arXiv:math/0204255v1 [math.LO] 20 Apr 2002

HILBERT'S "*VERUNGLÜCKTER BEWEIS*," THE FIRST EPSILON THEOREM, AND CONSISTENCY PROOFS

RICHARD ZACH

Abstract. On the face of it, Hilbert's Program was concerned with proving consistency of mathematical systems in a finitary way. This was to be accomplished by showing that that these systems are conservative over finitistically interpretable and obviously sound quantifier-free subsystems. One proposed method of giving such proofs is Hilbert's epsilon-substitution method. There was, however, a second approach which was not refelected 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." An analysis of this so-called "failed proof" lends further support to an interpretation of Hilbert according to which he was expressly concerned with conservatitvity proofs, even though his publications only mention consistency as the main question.

§1. Introduction. The aim of Hilbert's program for consistency proofs in the 1920s is well known: to formalize mathematics, and to give finitistic 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,

> 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 finitistic 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 finitistic 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 finitistically 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 program 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 very beginning of Hilbert's Program. 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 never made it into 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 proof of the first epsilon theorem by Hilbert and Ackermann a decade later. 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 to an understanding 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, the first published correct proof of Herbrand's theorem is that given by Bernays in *Grundlagen der Mathematik II* [Hilbert and Bernays 1939] based on the first ε-theorem. 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 $B_1 \vee \ldots \vee B_n$ 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").[4]

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).

The proof of the conservativity of the ε-calculus takes the same form as other "direct" consistency proofs given by Hilbert and his students in the 1920s (e.g., Ackermann [1924]). Suppose F proved a contradiction. We may assume the contradiction is a variable-free formula (in arithmetic, e.g., $0 \neq 0$). By the first epsilon-theorem, there would be such a proof already in the elementary calculus of free variables, but this can be easily shown to be consistent.

Bernays's "general consistency result" consists in extending the consistency proof for the pure ε-calculus to certain axiomatic theories for which the first ε-theorem also applies. Let us first outline the 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:

$$0 \neq x + 1$$
$$x = \delta(x + 1)$$

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.[5]

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

$$0 \neq \mathfrak{n} + 1$$
$$\mathfrak{n} = \mathfrak{n}$$

(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 $0 \neq 0$.

This proof was already presented by Hilbert in his courses on the foundations of mathematics of 1921/22 and 1922/23, and was extended there to axioms for primitive recursive functions. Ackermann then extended it further to include second-order primitive recursive functions.[6] 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 them *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 can 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.[7]

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 finitistic proof: It can be finitistically verified that any *given* proof in fact has true initial formulas (and hence, a true end formula).

§3. Hilbert's *Verunglückter Beweis*. The "general consistency result" is first formulated in print in Hilbert and Bernays [1939], but Hilbert had something like this in mind already in the early/mid 1920s. When working on *Grundlagen der Mathematik* in the late 1920s, Bernays revisits the idea, which had been abandoned in favour of the ε-substitution method. In correspondence with Ackermann in 1929 (discussed below), Hilbert 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* [Elements and Principle of Mathematics], taught in the Summer Semester 1910 in Göttingen.[8] Although it bears a note by Hilbert "Insertion in WS [Winter Semester] 1920", the notation used in the sketch suggests that it was written after sometime in 1923, and in any event after 1922, when the ε notation was first introduced. It bears the title "Consistency proof for the logical axiom of choice $Ab \to A\varepsilon_a Aa$, in the simplest case."

In the sketch 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

$$\begin{gathered}\mathfrak{A}(\mathfrak{t}_1) \to \mathfrak{A}(\varepsilon_x \mathfrak{A}(x)) \\ \vdots \\ \mathfrak{A}(\mathfrak{t}_n) \to \mathfrak{A}(\varepsilon_x \mathfrak{A}(x))\end{gathered}$$

are all the critical formulas involving $\mathfrak{A}$ in a proof of $\mathfrak{B}$. 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

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

The premises of this inference are propositional axioms. Thus we obtain a proof of $\overline{\mathfrak{A}}(\mathfrak{t}_1) \to \mathfrak{B}$ 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 $\mathfrak{A}(\mathfrak{t}_1) \to \mathfrak{B}$ without critical formulas. Putting the two proofs together and applying the law of excluded middle, we have found a proof of $\mathfrak{B}$ using only critical formulas for $\mathfrak{t}_2, \ldots, \mathfrak{t}_n$. By induction on n, there is a proof of $\mathfrak{B}$ using no critical formulas at all. In the resulting proof, we can replace $\varepsilon_x \mathfrak{A}(x)$ everywhere by 0.[9]

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 rethink 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.[10]

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

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

a formula

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

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:

$$\mathfrak{A}(\mathfrak{s}, \varepsilon_y \mathfrak{B}(\mathfrak{s}, y)) \to \mathfrak{A}(\varepsilon_x \mathfrak{A}(x, \varepsilon_y \mathfrak{B}(x,y)), \varepsilon_y \mathfrak{B}(\varepsilon_x \mathfrak{A}(x, \varepsilon_y \mathfrak{B}(x,y)), y)), \text{ i.e.,}$$
$$\mathfrak{A}(\mathfrak{s}, \varepsilon_y \mathfrak{B}(\mathfrak{s}, y)) \to \mathfrak{A}(\mathfrak{e}, \varepsilon_y \mathfrak{B}(\mathfrak{e}, y))$$

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

$$\mathfrak{A}(\mathfrak{s}, \mathfrak{t}) \to \mathfrak{A}(\varepsilon_x \mathfrak{A}(x, \varepsilon_y \mathfrak{B}(x, y)), \varepsilon_y \mathfrak{B}(\varepsilon_x \mathfrak{A}(x, \varepsilon_y \mathfrak{B}(x, y)), y)), \text{ i.e.,}$$
$$\mathfrak{A}(\mathfrak{s}, \mathfrak{t}) \to \mathfrak{A}(\mathfrak{e}, \varepsilon_y \mathfrak{B}(\mathfrak{e}, y))$$

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

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

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, use 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.

§4. The relevance of Hilbert's "failed proof". As I have argued in Zach [2002], a complete understanding of Hilbert's philosophy of mathematics requires an analysis of what I have called "the practice of finitism." Hilbert was, unfortunately, not always as clear as one would like in the exposition of his ideas about the finitist 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 program can we hope to get a complete picture of the principles and reasoning patterns he accepted as finitist, and about his views on the nature of logic and axiomatics. The ε-substitution method, of course, was considered the most promising avenue in the quest for a consistency proof. The perhaps surprising historical details outlined above, showing that an alternative approach was, to a certain degree, pursued in parallel to the more well-known substitution method, adds significantly to the understanding we have of Hilbert's approach to logic and proof theory.

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. It also provides another piece of evidence that when Hilbert spoke of consistency proof (in his publications) he really was interested in certain conservativity results (conservativity of the "ideal" over the "real" parts of mathematics). Such an interpretation is not uncommon among those writing on Hilbert's program,[11] but not explicit to a comparable degree in Hilbert's own publications. He only speaks of consistency, not of conservativity. However, not only did the consistency proofs in his school *establish* conservativity, but they were clearly specifically intended to.

This indicates that a reading of Hilbert's program as aiming for conservativity is not just a reconstruction, but reflects Hilbert's own intentions.

In addition to the light these results shed on the conceptual framework of Hilbert's program, the genesis of the ε-calculus is of independent and genuine importance. Interest in the historical development of Hilbert's program has seen a marked increase in the last decade or so, and naturally the ε-calculus takes center stage in the development of logic in Hilbert's school. Independently of Hilbert studies, renewed interest in the theory and applications of the ε-calculus[12] warrant a closer look at the foundations and origins of the epsilon calculus—the "failed proof" is a rather important piece of the puzzle.

NOTES

[1]Hilbert and Bernays [1923a, 30–31]. In accordance with the notation in Hilbert and Bernays [1934], [1939], we use the following notation: $a, b, \ldots$ stand for free variables, whereas $x, y, \ldots$ are bound variables. $A, B, \ldots$ are formula variables. $\mathfrak{A}, \mathfrak{B}, \ldots$, indicate actual formulas—not formula variables—and $\mathfrak{n}, \mathfrak{z}$, denote numerical terms. For uniformity, we 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]. The change to the dual ε-operator was carried out in a course given in Winter 1922/23 Hilbert and Bernays [1923a], [1923b].

[3]The basic idea was presented in Hilbert [1923] and in the course mentioned Hilbert and Bernays [1923a], [1923b], for discussion, see Zach [2002]. 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, nested ε-terms occur in the proof.

[4]Hilbert and Bernays [1939, 18].

[5]Hilbert and Bernays [1934], p. 225.

[6]Hilbert [1922a], [1922b], Hilbert and Bernays [1923a], [1923b], Ackermann [1924]; for discussion see Zach [2002].

[7]Hilbert and Bernays [1939], p. 36.

[8]Bibliothek, Mathematisches Institut, Universität Göttingen, 16.206t14.

[9]This is essentially the same proof as the one presented as the "Hilbertsche Ansatz" by Hilbert and Bernays [1939, 21]. The only difference is that instead of using induction on n, Bernays constructs one derivation of $\overline{\mathfrak{A}}(\mathfrak{t}_1) \wedge \ldots \wedge \overline{\mathfrak{A}}(\mathfrak{t}_n) \to \mathfrak{F}$ and n derivations of $\mathfrak{A}(\mathfrak{t}_i) \to \mathfrak{F}$, and then applies n-fold case distinction.

[10]"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

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

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].

[11]Kreisel [1960], for instance, stresses Hilbert's aim of not just proving consistency, but of proving conservativity by removing transfinite, "ideal" elements from proofs of "real" propositions. Smoryński [1977] suggests that Hilbert's motivation for proving consistency is the aim of establishing conservativity, since consistency establishes conservativity for Π_1-sentences. It is doubtful, however, that Hilbert was aware of this consequence. He *was* expressly interested in conservativity (for quantifier free sentences), *because* it implies consistency. Conservativity is also emphasized by those giving an instrumentalist reading of Hilbert's project, e.g., Detlefsen [1986] and Sieg [1990].

[12]For an overwiew, see Avigad and Zach [2002].

References

Hans Richard Ackermann [1983], *Aus dem Briefwechsel Wilhelm Ackermanns*, ***History and Philosophy of Logic***, vol. 4, pp. 181–202.

Wilhelm Ackermann [1924], *Begründung des "tertium non datur" mittels der Hilbertschen Theorie der Widerspruchsfreiheit*, ***Mathematische Annalen***, vol. 93, pp. 1–36.

Jeremy Avigad and Richard Zach [2002], *Epsilon calculus*, Stanford Encyclopedia of Philosophy, http://plato.stanford.edu/entries/epsilon-calculus/.

Michael Detlefsen [1986], ***Hilbert's program***, Reidel, Dordrecht.

William Bragg Ewald (editor) [1996], ***From Kant to Hilbert. A source book in the foundations of mathematics***, vol. 2, Oxford University Press, Oxford.

Jaques Herbrand [1930], *Recherches sur la théorie de la démonstration*, ***Doctoral dissertation***, University of Paris, English translation in Herbrand [1971, 44–202].

Jaques Herbrand [1971], ***Logical writings***, Harvard University Press.

David Hilbert [1922a], *Grundlagen der Mathematik*, Vorlesung, Winter-Semester 1921–22. Lecture notes by Paul Bernays. Unpublished typescript. Bibliothek, Mathematisches Institut, Universität Göttingen.

David Hilbert [1922b], *Grundlagen der Mathematik*, Vorlesung, Winter-Semester 1921–22. Lecture notes by Helmut Kneser. Unpublished manuscript, three notebooks.

David Hilbert [1922c], *Neubegründung der Mathematik: Erste Mitteilung*, ***Abhandlungen aus dem Seminar der Hamburgischen Universität***, vol. 1, pp. 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].

David Hilbert [1923], *Die logischen Grundlagen der Mathematik*, ***Mathematische Annalen***, vol. 88, pp. 151–165, Lecture given at the Deutsche Naturforscher-Gesellschaft, September 1922. Reprinted in Hilbert [1935, 178–191]. English translation in Ewald [1996, 1134–1148].

David Hilbert [1935], ***Gesammelte Abhandlungen***, vol. 3, Springer, Berlin.

David Hilbert and Paul Bernays [1923a], *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.

David Hilbert and Paul Bernays [1923b], *Logische Grundlagen der Mathematik*, Winter-Semester 1922–23. Lecture notes by Helmut Kneser. Unpublished manuscript..

David Hilbert and Paul Bernays [1934], ***Grundlagen der Mathematik***, vol. 1, Springer, Berlin.

David Hilbert and Paul Bernays [1939], ***Grundlagen der Mathematik***, vol. 2, Springer, Berlin.

Georg Kreisel [1960], *Ordinal logics and the characterization of informal notions of proof*, ***Proceedings of the international congress of mathematicians. edinburgh, 14–21 august 1958*** (Cambridge) (J. A. Todd, editor), Cambridge University Press, pp. 289–299.

A. C. Leisenring [1969], ***Mathematical logic and hilbert's ε-symbol***, MacDonald, London.

Paolo Mancosu (editor) [1998], ***From Brouwer to Hilbert. The debate on the foundations of mathematics in the 1920s***, Oxford University Press, Oxford.

Wilfrid Sieg [1990], *Reflections on Hilbert's program*, ***Acting and reflecting*** (Wilfrid Sieg, editor), Kluwer, Dordrecht, pp. 171–82.

Craig Smoryński [1977], *The incompleteness theorems*, ***Handbook of mathematical logic*** (Jon Barwise, editor), North-Holland, Amsterdam, pp. 821–865.

Johann von Neumann [1927], *Zur Hilbertschen Beweistheorie*, ***Mathematische Zeitschrift***, vol. 26, pp. 1–46.

Richard Zach [2002], *The practice of finitism. Epsilon calculus and consistency proofs in Hilbert's Program*, ***Synthese***, forthcoming.

DEPARTMENT OF PHILOSOPHY
UNIVERSITY OF CALGARY
2500 UNIVERSITY DRIVE N.W.
CALGARY, ALBERTA T2N 1N4, CANADA
E-mail: rzach@ucalgary.ca
URL: http://www.ucalgary.ca/~rzach/