Notre Dame Journal of Formal Logic
Volume 31, Number 1, Winter 1990

Characters and Fixed Points in Provability Logic

ZACHARY GLEIT and WARREN GOLDFARB*

Abstract Some basic theorems about provability logic—the system of modal logic that reflects the behavior of formalized provability predicates in theories such as arithmetic—are given simplified, model-theoretic proofs. The theorems include the Fixed Point Theorem of de Jongh and Sambin, the Craig Interpolation Theorem, and the Beth Definability Theorem. Attention is also paid to the complexity of models for formulas in this logic.

Provability logic is the modal logic whose axioms are the tautologies and all formulas of the forms $\Box(A \to B) \to (\Box A \to \Box B)$ and $\Box(\Box A \to A) \to \Box A$, and whose inference rules are modus ponens and necessitation. This system has been known variously as K4W, L, G, GL, and PRL; we adopt the next to last of these monikers, and use "$\vdash$" for provability in GL. GL is of interest since it reflects the behavior of formalized provability predicates in such theories as Peano Arithmetic (for details see [1] or [3]).

A formula A is said to be *modalized* in a sentence letter p iff every occurrence of p in A lies in the scope of a $\Box$. The Fixed Point Theorem, due to de Jongh and Sambin, states that if A is modalized in p then there exists a formula H containing only sentence letters of A aside from p such that $\vdash \Box(p \leftrightarrow A) \leftrightarrow \Box(p \leftrightarrow H)$. Via the connection of GL with provability in formal theories, this theorem implies that sentences in such formal theories constructed from formalized provability predicates by "self-referential" techniques are provably equivalent to sentences involving no self-reference.

In Section 1, we give a purely model-theoretic proof of the Fixed Point Theorem, which we think to be more perspicuous than the extant proofs. The

*We would like to thank George Boolos for much generous assistance and many helpful suggestions, and in particular for simplifying the counterexample found in the fourth paragraph of Section 3. We are also grateful to Lisa Reidhaar-Olson for catching several errors in the penultimate draft.

Received January 20, 1988; revised July 26, 1988

same basic apparatus is exploited in Section 2 to prove a theorem on the joint satisfiability of formulas, which then quickly yields the Craig Interpolation Lemma and the Beth Definability Theorem for GL. A final section is devoted to bounds on the complexity of models.

We assume that the basic concepts of the semantics of GL are known; here we review them quickly (using the notation of [1]). A model is a triple $M = \langle W, R, P\rangle$ where W is a nonempty finite set of objects called "worlds", R is an irreflexive, transitive relation on W called the accessibility relation, and P is a function assigning a truth value to each pair $\langle w, q\rangle$, where $w \in W$ and q is a sentence letter. The notion of a sentence F being true at a world w of M, in symbols $M \vDash_w F$, is defined in the usual modal manner; in particular, $M \vDash_w \Box F$ iff $M \vDash_v F$ for each world v accessible from w, that is, each v such that wRv. It should be noted that if $M \vDash_w \Box F$ and v is accessible from w, then $M \vDash_v \Box F$; for, by the transitivity of R, if v is accessible from w then all worlds accessible from v are accessible from w.

A formula is valid in a model iff it is true at each world of the model. The modal completeness theorem for GL, due to Segerberg, states that $\vdash F$ iff F is valid in all models. In what follows we assume this result, and deal only with models, often passing from validity to provability without special notice.

If $M = \langle W, R, P\rangle$ is a model and w a world of M, the *submodel of M generated by w* is the model $M' = \langle W', R', P'\rangle$ where $W' = \{w\} \cup \{x \in W \mid wRx\}$, and R' and P' are restrictions of R and P to W'. It is easily shown that $M \vDash_w F$ iff $M' \vDash_w F$ for any formula F. Hence, if u and v are worlds of models M and N, and the submodel of M generated by u is isomorphic to that of N generated by v, then $M \vDash_u F$ iff $N \vDash_v F$ for all formulas F. We refer to this principle as "continuity".

We shall need the following simple fact: If a formula G is valid in a model N whenever F is valid in N, then $\vdash \Box F \to \Box G$. For let w be a world in a model M such that $M \vDash_w \Box F$. Let N be the model that is the restriction of M to the worlds accessible from w. By continuity, F is valid in N; hence G is valid in N. By continuity again, $M \vDash_w \Box G$. Thus $\Box F \to \Box G$ is valid in all models; so by completeness, $\vdash \Box F \to \Box G$.

Let $M = \langle W, R, P\rangle$ be a model. The M-rank of the worlds of M is defined thus: $M\text{-rank}(w) = 0$ iff no world is accessible from w; otherwise, $M\text{-rank}(w) = 1 + \max\{M\text{-rank}(v) \mid wRv\}$. The finitude of W and the irreflexivity and transitivity of R ensure that every world of M has a unique M-rank.

Let S be a finite set of sentence letters. We define the *n-S-characters* by induction on n. The 0-S-characters are all the conjunctions $\pm q_1 \cdot \pm q_2 \cdot \ldots \cdot \pm q_k$, where $q_1, \ldots, q_k$ is a listing of S and $\pm$ indicates the presence or absence of a negation sign. (If S is empty, $\top$ is the sole 0-S-character.) If $K_1, \ldots, K_m$ are all the $(n-1)$-S-characters, then the n-S-characters are all the conjunctions $\pm q_1 \cdot \pm q_2 \cdot \ldots \cdot \pm q_k \cdot \pm \Diamond K_1 \cdot \pm \Diamond K_2 \cdot \ldots \cdot \pm \Diamond K_m$, where $\Diamond$ is defined as $\sim \Box \sim$. By induction, it may be easily shown that each satisfiable n-S-character implies a unique $(n-1)$-S-character. If u is a world of a model M, the *n-S-character of u* is the unique n-S-character K such that $M \vDash_u K$. If u and v have the same n-S-character for $n > 0$, then for every x accessible from u there is a y accessible from v such that x and y have the same $(n-1)$-S-character. For if K is the $(n-1)$-S-character of x then $\Diamond K$ is a conjunct of the n-S-character of u; hence

$\Diamond K$ is true at v, so that there is a y accessible from v at which K is true. (The term "character" comes from [1]. Our definition differs slightly from Boolos's, in order to avoid redundant conjuncts. Our definition is identical to that in [2], p. 34, for what Fine calls "normal form formulas for n".)

1 For the remainder of this section, let A be a formula modalized in p, S the set of sentence letters of A aside from p, and n the number of boxed subformulas of A, that is, the number of distinct subformulas of A having the form $\Box F$.

Since A is modalized in p, A is a truth-function of members of S and boxed formulas. Hence the truth values at any world w of the members of S and the boxed subformulas of A determine the truth value of A at w; thus if $M \vDash_w (p \leftrightarrow A)$ these truth values determine the truth value of p at w, and so also determine the truth values at w of all subformulas of A. That is, two worlds at which $p \leftrightarrow A$ is true will differ on some subformula of A only if they differ either on a sentence letter in S or on a boxed subformula of A.

This simple observation underlies our proof of the Fixed Point Theorem. Truth values at w of all subformulas of A are determined by truth values at w of members of S and subformulas $\Box E$. The truth value at w of $\Box E$ is determined by truth values of E at worlds v accessible from w. If $p \leftrightarrow A$ is true at all such v, the truth value at v of E is determined by truth values at v of members of S and the boxed subformulas of A. By iteration, if we assume $p \leftrightarrow A$ is valid in the model, truth values of all subformulas of A at any world w are uniquely determined by the truth values of members of S at w and at the worlds accessible from w.

Of course, the Theorem requires somewhat more: a specification of a fixed amount of information about the truth values of members of S at w and at the worlds accessible from w that suffices to determine the truth value of p at w. As we shall see, the n-S-character of w suffices for this.

Fixed Point Lemma *Suppose M and N are models in which $p \leftrightarrow A$ is valid, and let u_0 and v_0 be worlds of M and N, respectively, which have the same n-S-character. Then u_0 and v_0 agree on p.*

Proof: Suppose not. We construct a sequence $u_0, u_1, \ldots, u_n$ of worlds of M, a sequence $v_0, v_1, \ldots, v_n$ of worlds of N, and a sequence $\Box D_0, \ldots, \Box D_n$ of subformulas of A, with the following property for each i, $0 \leq i \leq n$:

(P_i) u_i and v_i have the same $(n - i)$-S-character; for each $j < i$, $\Box D_j$ is true at both u_i and v_i, but u_i and v_i differ on $\Box D_i$.

Basis: The worlds u_0 and v_0 are given. By supposition, u_0 and v_0 differ on p. So u_0 and v_0 must differ either on a sentence letter in S or on a boxed subformula of A. But u_0 and v_0 agree on all sentence letters in S, since they have the same n-S-character. Let $\Box D_0$ be a boxed subformula of A on which u_0 and v_0 differ. Thus (P_i) holds for $i = 0$.

Induction: Let $i < n$, and suppose $u_0, \ldots, u_i$, $v_0, \ldots, v_i$ and $\Box D_0, \ldots, \Box D_i$ are given and possess property (P_i). To fix ideas, suppose $\Box D_i$ is false at u_i and true at v_i; the other case is treated symmetrically. Let u_{i+1} be a world of min-

imal M-rank among those worlds u accessible from u_i such that D_i is false at u. By minimality, D_i is true at every world accessible from u_{i+1}, so that $\Box D_i$ is true at u_{i+1}. In addition, since u_{i+1} is accessible from u_i, every boxed subformula true at u_i is true at u_{i+1}.

Since u_i and v_i have the same $(n-i)$-S-character and $n-i>0$, there is a world v of N accessible from v_i such that v and u_{i+1} have the same $(n-i-1)$-S-character. Let v_{i+1} be any such v. Then every necessitation true at v_i is true at v_{i+1}. Thus $\Box D_0, \ldots, \Box D_i$ are all true at both u_{i+1} and v_{i+1}.

It remains to select $\Box D_{i+1}$. Since $\Box D_i$ is true at v_i, D_i is true at v_{i+1}. But D_i is false at u_{i+1}. Hence u_{i+1} and v_{i+1} must differ on either a sentence letter in S or a boxed subformula of A. Now u_{i+1} and v_{i+1} agree on all sentence letters in S, since they have the same $(n-i-1)$-S-character. Let $\Box D_{i+1}$ be any boxed subformula of A on which u_{i+1} and v_{i+1} differ.

This concludes the construction of the sequence. Now property (P_i) for each $i \leq n$ implies that $\Box D_0, \ldots, \Box D_n$ are all distinct. This is a contradiction, since by hypothesis there are n boxed subformulas of A. Hence the supposition that u_0 and v_0 differ on p is false.

Fixed Point Theorem *There exists a sentence H containing only sentence letters from S such that* $\vdash \Box(p \leftrightarrow A) \leftrightarrow \Box(p \leftrightarrow H)$.

Proof: Let H be the disjunction of all n-S-characters K with the following property: There exists a model M and a world w of M such that $p \leftrightarrow A$ is valid in M, p is true at w, and w has n-S-character K.

Suppose $p \leftrightarrow A$ is valid in a model N; we show that $p \leftrightarrow H$ is also valid in N. Let u be a world of N, and let K be the n-S-character of u. If p is true at u, then K is among the n-S-characters disjoined to form H; hence H is true at u. If H is true at u then K must be a disjunct of H; hence there exist a model M in which $p \leftrightarrow A$ is valid and a world v of M at which p is true and which has n-S-character K. Thus u and v have the same n-S-character. By the Fixed Point Lemma, u and v agree on p. Thus p is true at u.

Now suppose $p \leftrightarrow H$ is valid in a model N; we show that $p \leftrightarrow A$ is valid in N. Suppose $p \leftrightarrow A$ is not valid in N. Let w be a world of lowest N-rank at which $p \leftrightarrow A$ is false, and let M be the submodel of N generated by w. Then $M \models_v (p \leftrightarrow A)$ for each world v accessible from w. Let M' be the model like M except that $M' \models_w p$ iff $M \models_w \sim p$. By continuity, if v is accessible from w and D is a subformula of A, $M' \models_v D$ iff $M \models_v D$. Moreover, $M' \models_w H$ iff $M \models_w H$ and $M' \models_w A$ iff $M \models_w A$, since H does not contain p and A is modalized in p, so that the truth value of p at w does not affect the truth values of H and A at w. Consequently, $p \leftrightarrow A$ is true at w in M', and hence valid in M', but $p \leftrightarrow H$ is false at w in M'. This contradicts what was proved in the previous paragraph.

We have shown that $p \leftrightarrow A$ is valid in a model iff $p \leftrightarrow H$ is valid in the model. Hence $\vdash \Box(p \leftrightarrow A) \leftrightarrow \Box(p \leftrightarrow H)$.

Note 1 The argument given two paragraphs above can easily be extended to show that any M can be transformed into a model in which $p \leftrightarrow A$ is valid by altering only truth values assigned to p. Thus any satisfiable formula F lacking p can be made true at some world of a model in which $p \leftrightarrow A$ is valid; equivalently, a formula F lacking p is valid if it is valid in all models in which $p \leftrightarrow A$ is valid. Formulated in this last way, this result was first shown by de Jongh.

Note 2 Our proof of the Fixed Point Theorem is at bottom a model-theoretization of the proof in [1], Chapter 11. Boolos defines his fixed point as the disjunction of n-S-characters K such that $\vdash (A \leftrightarrow p) \cdot \Box (A \leftrightarrow p) \cdot K \rightarrow p$. If an n-S-character K is a disjunct of our fixed point H, then the Fixed Point Lemma and completeness suffice to show that it meets Boolos's condition, and so is a disjunct of his fixed point. Conversely, let K be an n-S-character that meets Boolos's condition. If K is unsatisfiable then it will not be a disjunct of H, but if K is satisfiable then the result given in Note 1 shows that K is a disjunct of H. Thus, Boolo's fixed point and ours are the same except for unsatisfiable n-S-characters included in Boolos's.

Note 3 Our fixed point, like Boolos's, is a disjunction of n-S-characters, where n is the number of boxed subformulas of A. Hence, as Boolos noted, it contains at most n nested $\Box$'s. Of course, such a disjunction may be equivalent to a formula with less nesting, but as a general bound this is best possible. For example, if A is $\sim\Box^n p$ then H is $\sim\Box^n \perp$ ($\Box^n$ denotes a string of n consecutive $\Box$'s). Note that the bound on the nesting in the fixed point is the number of boxed subformulas of A, not the nesting of A. Thus, for example, if A is $\Box (q \leftrightarrow p) \rightarrow \Box (\sim q \leftrightarrow p)$, then A has nesting of one, but A has no fixed point of nesting < 2 (see below).

The syntactic computation of fixed points given in [3], p. 80, does not yield this bound on the modal complexity of fixed points. Thus our proof provides more information in this regard. However, our proof provides a far less efficient algorithm for calculating fixed points than does Smoryński's. For if $|S| = s$ then the number $c(n,s)$ of n-S-characters obeys the recursion $c(0,s) = 2^s$; $c(n+1,s) = 2^s 2^{c(n,s)}$; thus, if s and n are anything but very small, a procedure that requires checking each n-S-character for inclusion as a disjunct of the fixed point will be impractical. (Smoryński's discussion of this point in [3], pp. 122–124, contains an error, on which we shall elaborate in Section 3.) Our proof does suggest a heuristic method for seeking fixed points that, in many cases, is reasonably expeditious: search for an exhaustive list of properties of a world in a model, expressible using sentence letters from S and at most n nested $\Box$'s, that imply the truth of p at the world, given that $A \leftrightarrow p$ is valid in the model; then disjoin the formulas expressing those properties. For example, let A be $\Box (q \leftrightarrow p) \rightarrow \Box (\sim q \leftrightarrow p)$, let $p \leftrightarrow A$ be valid in M, and let w be a world of M. Either of two simple properties of w yield the truth of p at w:

(1) M-rank$(w) = 0$
(2) There exists a v accessible from w with M-rank$(v) = 0$ and $M \vDash_v \sim q$.

For if (1) holds then $M \vDash_w \Box (\sim q \leftrightarrow p)$, so that $M \vDash_w A$ and hence $M \vDash_w p$. If (2) holds, then $M \vDash_w \sim\Box (q \leftrightarrow p)$, so that again $M \vDash_w A$ and $M \vDash_w p$.

Now suppose w has neither property; we find a third property that holds if and only if p is true at w. Since (1) and (2) fail, M-rank$(w) > 0$ and there exists a u of M-rank 0 accessible from w such that $M \vDash_u q$, so that $M \vDash_u p \leftrightarrow q$; hence $M \vDash_w \sim\Box (\sim q \leftrightarrow p)$. Thus $M \vDash_w p$ iff $M \vDash_w \sim\Box (q \leftrightarrow p)$. If $M \vDash_w \sim\Box (q \leftrightarrow p)$ then there exists a v of lowest M-rank accessible from w such that $M \vDash_v \sim q \leftrightarrow p$. Since (2) fails, M-rank$(v) > 0$; hence $M \vDash_v \sim\Box (\sim q \leftrightarrow p)$. But, by the choice of v, $M \vDash_v \Box (q \leftrightarrow p)$. Hence $M \vDash_v \sim p$, so that $M \vDash_v q$. Con-

versely, if $M \models_v q$, v is accessible from w and M-rank$(v) > 0$, then either $M \models_v \sim\Box(q \leftrightarrow p)$, so that $M \models_w \sim\Box(q \leftrightarrow p)$; or $M \models_v \Box(q \leftrightarrow p)$, so that $M \models_v \sim A$, whence $M \models_v \sim p$, which yields $M \models_v \sim q \leftrightarrow p$, and thus again $M \models_w \sim\Box(q \leftrightarrow p)$. In sum, if neither (1) nor (2) holds, then $M \models_w p$ iff there exists a v accessible from w of M-rank > 0 at which q is true. We thus obtain the disjunction $\Box\bot \vee \Diamond(\sim q \cdot \Box\bot) \vee \Diamond(q \cdot \Diamond\top)$ as the fixed point H.

2 Two formulas A and B will be jointly satisfiable only if they do not make conflicting demands on their common vocabulary, that is, on the sentence letters they share. Conversely, one might expect that if A and B can be satisfied separately in ways that treat their common vocabulary suitably alike, then A and B should be jointly satisfiable. The theorem below shows this to be the case; in it, n-S-characters are used to make precise the requisite suitable likeness. The theorem yields as corollaries effective versions of the Interpolation Theorem and Implicit Definability Theorem for GL. For any formula C let $\nu(C)$ be the number of boxed subformulas of C.

Joint Satisfiability Theorem *Let A and B be formulas of* GL, *S the set of sentence letters common to A and B, and $n = \nu(A) + \nu(B)$. If there exists an n-S-character K such that $A \cdot K$ and $B \cdot K$ are each satisfiable, then $A \cdot B$ is satisfiable.*

Proof: For $C = A$ or $C = B$ and x a world in a model, let $\boldsymbol{p}(C,x)$, the *C-profile of x*, be the conjunction of all sentence letters of C, negations of sentence letters of C, boxed subformulas of C, and negations of boxed subformulas of C that are true at x. Note that if two worlds have the same C-profile then they agree on all subformulas of C. Let $\pi(C,x)$ be the number of boxed subformulas of C that are false at x.

Lemma *Let u and v be worlds in models M and N, respectively, that have the same k-S-character, where $k = \pi(A,u) + \pi(B,v)$. Then $\boldsymbol{p}(A,u) \cdot \boldsymbol{p}(B,v)$ is satisfiable.*

Proof: By induction on k. If $k = 0$, $\pi(A,u)$ and $\pi(B,v)$ are each 0. Thus, no conjunct of $\boldsymbol{p}(A,u) \cdot \boldsymbol{p}(B,v)$ is a negated necessitation. Since u and v have the same 0-S-character, they agree on all sentence letters in S. The following $M^* = \langle W,R,P\rangle$ is then a model for $\boldsymbol{p}(A,u) \cdot \boldsymbol{p}(B,v)$: W contains one world w; the accessibility relation R is null; and, for each sentence letter p, $P(w,p) = \top$ iff p is a conjunct of $\boldsymbol{p}(A,u) \cdot \boldsymbol{p}(B,v)$.

Let $k > 0$; suppose the lemma holds for all integers $< k$ and that $\pi(A,u) + \pi(B,v) = k$. Let $\Box D_1, \ldots, \Box D_j$ be the boxed subformulas of A that are false at u and let $\Box D_{j+1}, \ldots, \Box D_k$ be those of B that are false at v. For each i, $1 \leq i \leq k$, we show that there is a model M_i and a world w such that:

(a) $M_i \models_w \sim D_i$
(b) E is valid in M_i whenever $\Box E$ is a conjunct of $\boldsymbol{p}(A,u) \cdot \boldsymbol{p}(B,v)$.

To fix ideas, let $i \leq j$; the case $j < i \leq k$ is treated symmetrically. Let x be a world of M of lowest rank that is accessible from u and at which D_i is false. Then $\Box D_i$ is true at x. Since u and v have the same k-S-character, there exists a world

a world y of N accessible from v such that x and y have the same $(k-1)$-S-character. Since y is accessible from v, every boxed subformula true at v is true at y; hence $\pi(B,y) \leq \pi(B,v)$. Similarly, $\pi(A,x) \leq \pi(A,u)$; in fact, since $\Box D_i$ is true at x but false at u, $\pi(A,x) < \pi(A,u)$. Let $m = \pi(A,x) + \pi(B,y)$; thus $m < k$. Since x and y have the same $(k-1)$-S-character, they have the same m-S-character. By the induction hypothesis, $\boldsymbol{p}(A,x) \cdot \boldsymbol{p}(B,y)$ is satisfiable. Let M_i be a model and w a world such that $M_i \vDash_w \boldsymbol{p}(A,x) \cdot \boldsymbol{p}(B,y)$. By passing, if necessary, to the submodel of M_i generated by w, we may assume that every world of M_i except w is accessible from w.

Since $M_i \vDash_w \boldsymbol{p}(A,x) \cdot \boldsymbol{p}(B,y)$, w agrees with x on all subformulas of A, and agrees with y on all subformulas of B. In particular, D_i is false at w, which is property (a). If $\Box E$ is a conjunct of $\boldsymbol{p}(A,u)$ or $\boldsymbol{p}(B,v)$, then E and $\Box E$ are both true at x or at y, respectively. Thus E and $\Box E$ are both true at w. Since every world of M_i except w is accessible from w, E must be true at every world of M_i. That is, E is valid in M_i, which is property (b).

We may assume without loss of generality that the models $M_i = \langle W_i, R_i, P_i \rangle$ are disjoint. We shall obtain a model for $\boldsymbol{p}(A,u) \cdot \boldsymbol{p}(B,v)$ by amalgamating the M_i, and adding a new world at the top. That is, let z be an object not in any W_i, and let $M^* = \langle W, R, P \rangle$, where $W = W_1 \cup \ldots \cup W_k \cup \{z\}$, $R = R_1 \cup \ldots \cup R_k \cup \{\langle z,x \rangle | x \in W_1 \cup \ldots \cup W_k\}$, and, for each sentence letter p, $P(x,p) = P_i(x,p)$ for x in W_i, and $P(z,p) = \top$ iff p is a conjunct of $\boldsymbol{p}(A,u) \cdot \boldsymbol{p}(B,v)$.

By continuity, if $x \in W_i$, then anything true at x in M_i is true at x in M^*. We show that $M^* \vDash_z \boldsymbol{p}(A,u) \cdot \boldsymbol{p}(B,v)$. Let p be a sentence letter. If p is a conjunct of $\boldsymbol{p}(A,u) \cdot \boldsymbol{p}(B,v)$, then p is true at z by the definition of $P(z,p)$. If $\sim p$ is a conjunct of $\boldsymbol{p}(A,u) \cdot \boldsymbol{p}(B,v)$, then, by the definition of $P(z,p)$, p is false at z iff p is not a conjunct of $\boldsymbol{p}(A,u) \cdot \boldsymbol{p}(B,v)$. But were p and $\sim p$ both conjuncts of $\boldsymbol{p}(A,u) \cdot \boldsymbol{p}(B,v)$, p would be in S and u and v would differ on p. But this is impossible, since u and v have the same k-S-character. Let $\Box E$ be a subformula of A or B. If $\Box E$ is a conjunct of $\boldsymbol{p}(A,u) \cdot \boldsymbol{p}(B,v)$, then E is valid in each M_i, so that $M^* \vDash_x E$ for each x in each W_i. Hence $\Box E$ is true at z. If $\sim \Box E$ is a conjunct of $\boldsymbol{p}(A,u) \cdot \boldsymbol{p}(B,v)$, then $\Box E$ is false at u or at v, so that $\Box E$ is $\Box D_i$ for some i, $1 \leq i \leq k$. Hence there exists a w in W_i such that $M^* \vDash_w \sim D_i$; hence $\sim \Box E$ is true at z.

The Joint Satisfiability Theorem follows quickly from the Lemma. Suppose K is an n-S-character, where $n = \nu(A) + \nu(B)$, and $A \cdot K$ and $B \cdot K$ are each satisfiable. Then there exist models M and N and worlds u and v such that $M \vDash_u A \cdot K$ and $N \vDash_v B \cdot K$. Let $k = \pi(A,u) + \pi(B,v)$. Then $n \geq k$, so that u and v have the same k-S-character. By the Lemma, $\boldsymbol{p}(A,u) \cdot \boldsymbol{p}(B,v)$ is satisfiable. But since A is true at u, $\boldsymbol{p}(A,u)$ truth-functionally implies A; similarly, $\boldsymbol{p}(B,v)$ truth-functionally implies B. Hence $A \cdot B$ is satisfiable.

Interpolation Lemma *Suppose* $\vdash A \rightarrow C$. *Then there is a formula* B *whose sentence letters are common to* A *and* C *such that* $\vdash A \rightarrow B$ *and* $\vdash B \rightarrow C$.

Proof: Let S be the set of sentence letters common to A and C, $n = \nu(A) + \nu(C)$, and B the disjunction of all n-S-characters K such that, for some model M and world w, $M \vDash_w A \cdot K$. Obviously $\vdash A \rightarrow B$, since by definition at any world at which A is true there is a disjunct of B that is true. Suppose it is not

the case that $\vdash B \to C$. Then $B \cdot \sim C$ is satisfiable, so $K \cdot \sim C$ is satisfiable for some disjunct K of B. Now, by definition, $A \cdot K$ is satisfiable. Hence, by the Joint Satisfiability Theorem, $A \cdot \sim C$ is satisfiable, contrary to hypothesis.

Beth Definability Theorem *Let* $A(p)$ be a formula of GL, *and let* $A(q)$ *result from* $A(p)$ *by replacing* p *by* q. *Suppose* $\vdash A(p) \cdot A(q) \to (p \leftrightarrow q)$. *Then there exists a formula* H *whose sentence letters are among those of* $A(p)$ *other than* p, *such that* $\vdash A(p) \to (p \leftrightarrow H)$.

Proof: Let S be the sentence letters of $A(p)$ other than p, $n = 2\nu(A(p))$, and H the disjunction of n-S-characters K such that, for some model M and world w, $M \vDash_w A(p) \cdot p \cdot K$. Then $\vdash A(p) \to (p \to H)$, since, by definition, in any world at which $A(p)$ and p are true some disjunct of H is true. Suppose it is not the case that $\vdash A(p) \to (H \to p)$. Thus $A(p) \cdot \sim p \cdot H$ is satisfiable. Hence $A(p) \cdot \sim p \cdot K$ is satisfiable for some disjunct K of H. Let q be a sentence letter foreign to $A(p)$; by substitution for p, $A(q) \cdot \sim q \cdot K$ is satisfiable (recall that K does not contain p). Now, by the definition of H, $A(p) \cdot p \cdot K$ is satisfiable. Since K is an n-S-character, S is the set of sentence letters common to $A(p) \cdot p$ and $A(q) \cdot \sim q$, and $n = 2\nu(A(p)) = \nu(A(p) \cdot p) + \nu(A(q) \cdot \sim q)$, the Joint Satisfiability Theorem applies. We may conclude that $A(p) \cdot p \cdot A(q) \cdot \sim q$ is satisfiable. But this contradicts the hypothesis that $\vdash A(p) \cdot A(q) \to (p \leftrightarrow q)$.

Note The Beth Definability Theorem yields the Fixed Point Theorem as a fairly direct consequence (see [1], p. 173, or [3], p. 110), although this was not noticed—nor was the Beth Theorem proved for GL—until several years after de Jongh and Sambin first proved the Fixed Point Theorem. (The Beth Theorem and Interpolation Lemma for GL were originally proved, independently, by Boolos and Smoryński.) We prefer the direct proof of the Fixed Point Theorem, since it is clearer, and it provides the best possible bound on the number of nested $\Box$'s mentioned at the end of Section 1.

3 In this section, we investigate the complexity of models for n-S-characters, as measured by the rank of worlds at which n-S-characters can be made true. Now the usual proof of the modal completeness theorem shows that if F is satisfiable then there exist M and w with $M \vDash_w F$, and M-rank(w) at most the number of boxed subformulas of F; but that result is far too crude when applied to n-S-characters.

If n or S is extremely small, bounds on rank are easily obtained. Any 0-S-character can be satisfied by some M and w with w the only world of M, and hence M-rank$(w) = 0$. A 1-S-character is a conjunction of a 0-S-character and formulas $\pm\Diamond C$, where C is a 0-S-character; any 1-S-character can be satisfied by some M and w with M-rank$(w) \leq 1$. If S is empty, then the satisfiable n-S-characters are equivalent to one of the formulas $\sim\Box^k\bot \cdot \Box^{k+1}\bot$, for $0 \leq k < n$, and $\sim\Box^n\bot$. Now $M \vDash_w \sim\Box^k\bot \cdot \Box^{k+1}\bot$ iff M-rank$(w) = k$, and $M \vDash_w \sim\Box^n\bot$ iff M-rank$(w) \geq n$. Thus, for any n, every satisfiable n-$\varnothing$-character can be satisfied by some M and w with M-rank$(w) \leq n$.

These simple cases might make it seem plausible that any satisfiable n-S-character can be satisfied by some M and w with M-rank$(w) \leq n$. However, such a claim is false. A simple counterexample can be constructed with $n = 2$

and S containing one sentence letter, say q. Consider any satisfiable 2-$\{q\}$-character K containing the conjuncts $\Diamond(q \cdot \Diamond q \cdot \Diamond \sim q)$ and $\sim\Diamond(\sim q \cdot \sim\Diamond q \cdot \sim\Diamond \sim q)$. If $M \models_w K$, then the truth of the former conjunct at w requires the existence of a world v accessible from a world u accessible from w such that $M \models_v \sim q$. The truth of the latter conjunct at w requires that for all x accessible from w of M-rank 0, $M \models_x q$. Hence M-rank$(v) > 0$, so that M-rank$(w) > 2$. An example of a satisfiable 2-$\{q\}$-character with these conjuncts is the 2-$\{q\}$-character of the world w_3 of the following model $\langle W, R, P\rangle$: $W = \{w_0, w_1, w_2, w_3\}$, $w_j R w_i$ iff $i < j$, and P so defined that q is true at w_0, w_2, w_3 but not at w_1.

(This example shows Smoryński ([3], p. 122) to be mistaken in his account of a Boolos-style computation of fixed points. Given a formula A modalized in p, containing n boxed subformulas, and with S the set of sentence letters of A aside from p, Smoryński proceeds thus: construct all models M in which $p \leftrightarrow A$ is valid and in which all worlds have M-rank $\leq n$; then disjoin the n-S-characters of all worlds of such models at which p is true. This procedure can fail to yield correct results. For suppose A contains two boxed subformulas, $S = \{q\}$, and A is tautologous. Obviously, any fixed point H must be equivalent to $\top$; yet, as the above example shows, Smoryński's method will yield a disjunction of 2-$\{q\}$-characters that omits some satisfiable ones, and hence will not be equivalent to $\top$.)

More generally, there are satisfiable 2-S-characters K that cannot be made true at worlds of rank $\leq 2^s$, where $s = |S|$. Let $S = \{q_1, \ldots, q_s\}$, and let C_0, $\ldots, C_m$ be the different conjunctions $\pm q_1 \cdot \pm q_2 \cdot \ldots \cdot \pm q_s$, so that $m = 2^s - 1$. Let $M = \langle W, R, P\rangle$, where $W = \{u_0, \ldots, u_{m+2}\}$, $u_i R u_j$ iff $j < i$, and P is such that $M \models_{u_i} C_i$ for $0 \leq i \leq m$, and $M \models_{u_i} C_0$ for $i > m$. Let K be the 2-S-character of u_{m+2}. Let $N \models_w K$; we show N-rank$(w) \geq m + 2$. Let x be a world of N accessible from w whose 1-S-character is the same as that of u_{m+1}. That 1-S-character includes $\Diamond C_m$ as a conjunct; hence there exists a world z of N accessible from x such that $N \models_z C_m$. Note that N-rank$(w) \geq N$-rank$(z) + 2$. Thus it suffices to show that, for all v accessible from w and each i, $0 \leq i \leq m$, if $N \models_v C_i$ then N-rank$(v) \geq i$. This is trivial for $i = 0$. Suppose it is true for i, let v be accessible from w and $N \models_v C_{i+1}$. Since w and u_{m+2} share the same 2-S-character, there is a world u_j of M with the same 1-S-character as v. Clearly $j = i + 1$, since u_{i+1} is the only world of M at which C_{i+1} is true. Now the 1-S-character of u_{i+1} contains $\Diamond C_i$ as a conjunct, since C_i is true at u_i and $u_{i+1} R u_i$. Hence $N \models_v \Diamond C_i$, so there exists a v' in N accessible from v such that $N \models_{v'} C_i$. By the induction hypothesis, N-rank$(v') \geq i$; hence N-rank$(v) \geq i + 1$, and the claim follows by induction.

The example just given is best-possible: any satisfiable 2-S-character can be satisfied by some M and w with M-rank$(w) \leq 2^s + 1$. That bound is implied by the following general result.

Theorem *Let S be a finite set of sentence letters, and let $n \geq 2$. If K is a satisfiable n-S-character, then there exist M and w such that $M \models_w K$ and M-rank$(w) \leq \alpha + 1$, where α is the number of satisfiable $(n-2)$-S-characters.*

Proof: A model $M = \langle W, R, P\rangle$ is a tree iff, for every $x \in W$, the set of v such that vRx is linearly ordered by R. Any model can be transformed into a tree,

with the same maximal rank (see [1], p. 106, or [3], p. 108), so there is no loss of generality in restricting attention to tree-models.

Suppose $M \vDash_w K$, where $M = \langle W, R, P\rangle$ is a tree and w has rank $> \alpha + 1$. For each world u, let $\varphi(u) = \{J \mid J \text{ an } (n-2)\text{-}S\text{-character such that } M \vDash_u \Diamond J\}$. Note that if uRv then $\varphi(v) \subseteq \varphi(u)$. By the pigeonhole principle, there exist worlds x and y such that wRx and xRy, x has rank at most $\alpha + 1$, and $\varphi(x) = \varphi(y)$. Let $Z = \{u \mid xRu \text{ and not } yRu\}$. Let $W' = W - Z$, R' and P' restrictions of R and P to W', and $M' = \langle W', R', P'\rangle$.

We show first, by induction on M-rank, that every v in W' has the same $(n-1)$-S-character in M' as in M. Suppose this is true for all worlds of W' of lower M-rank than v.

Case 1: $v \neq x$ and not vRx. Then, since M is a tree, in M no world in Z is accessible from v. Hence the submodel of M generated by v is identical to the submodel of M' generated by v, so that v in M and v in M' agree on all formulas.

Case 2: $v = x$ or vRx. It suffices to show, for every $(n-2)$-S-character J, that $M \vDash_v \Diamond J$ iff $M' \vDash_v \Diamond J$. Suppose $M \vDash_v \Diamond J$. Then there exists a z in W such that vRz and $M \vDash_z J$. If $z \in W'$ then, by the induction hypothesis, since z has lesser rank than v, z has the same $(n-1)$-S-character in M and M'; hence it has the same (n − 2)-S-character in M and M', whence $M' \vDash_z J$, so that $M' \vDash_v \Diamond J$. If $z \in Z$, by the choice of x and y there exists a z' such that yRz' and $M \vDash_{z'} J$; z' is then in W' and has lower M-rank than v, and, as in the previous case, we obtain $M' \vDash_v \Diamond J$. Conversely, suppose $M' \vDash_v \Diamond J$. Then there exists a z in W' accessible from v with $M' \vDash_z J$, and z has lower rank than v. By the induction hypothesis, z has the same $(n-1)$-S-character in M and in M'; hence it has the same $(n-2)$-S-character in the two models. Hence $M \vDash_z J$, whence $M \vDash_v \Diamond J$.

Now we should like to show that the world w has the same n-S-character in M and in M'; but this may not be the case. Hence we amplify M' by adjoining a copy of what was in M accessible from x. That is, let $X = \{u \in W \mid xRu\}$, let R'' and P'' be the restrictions of R and P to X, and let $\langle X^*, R^*, P^*\rangle$ be an isomorphic copy of $\langle X, R'', P''\rangle$ such that X^* and W' are disjoint. Finally, let $M^\dagger = \langle W' \cup X^*, R^\dagger, P' \cup P^*\rangle$, where $R^\dagger = R' \cup R^* \cup \{\langle w, v\rangle \mid v \in X^*\}$. By continuity, if $v \in W'$ and v is accessible from w, then v has the same $(n-1)$-S-character in M' and in $M^\dagger$, hence the same $(n-1)$-S-character in M and in $M^\dagger$. Also by continuity, if $v \in X^*$, then v has the same $(n-1)$-S-character in $M^\dagger$ as its isomorphic image has in M. It follows that, for all $(n-1)$-S-characters J, $M \vDash_w \Diamond J$ iff $M^\dagger \vDash_w \Diamond J$, whence w has the same n-S-character in M and in $M^\dagger$.

Thus, in $M^\dagger$, w has n-S-character K. Let an M- or $M^\dagger$-path be any sequence $\langle v_1, \ldots, v_k\rangle$ of worlds of M or $M^\dagger$ such that each v_{i+1} is accessible from v_i. We claim that the number of $M^\dagger$-paths that begin with w and have length $> \alpha + 2$ is less than the number of such M-paths. For the maximal $M^\dagger$-rank of worlds in X^* is α, so any $M^\dagger$-path beginning with w and of length $> \alpha + 2$ must be composed of worlds in W', and hence is an M-path as well; while clearly some M-paths beginning with w and of length $> \alpha + 2$ are not $M^\dagger$-paths, since they contain elements of Z. It follows that, by iterating the process by which $M^\dagger$ is constructed from M, we eventually obtain a model in which w has n-S-character K and has rank at most $\alpha + 1$.

REFERENCES

[1] Boolos, G., *The Unprovability of Consistency*, Cambridge University Press, Cambridge, 1979.

[2] Fine, K., "Logics containing K4. Part I," *The Journal of Symbolic Logic*, vol. 39 (1974), pp. 31–42.

[3] Smoryński, C., *Self-reference and Modal Logic*, Springer-Verlag, New York, Berlin, Heidelberg, Tokyo, 1985.

Department of Philosophy
Harvard University
Cambridge, Massachusetts 02138