The metamathematics of very weak arithmetics

Curtis Franks

October 30, 2005

Abstract

Feferman introduced intensional arithmetizations in [1960] and stressed that some metatheoretic results must be executable thus if they are to be semantically contentful. This program is particularly interesting in weak arithmetics where formal results like Gödel's second incompleteness theorem are provable only in nonstandard ways, skirting the conditions for intensional adequacy. We examine some recent attempts to explain how one might get meaningful metatheoretic results for weak theories through interpretations of stronger theories where these formal results have straightforward proofs. These attempts fail doubly by assuming incorrectly (1) that the proper conditions for intensionality are independent of the theory one is investigating and (2) that the interpretability of one theory in another enjoins the latter theory with the former's semantic richness. The significance of the result for philosophical foundationalism is that metamathematical investigations must be sensitive to the strength of the system one is studying, and in particular the formal notion of relative interpretability has in many cases no semantic significance.

1 Introduction

Paul Young [1985] improved slightly Fischer and Rabin's proof that the decision procedure for Presburger Arithmetic "A" (for addition) is doubly-exponentially difficult. His principal aim was pedagogical: to demonstrate with his new proof that Gödel's undecidability theorem was essentially the same as Fischer and Rabin's theorem and to urge that instructors who teach the one theorem teach also the other. His aim has not been realized in very many curricula in the last twenty years. That A is decidable, however, is a standard point of emphasis in courses on incompleteness and undecidability. Students consequently know not to overstate Gödel's theorems. For example they are careful to paraphrase the second theorem thus: "Consistent arithmetic theories *with sufficient expressive power* do not prove their own consistency". The impression one is left with after brief reflection is that since the weaker an arithmetic theory is the simpler should be a demonstration of its consistency, examples of weak enough consistent theories do prove their own consistency.

A is not a good model to base this impression on, though, because while Gödel's second theorem is not provable for A, neither is any proof of A's consistency formalizable in A. So keeping with this impression one might joke that Presburger's theory is too weak even to prove its own consistency. Better grounding for this impression can be found in Herbrand's [1931] "On the consistency of arithmetic" where he proves the consistency of his weak arithmetic H by way of an argument formalizable in H. H certainly is a stronger theory than A, but as Herbrand points out it too fails to arithmetize its own metamathematical description. For this reason Gödel's second theorem does not contradict Herbrand's result: The theorem cannot be formulated with respect to H. Herbrand fails to point out, though, that for the same reason the formalizability in H of the proof of the consistency of H cannot literally be seen as "a proof in H of the consistency of H". It takes some reasoning outside of H to interpret the formal proof as a consistency proof because in H there is no formula expressing *in any sense* the consistency of H.

On the other hand, for consistent theories that do adequately arithmetize their own metamathematics, one expects weaker theories to fall even shorter of proving their own consistency. Consistency is after all a metatheoretic property, and since computational strength is needed for any metatheoretic description we expect weaker theories to say less about their own metatheory than stronger systems say about theirs. Indeed these expectations are realized. We have, for example, not only not $I\Delta_0 + Exp \vdash Con(I\Delta_0 + Exp)$ but also not $I\Delta_0 + Exp \vdash Con(Q)$.

One has, then, a tension between two trends–a decrease in the complexity of consistency statements, making these more easily provable, and a decrease in deductive strength, making everything less easily provable–as one investigates increasingly weak fragments of arithmetic. A natural question is how this tension resolves at various points along the arithmetic hierarchy. The consistency statements of theories at the top of the scale, like PA, are equivalent to those theories' Gödel sentences, and therefore proofs of those theories' consistency cannot be formalized in the theories if the theories are consistent. Moreover, these theories can be said to interpret the formulas expressing their consistency as such. Meanwhile theories at the bottom of the hierarchy are not diagonalizable and therefore have no Gödel sentences. These theories very well may formalize proofs of their own consistency but as a rule cannot "see" these proofs as such. A third possibility is for Gödel's Second Theorem formally to hold for a theory, but for this not to be sensibly interpretable as a demonstration of the unprovability in T of T's consistency. This seems to be the case with induction-free and bounded arithmetics.

Recently Professor Pudlák has argued that a version of Gödel's second theorem for Robinson Arithmetic (Q) can be seen as a demonstration of the unprovability in Q of Q's consistency. This is contrary to roughly half a century of speculation that begins with Kreisel and includes among its speculators Bezboruah and Shepherdson, who cautioned against interpreting the theorem in this way alongside their [1976] proof of the result:

> We must agree with Kreisel that this [result] is devoid of any philosophical interest and that in such a weak system this formula [Con_Q] cannot be said to express consistency but only an algebraic property which in a stronger system (e. g. Peano arithmetic P) could reasonably be said to express the consistency of Q. (pg. 504)

Pudlák's argument is based on a strengthening of Shepherdson and Bezboruah's theorem. According to his result, Q proves that it is consistent to assume that there is a contradiction encoded already by a number in any proper syntactic cut of its numbers. Such a syntactic cut can be shortened using a method of Solovay so that it is a definition in Q of a model of a stronger arithmetic theory ($I\Delta_0 + \Omega_1$) that "interprets" consistency unproblematically. Pudlák argues that this allows a sort of semantic bootstrapping from Q to $I\Delta_0 + \Omega_1$. This would block the third possibility described above, since it puts Q alongside PA in the first class of theories and since theories weaker than Q fall uniformly in the second class. It is therefore illuminating to study the argument. The argument is fallacious though, and so we conclude that Kreisel's original intuition was correct.

2 Intensionality of arithmetization

The theory Q was defined by Tarski, Mostowski, and Robinson in [1953]. It's language adds to the standard logical symbols one one-place function symbol S, two two-place function symbols $+$ and $\cdot$, and a constant symbol 0. It has six axioms:

1. $\forall x S(x) \neq 0$
2. $\forall x \forall y (S(x) = S(y) \rightarrow x = y)$
3. $\forall x (x \neq 0 \rightarrow \exists y (S(y) = x))$
4. $\forall x (x + 0 = x)$
5. $\forall x \forall y (x + S(y) = S(x + y))$
6. $\forall x (x \cdot S(y) = x \cdot y + x)$

Extended conservatively with a definition for inequality

$$x \leq y \Leftrightarrow \exists x (x + y = y)$$

Q is the standard base subtheory of most arithmetic systems studied today. It is strengthened usually by the inclusion of an axiom schema or rule for some combinatorial principle, or by adding an axiom that stipulates that certain functions are total. For example Peano Arithmetic is Q together with the rule

$$\frac{A(b), \Gamma \quad \longrightarrow \quad \Delta, A(b+1)}{A(0), \Gamma \quad \longrightarrow \quad \Delta, A(t)} \qquad \text{where the } \textit{eigenvariable } b \text{ does not occur except as indicated}$$

for induction on all formulas. Elementary Arithmetic ($I\Delta_0 + Exp$) is Q with the induction rule restricted to bounded formulas and the axiom

$$Exp \Leftrightarrow \forall x \forall y \exists z(z = x^y)$$

saying that exponentiation is a total function[1].

By itself Q is quite a weak system, though. For example it does not prove the commutativity of addition, nor even the transitivity of $\leq$ once this relation is defined. The claim that it can prove *facts* about it's own metatheory (and not just *formulas* that from outside the theory one can verify as formulations of these facts) is initially very surprising.

Let us turn to our task of resolving the tension that arises from the direct relationship that holds between the complexity of a theory's consistency statement and the theory's deductive strength. Already we have suggested that the task is delicate: One cannot automatically interpret the provability of Gödel's second theorem for a theory as a meaningful metamathematical result. One must also say something about how such an interpretation embeds in the theory being studied. The formula Con_x that appears in the theorem is only a formula. If we replace the free variable x with an enumeration of a mathematical system, then the resulting closed formula will in some systems express the consistency of the enumerated system and in other systems remain uninterpreted. Which of these possibilities in fact happens when the question of the meaningfulness of Con_τ is put to the theory T that τ encodes has no uniform answer. From this we may conclude that the provability of Gödel's second theorem does not decide the question that it was designed for in every system. In the words of Georg Kreisel "Gödel's work on formulae expressing the consistency of classical arithmetic goes beyond arithmetic concepts because it uses metamathematical interpretation" ([1958] pg. 177). For foundational purposes one would like to be able to arithmetize that metamathematical interpretation itself, and the prospects for doing so successively diminish as one considers increasingly weak arithmetic systems.

This does not preclude the possibility that some other result might decide the metatheoretic question for theories that fail to interpret Gödel's second theorem. The metatheoretic content of Gödel's second theorem is that the consistency of a theory is not provable in that theory unless the theory is inconsistent. But the standard form of the theorem is an uninterpretable sentence when relativized to sufficiently weak theories. Intuitively there are two principal sources of this sort of inscrutability where it arises, so establishing the intended content of the theorem for such theories can be reduced to reformulating the theorem in a form immune from these two sources of inscrutability.

One way for the standard form of Gödel's second theorem to be provable but uninterpretable is for the theory to be unable to decipher the coding scheme

[1] Perhaps the most well known fact about bounded arithmetic is that the exponential function is not definable in any bounded theory. This presents an obstacle even for writing the axiom *Exp*. The undefinability of exponentiation is due to Parikh [1971], and remarkably in the same report Parikh showed a way around this obstacle by explaining that the predicate $P(x, y, z) \Leftrightarrow x^y = z$ defining the graph of the exponential function is concrete. Following suite, one may write the axiom *Exp* based on this predicate.

used to construct the binumeration. In this case we say following Feferman that the arithmetization is not intensional. The appearance of τ in a theorem of T is uninteresting from within T if T does not recognize itself in τ. *Outside* of T the veracity of the coding may be verifiable, but the arithmetization does not admit genuine self-reference phenomena if the verification is essentially extensional in that it cannot be executed in T.

Alternatively, the open formula $Con(x)$ may be uninterpretable in T. This could be because the provability predicate from which it is built is not recognizable within T as a representation of theoremhood. This is a plausible obstacle in weak theories because standard Hilbert style proofs or sequent calculus proofs are syntactically complex. From the point of view of T, if T does not prove an elimination theorem for one of its underlying logic's inference rules, then a provability predicate that verifies only encodings of proofs that do not employ that rule will represent something essentially different than will the standard provability predicate, even if the rule's conservativity is something one can readily verify outside of the theory. Once again, theory-independent facts turn out not to be relevant to meaningful metamathematical results.

For example, the cut rule

$$\frac{\Gamma \longrightarrow \Delta, A \qquad\qquad A, \Gamma \longrightarrow \Delta}{\Gamma \quad\longrightarrow\quad \Delta}$$

can be shown to be redundant in the standard sequence calculus presentation of first order logic by a routine semantic argument, but this argument is only as good as the semantic theory on which it is based. One might argue that this theory is as good as one could ask for, but in so doing one would essentially dodge certain metatheoretic questions. For if one cannot execute the argument within an arithmetic theory, then the argument is useless towards gaining ground on questions of that theory's consistency.

The uselessness of the argument towards settling the consistency question is an epistemological uselessness. True, one may base a consistency proof of a theory on such semantic techniques that extend that theory's strength. This would be to *assume* from the outset the reliability of those techniques, though, which is an unwelcome assumption when the consistency of the theory under investigation is a weaker claim than the soundness of the semantic techniques. This "epistemological uselessness" has been cited repeatedly in the philosophical literature to suggest that the consistency of various arithmetic theories is quite open. The suggestion obviously is controversial, and this study advances a far weaker claim. Here the "epistemological uselessness" of an argument draws into question whether certain formal constructions are appropriate representations of metatheoretic facts. That is, whether or not an alleged consistency proof succeeds actually in securing the consistency of a system or whether or not a certain rule-elimination is admissible, if the proofs of these results cannot be carried out in a particular system, then *that system* neither "proves its own consistency" nor even recognizes as statements of consistency all the formulas that from outside the theory look like consistency statements[2]. Short of casting doubt on a theory's consistency, this casts doubt only on whether a theory proves its own consistency or on whether a theory can be said conclusively not to prove its own consistency.

Even in this modest form the point is one of "epistemological uselessness" because foundational investigations are not restricted to the discovery of metatheoretic facts but also are concerned with the amount of formal apparatus one needs in order to prove facts. The example from this study's introduction is illustrative: One might want to investigate the purely logical relationship between self-reference and the unprovability of consistency by deciding whether all mathematical systems that compute self-referential statements can be shown not to prove their own consistency. But to decide this question for any particular system, the self-reference and consistency claims must be represented in a way recognizable as such by the system itself.

Something like this insight originally motivated Gentzen's research on constructive proofs of the eliminability of the cut rule. The semantic account of the redundancy of the cut rule already was known before 1936, but Gentzen deemed a consistency proof relative to this semantic argument unhelpful for foundational research. By virtue of its constructive proof-transformation nature, Gentzen's *Hauptsatz* gives more foundational insight. It is well known that the arithmetic consistency result built onto the *Hauptsatz* is relative to the consistency of an associated transfinite induction. But even short of applying the cut elimination theorem to the consistency question, one may ask what the cut elimination theorem itself teaches us. If the *Hauptsatz* cannot be arithmetized in a theory T, then from a certain perspective the theorem says nothing about T-provability. In particular, the formulas $\exists x Prf(x, y)$ and $\exists x CFPrf(x, y)$ expressing in turn "there is a proof of y" and "there is a cut-free proof of y" are not equivalent in such a theory. Hence their corresponding consistency formulas $Con_{\tau} \Leftrightarrow \neg\exists x Prf(x, \bot)$ and $CFCon_{\tau} \Leftrightarrow \neg\exists x CFPrf(x, \bot)$ are not equivalent either.

In the light of these two types of inscrutability, two new questions emerge: (1) What are the correct arithmetization schemes relative to a theory? and (2) What is the correct formulation of provability relative to a theory? In both cases, correctness is the ability to capture the metatheoretic content of a formula within the theory.

The first question is the subject of Feferman's [1960]. He found that there is no uniform answer and submits this finding as a cautionary note about reading too much into standard results for theories like PA. Specifically he found (corollary 5.10) a binumeration α^* of Peano Arithmetic that is extensionally correct, but whose corresponding consistency statement is provable in PA. One might hastily conclude that the arithmetization that gives rise to such an enumeration of PA blocks any generalization of Gödel's second theorem from the specific arithmetization of Gödel [1931]. By contrast Feferman concludes that while the

[2] And conversely, a theory might recognize as a statement of consistency some formula that from outside the system *does not* appear to be a consistency statement.

generalization of the technical result is restricted[3] by this variant binumeration, the generalization of the "unprovability of consistency" is not:

> We have maintained that insofar as a formula α expresses membership in A, the formula Pr_α expresses provability of $\mathcal{A}$ in $\mathcal{M}(\mathcal{P})$ and the sentence Con_α expresses the consistency of $\mathcal{A}$ in (M) and (P). Thus, one particular conclusion we can draw is that the formula α^*, although it extensionally corresponds to A, does not properly express membership in $\mathcal{A}$. (pg. 69)

The expression of membership in a theory must be intensional in order for it to be a proper expression on top of which one may define formally metatheoretic properties. That is, the failure of Gödel's second theorem in some general settings is independent of the stability of the result on the unprovability of consistency when the settings in question are ones where intensional arithmetization fails.

An arithmetization τ of a theory T in T is intensional to the extent that salient metatheoretic properties of predicates built on τ are provable in T. For example Feferman requires an intensionally correct arithmetization of T-provability to satisfy

$$\begin{array}{lcl}
T & \vdash & \forall u \forall v \forall w (Fmla(u) \land Term(v) \rightarrow Fmla(sub(u,v))) \\
T & \vdash & \forall A \forall B (Thm_\tau(A) \land Thm_\tau(A \rightarrow B) \rightarrow Thm_\tau(B)) \\
T & \vdash & \forall u (Proof_\tau(u) \rightarrow Thm_\tau(Proof_\tau(u))) \\
T & \vdash & \forall A \forall u (Prf_\tau(u,A) \rightarrow Thm_\tau(Prf(u,A))) \\
T & \vdash & \forall A (Thm_\tau(A) \rightarrow Thm_\tau(Thm_\tau(A)))
\end{array}$$

and the variant definition of PA just discussed fails to admit an intensionally correct arithmetization of PA provability. One might take this argument as a defense of the Hilbert-Bernays-Löb derivability conditions for the second incompleteness theorem. For if only intensionally correct arithmetizations are admitted, then there is no question about the meaning of Gödel's second theorem for PA. Any candidate arithmetization that skirts the derivability conditions does so by failing to produce an intensionally correct arithmetization of theoremhood. Therefore even if the resulting consistency statement were a theorem of PA, this could not count as a proof in PA of the theory's consistency. Intensionality criteria are a set of necessary conditions for an arithmetization to meaningfully represent metatheory in a purely arithmetic environment. Derivability conditions are a set of conditions on a provability predicate sufficient for Gödel's second theorem. If derivability conditions are consequences of intensionality criteria, then one may infer from Gödel's second theorem the unprovability of a theory's consistency. Otherwise Gödel's second theorem by itself tells one nothing about a theory's consistency. This study does not investigate the question whether Feferman's account of intensional correctness is satisfactory for strong

[3] Feferman's study revealed that the result must be restricted to recursively enumerable representations of theories, though it holds universally for those arithmetizations (theorem 5.6)

arithmetic systems like PA, but it is worth noting that there is no consensus about what might be proper conditions for intensionality even in such strong settings[4]. The treatment presented below is meant to suggest only that deciding what the proper conditions are for the intensionality of an arithmetization itself depends on the strength of the theory one is studying.

The standard arithmetization of Q is not intensional according to these criteria. In fact no formula κ represents Q so that the predicate $Thm_\kappa(x)$ satisfies Feferman's intensionality criteria. (This is readily seen since Q is finitely axiomatized. For if $q(x)$ is the formula that defines Q simply by listing its axioms, then for any other definition $\kappa(x)$ of Q in Q, $Q \vdash Con_\kappa \rightarrow Con_q$. Now since the above criteria fail for the straightforward arithmetization $q(x)$, they fail also for $\kappa(x)$.) From a logical point of view, this tells us immediately that the standard proof of Gödel's second theorem does not apply to Q, since the Hilbert-Bernays-Löb derivability conditions fail alongside the intensionality criteria. Philosophically speaking one can say more than this, specifically that the theorem of Bezboruah and Shepherdson, since its proof does not utilize the derivability conditions, is intensionally inadequate.

Let us focus now on the second question. One might say that when provability with all the inference rules of a theory's underlying logic separates from a restricted notion of provability, the unrestricted version has a natural claim to correctness. Of what consequence is it that a theory doesn't prove a variant of Gödel's second theorem with a seemingly contrived, inefficient version of provability in place of the standard one? If the theory proves the standard formulation of Gödel's second theorem, then the unprovable, nonstandard version is the anomalous one. The failure to prove the nonstandard version of the theorem looks analogous to Feferman's example of the provability of the "theory's" consistency when the theory is arithmetized inefficiently. The burden of correctness would seem to be on the contrived variant, and the effective separation of the variant arithmetization or formula from the standard one would appear to be evidence against it.

One must distinguish restricted and unrestricted provability in a metatheoretic sense from their formal counterparts in a theory, however. For when the notionsdo separate, so too must at least one formulation separate from intensional correctness. In the present case neither formula has a very strong claim to genuine expression of Q-theoremhood, because of the gap in complexity between first order derivability (with or without a cut rule) and the expressive strength of Q. Correctness could be earned by investing one of the formulas with meaning by proving in Q the equivalence between that formula and a more wieldy combinatorial sentence. When this is possible for one formula but not for the other, then the first actually has better claim to correctness for meaningful foundational results regardless of what looks appropriate extensionally.

The cut-elimination theorem for first order logic cannot be formalized in Q. One suspects then that "provability in Q" and "cut-free provability in Q"

[4] In [1988] Detlefsen argues that Hilbert's foundational project, because of it's underlying "instrumentalist" epistemology, is in theory realizable by proving in strong arithmetics formulas that fail to satisfy the derivability conditions.

should have different "meanings" from the point of view of Q. Indeed, the standard arithmetizations of these predicates are inequivalent in Q. Similarly, since Herbrand's theorem does not hold in Q, "provability in Q" and "the construction of a Herbrand disjunction of the quantifier free part of a sentence" provably separate in Q.

The parallel in these two cases is strong. Gentzen argued that provability without the cut rule corresponds in a certain sense with the construction of a finite Herbrand disjunction. The reason is the analogy between Gentzen's *Hauptsatz* and Herbrand's theorem. The "sense" in which the two types of provability correspond is a slippery one. For example, Gentzen argued that Herbrand's theorem was a special case of his *Hauptsatz* for sequents with empty antecedents. From another perspective Herbrand's result seems more general though, for it treats formulas of arbitrary quantifier complexity. But from either perspective the notion of correspondence under scrutiny appears heuristic, because both fundamental theorems demonstrate the equivalence between the restricted type of provability they treat and standard provability. It follows that cut-free provability and Herbrand disjunction construction are equivalent. The heuristic correspondence is highlighted by an application that Herbrand made with his theorem: an elimination theorem for *modus ponens* with respect to a redundant set of inference rules for the predicate calculus. Such heuristic observations play a pedagogical role but seem mathematically empty when the proof systems under consideration all are formally equivalent.

From the perspective of this study, however, the correspondence between the construction of Herbrand disjunctions and cut-free provability has definite mathematical content. The equivalence or inequivalence of two types of provability is not theory-independent. For example, Herbrand's theorem is provable in $I\Delta_0 + \mathit{Superexp}$ but not in any bounded theory. At least some arithmetic theories[5] that prove neither cut-elimination nor Herbrand's theorem do prove the equivalence between the construction of Herbrand disjunctions and cut-free provability, however, so the "correspondence" between these two proof systems holds even in places where Gentzen and Herbrand's fundamental theorems do not.

On the other hand, even the correspondence between cut-free provability and Herbrand provability fails in Q. Since this failure reflects also in the inequivalence of the consistency statements associated with each type of provability, the question about the meaningfulness of Gödel's second incompleteness theorem can be made precise. If one of these variant types of provability is provable in Q, might it have more claim to an adequate representation of Q's consistency than has the standard consistency statement? Or if theorems analogous to Gödel's can be proven for Q with each of these variant formulations of consistency in the place of the standard one, can this answer our reservations about the meaningfulness of Gödel's theorem for this theory?

In a certain sense the two intuitive sources of inscrutability in the appro-

[5] Theorem 5.19 on pg. 379 of (Hájek and Pudlák [1993]) is the equivalence of Herbrand and cut-free provability in elementary arithmetic, and it is an open question whether the equivalence can be demonstrated in a bounded theory.

priateness of an arithmetization are not entirely distinct. For when arithmetic theories are presented, as they are above, with deduction rules for combinatorial principles like induction over some formulas, the theory's axioms have a canonical arithmetization, since only finitely many axioms ever are present[6]. Arithmetizing, for instance, the validity of induction over bounded formulas in $I\Delta_0 + \Omega_1$ is built in to the choice of formulation of the predicate "x is a proof" rather than "x is an axiom". This study presents arithmetic theories with deduction rules, rather than axioms, for induction precisely to make clear that questions of intensionality are most properly thought of in terms of how best to formulate metatheoretic properties, rather than–as has become customary–merely how exactly to define the theory. Another suggestion of the priority of the second source of inscrutability over the first is evident in the proposed intensionality conditions themselves, which present the definition τ never in isolation, but rather always embedded in another formula like $Proof_\tau(x)$ or $Thm_\tau(x)$.

Feferman [1960] actually discusses both sources of inscrutability in his study. Considerations of the appropriate manner to arithmetize a theory he describes as considerations "from the inside", while those about the arithmetization of theoremhood and the like he calls considerations "from the outside". Calling a change in one's arithmetization a change "from the outside" is meant to connote artificiality, and Feferman dismisses these sorts of considerations as technical tricks that can be useful for solidifying formal results but not for generalizing metatheoretic findings. The reason, he says, is that changes from the outside essentially are "changes in the notion of logical derivation" (pg. 39). In the strong arithmetics Feferman studied, "changes in the notion of logical derivation" from standard provability are departures from meaningfulness and therefore are perhaps rightly dismissed as irrelevant for intensional purposes. But in weak arithmetics care in the arithmetization of syntactic properties like the notion of derivation is called for, since only some of the many extensionally equivalent notions may be interpretable in the theory. Questions of intensionality accordingly extend in these theories to choices at every stage of arithmetization, from the representation of axiomhood to more complex constructions. The optimal arithmetization of axiomhood for Q and the extendability of Q to stronger arithmetic systems through the introduction of rules in the sequent calculus, moreover, dissolves the "inside/outside" distinction and exposes the activity of arithmetizing metatheoretic predicates as the primary stage for questions of intensionality.

However, the axiomatic framework can be used to distinguish notions of provability that cannot be formulated in the presentation of a theory with nonlogical deduction rules. The next section presents Herbrand provability and consistency and develops a simple, informal semantics for these notions. Of especial significance is that the semantic interpretation presented for provability and unprovability based on the construction of Herbrand disjunctions is, unlike their classical counterparts, relativizable to weak theories without any apparent loss of meaningfulness. But in order to pursue these notions, arithmetic theories

[6] See the discussion about the canonical axiomatization of Q above.

must be reformulated. For the claim that ϕ is a theorem of a theory T will be construed as their being a tautology

$$\bigvee_{i=1}^{k}[He^*(\bigwedge_{j=1}^{l} T_j \to \phi)(t_{i1}, f_1(t_{i1}), t_{i2}, f_2(t_{i1}, f_1(t_{i1}), t_{i2}), \ldots)]$$

for some finite set $\{T_j\}_{j=1}^{l}$ of the axioms of T, where $He^*(\phi)$ is the quantifier-free part of the Herbrand form of ϕ. For Robinson's theory the canonical arithmetization $q(x)$ of Q-axiomhood allows one to replace the subformula $\bigwedge_{j=1}^{l} T_j$ with the simple conjunction of Q's axioms, but in order for this formulation to make sense for theories of even modest strength, combinatorial principles like induction and collection must be reformulated as axioms instead of deduction rules. In the remaining sections, therefore, combinatorial principles in arithmetic theories are assumed to be presented in terms of axiom schemata. (Induction over the set of Ψ formulas will be axiomatized by the familiar schema $A(0) \wedge \forall x(A(x) \to A(S(x)) \to \forall x(A(x))$ for $A \in \Psi$). One may keep in the back of one's mind the fact that presentations in terms of deduction rules actually are standard in many proof theoretic investigations and also the lesson just learned about how the choice of defining a theory, which Feferman emphasizes in his pioneering study of arithmetization, is only one parameter to consider when sorting out questions of intensionality[7].

3 A theory-dependent interpretation of Herbrand consistency

Kreisel [1958] argues that since the no-counterexample interpretation of provability is in a sense simple, it is a more primitive interpretation than the standard one and therefore is more appropriate when one restricts one's attention to constructive methods. He has in mind specifically recursive methods, but his sentiment can be extended to the kinds of restrictions one encounters when investigating arithmetization in weak theories. The no-counterexample interpretation is based on Herbrand's theorem in such a way that theories not proving Herbrand's theorem will not prove that the no-counterexample interpretation is equivalent to standard theoremhood. Kreisel's intuition, translated into this arena, is that this testifies against the meaningfulness of the standard formulation of theoremhood for these theories. Since Herbrand proofs are propositional proofs, they are combinatorially very simple. One might say that weak theories are able to make sense of them when they are not able to make sense of the combinatorially more complex sequent or predicate calculus proofs. Herbrand proofs are at the same time generally much longer than standard proofs, and therefore more difficult to execute. Statman [1978] proved in fact that no

[7] In [1985] Sieg demonstrates that the two styles of presenting arithmetic theories are equivalent. His demonstration shows that the induction rule is strong enough only with the inclusion of the side formulas A. For a detailed description of the proof-theoretic benefits of the rule-based presentation, his report is an excellent resource, as is section 1.4 of (Buss [1998]).

Kalmar elementary procedure transforms standard proofs into these combinatorially simpler "direct proofs". The intuition that seems appropriate is that meaningful results should be more difficult to prove than meaningless ones.

What concretely can be said about the comparative meaningfulness of standard formulations of consistency and the nonstandard one based on Herbrand's theorem? In weak theories where the two notions separate, it seems that there is quite a lot to say. The formula Con_T says that there are not proofs from the axioms of Q of a sentence and of its negation. The formula $HCon_T$ says that there is not a propositional proof of the quantifier free matrix of a Herbrand disjunction:

$$\bigvee_{i=1}^{k}[He^*(\bigwedge_{j=1}^{l} T_j \rightarrow \bot)(t_{i1}, f_1(t_{i1}), t_{i2}, f_2(t_{i1}, f_1(t_{i1}), t_{i2}), \ldots)] \tag{1}$$

where $He^*(\phi)$ is the open part of the Herbrand form of ϕ. To analyze this construction, let some finite conjunction $\bigwedge_{j=1}^{l} T_j$ of T's axioms be given so that

$$\bigwedge_{j=1}^{l} T_j \iff \forall x_1 \exists y_1 \forall x_2 \exists y_2 \cdots \Phi(x_1, y_1, x_2, y_2, \ldots).$$

and consider the negation of the right hand side of the above sentence:

$$\exists x_1 \forall y_1 \exists x_2 \forall y_2 \cdots \neg\Phi(x_1, y_1, x_2, y_2, \ldots).$$

A natural way to interpret formulas with alternating quantifiers is in terms of a two player "adversary game". In this case, the above formula says that there is a strategy for Eloise ("playing the existential quantifiers" against Abelard) to falsify the open formula $\neg\Phi$. This is interpreted as Eloise demonstrating the inconsistency of Q. According to Herbrand's theorem, the formula is provable in the predicate calculus just in case there is a propositional tautology of the form:

$$\bigvee_{i=1}^{l} \neg\Phi(t_{i1}, f_1(t_{i1}), t_{i2}, f_2(t_{i1}, f_1(t_{i1}), t_{i2}), \ldots)$$

where the t_{ij} are terms in the language of Φ expanded to include the function symbols f_j.

This gives the original formula (1) with Φ standing in place of the finite conjunction of T's axioms. One would like to recover the adversary game semantics for this object. Intuitively the functions $\{f_j\}_j$ should compute Abelard's moves based on the previous moves, and the terms $\{t_{ij}\}_{i,j}$ represent Eloise' moves. But on which previous moves are Abelard's moves based, and on what grounds does Eloise choose her moves? According to the way Herbrand functionals are introduced in the construction of the disjunction, while Abelard's moves are defined by functions over shorter terms (earlier moves) in a single disjunct, Eloise' moves are not. *Her moves* in disjunct i may depend on what she knows of Abelard's move-computing functions from other disjuncts, because nothing prevents terms computed by functions from disjuncts i from appearing as subterms of terms t_{jk} from other disjuncts. That is, once a function symbol has been introduced into the language (by a play of Abelard), Eloise may use it in the terms that make up her future plays in *any* of the disjuncts. So in the adversary game, the individual disjuncts are constructed in parallel rather than sequentially. In typical cases in fact the parallel construction is necessary in order to arrive at the disjunction guaranteed by Herbrand's theorem[8].

The semantic interpretation that arises is this. Abelard and Eloise play the game $\mathbf{\Phi}$ given by the formula $\exists x_1 \forall y_1 \exists x_2 \forall y_2 \cdots \neg\Phi(x_1, y_1, x_2, y_2, \ldots)$ on several boards (indexed by i), with Abelard boasting a winning strategy for $\mathbf{\Phi}$ (given by the functions f_j). Eloise's goal is just to win on at least one board, thereby disproving Abelard's boast. If Eloise succeeds, then she will have proven T inconsistent. In that case the terms she plays define a winning "superstrategy"– that is, these terms define a way to reply to Abelard's strategy for the game $\mathbf{\Phi}$ either by replying directly to a move of Abelard on the board he just played on, or by using the information learned about a position from one of Abelard's moves by making a different move on another board, or by beginning a game on a new board, adding thereby to the number of disjuncts[9].

A tautological Herbrand disjunction of this form would be "a counterexample to Q". We may formalize the claim that there is no counterexample as

$$\forall\Delta(HD(\neg\Phi, \Delta) \rightarrow Tr_{\exists}(\exists z_1 \cdots \exists z_m \neg\Delta(z_1, \ldots, z_m)))$$

where $Tr_{\exists}(x)$ is a truth definition for existential formulas and $HD(x, y)$ is the relation that says that y is a Herbrand disjunction for x. In our informal semantics, this last formula says that Abelard has a winning superstrategy for any Herbrand disjunction that would be, if Eloise had a winning strategy for it, a counterexample to T's consistency.

One might think in this case of Herbrand's theorem as showing us that the claim that no such formula $\Delta(z_1, \ldots z_m)$ is a tautology is a claim about the

[8] This discussion of Herbrand disjunctions is essentially due to Adamowicz [2005]. See there or (Pudlák [2004]) for the necessity of the parallel construction. Pudlák in fact presents Adamowicz' analysis in terms of games, similar to the presentation in this study. They use this analysis to define combinatorial principles independent of bounded arithmetic but do not discuss questions of intensionality. Pudlák, interestingly, does emphasize that although the combinatorial principles he treats are essentially Π_2 and consequently in a sense less interesting than other, well-known independent sentences, the game semantics suggests that they are more meaningful in weak settings than the familiar, arithmetically less-complex sentences.

[9] The scenario is analogous to a human player (Eloise) playing chess against a good chess program (Abelard) with an option to take back her moves. The computer claims to have a strategy making it the Chess Master©and always plays this strategy. The human player, on the other hand, can at any point rethink a particular move, return to that stage of the game, and substitute a different move, all the while keeping the board with the original move on it "alive" in case later it appears preferable after all to play from that position. Since human time-resources are valuable, most chess players are aware of how difficult it is to beat very good programs even with the option to take back moves. On the other hand, since we do not have access to the best strategies due to the combinatorial explosion in chess positions, an option to take back moves increases our chances substantially.

consistency of T. We want to suggest the strict converse–that the provability of Herbrand's theorem in a weak arithmetic would invest the standard consistency statement with meaning, by virtue of the meaning of the claim that no such Herbrand disjunction exists, but that the theorem's failure in a theory renders the standard formula meaningless even if the claim about Herbrand disjunctions is interpretable. In other words, the failure of Herbrand's theorem in an arithmetic system casts doubt on the meaningfulness in that setting of the standard consistency claim. In this case the "no counterexample" construction is the only one interpretable in T as a consistency statement.

Consider the formula embedded in the consequent of the above "no counterexample" claim:

$$\exists z_1 \cdots \exists z_m \neg \Delta(z_1, \ldots z_m)$$

This formula says not only that $\Delta(z_1, \ldots z_m)$ is not a tautology but also that Abelard can substitute values for the variables z_i in such a way that he beats Eloise' strategy (defined by the terms t_{ij} in Δ) "on every board". Extensionally, this sentence is stronger than the mere claim that $\Delta(z_1, \ldots z_m)$ is not a tautology, because its truth depends on what substitutions are effective in T (The substitution functions must be provably total in T). Likewise $\bigvee_{i=1}^{l} \neg\Phi(t_{i1}, f_1(t_{i1}), t_{i2}, f_2(t_{i1}, f_1(t_{i1}), t_{i2}), \ldots)$ is stronger than a mere claim of inconsistency, because Eloise' counterexample construction strategy must be computable by terms t_{i_j} of the language of T. In all bounded theories, however, these strengthenings are unavoidable in meaningful talk about strategies, for if moves require for their determination the results from a function that one cannot even prove is total on the natural numbers, then this requirement undercuts the effectiveness of a proposed strategy. One has in that case no strategy at all.

In particular, in bounded theories like $I\Delta_0 + \Omega_1$ or Buss' S^1_2, one cannot define the terms needed to construct Herbrand disjunctions for provable formulas, because the terms needed are values of functions that those theories cannot prove total. The failure of Herbrand's theorem for these theories in fact amounts just to this. The tautological Herbrand disjunction one gets from the theorem applied to provable formulas often involves $I\Delta_0 + \Omega_1$ or S^1_2-uncomputable functions. In particular the consistency statements for these theories are equivalent, not to the claim that Abelard can beat any superstrategy that Eloise plays, but to the non-existence of tautological Herbrand disjunctions that are not meaningfully interpretable. One can then bifurcate the notion of "Herbrand consistency" into two distinct claims, neither provably equivalent in bounded arithmetic to the standard consistency statement. For the first the existence of all possible counterexample disjunctions is considered. Though inequivalent to the consistency statement expressed in terms of provability in the sequent calculus, a Gödel-like theorem is provable for this statement[10]. A second statement allows in the disjunctions only terms whose values are polynomial in the size

[10] ...in many weak arithmetics. Willard [2002] discusses how strong a theory one needs in order to prove Gödel-like theorems for direct proof systems. Among his results is a proof that $I\Delta_0$ suffices, as conjectured by Wilkie and Paris [1981].

of their subterms so that the existence of a counterexample depends on plays decidable with the resources of T itself. However, it is not known how to prove a Gödel-like theorem with this restriction, or even whether such a theorem can be proven.

Kreisel's intuition can now be reformulated in terms of these constructions. The straightforward two-player game semantics for the predicate calculus can be adapted to a weak notion of provability based on the construction of Herbrand disjunctions. Game strategies become match strategies, where the same game is played on several boards. Abelard plays on all boards the unique move that he deems best in that position, while Eloise is free to play differently on every board. (This restriction on Abelard and license for Eloise is the exact reverse of the scenario from the standard quantifier semantics. This results from Eloise trying to falsify the originally quantified predicate of interest rather than trying to confirm it.) Both players' search strategies, however, are bounded by the terms of the theory. From the "point of view" of the theory under consideration, though, this restriction is perfectly natural. We experience not a privation of strategies for the game at hand, only a realistic focus on strategies that one actually can find (as opposed to a hypothesized realm of ideal strategies, the existence of which is more dubious that the consistency of the theory in question).

The no-counterexample interpretation applied to the question of consistency, then, results in a semantics relativized to the computational strength of the theory in question. The same cannot be said about the standard formulation of consistency, which relies still on the arithmetization of combinatorially complex proof calculi. The significance of the respective ability and inability of these two formulations of consistency to relativize is evident in weak theories that do not prove those formulations' equivalence. If the no-counterexample interpretation, based on Herbrand's theorem, is interpretable still in such contexts, its inequivalence with the standard construction draws into question the meaning of metamathematical results based on the standard construction. In this case we would say that the arithmetization of the notion of consistency is ambiguous in the theory, because while a weak notion (Herbrand consistency) has a definite combinatorial meaning, this meaning is not synonymous with the claim that a contradiction could not be derived. In fact in such a weak setting one may not be able sensibly to speak about the possibility of deriving a contradiction, because that claim is equivalent to a claim about "ideal" match strategies that are not computable in the system.

This suggests a limitation to the intensionality criteria from the previous section. Those criteria express facts that seem closely tied in to the notion of provability, but since sufficiently weak theories don't prove that proof by natural deduction is equivalent to proof via construction of Herbrand disjunctions, the evident interpretability of the latter (in theories strong enough to interpret *it*) suggests that arithmetization schemes satisfying the "intensionality criteria" need not capture adequately the notions of provability and consistency after all. Since in theories of bounded arithmetic, the standard consistency statement is provably equivalent to the theory's Gödel sentence, but the Herbrand consistency statement is not, "the consistency of bounded arithmetic" construed informally but meaningfully, is not equivalent to such theories' Gödel sentence. Buss ([1985] pg. 145) has shown that for theories containing S_2^1 provable equivalence to a the Gödel sentence is unnecessary for the second Gödel theorem, since the verification of the conditional from a "consistency formula" to the Gödel sentence suffices for the proof (he proves the theorem for a sequent calculus version of free cut-free consistency). However, the next section demonstrates how the establishment of Gödel's second theorem by alternative means typically compromises the metatheoretic meaning of the result.

4 The ambiguity of metatheory in weak arithmetics

In [1996] Pudlák writes:

> Bezboruah and Shepherdson proved the second incompleteness theorem in Q, which is one of the weakest arithmetic theories. The result was in a certain sense problematic: if the theory is so weak, does the particular formulation of Con_Q really mean what was intended? A solution, which will be presented below, is to define an initial segment J of the numbers in Q, which is an inner model of a stronger theory T and prove that it is consistent to assume that a proof of contradiction from axioms of Q is encoded by a number which is already in J. Since T is strong, the meaning of Con_Q is not so ambiguous.

After presenting the proof that it is consistent with the axioms of any theory containing Q to assume that there is a proof of contradiction from those axioms encoded in any inductive cut of the theory's numbers, he adds: "A corollary of this result [from (Pudlák [1985])] is that the second Gödel's incompleteness theorem holds in weak theories, in particular in Q, without any doubts about what Con_T really means there. This is because by [Wilkie's interpretation of $I\Delta_0 + \Omega_1$ in Q] there is a cut in Q which is a model of $I\Delta_0 + \Omega_1$. In such a cut all reasonable definitions of Con_Q are equivalent."

The argument seems to be that since

1. in the theory $I\Delta_0 + \Omega_1$ the consistency of Q is unambiguous,
2. this theory proves Gödel's second incompleteness theorem for Q, and
3. $I\Delta_0 + \Omega_1$ is interpretable in Q,

Q proves not only Gödel's second incompleteness theorem but also "the unprovability of Q's consistency". In the introductory section of this essay we described this argument as a bootstrapping argument. There is an implicit concession by Pudlák that Bezboruah and Shepherdson's proof of Gödel's theorem in Q is not metamathematically meaningful, that Robinson Arithmetic is too semantically impoverished to recover the meaning of Con_Q intensionally. But in $I\Delta_0 + \Omega_1$ an intensional arithmetization is possible. That is, assuming we adopt Feferman's criteria directly for this system, $I\Delta_0 + \Omega_1$ proves facts about the predicate Con_Q that suffice for it to be an actual statement of Q's consistency. Pudlák claims that Q can in a sense borrow this stronger theory's semantic richness by proving theorems whose quantifiers all are relativized to a formula that models the stronger theory. We call this a bootstrapping *of semantics* because ordinary bootstrapping results show how to emulate certain purely formal constructions in a weak theory through such relativizing of quantifiers or conservative extensions by definition, while in the present case, the very same formal result–Gödel's second incompleteness theorem–is being proved, and only the stronger theory's ability to make sense of it is supposedly emulated by the weaker theory.

Pudlák's argument fails on two points. First, the "strong" theory $I\Delta_0 + \Omega_1$ does not interpret consistency so unambiguously. It distinguishes, in fact, several extensionally equivalent formulations of consistency and does not prove a "Gödel theorem" relative to each of them. Second, the bootstrapping of semantics through interpretability is incoherent from the point of view that we have presented in this study. The intensional arithmetization of syntax was meant to invest the metamathematical results following out of this arithmetization with an epistemologically secure meaning. But if the meaning of a sentence about the consistency of one theory depends on the interpretation in that theory of a stronger theory and the properties that *it* recognizes in that sentence, then there is no epistemological gain. This is because the adequacy of intensionality criteria, just like the meaningfulness of specific arithmetization schemes, is theory dependent.

To see the first failure, consider again the notion of Herbrand consistency from the previous section. We saw there that since Robinson Arithmetic fails to prove Herbrand's theorem, it distinguishes the formal expressions of Herbrand consistency and standard consistency. Of course we are not considering realistic the prospects that either of these expressions is meaningful in the context of Robinson's theory, because even the combinatorially simpler notion of Herbrand consistency is complex relative to this setting. Our formulation of Herbrand consistency is a Π_2 sentence about the existence of a counter strategy for Abelard to every counterexample disjunction Eloise might construct by her moves. This formally is a statement about the totality of a very fast growing function (a counter strategy finder) that is beyond the power of Q to interpret at all (Since the function is not provably total in Q).

In fact, Herbrand's theorem fails also in $I\Delta_0 + \Omega_1$. So while extensionally the formal expression of this theory's Herbrand consistency and of its standard consistency are equivalent, intensionally they separate. The separation of purportedly "metamathematical" formulas in a theory like Q is not very interesting, because those formulas can be seen directly not to have meanings from the theory's perspective. The same separation can be more informative when it arises in theories that arithmetize syntax intensionally, because the separation may be of two formulas that correspond with some metamathematical properties. In the present case formulas that do provably correspond with some of our intuitions about the consistency of $I\Delta_0 + \Omega_1$ are distinct. When one is interested in the consistency of a theory, it seems appropriate only to investigate the question in a setting free from the assumption of the theory's consistency. Specifically, one cannot make more sense of the consistency of $I\Delta_0 + \Omega_1$ than what can be said about it meaningfully in the theory, because beyond this one would be saying what necessarily only makes sense under the assumption that $I\Delta_0 + \Omega_1$ is consistent. It seems therefore correct to say not only that $I\Delta_0 + \Omega_1$ fails to identify its consistency with its Herbrand consistency, but also that its consistency and its Herbrand consistency really are distinct properties. Here Kreisel's point is salient: $I\Delta_0 + \Omega_1$ does not associate the lack of a derivation of contradiction with its Herbrand consistency. Since the no-counterexample demonstration based on Herbrand provability is epistemologically primitive due to its simplicity, it seems appropriate to say that $I\Delta_0 + \Omega_1$ does not see the underivability of contradiction as an essential feature of consistency. To be sure, this reveals a serious weakness in the operative notion of consistency for this theory, but it is one we are forced to by the theory's low computational strength.

As for the intensionality criteria, the $I\Delta_0 + \Omega_1$-separation of Con_δ and $HCon_\delta$ together with the criteria's failure for the latter formula suggest that they are not necessary conditions on an arithmetization of syntax for theorems related to this arithmetization to be meaningful. The adequacy of these criteria in the setting of strong arithmetics is derivative from the neat identification there of several notions of provability one might consider. However one chooses informally to formulate provability, the notion is sure to meet these criteria by virtue of its equivalence with Hilbert-style provability. But in a weak setting, where any informal notions of provability that are interpretable by the theory itself provably separate from any notion satisfying these criteria, the criteria no longer seem adequate as conditions of intensionality.

Let us turn to the second failure in Pudlák's argument.

The fact that Q interprets $I\Delta_0 + \Omega_1$ does not make the content of the latter theory's theorems or the intensionality of its arithmetization hold also in Q. If Pudák's argument were simply this, it would apply also in familiar settings with shocking results. For example if we add to PA an axiom $\neg Con_{PA}$ saying that PA is inconsistent, the resulting theory is interpretable in PA[11]. By the "simple" argument we are considering, one would conclude that PA proves its own inconsistency!

Pudlák's claim must, then, not be simply a heuristic point about the significance of formal interpretability. One can do it more justice by understanding it rather as a prescriptive point about how to construct a predicate that unlike Con_Q is an intensionally adequate representation of provability in Q. The prescription depends not only on the interpretability of $I\Delta_0 + \Omega_1$ in Q but on the specific method of interpretation.

[11] This follows immediately from Theorem 6.6 of (Feferman [1960]).

The interpretation of $I\Delta_0 + \Omega_1$ in Q is via a proper syntactic cut, a formula that Q proves is closed under successor and addition, but not induction. Formally a syntactic cut of Q's numbers is a formula $J(x)$ such that

$$\begin{array}{lll} Q & \vdash & J(0) \\ Q & \vdash & \forall x(J(x) \rightarrow J(x+1)) \\ Q & \vdash & \forall x \forall y(y < x \wedge J(x) \rightarrow J(y)). \end{array}$$

If additionally not $Q \vdash \forall x J(x)$, then J is a *proper* cut. PA, since it has a rule for induction on all formulas, does not have any proper syntactic cuts, but fragments of PA with induction restricted to formulas beneath a certain quantifier complexity may have syntactic cuts. Intuitively we think of a cut of an arithmetic theory's numbers as an initial segment of a model of that theory, but not all initial segments of a theory's models are necessarily definable by a predicate, and it is possible for a theory to have a syntactic cut that does not define any tangible geometric structure treatable outside the theory.

Syntactic cuts can be used to interpret relatively strong theories in weaker ones. Earlier it was pointed out that Q does not prove the commutativity of addition. It may, however, prove the commutativity of addition for all numbers that fall under a predicate J, i. e.

$$Q \vdash \forall x \forall y(J(x) \wedge J(y) \rightarrow x + y = y + x).$$

Of course the above formula is trivially provable in Q when J is empty. There are, however, predicates $J(x)$ that contain 0, are closed under successor and addition, and for which the above formula is provable. Since Q doesn't prove the commutativity of addition for all natural numbers, Q must not prove induction over $J(x)$ if Q is consistent, so they define proper syntactic cuts for Q's numbers. In such a case one says that Q proves the commutativity of addition "with quantifiers restricted to J". If we define the commutativity of addition by the formula

$$Commute_+ \Leftrightarrow \forall x \forall y(x + y = y + x),$$

then the above formula with quantification restricted to J is abbreviated $Commute_+^J$.

It is possible for a theory to prove, not just properties of arithmetic operations that it otherwise cannot prove, but also axioms of stronger arithmetic theories by restricting quantifiers to suitable predicates. For example, Wilkie and Paris [1987] and Nelson [1986] independently showed that there is a cut I for which

$$Q \vdash \Phi^I \quad \text{for all theorems} \quad \Phi \quad \text{of} \quad I\Delta_0.$$

In this case we say that the predicate $I(x)$ is an interpreting domain for $I\Delta_0$ in Q: Q can emulate $I\Delta_0$ by quantifying only over numbers in J. This relation is expressed $Q \overset{I}{\succeq} I\Delta_0$. Wilkie has also shown [1987] that there is a subcut J of I (i. e. $J(x)$ is a cut and $Q \vdash \forall x(J(x) \rightarrow I(x))$) for which

$$Q \vdash \forall x \forall y \exists z (J(x) \wedge J(y) \rightarrow J(z) \wedge z = x^{(logy)}).$$

This last theorem of Q is the totality of the ω_1 function with quantifiers restricted to J $(\omega_1(x, y) = x^{(\log y)})$. The formula for the totality of ω_1 is abbreviated Ω_1. Since $Q \vdash \forall x(J(x) \rightarrow I(x)))$, it follows that $Q \vdash \Phi^J$ whenever $Q \vdash \Phi^I$. Hence

$$Q \overset{J}{\succeq} I\Delta_0 + \Omega_1.$$

Now Pudlák's theorem is that for no cut J does $Q \vdash Con_Q^J$. This means, in particular, that we not only have not $Q \vdash Con_Q$ but also not $Q \vdash Con_Q^J$ for J such that $Q \overset{J}{\succeq} I\Delta_0 + \Omega_1$. That is, Q cannot prove even that there is no Q-proof of contradiction whose Gödel number lies in an interpreting domain for $I\Delta_0 + \Omega_1$.

This surely is a strengthening of the theorem of Shepherdson and Bezboruah. Their result says that Q cannot rule out the possibility that one of its theorems is $\perp$. Wilkie and Pudlák's results together show that Q cannot rule out $\perp$ even from among just those theorem's whose Gödel numbers are in a model of $I\Delta_0 + \Omega_1$. Intuitively Pudlák's argument is that Gödel's theorem in its classical form might be meaningless for Q because such a weak theory does such a poor job of defining the natural numbers that it is no wonder it cannot rule out a proof of $\perp$. It after all has to consider a rather wild array of nonstandard "pseudonumbers". But if it is consistent with the axioms of Q even for there to be a proof of $\perp$ coded up in a model that Q defines of $I\Delta_0 + \Omega_1$, we might be inclined to understand this as "the unprovability in Q of Q's consistency" because models of this stronger theory are so much more well behaved.

Exactly how well behaved must models of $I\Delta_0 + \Omega_1$ be for this strengthening to be meaningful? Pudlák says "In such a cut all reasonable definitions of Con_Q are equivalent". Above it was questioned whether this is true, i. e. if the idea is that all definitions of "the consistency of Q" are equivalent, the separation in bounded arithmetic of Con_Q and $HCon_Q$ draws this into question. One might ask further, even if this were true, whether it would be relevant to the metatheoretic meaning for Q of Wilkie and Pudlák's theorems. Pudlák's suggestion can be made precise by pointing out that models of $I\Delta_0 + \Omega_1$ are sufficiently well behaved because the theory proves all the intensionality criteria for theoremhood. His implicit argument is prescriptive in this sense: Since the predicate Thm_Q is intensionally adequate in $I\Delta_0 + \Omega_1$, one need only formulate provability with the variant predicate $Thm_Q^J \Leftrightarrow J(x) \wedge Thm_Q$ to attain intensional adequacy in Q.

The intensionality of an arithmetization of syntax for Q, even with the variant predicate, is not so simple, though. To see this it is instructive to review the proof of Gödel's second theorem. With the intensionality criteria for a theory $T \supseteq I\Delta_0 + \Omega_1$ and the existence of a formula ϕ for which

$I\Delta_0 + \Omega_1 \vdash \phi \leftrightarrow \neg Thm_T(\overline{\phi})$ [12] the theorem for T proceeds in a standard way:

Theorem 1 (Gödel) *If $T \supseteq I\Delta_0 + \Omega_1$ is decidable and consistent, then not $T \vdash Con_T$.*

Proof. From the first incompleteness theorem for T, not $T \vdash \phi$. $I\Delta_0 + \Omega_1 \vdash \neg\phi \to Thm_T(\overline{\phi})$. $I\Delta_0 + \Omega_1 \vdash Thm_T(\overline{\phi}) \to Thm_T(\overline{Thm_T(\overline{\phi})})$ by the fifth intensionality criterion. $I\Delta_0 + \Omega_1 \vdash Thm_T(\overline{Thm_T(\overline{\phi})}) \to Thm_T(\overline{\neg\phi})$ by the first intensionality criterion and choice of ϕ. It follows that $I\Delta_0 + \Omega_1 \vdash \neg\phi \to Thm_T(\overline{\neg\phi})$. Hence $I\Delta_0 + \Omega_1 \vdash \neg\phi \to (Thm_T(\overline{\phi}) \wedge Thm_T(\overline{\neg\phi}))$.

Now by the fifth intensionality criterion $I\Delta_0 + \Omega_1 \vdash Thm_T(\overline{\psi})$ for any tautology ψ, and in particular $I\Delta_0 + \Omega_1 \vdash Thm_T(\overline{\phi \to (\neg\phi \to \bot)})$. Also by two applications of the second intensionality criterion $I\Delta_0 + \Omega_1 \vdash \neg\phi \to Thm_T(\overline{\neg\phi \to \bot}))$ and $I\Delta_0 + \Omega_1 \vdash \neg\phi \to Thm_T(\overline{\bot}))$. Hence $I\Delta_0 + \Omega_1 \vdash Con_T \to \phi$. The theorem now follows by the first incompleteness theorem, since $T \supseteq I\Delta_0 + \Omega_1$.

$\dashv$

It is obvious that one cannot readily generalize this proof for arithmetics not extending $I\Delta_0 + \Omega_1$ because even if one had Gödel's first theorem for such a weak theory, one would not be able to infer ϕ from Con_T, since the conditional $Con_T \to \phi$ would not necessarily be a theorem of T. There is a way around this obstacle, though, as the Shepherdson and Bezboruah's proof reveals. A solution that is evident from this study is to recast provability in $I\Delta_0 + \Omega_1$ as provability in a weaker theory T relativized to a syntactic cut J for which $T \overset{J}{\succeq} I\Delta_0 + \Omega_1$. Still another difficulty remains, however, which is that the intensionality criteria used throughout the proof might not hold for arithmetizations of weak theories, even with the relativized predicate $Thm_T^J(x)$. This is an obstacle to generalizing the proof under consideration to the setting of very weak arithmetics, since these criteria are cited throughout the proof. More importantly (since we know the theorem *can* after all be proven by a different method) this jeopardizes the meaningfulness of the theorem for theories like Q, in so far as the intensionality criteria were supposed to secure the theorem's semantic content.

For example in order to arrive at $I\Delta_0 + \Omega_1 \vdash \neg\phi \to Thm_T(\overline{\neg\phi})$, the above proof relies on $I\Delta_0 + \Omega_1 \vdash Thm_T(\overline{\phi}) \to Thm_T(\overline{Thm_T(\overline{\phi})})$ which is an instance of intensionality criterion five (and is one of the Hilbert-Bernays-Löb derivability conditions). To relativize the proof one would like to show

$$Q \vdash \neg\phi \to (J(\overline{\neg\phi^J}) \wedge Thm_Q(\overline{\neg\phi^J})),$$

but one cannot so easily establish

$$Q \vdash J(\overline{\phi^J}) \wedge Thm_Q(\overline{\phi^J}) \to J(\overline{\phi^J}) \wedge Thm_Q(\overline{Thm_Q^J(\overline{\phi^J})})^{13}.$$

[12] See (Buss [1998]) sections 2.1-2.2 for demonstrations.

[13] In the proof above, we are able to conclude $I\Delta_0 + \Omega_1 \vdash Thm_T(\overline{\phi}) \to Thm_T(\overline{Thm_T(\overline{\phi})})$ because by the fifth intensionality condition $I\Delta_0 + \Omega_1 \vdash Thm_T(\overline{\phi}) \to Thm_\delta(\overline{Thm_T(\overline{\phi})})$ and also $I\Delta_0 + \Omega_1 \vdash \forall u Fmla(u) \to (Thm_\delta(u) \to Thm_T(u))$ since $T \supseteq I\Delta_0 + \Omega_1$.

This last sentence is a derivability condition for Thm^J_Q. It does not follow from a straightforward relativization of the proof because substituting a Δ^b_1 definition of Q for T does not yield $I\Delta_0+\Omega_1 \vdash \forall\overline{A}(Thm_Q(\overline{A}) \to Thm_Q(\overline{Thm_Q(\overline{A})}))$ but only $I\Delta_0 + \Omega_1 \vdash \forall\overline{A}(Thm_Q(\overline{A}) \to Thm_\delta(\overline{Thm_Q(\overline{A})}))$ where δ is a Δ^b_1 definition in $I\Delta_0 + \Omega_1$ of $I\Delta_0 + \Omega_1$. Furthermore, even if one could show $I\Delta_0 + \Omega_1 \vdash \forall\overline{A}(Thm_Q(\overline{A}) \to Thm_Q(\overline{Thm_Q(\overline{A})}))$, a straightforward relativization of this would not result in the desired formula, for

$$Q \vdash \forall\overline{A}(Thm_Q(\overline{A}) \to Thm_Q(\overline{Thm_Q(\overline{A})}))^J. \tag{2}$$

abbreviates

$$Q \vdash \forall\overline{A}(Thm^J_Q(\overline{A}) \to Thm^J_Q(\overline{Thm_Q(\overline{A})})). \tag{3}$$

but from this one may not derive

$$Q \vdash \forall\overline{A}(Thm^J_Q(\overline{A}) \to Thm^J_Q(\overline{Thm^J_Q(\overline{A})})).$$

because the embedded occurrence of "$Thm_Q(\overline{A})$" in (2) and (3) is a numeral rather than a formula and consequently doesn't change in the relativization.

In fact, the proof in (Pudlák [1996]) of not $Q \vdash Con^J_Q$ for $Q \overset{J}{\succeq} I\Delta_0 + \Omega_1$ is entirely different from the proof of Gödel's second theorem just presented, as it relies on bounds on proof length rather than proceeding directly from derivability conditions for the relativized predicate $Thm^J_Q(x)$. Without an appeal to the intensionality of the formula $Thm^J_Q(x)$, however, it is unclear on what grounds one is to find Gödel's second theorem semantically contentful for Robinson's theory. It *is* possible to prove the intensionality criteria for this formula, though as we have seen the interpretability of $I\Delta_0+\Omega_1$ in Q and the fact that the criteria hold in $I\Delta_0+\Omega_1$ alone do not suffice. Since it has been suggested [loc. cite and [1998] pg. 118] that the intensionality in Q of $Thm^J_Q(x)$ as a representation of theoremhood is immediate from these two facts, we now show how actually to prove the criteria in Q.

The proof relies on two additional facts about what $I\Delta_0 + \Omega_1$ proves about Q. The first is that $I\Delta_0 + \Omega_1$ proves that all numbers fall, provably in Q, under all syntactic cuts in Q.

Lemma 2 *For every formula $I(x)$ for which $Q \vdash I(0) \wedge \forall x(I(x) \to I(x+1))$,*

$$I\Delta_0 + \Omega_1 \vdash \forall x Thm_Q(\overline{I(x)}).$$

The next fact says that the Σ-completeness of Q is provable in $I\Delta_0 + \Omega_1$.

Lemma 3 *For all Σ^b_1 formulas $\psi(x,y)$,*

$$I\Delta_0 + \Omega_1 \vdash \forall x(\exists y\psi(x,y) \to Thm_Q(\overline{\exists y\psi(x,y)})).$$

The proofs of these two facts involve techniques in bounded arithmetic not developed in this study. They are Lemmata 5.21 and 5.24(ii) of Hájek and Pudlák [1993].

We shall prove only the fifth of Feferman's intensionality criteria for $Thm_Q^J(x)$ for two reasons. (i) The first two criteria do in fact follow from the intensionality of the arithmetization of $I\Delta_0 + \Omega_1$ and the interpretation $Q \overset{J}{\succeq} I\Delta_0 + \Omega_1$, and it is evident how to prove the third and fourth criteria from the proof of the fifth. (ii) The fifth criteria entails the derivability conditions that make the above proof of Gödel's second theorem possible.

Theorem 4 $Q \vdash Thm_Q^J(\phi) \rightarrow Thm_Q^J(\overline{Thm_Q^J(\overline{\phi})})$.

Proof. First note that since Q has an explicit finite axiomatization, the formulas $Fmla_Q(x)$, $Proof_Q(x)$, $Prf_Q(x,y)$, etc. all are Δ_1^b in Q. since the finite axiomatization of Q ensures a polynomial time decision procedure for them. (On the other hand $Thm_Q(x)$ is not even decidable, of course, by Gödel's first theorem.) One cannot say the same about the relativizations $Prf_Q^J(x,y)$, etc. of these predicates because as the cut J is *proper* in Q it is relatively computationally complex. However, by the $I\Delta_0+\Omega_1$-formalized Σ-completeness of Q,

$$I\Delta_0 + \Omega_1 \vdash Prf_Q(p, \overline{\phi}) \rightarrow Thm_Q(\overline{Prf_Q(p, \overline{\phi})}).$$

Now let J be a cut in Q for which $Q \overset{J}{\succeq} I\Delta_0 + \Omega_1$. By the first lemma,

$$I\Delta_0 + \Omega_1 \vdash Thm_Q(\overline{J(p) \wedge J(\overline{\phi})}).$$

From these two formulas it follows that

$$I\Delta_0 + \Omega_1 \vdash Prf_Q(p, \overline{\phi}) \rightarrow Thm_Q(\overline{J(p) \wedge J(\overline{\phi}) \wedge Prf_Q(p, \overline{\phi})}),$$

which with the $\exists$ rules yields

$$I\Delta_0 + \Omega_1 \vdash \exists u Prf_Q(u, \overline{\phi}) \rightarrow Thm_Q(\overline{\exists u(J(u) \wedge J(\overline{\phi}) \wedge Prf_Q(u, \overline{\phi}))}),$$

or

$$I\Delta_0 + \Omega_1 \vdash Thm_Q(\overline{\phi}) \rightarrow Thm_Q(\overline{Thm_Q^J(\overline{\phi})}).$$

And only now, by choice of J,

$$Q \vdash Thm_Q^J(\overline{\phi}) \rightarrow Thm_Q^J(\overline{Thm_Q^J(\overline{\phi})}).$$

$\dashv$

The second incompleteness theorem for the pair (Q, Con_Q^J) is now immediate. What can be said, however, of the semantic content of this theorem? The argument above was that the meaning of consistency in $I\Delta_0 + \Omega_1$ is more ambiguous than it first appears. Not withstanding the provability of the intensionality criteria for the standard provability formula $Thm_\delta(x)$, one may ask whether these criteria are appropriate in such a weak setting. It has been suggested that the criteria might be too strong to be necessary conditions on an arithmetization for it to be semantically contentful, among other reasons because they imply not only that ϕ is provable only if $Thm_T(\overline{\phi})$ also is, but also that "they assume certain ontological principles (e.g. that there must exist numbers $\overline{A}$, $\overline{Prov(\overline{A})}$, $\overline{Prov(\overline{Prov(\overline{A})})}$, etc. ad infinitum, given the provability of A) which do not seem to be inherent in the notion of A's provability" and which therefore seem unrelated to common intuitions about provability even for relatively strong theories[14]. This study suggested that this is especially true in bounded arithmetic where concepts like Herbrand consistency, which are combinatorially very simple and more readily treatable, provably separate from the standard consistency statement. In the present case, where arithmetization is achieved via interpretation of a stronger theory, this type of ambiguity is only compounded. For through interpretation, the separation of these formulations of consistency sharpens. A stronger separation than the T-provable inequivalence of two formulas is the T-provability of one and T-unprovability of the other. In the case where this happens the formulas will be called "sharply T-separated". The following two results relating Q with the unbounded theory $I\Delta_0 + Exp$ show that the formulas Con_Q and $HCon_Q$ are sharply Q-separated when relativized to a sufficiently short cut. The first is a result of Wilkie for which he gave a model-theoretic proof:

Theorem 5 (Wilkie) *For bounded formulas $\psi(x)$ the following are equivalent.*

1. $Q \preceq Q + \forall x \psi(x)$
2. $I\Delta_0 + Exp \vdash \forall x \psi(x)$

Proof. We need a proof-theoretic version of this theorem since it is more useful for our analysis. Accordingly we will prove actually the equivalence between 2 and the following condition:

3. There is a syntactic cut $I(x)$ in Q such that $Q \vdash \forall x(I(x) \rightarrow \psi(x))$.

This condition is equivalent to 1. Moreover, we shall only use the direction 2 ⇒ 3, so we prove only this entailment. For our proof we will need to cite Wilkie's result that if $I\Delta_0 + Exp \vdash \forall x \psi(x)$ then there is a k such that $I\Delta_0 \vdash \forall x(\exists y(y = 2_k^x) \rightarrow \psi(x))$. We assume this result, since we cannot replace Wilkie's original proof with one using only techniques from this essay. We also need a lemma of Solovay known as the technique for shortening cuts:

[14] Andrew Boucher *FOM Digest* **Vol 30** Issue 1

Lemma 6 (Solovay) *Lemma: Let $T \supseteq I\Sigma_1$. For each $n \in \omega$ and each T-cut I, there is a T-cut J_n such that $T \vdash \forall x(J_n(x) \to I(2^x_n)$.*

1. Given I, there is a T-cut J such that $J \subseteq I$ and J is closed under addition.

   Proof: Define $J(x) \Leftrightarrow I(x) \wedge \forall y(I(y) \to I(x+y))$. It is easy to check that $J \subseteq I$ and J is a T-cut.

   If $x, z \in J$, then for each $y \in I$, $z + y \in I$. Therefore $x + z \in J$. So $x + z + y \in I$ for each $y \in I$. In particular $x + z + 0 \in I$. So $I(x+y) \wedge \forall y(I(y) \to (I(x+z+y))$, i. e. $J(x+y)$. Therefore, J is closed under addition.

2. For each $n \in \omega$ there is a T-cut J_n such that $T \vdash J_n \subseteq I$ and $T \vdash \forall x(J_n(x) \to I(2^x_n))$.

   Proof (by induction on n):

   Let $J_0 = J$ from 1. Then $T \vdash \forall x(J_0(x) \to I(x))$, so $T \vdash \forall x(J_0(x) \to I(2^x_0))$.

   Now suppose J_n is given. From 1 we know that we may assume J_n is closed under addition, so long as we can show J_{n+1} is.

   Define I_{n+1} in T by $I_{n+1}(x) \leftrightarrow J_n(x)$. Then $T \vdash \forall x(J_n(2^x) \to I(2^x_{n+1}))$ by hypothesis. So $T \vdash \forall x(I_{n+1}(x) \to I(2^x_{n+1}))$.

   I_{n+1} is a T-cut:

   (a) $T \vdash I_{n+1}(0)$ because $T \vdash J_n(1)$.

   (b) $T \vdash \forall x(I_{n+1}(x) \to I_{n+1}(x+1))$ because if $x \in I_{n+1}$, then $2^x \in J_n$. And J_n is closed under addition, thus $2^x + 2^x \in J_n$. So $2^{x+1} \in J_n$, and $x + 1 \in I_{n+1}$.

   (c) $T \vdash \forall x \forall y(I_{n+1}(x) \wedge y < x \to I_{n+1}(y))$ because $T \vdash \forall x \forall y(J_n(2^x) \wedge y < x \to J_n(2^y))$.

   Now I_{n+1} is not necessarily closed under addition. But we repeat the procedure from 1 and define: $J_{n+1}(x) \Leftrightarrow I_{n+1}(x) \wedge \forall y(I_{n+1} \to I_{n+1}(x+y))$. Since $T \vdash \forall x(J_{n+1}(x) \to I(2^x_{n+1}))$ and also $T \vdash J_{n+1}(x) \to I_{n+1}(x)$, one has $T \vdash \forall x(I_{n+1}(x) \to I(2^x_{n+1}))$. Moreover J_{n+1} is easily seen to be a T-cut. But unlike I_{n+1}, J_{n+1} is provably closed under addition, as needed.

$\dashv$

Now we may prove 2 ⇒ 3 of Theorem 5. Suppose $I\Delta_0 + Exp \vdash \forall x \psi(x)$. Wilkie showed with a model-theoretic argument that there is a k such that $I\Delta_0 \vdash \forall x(\exists y(y = 2^x_k) \to \psi(x))$ (we use the definition of the graph of exponentiation in $I\Delta_0$). Let us rewrite this last sentence $I\Delta_0 \vdash \forall x(ANT \to CONS)$. Now let J be a syntactic cut such that $Q \overset{J}{\succeq} I\Delta_0$. Then we have

$$\begin{array}{rll} Q & \vdash & \forall x(J(x) \to (ANT \to CONS)^J) \\ & \vdash & \forall x(J(x) \to (ANT^J \to CONS^J)) \\ & \vdash & \forall x(J(x) \to (ANT^J \to \psi^J)) \\ & \vdash & \forall x(J(x) \to (ANT^J \to \psi)) \\ & \vdash & \forall x(J(x) \to ((\exists y(y = 2^x_k))^J \to \psi)) \\ & \vdash & \forall x(J(x) \to (\exists y(J(y) \wedge y = 2^x_k) \to \psi)) \end{array}$$

Now we wish to adapt Lemma 6 to shorten J appropriately. However, the lemma holds only for arithmetics containing $I\Sigma_1$. Inspection of the proof reveals that Σ_1-induction is never used, though, since the induction is in the metatheory and never formalized. The restriction to theories extending $I\Sigma_1$ was only to ensure the existence of the exponential function. However, since exponentiation can be defined even in $I\Delta_0$ via its graph, the technique for shortening cuts is applicable for these theories also. Since exponentiation is not provably total in $I\Delta_0$, the proof serves also to guarantee the existence of appropriate numerals for exponential instances: For each $n \in \omega$ and each syntactic cut K of $I\Delta_0$, there is a cut J_n of $I\Delta_0$ such that $I\Delta_0 \vdash \forall x(J_n(x) \to \exists y(y = 2^x_n \wedge K(y)))$.

Consider the case where K is improper in $I\Delta_0$ (i. e. $I\Delta_0 \vdash \forall x K(x)$) and $n = k$. Then

$$I\Delta_0 \vdash \forall x(J_k(x) \to \exists y(y = 2^x_k)).$$

Since $Q \overset{J}{\succeq} I\Delta_0$, it follows that $Q \vdash \forall x(J(x) \to (J_k(x) \to (\exists y(y = 2^x_k))^J))$, or

$$Q \vdash \forall x(J(x) \wedge J_k(x) \to \exists y(J(y) \wedge y = 2^x_k)).$$

Define the Q-cut I by $I(x) \leftrightarrow J(x) \wedge J_k(x)$ so that $Q \vdash \forall x(I(x) \to \exists y(J(y) \wedge y = 2^x_k))$. Since $Q \vdash I(x) \to J(x)$ and $Q \vdash \forall x(J(x) \to \exists y(J(y) \wedge y = 2^x_k) \to \psi)$, it follows that

$$Q \vdash \forall x(I(x) \to \psi(x)).$$

$\dashv$

Theorem 7 (Sheperdson) *$I\Delta_0 + Exp \vdash HCon_Q$.*

Proof. From our axiomatization of Q construct the open theory Q_{open} by removing all the universal quantifiers from Q's axioms, and replacing the axiom (3) for the successor function with an equivalent (open) axiom for a predecessor function. Clearly, $Q_{open} \succeq Q$. If $I\Delta_0 + Exp \vdash \neg HCon_{Q_{open}}$, then $I\Delta_0 + Exp$ proves that there is some tautology

$$\bigvee_{i=1}^{k} [He^*(\bigwedge Q_{open} \to \bot)(t_{i1}, f_1(t_{i1}), t_{i2}, f_2(t_{i1}, f_1(t_{i1}), t_{i2}), \ldots)].$$

This is impossible, though, since a truth predicate for open formulas is definable in $I\Delta_0 + Exp$. Thus $I\Delta_0 + Exp \vdash HCon_{Q_{open}}$.

Define in $I\Delta_0 + Exp$ the formula $RCon_T(n)$ saying "contradiction is not provable in T with a proof of cut-rank n" so that $RCon_T(0) \Leftrightarrow CFCon_T$. By the interpretability of Q in Q_{open}, for every n there is an m such that

$$I\Delta_0 + Exp \vdash RCon_{Q_{open}}(m) \to RCon_Q(n).$$

Then by a partial cut-elimination theorem in $I\Delta_0 + Exp$ (Theorem 5.17(ii) of (Hájek and Pudlák [1993]),

$$I\Delta_0 + Exp \vdash CFCon_{Q_{open}} \to RCon_{Q_{open}}(m).$$

Thus, for all n, $I\Delta_0 + Exp \vdash RCon_Q(n)$. In particular, $I\Delta_0 + Exp \vdash CFCon_Q$. Finally, by the provability in $I\Delta_0 + Exp$ of the equivalence of Herbrand provability and cut-free provability,

$$I\Delta_0 + Exp \vdash HCon_Q.$$

⊣

Corollary 8 *There is a cut K such that $Q \vdash HCon_Q^K$ and not $Q \vdash Con_Q^K$ for $Q \overset{K}{\succeq} I\Delta_0 + \Omega_1$.*

Proof. The formula $HPrf_Q(x,y)$ is bounded (it is Δ_1^b-definable in $I\Delta_0 + \Omega_1$, so $HCon_Q$ is a $\forall$-theorem of $I\Delta_0 + Exp$. Therefore by 2 $\Rightarrow$ 3 of Theorem 5, there is a cut $I(x)$ in Q such that $Q \vdash \forall x(I(x) \to \neg HPrf(x, \overline{\bigwedge Q \to \bot}))$. This last formula is just $Q \vdash HCon_Q^I$. Since subcuts preserve interpretability, choose $K \subseteq I \bigcap J$ for $Q \overset{J}{\succeq} I\Delta_0 + \Omega_1$. One then has $Q \vdash HCon_Q^K$ and not $Q \vdash Con_Q^K$ for $Q \overset{K}{\succeq} I\Delta_0 + \Omega_1$.

⊣

Thus there is a very sharp separation between the formulas Con_Q and $HCon_Q$ when the domain is duly restricted. In particular, the arithmetization of metatheory in Q via relativization is less stable than the arithmetization of metatheory in the theories that Q thereby interprets: Whatever intensionality can be recovered depends arbitrarily on the cut one chooses for one's interpretation. If, as Pudlák suggests, restricting of the domain of quantification increases the metatheoretic meaning of formulas, then there arises the paradoxical situation where an increase in semantic content results in a precisification of how nebulous a notion theoretic consistency is: In particular in Robinson Arithmetic if one relativizes one's arithmetization so as to maximize interpretive strength, whether or not "consistency" even is provable or unprovable depends on exactly how one phrases the question.

5 Conclusions

This study began with the question whether arithmetic theories in general strong enough to arithmetize their own syntax prove the unprovability of their consistency. We have found that the question requires precisification. For there are different conceptions of adequate arithmetization, and if adequacy entails the ability to verify purely from within the arithmetic one is studying salient properties of the arithmetization, then it is possible for an arithmetic to numeralwise represent its entire classical metatheory without doing so adequately. For instance the result of Bezboruah and Shepherdson is not intensionally adequate and thus is not a proper demonstration of the unprovability of the consistency of Q.

The question must be made further precise by an additional criterion for the proper metatheoretic properties to arithmetize. Since some combinatorial principles are needed to associate standard sequent calculus or Hilbert-style proofs with the construction of Herbrand disjunctions, or to determine that *modus ponens* or the cut rule are conservative over direct derivation techniques, arithmetics lacking such principles do not "understand" provability with indirect techniques and might be able to rule out the possibility of a Herbrand proof of contradiction despite proving Gödel's theorems in their standard form. For bounded theories like $I\Delta_0 + \Omega_1$ this might mean that combinatorially simpler predicates like Herbrand provability are more appropriate carriers of metatheoretic content. To motivate this possibility, it was shown in §3 that the notion of provability based on the the no-counterexample interpretation relativizes immediately to the theory it is formulated in, since only construction techniques provably feasible from that theory's perspective are considered when the question of the provability of a formula is put to the theory. This suggests that the intensionality criteria proposed by Feferman might only be adequate for theories that view at least all primitive recursive functions as feasible. For a bounded arithmetic T could conceivably prove such criteria for a predicate Thm_τ without being able to prove the equivalence $Thm_\tau(x) \leftrightarrow \exists y HPrf_\tau(y,x)$–evidence that the criteria express properties that, while traditionally seen as central to or even constitutive of the notion of theoremhood, are not even essential to the theoremhood as treatable with the resources of T.

In fact, precisely this occurs in the case of the arithmetic $I\Delta_0 + \Omega_1$: The theory proves the intensionality criteria for Thm_δ, but doesn't prove either Herbrand's theorem or a Gödel-like theorem for $HCon_\delta$. The reason for this is that the propositional proofs that would be needed to construct the Herbrand disjunctions for some of this theory's theorems are not sufficiently concrete, feasibly computable, or rather–speaking from the theory's point of view–they don't exist. At the very least, this casts doubt on the adequacy of those conditions in bounded arithmetic. On the other hand, since those same conditions fail in bounded arithmetic for arithmetizations of Herbrand provability, it is not entirely clear that this alternative notion captures theoremhood at all satisfactorily either. Saying that bounded theories cannot answer or even formulate questions about their own theoremhood and consistency seems the most accurate synopsis.

Finally, however compelling the above analysis of the prospects for an intensional metatheory for bounded arithmetic in general is, the situation seems bleaker yet in the weakest arithmetics like Robinson's Q. Since Q doesn't prove the proposed intensionality criteria for the standard formulation of provability, the only hope is to adopt an alternative formulation. The best attempt at this seems to be the one suggested by Pudlák: relativize the standard formula to a definition in Q of a model of a theory that proves the criteria. Theorem 4 shows that it is possible to do this in such a way that the criteria apply also in Q for the relativized formula. However this technique in the end reveals that the ambiguity of metatheoretic notions in bounded theories is only more rampant in Q, for the arithmetization of Herbrand consistency is not only inequivalent in suitable relativizations to the standard consistency statement–it is also provable.

This answers the question posed in the introduction. It is possible for a first order theory to be strong enough to generate self-reference phenomena without being able meaningfully to prove the unprovability of its own consistency (or even meaningfully to pose the question). This suggests that the fine structure of the lower reaches of the arithmetic hierarchy might prove to be a valuable resource for discovering further interrelationships among the various metatheoretic notions investigated first in the foundational studies of the 1930's which in strong arithmetics can only be glossed over.

References

[2005] Adamowicz, Z. and L. Kolodziejczyk. "Well-behaved principle alternative to bounded induction" forthcoming in *Theoretical Computer Science.*

[1976] Bezboruah, A. and J. C. Shepherdson. "Gödel's Second Incompleteness theorem for *Q*" in *The Journal of Symbolic Logic* **41**, pgs. 503-512.

[1985] Buss, S. R. *Bounded Arithmetic.* PhD. Thesis for the Princeton University Department of Mathematics.

[1998] Buss, S. R. "Proof theory of arithmetic" in *Handbook of Proof Theory* edited by Samuel R. Buss. North Holland. New York. 1998.

[1998] Buss, S. R. "Introduction to proof theory" in *Handbook of Proof Theory* edited by Samuel R. Buss. North Holland. New York. 1998.

[1988] Detlefsen, Michael. *Hilbert's Program: An Essay on Mathematical Instrumentalism.* 1988.

[1960] Feferman, S."Arithmetization of metamathematics in a general setting" in *Fundamenta Mathematica* **XLIX**, pgs. 37-92.

[1931] Gödel, K. "On formally undecidable propositions of *Principia Mathematica* and related systems I" reprinted in *From Frege to Gödel: A sourcebook* *in mathematical logic 1879-1931* edited by Jean van Heijenoort. Harvard University Press. Cambridge. 1967.

[1993] Hájek, P. and P. Pudlák. *Metamathematics of First-Order Arithmetic.* Perspectives in Mathematical Logic, Springer-Verlag, Santa Clara. 1993.

[1931] Herbrand, J. "On the consistency of arithmetic" reprinted in *From Frege to Gödel: A sourcebook in mathematical logic 1879-1931* edited by Jean van Heijenoort. Harvard University Press. Cambridge. 1967.

[1958] Kreisel, G. "Mathematical significance of consistency proofs" in *The Journal of Symbolic Logic* **23**, pgs. 159-182.

[1986] Nelson, E. *Predicative Arithmetic.* Princeton University Press. Princeton. 1986.

[1971] Parikh, R. "Existence and feasibility in arithmetic" in *The Journal of Symbolic Logic* **36**, pgs. 494-508.

[1985] Pudlák, P. "Cuts, consistency statements, and interpretations" in *Journal of Symbolic Logic* **50**, pgs. 423-441.

[1996] Pudlák, P."On the length of proofs of consistency" in *Collegium Logicum, Annals of the Kurt-Godel-Society* **2**, pgs. 65-86.

[2004] Pudlák, P. "Consistency and games: in search of new combinatorial principles" unpublished manuscript.

[1985] Sieg, W. "Fragments of arithmetic" in *Annals of Pure and Applied Logic* **28**, pgs. 33-71. 1985.

[1978] Statman, R. "Bounds for proof search and speed-up in the predicate calculus" in *Annals of Mathematical Logic* **15**, pgs. 225-287. 1978.

[1953] Tarski, A., A. Mostowski and R. M. Robinson. *Undecidable Theories.* North Holland. Amsterdam. 1953.

[1981] Wilkie, A. J. and J. B. Paris. "Δ_0 sets and induction" in *Proceedings from the Jadswin Logic Conference.* Leeds University Press, pgs. 237-248.

[1987] Wilkie, A. J. and J. B. Paris. "On the scheme of induction for bounded arithmetic formulas" in *Annals of Pure and Applied Logic* **35**, pgs. 261-302. 1987.

[2002] Willard, D. "How to extend the semantic tableaux and cut-free versions of the second incompleteness theorem almost to Robinson's arithmetic Q" in *The Journal of Symbolic Logic* **67**, pgs. 465-496.

[1985] Young, P. "Gödel theorems, exponential difficulty and undecidability of arithmetic theories: an exposition" in *Recursion Theory.* Proceedings of Symposia in Pure Mathematics **42**, American Mathematical Society. Providence. 1985.