}(0)$. If this formula is false, $\mathfrak{A}(\mathfrak{n})$ is true. In the next iteration of the procedure, replace $\varepsilon_x \mathfrak{A}(x)$ by $\mathfrak{n}$. The difficulty is to extend this idea to the case where more than one ε-term, and in particular, when nested ε-terms occur in the proof.

[4] Gentzen (1934) mentions that a version of Herbrand's Theorem is a consequence of his 'Verschärfter Hauptsatz'. He does not, however, spell out the details.

[5] Ackermann to Bernays, June 25, 1925. 9pp. ETH Zürich Library, Hs. 975.96. Ackermann writes: '[Critical formulas] are eliminated in the way in which Hilbert wanted to earlier [*früher*]' (p. 3). Ackermann here tries to use the idea of the 'failed proof ' to fix his own faulty ε-substitution proof.

[6] This is essentially the same proof as the one presented as the 'Hilbertsche Ansatz' by Hilbert and Bernays (1939, 21). There, the proof is carried out for end formulas $\mathfrak{B}$ instead of the specific $0 \neq 0$. The only other difference is that instead of using induction on n, Bernays constructs one derivation of $\overline{\mathfrak{A}}(\mathfrak{t}_1) \land \ldots \land \overline{\mathfrak{A}}(\mathfrak{t}_n) \to \mathfrak{F}$ and n derivations of $\mathfrak{A}(\mathfrak{t}_i) \to \mathfrak{F}$, and then applies n-fold case distinction.

[7] 'Anlässlich der Arbeit für das Grundlagenbuch sah ich mich dazu angetrieben, den zweiten Hilbertschen Wf.-Beweis für das ε-Axiom, den sogenannten „verunglückten" Beweis, nochmals zu überlegen, und es scheint mir jetzt, dass dieser sich doch richtig stellen lässt.

Da ich weiss, dass man sich im Gebiete dieser Beweise äusserst leicht versieht, so möchte ich Ihnen meine Überlegung zur Prüfung vorlegen.

Die bisherigen Hindernisse für die Durchführung des Beweises bestanden in dreierlei:

1. Es konnte vorkommen, dass durch die Ersetzungen, die bei der Behandlung einer kritischen Formel auszuführen waren, eine andere kritische Formel ihre characteristische Gestalt verlor, ohne doch in eine beweisbare Formel überzugehen.

2. Die Berücksichtigung des zweiten Gleichheits-Axioms, das ja in seiner Anwendung auf die ε-Funktion [Footnote: abgesehen von der harmlosen Anwendung, bestehend in d. Einsetzung eines ε-Funktionals für eine Grundvariable im Gleichheits-Axiom.] — es handelt sich hier immer nur um $\varepsilon_a \mathfrak{A}(x)$, wobei x eine *Grund*variable ist—durch das Axiom

$$(G) \qquad a = b \rightarrow (\varepsilon_x A(x,a) = \varepsilon_x A(x,b))$$

vertreten werden kann, machte Schwierigkeiten.

3. Es kam vor, dass nach gelungener Ausschaltung eines ε-Funktionals ein anderes ε-Funktional hinzurat, sodass im ganzen keine Reduktion nachweisbar war.'

Bernays to Ackermann, October 16, 1929. Manuscript, 13 pages. In the possession of Hans Richard Ackermann. See also Ackermann (1983).

[8] The proof of the first epsilon theorem, the 'general consistency result', and Herbrand's theorem of Hilbert and Bernays (1939) are also contained in lectures Bernays gave at Princeton (1935–36).

[9] 'Diese so entstandenen Formeln wie $a+b=b+a$ bedeuten an sich nichts, so wenig wie die Zahlzeichen, sie sind nur Abbilder unserer Gedanken; wohl aber können aus ihnen Aussagen abgeleitet werden, wie $2+3=3+2$ und wir kommen so dazu, diese elementaren Aussagen auch als Formeln aufzufassen und zu bezeichnen; es sind das dann Formeln, die jene elementaren unproblematischen Aussagen bedeuten. Diese Formeln, die etwas bedeuten, sind die alten Gebilde, nur in neuer Auffassung; alle die hinzugefügten Formeln hingegen, die an sich nichts bedeuten, sind die idealen Gebilde unserer Theorie.' (Hilbert, 1924–25, 126–127).

[10] Smoryński (1989) also holds that the real propositions are variable-free formulas, and introduces a three-fold distinction between real propositions (variable-free), finitary general propositions (quantifier free, but containing free variables) and ideal propositions (containing quantifiers). I agree with Detlefsen (1990) that Smorynski's reading of Hilbert is flawed; but do not agree with Detlefsen's assessment of the real/ideal distinction. This distinction is orthogonal to the finitary/infinitary distinction. The real/ideal distinction is a purely syntactic distinction between those formulas which do not contain variables and those that do. Every real formula can be immediately interpreted as a particular finitary propositions, and its truth is decidable. The ideal formulas cannot immediately be interpreted finitarily, and when they do *admit* of a finitary interpretation (e.g., free-variable formulas as statements about 'any given numeral'), these interpretations are problematic because they do not obey the usual logical laws (e.g., they are 'incapable of being negated').

[11] There is no reason to think that 'verification' as used here by Hilbert is the same notion as '*verifizierbar*' in (Hilbert and Bernays, 1934, 238), which does apply to free-variable formulas and indeed even to formulas with bound variables. See below.

[12] This interpretation is not undermined by the fact that people outside the Hilbert school did not always understand 'real proposition' as variable-free; e.g., Weyl (1928). In his remarks at the Königsberg conference (1986, 200–202), Gödel suggests that he understands the real (in his terminology: 'meaningful') propositions as including not only the variable-free formulas, but also formulas of the form $(Ex)F(x)$. In his review of von Neumann (1931), however, he specifically reports the aim of Hilbert's Programme as 'showing that every numerical formula verifiable (calculable) in finitely many steps that can be derived according to the rules of the game by which classical mathematics is played must turn out to be correct when actually calculated' (Gödel, 1986, 249).

[13] Proof theorists will realize that Gentzen's (1936) and Ackermann's (1940) consistency proofs yield conservativity result even for Π_2 sentences, in the follwing sense: If *PA* proves $\forall x \exists y A(x,y)$ then there is a $< \varepsilon_0$-recursive function f so that a suitable extension of *PRA* proves $A(x, f(x))$. Gentzen (1936) only vaguely hinted at this consequence in his paper, saying that the reduction rules in his consistency proof provide a finitist sense to actualist, i.e., infinitary propositions. Again it was Kreisel (1951) who expanded on this idea.

References

Ackermann, H. R. 1983. 'Aus dem Briefwechsel Wilhelm Ackermanns', *History and Philosophy of Logic* **4**, 181–202.

Ackermann, W. 1924. 'Begründung des "tertium non datur" mittels der Hilbertschen Theorie der Widerspruchsfreiheit', *Mathematische Annalen* **93**, 1–36.

Ackermann, W. 1940. ‘Zur Widerspruchsfreiheit der Zahlentheorie’, *Mathematische Annalen* **117**, 162–194.

Avigad, J. and Zach, R. 2002. ‘The epsilon calculus’, Stanford Encyclopedia of Philosophy, http://plato.stanford.edu/entries/epsilon-calculus/.

Benacerraf, P. and Putnam, H., eds. 1983. *Philosophy of Mathematics*, Cambridge: Cambridge University Press, 2nd edition.

Bernays, P. 1935–36. ‘Logical calculus’, Unpublished typescript, Institute for Advanced Study, Princeton.

Detlefsen, M. 1986. *Hilbert’s Program*, Dordrecht: Reidel.

Detlefsen, M. 1990. ‘On an alleged refutation of Hilbert’s program using Gödel’s first incompleteness theorem’, *Journal of Philosophial Logic* **19**, 343–377.

Ewald, W. B., ed. 1996. *From Kant to Hilbert. A Source Book in the Foundations of Mathematics*, volume 2, Oxford: Oxford University Press.

Gentzen, G. 1934. ‘Untersuchungen über das logische Schließen I–II’, *Mathematische Zeitschrift* **39**, 176–210, 405–431, English translation in Gentzen (1969, 68–131).

Gentzen, G. 1936. ‘Die Widerspruchsfreiheit der reinen Zahlentheorie’, *Mathematische Annalen* **112**, 493–565, English translation in Gentzen (1969, 132–213).

Gentzen, G. 1969. *The Collected Papers of Gerhard Gentzen*, Amsterdam: North-Holland.

Giaquinto, M. 1983. ‘Hilbert’s philosophy of mathematics’, *British Journal for Philosophy of Science* **34**, 119–132.

Gödel, K. 1986. *Collected Works*, volume 1, Oxford: Oxford University Press.

Herbrand, J. 1930. *Recherches sur la théorie de la démonstration*, Doctoral dissertation, University of Paris, English translation in Herbrand (1971, 44–202).

Herbrand, J. 1971. *Logical Writings*, Harvard University Press.

Hilbert, D. 1910. ‘Elemente und Prinzipienfragen der Mathematik’, Vorlesung, Sommer-Semester 1910. Lecture notes by Richard Courant. Unpublished manuscript, 163 pp. Bibliothek, Mathematisches Institut, Universität Göttingen. 16.206t14.

Hilbert, D. 1922a. ‘Grundlagen der Mathematik’, Vorlesung, Winter-Semester 1921–22. Lecture notes by Paul Bernays. Unpublished typescript. Bibliothek, Mathematisches Institut, Universität Göttingen.

Hilbert, D. 1922b. ‘Grundlagen der Mathematik’, Vorlesung, Winter-Semester 1921–22. Lecture notes by Helmut Kneser. Unpublished manuscript, three notebooks.

Hilbert, D. 1922c. ‘Neubegründung der Mathematik: Erste Mitteilung’, *Abhandlungen aus dem Seminar der Hamburgischen Universität* **1**, 157–77, series of talks given at the University of Hamburg, July 25–27, 1921. Reprinted with notes by Bernays in Hilbert (1935, 157–177). English translation in Mancosu (1998, 198–214) and Ewald (1996, 1115–1134).

Hilbert, D. 1923. ‘Die logischen Grundlagen der Mathematik’, *Mathematische Annalen* **88**, 151–165, lecture given at the Deutsche Naturforscher-Gesellschaft, September 1922. Reprinted in Hilbert (1935, 178–191). English translation in Ewald (1996, 1134–1148).

Hilbert, D. 1924–25. ‘Über das Unendliche’, Vorlesung, Winter-Semester 1924–25. Lecture notes by Lothar Nordheim. Unpublished typescript. Bibliothek, Mathematisches Institut, Universität Göttingen.

Hilbert, D. 1926. ‘Über das Unendliche’, *Mathematische Annalen* **95**, 161–90, lecture given Münster, 4 June 1925. English translation in van Heijenoort (1967, 367–392).

Hilbert, D. 1928. ‘Die Grundlagen der Mathematik’, *Abhandlungen aus dem Seminar der Hamburgischen Universität* **6**, 65–85, English translation in van Heijenoort (1967, 464-479).

Hilbert, D. 1935. *Gesammelte Abhandlungen*, volume 3, Berlin: Springer.

Hilbert, D. and Bernays, P. 1923a. ‘Logische Grundlagen der Mathematik’, Winter-Semester 1922–23. Lecture notes by Helmut Kneser. Unpublished manuscript.

Hilbert, D. and Bernays, P. 1923b. ‘Logische Grundlagen der Mathematik’, Vorlesung, Winter-Semester 1922–23. Lecture notes by Paul Bernays, with handwritten notes by Hilbert. Hilbert-Nachlaß, Niedersächsische Staats- und Universitätsbibliothek, Cod. Ms. Hilbert 567.

Hilbert, D. and Bernays, P. 1934. *Grundlagen der Mathematik*, volume 1, Berlin: Springer.

Hilbert, D. and Bernays, P. 1939. *Grundlagen der Mathematik*, volume 2, Berlin: Springer.

Kitcher, P. 1976. ‘Hilbert’s epistemology’, *Philosophy of Science* **43**, 99–115.

Kreisel, G. 1951. ‘On the interpretation of non-finitist proofs. Part I’, *Journal of Symbolic Logic* **16**, 241–267.

Kreisel, G. 1958. ‘Hilbert’s programme’, *Dialectica* **12**, 346–372, reprinted as Kreisel (1983).

Kreisel, G. 1960. ‘Ordinal logics and the characterization of informal notions of proof’, in J. A. Todd, ed., ‘Proceedings of the International Congress of Mathematicians. Edinburgh, 14–21 August 1958’, 289–299, Cambridge: Cambridge University Press.

Kreisel, G. 1968. ‘A survey of proof theory’, *Journal of Symbolic Logic* **33**, 321–388.

Kreisel, G. 1983. ‘Hilbert’s programme’, in (Benacerraf and Putnam, 1983), 207–238.

Leisenring, A. C. 1969. *Mathematical Logic and Hilbert’s ε-Symbol*, London: MacDonald.

Mancosu, P., ed. 1998. *From Brouwer to Hilbert. The Debate on the Foundations of Mathematics in the 1920s*, Oxford: Oxford University Press.

Prawitz, D. 1981. ‘Philosophical aspects of proof theory’, in G. Fløistad, ed., ‘Contemporary Philosophy. A New Survey’, volume 1, 235–77, The Hague: Nijhoff.

Smoryński, C. 1977. ‘The incompleteness theorems’, in J. Barwise, ed., ‘Handbook of Mathematical Logic’, 821–865, Amsterdam: North-Holland.

Smoryński, C. 1989. ‘Hilbert’s programme’, *CWI Quarterly* **1**, 3–59.

van Heijenoort, J., ed. 1967. *From Frege to Gödel. A Source Book in Mathematical Logic, 1897–1931*, Cambridge, Mass.: Harvard University Press.

von Neumann, J. 1927. ‘Zur Hilbertschen Beweistheorie’, *Mathematische Zeitschrift* **26**, 1–46.

von Neumann, J. 1931. ‘Die formalistische Grundlegung der Mathematik’, *Erkenntnis* **2**, 116–34, English translation in: Benacerraf and Putnam (1983, 61–65).

Weyl, H. 1928. ‘Diskussionsbemerkungen zu dem zweiten Hilbertschen Vortrag über die Grundlagen der Mathematik’, *Abhandlungen aus dem Mathematischen Seminar der Universität Hamburg* **6**, 86–88, English translation in van Heijenoort (1967, 480-484).

Zach, R. 2003. ‘The practice of finitism. Epsilon calculus and consistency proofs in Hilbert’s Program’, *Synthese* **137**, 211–259.