**Title**
Three Essays on the Foundations of Science



**Author**
Johansson, Rolf Henry

**Publication Date**
2014

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA,
IRVINE

# THREE ESSAYS ON THE FOUNDATIONS OF SCIENCE

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Social Science – Mathematical Behavioral Sciences

by

Rolf Henry Johansson

Dissertation Committee:
Professor Louis Narens, Chair
Professor Donald Saari
Professor Kent Johnson

2014

# DEDICATION

For my mother Lalla and my father Kurt for their continuing support.

# TABLE OF CONTENTS

# ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my committee chair, Louis Narens, for many stimulating conversations over the years and for guidance throughout my graduate education. Louis' ability to weave together sophisticated mathematical, philosophical, and scientific knowledge is without compare, and I consider myself extremely fortunate to have worked so closely with such a creative and profound thinker.

In addition, I would like to thank Duncan Luce, Jean-Claude Falmagne, and my other committee members Brian Skyrms and Don Saari. Both Duncan and Jean-Claude were of great help in shaping the research in ESSAY 1, and Duncan in particular gave the essay a very close reading and provided many helpful comments and criticisms. My other committee members Brian Skyrms and Don Saari also provided helpful comments on other parts of the research as it was developing.

Financial support was provided by the University of California, Irvine, in the form of three Summer Research Fellowships and a Regents Dissertation Writing Fellowship.

# CURRICULUM VITAE

ROLF JOHANSSON

## Education

- University of California, Irvine
  - Ph.D. in Social Science – Mathematical Behavioral Sciences 2014
  - M.A. in Social Science – Mathematical Behavioral Sciences 2006
- Columbia University
  - Ph.D. Program in Philosophy 1994-95
- M.I.T.
  - Ph.D. Program in Linguistics and Philosophy 1986-88
- University of Pennsylvania
  - A.B. in Philosophy 1986

## Areas of Specialization

- Philosophy of Language
- Philosophy of Religion
- Philosophy of Science

## Areas of Competence

- Mathematical Logic
- Decision Theory
- Metaphysics
- History of Analytic Philosophy

## Working Papers

1. "Troubles with Convention T"
2. "On Formulas for Primes"
3. "Measurement of Finite Semiorders"
4. "Ross' Principle of Possible Explanation"
5. "The Cosmological Argument and Russell's Paradox"
6. "Chomsky on the Indeterminacy of Translation"

7. "Common Mistakes of Skepticism"
8. "Natural Languages are *Not* Vast"
9. "Sorites, Semiorders, and Individuation"
10. "Reduction, Ontology, Information"
11. "The Scotus-Ross Ontological Argument"

Talks

- "Troubles with Deflationism," *Central Valley Philosophy Conference, University of California, Merced*, (October, 2008)
- "Qualitative Semiorders and the Empirical Adequacy of Linear Representations." *INFORMS Conference*, Session on Consumer's Preference Modeling, San Jose, CA (November 17, 2002)

Courses Taught

*University of California, Merced*

Upper Division:

- PHIL 105: *Philosophy of Language*: Spring 2010, Fall 2011
- PHIL 101: *Metaphysics*: Fall 2007, Spring 2009, Spring 2012
- PHIL 160/MATH 171: *Mathematical Logic*: Spring 2008, Fall 2009, Spring 2013
- PHIL 107: *Philosophy of Religion*: Spring 2007, Fall 2008, Spring 2011, Fall 2012
- PHIL 195: *Undergraduate Research*: Spring 2009, Fall 2009, Fall 2011
- PHIL 199: *Individual Study*: Fall 2007, Spring 2008, Spring 2010

Lower Division:

- PHIL 001: *Introduction to Philosophy*: Fall 2006, Fall 2007, Fall 2008, Fall 2009, Fall 2010, Fall 2011, Fall 2012
- PHIL 005: *Introduction to Logic*: Spring 2007, Spring 2008, Spring 2009, Spring 2010 Spring 2011, Spring 2012, Spring 2013

Independent Studies Supervised:

- *Teleology in Systems Theory* (Fall 2011)
- *Ontology and Ontological Commitment* (Spring 2010)
- *Theories of Immortality* (Fall 2009)
- *Philosophy of Quantum Mechanics* (Spring 2009)

- *Set Theory* (Spring 2008)
- *Gödel's Ontological Argument* (Fall 2007)

Other Experience

Teaching Assistant, University of California, Irvine

*Problems of Philosophy, Inductive Logic, Decision Making, Probability and Statistics II, Probability and Statistics III, Basic Microeconomics, Intermediate Microeconomics, Fundamentals of Psychology, Social Psychology, Cultural Anthropology*

Teaching Assistant, Massachusetts Institute of Technology

*Introduction to Ethics* (2 times)

Research Assistant, University of Pennsylvania, Professor Paul Guyer

Honors and Awards

- Distinguished Undergraduate Teaching Award, $1,000, UC Merced (2012-2013)
- Regents Dissertation Writing Fellowship, $4,600, U.C. Irvine
- Dean's Commendations for Excellence in Teaching, U.C. Irvine (3 times)
- Summer Research Fellowships, $3,400, U.C. Irvine (3 times)
- Institute Fellowship, $8,800, Massachusetts Institute of Technology

Service to the Profession

- WASC Accreditation Reports (Spring 2010 and Spring 2011)
- Undergraduate Philosophy Club (Sponsor)

Professional Affiliations

- American Philosophical Association
- Philosophy of Science Association

# ABSTRACT OF THE DISSERTATION

THREE ESSAYS ON THE FOUNDATIONS OF SCIENCE

By

Rolf Henry Johansson

Doctor of Philosophy in Social Science – Mathematical Behavioral Sciences

University of California, Irvine, 2014

Professor Louis Narens, Chair

The general uniqueness problem for finite semiorders is still unsolved, and this has impeded their wider use in decision modeling. In Essay 1, I show that for semiorders that satisfy some relatively weak constraints, unique representation (and hence interval scalability) may be obtained.

In Essay 2, I discuss a type of counterexample to Tarski's Convention T that was originally discovered by Hintikka. I show that Hintikka's counterexample generalizes in quite unexpected ways, and that there are in fact a large number of unambiguous sentences that generate counterexamples of the same general type. I then show that various proposals for dealing with Hintikka's original counterexample are unsatisfactory, and that none of the proposed solutions can resolve all of the counterexamples presented in this essay.

In Essay 3 I present elementary formulas for the $n^{th}$ prime and for the number of primes up to a given limit, both of which improve upon existing formulas by avoiding the computation of factorials and the exponential growth of terms. The formulas are based on the idea of "embedding" characteristic functions – a characteristic function for non-divisibility is used to construct a characteristic function for primality, and no use is made of either Wilson's theorem or the inclusion-exclusion process.

# ABSTRACT OF ESSAY 1

THE UNIQUENESS PROBLEM FOR FINITE SEMIORDERS

By

Rolf Henry Johansson

Doctor of Philosophy in Social Science – Mathematical Behavioral Sciences

University of California, Irvine, 2014

Professor Louis Narens, Chair

Semiorders were introduced by Luce (1956) to account for the intransitivities found empirically in indifference judgments. In principle, they are superior to weak orders as descriptive models of choice behavior. However, the general uniqueness problem for finite semiorders is still unsolved, and this has impeded their wider use in decision modeling. First, we will discuss representational anomalies for semiorders in order to better understand the source of the difficulty in obtaining uniqueness. We will then show that for semiorders that satisfy some relatively weak constraints, unique representation (and hence interval scalability) may be obtained. This result follows by combining independent results of Suppes (1972) on equal-difference structures and Fishburn (1973b) on the construction of weak orders from fragmentary information. The finite semiorders for which unique representation may be obtained are "well-behaved" in the sense that they constitute partial information about an "underlying" equal-difference structure, and have a constant discrimination threshold. A very weak constraint on the size of the threshold enables the unique representation. Since most applications of utility models involve the comparison of alternatives within a limited range of utility values, over which discrimination thresholds are more or less constant, then well-behaved semiorders may have practical applications for qualitative modeling in such cases.

## 1. INTRODUCTION

In the classical formulation of expected utility theory by von Neumann and Morgenstern (1944), strong assumptions were made about the structure of preference and indifference. It was assumed that preferences induce a linear ordering of a set of goods, and that transitivity holds for both preference and indifference. These assumptions were also made by Marschak (1950), Debreu (1954), and even by Savage (1954) in his development of subjective expected utility (*SEU*). It is fair to assume that all of these theorists were well aware that the assumptions of transitivity and linear ordering did not necessarily hold for actual human decision making, but rather were meant to hold for *ideally rational* agents. Yet it still seems to have been believed that when assumed to hold empirically, these idealizations are relatively innocuous. It is, no doubt, partly because of this belief that the idealizations have now become standard assumptions. Although *SEU* and its variants are still best thought of as *normative* models of behavior, they are widely considered to be at least *approximately* correct as descriptive models.

In spite of their wide use, there has been mounting evidence that the assumptions of linear ordering and transitivity may not be so harmless after all. The many criticisms are now well-known, dealing with phenomena such as preference reversals, framing effects, violations of Savage's sure-thing principle, portfolio effects, and problems concerning the temporal resolution of uncertainty. Economists vary in their opinions about how important these effects are for economic theory.[1] Nevertheless, there seems to be a consensus that regardless of how idealized the standard assumptions may be, they are certainly beneficial from the standpoint of enabling us to derive workable mathematical representations. Another reason for their wide use is that there is no consensus that any other assumptions could easily take their place. Consequently, in most

[1] See Kreps (1990, p. 112-122) for a general discussion of the views of economists on this issue, and see the anthology edited by Hogarth and Reder (1986) for further details.

presentations of consumer behavior in economics textbooks today, one finds preference and indifference characterized as a weak order at the qualitative level, with a linearly ordered numerical utility representation.

Most of the criticisms of modeling choice behavior with weak and linear orders have focused on the problems that arise when using a transitive relation to model preference. But in both linear and weak orders, the symmetric relation – interpreted qualitatively in utility theory as *indifference* – is assumed to be transitive as well. Yet it was pointed out long ago by Armstrong (1939) and Georgescu-Roegen (1936, 1958) that the imperfect discrimination ability of humans leads to intransitivities of indifference in actual behavior. A person may be indifferent between goods *a* and *b*, and indifferent between goods *b* and *c*, but find that she prefers *a* to *c*. A similar intransitivity is observed in psychophysical contexts where perceptual discrimination is studied, and the symmetric relation in question is interpreted as *indiscriminability*. Judgments of comparative brightness, loudness, sweetness, etc. all give rise to intransitivities with respect to indiscriminability.

For simplicity, we will speak of all such symmetric complement relations as "indifference" relations, understanding that the specific interpretation of the relation may vary with context. One way of understanding the intransitivity of indifference is to think of it as arising from an underlying discrimination threshold, or just noticeable difference (jnd). This is defined as a distance that a pair of objects must be separated on the relevant attribute continuum (e.g., utility for goods, loudness for tones, etc.) in order for a person to discriminate between them.[2] In

[2] There is some equivocation in the literature between *thresholds*, which we think of as maximal indiscriminable differences, and the *jnd*, which we think of as the minimal discriminable difference. This equivocation is harmless in most contexts, so we will continue to use "threshold" to describe what is, strictly speaking, a jnd. The exact notion that is intended will be made clear in context. The reader should also note that actual subjects will not exhibit a precise threshold below which they *never* discriminate, and above which they *always* discriminate. Thresholds are normally determined probabilistically, as the interval above which the subject discriminates with, say, probability $P > 0.5$.

psychophysical cases, the threshold apparent in the data is due to the discrimination sensitivity of the relevant sensory system of the subject. With utility, thresholds arise because of higher order cognitive mechanisms. In most contexts they will arise because of the multiattribute nature of most decision making. The attributes that are seen as most salient for comparing goods *a* and *b* and goods *b* and *c* may not be the same attributes used for comparing *a* and *c*, and consequently intransitivities may result. Even though minimal differences in *price* may be discriminated unambiguously, the multiattribute nature of much decision making nevertheless yields intransitivities with respect to *utility*.

There are now a number of different orders that relax the transitivity assumption for indifference, and thus generalize weak orders. The first of these to be applied in the context of utility theory was a "semiorder," which was introduced by Luce (1956). Unlike some of the other orders that relax the transitivity assumption, semiorders enable numerical representations of discrimination thresholds, which more directly reflect the complications observed in actual behavior. At the *qualitative* level, they certainly provide a more descriptively adequate alternative to weak orders for modeling choice behavior.

Nevertheless, semiorders have not yet replaced weak orders in economics textbooks.[3] A partial explanation for this must certainly be the absence, thus far, of any adequate uniqueness theorem for the finite case. The lack of uniqueness allows several peculiar representation problems for semiorders that don't arise with weak orders. For example, Roberts and Franke (1976) showed that for a given semiorder, one may construct two separate numerical representations $f$ and $g$, where the two representations cannot be related by any transformation

[3] They have found particular applications, however. Examples of the use of semiorders by economists may be found in Jamison and Lau (1973, 1977), and in Vinke (1980). Although the earlier work by Armstrong (1939) and Georgescu-Roegen (1936, 1958) was prior to most semiorder research, it was motivated by the same problem of thresholds in utility discriminations.

$\phi$, and may even be of different scale types. In such cases, at least one of the representations must be, in their terminology, "irregular." These cases create problems for the scientist, since under the most widely used definition of meaningfulness, scientific claims are meaningful only if they can be shown to be invariant with respect to transformations of the employed measurement scale.[4] Without this invariance, any inferences or claims might turn out to be mere artifacts of the particular numerical representation chosen. For example, it is meaningful to infer "*a* is hotter than *b*" by simply looking at the numerical values on either a Fahrenheit or Celsius scale, since the relation "hotter than" is invariant under the affine transformations used to change Fahrenheit to Celsius and vice versa. Put differently, the empirical conclusions we draw using Fahrenheit can be translated via an algebraic transformation into different formulations of the same conclusions expressed in terms of the Celsius scale, and vice versa. If a statement were to *fail* the invariance constraint by changing its truth value depending on which numerical representation we were using for measurement, then the claim would be a mere artifact of the representation, and would not qualify as a meaningful scientific statement.

In the case of semiorders, the absence of a uniqueness theorem leaves open the possibility of exactly this kind of problem. Without a uniqueness theorem we have no guarantee that we can use numerical representations of semiorders to make meaningful scientific claims about choice behavior. And since most actual choices are made within finite sets of alternatives, what we would like to have is a uniqueness theorem for the finite case. More specifically, the type of invariance we would like to find for semiorders is uniqueness up to a positive *affine* transformation, and hence interval scalability.

[4] See Suppes and Zinnes (1963) for a basic discussion of meaningfulness, and Narens (2002) for an advanced, comprehensive treatment.

An additional representation problem was pointed out by Swistak (1980), who noted that many representations of semiorders have a "paradoxical" quality. His idea was that if one thinks of a semiorder as arising from an underlying linear order, as will generally be the case in empirical contexts, then many representations of the semiorder will not be consistent with additional information that may be obtained about the underlying order. Specifically, they may fail to preserve the underlying linearity of the semiorder.

In Section 3 we will discuss the problems raised in Roberts and Franke (1976) and Swistak (1980) in more detail. These representation problems for semiorders cast some light on the difficulties with obtaining uniqueness, and help explain the continued use of weak orders as the standard model. Suppes and Zinnes (1963, p. 34) commented that "The uniqueness problem for semiorders is complicated and appears to have no simple solution." The problem has had no improvement since, and the most recent comment on the problem that I am aware of is by Roberts (1989a, p. 28), who called the uniqueness problem "a difficult one ... [that] remains an open question." But even if the problem is not solved for the general case, it will be of interest to try to determine exactly what the difficulties are, since even a partial solution may be of interest for potential applications to decision problems. As we'll see below, a satisfactory solution *is* obtainable for a large class of finite semiorders.

In what follows, we will assume that the actual structure of preference and indifference for humans forms a qualitative semiorder, and we will consider what assumptions are needed to obtain a unique representation. We will show that independent results by Suppes (1972) and Fishburn (1973b) jointly entail uniqueness of representation for a very useful subclass of semiorders. Specifically, we will show that finite semiorders are uniquely representable and interval scalable provided that they satisfy two conditions: i) they are "well-behaved," in the

sense that they are derived from an "underlying" linear order of equally-spaced elements *and* have a constant discrimination threshold, and ii) the threshold is within a "reasonable" bound (more precisely, the threshold is no larger than ½ the length of the entire semiorder, where "length" denotes the span from the first to the last element of the semiorder, and "½" denotes the median point of the semiorder). For semiorders satisfying these conditions, there exists a numerical representation in an arithmetical progression of integers that is unique up to a positive affine transformation. As a corollary, there exists a representation in a convex set of integers. For these representations, the "irregular" and "paradoxical" cases mentioned above do not arise.

Hence, uniqueness is obtainable provided that we idealize the semiorder itself somewhat. Of course, this does not solve the uniqueness problem in general, but what is of interest is that the idealization is very weak – *far* weaker than the standard idealization of transitivity discussed earlier – and it is satisfied in many applications of utility models to decision problems. The equal spacing assumption does not affect the generality of the result, since such structures may either be chosen or approximated by the construction of standard sequences. The constant threshold assumption, in turn, is descriptive in most contexts, since most applications of utility models involve the comparison of alternatives that are within a restricted range of utility values, and over this limited range discrimination thresholds are more or less constant. Finally, as will be shown below, the size constraint on the threshold is so weak that it is difficult to imagine any cases that would fail to satisfy it. Hence, although this paper focuses on the foundations of utility from the standpoint of measurement theory, it suggests one route by which greater descriptive adequacy may be obtained over a wide range of decision problems.

## 2. PRELIMINARIES

We will now present most of the definitions that will be used in the remainder of the paper. The reader may wish to skip to the next section and refer back as needed.

DEFINITION 1: A *weak order* is any structure $\mathsf{S} = \langle A, \prec_w, \sim_w \rangle$ where $A$ is a set, and both $\prec_w$ and $\sim_w$ are binary relations on $A$ that satisfy the following (for all $a, b, c \in A$):

Axiom W1: Exactly one of the following holds: $a \prec_w b$, $b \prec_w a$, or $a \sim_w b$.

Axiom W2: If $a \prec_w b$ and $b \prec_w c$, then $a \prec_w c$.

Axiom W3: $\sim_w$ is an equivalence relation

In utility theory, $\prec_w$ is interpreted as preference and $\sim_w$ as indifference. It is immediate from the axioms that $a \sim_w b$ *iff* both $a \not\prec_w b$ and $b \not\prec_w a$, and consequently that $a \not\prec_w a$ (since $a \sim_w a$). It is also immediate that $\sim_w$ is transitive. Any weak order where the relation $\sim_w$ is the identity relation is called a *linear order* (or equivalently, a *total order*). We will use "linear" and "total" interchangeably. The essential difference between weak orders and linear orders is that in a weak order that is not also a linear order, we may have $a \sim_w b$ for *distinct* $a$ and $b$. For the sake of the occasional reference below, we may also define a *partial order* as any set ordered by an irreflexive, transitive relation.[5]

The following axioms for semiorders were presented in Scott and Suppes (1958), and are a slight modification of Luce's original axioms:[6]

[5] The essential property of all partial orders is transitivity. Different authors also require them to be either *irreflexive*, or both *reflexive* and *antisymmetric*, depending on the application. The former are sometimes also called *strict partial orders* and the latter *weak partial orders*.

[6] Luce (1956) used an additional primitive for the indifference relation, but Scott and Suppes introduced this relation by definition. We follow the latter method, which affords a slight streamlining of the axioms.

DEFINITION 2: A *semiorder* is any structure $\mathsf{S} = \langle A, \prec \rangle$, where $A$ is a set and $\prec$ is a binary relation on $A$, that satisfies the following three axioms (for all $a, b, c, d \in A$):

Axiom S1. $a \not\prec a$.

Axiom S2. If $a \prec b$ and $c \prec d$, then $a \prec d$ or $c \prec b$.

Axiom S3. If $a \prec b$ and $b \prec c$, then $a \prec d$ or $d \prec c$.

If we interpret "$\prec$" as meaning "is discriminated as lower on the attribute continuum than," then axioms S2 and S3 prevent a discriminated pair from being "captured" by a non-discriminated pair, as the reader may easily verify. A generalization of semiorders may be obtained by deleting S3, and the resulting structure is called an *interval order*. We may define the symmetric complement of $\prec$ as follows:

DEFINITION 3: $a \sim b$ *iff* $a \not\prec b$ and $b \not\prec a$.

We will call any pair $\{a, b\} \subset A$ such that $a \sim b$ an *incomparable pair*. If either $a \prec b$ or $b \prec a$, then we will sometimes write $(a, b)$ or $(b, a)$, respectively. Since semiorders implicitly capture the notion of a discrimination threshold, we will introduce this notion explicitly for clarity:

DEFINITION 4: For any semiorder $\mathsf{S} = \langle A, \prec \rangle$, and for any $a_i \in A$, let $j = \min\{k: a_i \prec a_k\}$. Then we will call $\delta_i = j - i$ the *size of the discrimination threshold* at $a_i$.

Since we will be focusing on semiorders with constant thresholds, we will usually drop the subscript on the threshold and write $\delta$ rather than $\delta_i$. The reader should also note that we have

defined thresholds so that the smallest possible threshold, or "perfect discrimination," is defined as $\delta = 1$ rather than $\delta = 0$. This affords some simplification below.

Semiorders, weak orders, and linear orders are easily seen to be special kinds of partial orders. To clarify the interrelation between these various orders we state the following inclusions, understood as holding between whole classes of orders:

$$\text{linear orders} \subset \text{weak orders} \subset \text{semiorders} \subset \text{partial orders.}$$

In a semiorder, the symmetric relation $\sim$ is not necessarily transitive, in a weak order it *is* transitive, and in a linear order it is the identity relation. Hence weak orders generalize linear orders, semiorders generalize weak orders, and partial orders form the most general class.

DEFINITION 5: $\mathsf{S}_L = \langle A_L, \prec_L \rangle$ is a *linear extension* of a semiorder $\mathsf{S}$

*iff*

(i) $\mathsf{S}_L$ is a linear order, and

(ii) $(a, b) \in \mathsf{S} \Rightarrow (a, b) \in \mathsf{S}_L$ (for all $a, b \in A$).

CONVENTION 1: (i) $\mathbf{Z}_k^+$ is the first $k$ positive integers $\{1, 2, 3, \ldots, k\}$ in the usual ordering.

(ii) $|A|$ is the cardinality of the set $A$.

This is merely a notational convention. Since we will be working only with finite sets of qualitative objects, we will also use the following:

CONVENTION 2: Every set $A$ is indexed by $\mathbf{Z}_{|A|}^+$.

By this convention, the index set will always be exactly the cardinality of the set indexed.

DEFINITION 6: For all semiorders $S = \langle A, \prec \rangle$, *the relation* $\prec_*$ *induced on* $A$ *by* $\prec$ is defined as follows: For all $a_i, a_j \in A$ (where $i \neq j$),

$$a_i \prec_* a_j$$

*iff*

1. $a_i \prec a_j$

or 2. $a_i \sim a_j$ and $[(a_i \prec a_k$ and $a_j \sim a_k)$ for some $a_k \in A]$

or 3. $a_i \sim a_j$ and $[(a_m \prec a_j$ and $a_m \sim a_i)$ for some $a_m \in A]$.

If neither $a_i \prec_* a_j$ nor $a_j \prec_* a_i$, then we will write $a_i \sim_* a_j$. It can easily be shown from the definition that if $\prec$ satisfies trichotomy, then so does $\prec_*$.

DEFINITION 6 was first introduced by Luce (1956), where he also proved that the induced relation $\prec_*$ forms a weak order on $A$.[7] In many cases, a semiorder relation $\prec$ will *only* induce a weak order through DEFINITION 6. But whether it also induces a linear order will depend on how much information is present in the semiorder, as will be seen in the next section.

[7] Any pairs where $a_i \sim a_j$ and $a_i \prec_* a_j$ are what Fishburn (1973b), speaking in terms of utility, referred to as cases where $a_j$ is "slightly preferred to" $a_i$. For the interested reader, DEFINITION 6 without line 1 is equivalent to Property (i) of Swistak (1980, p. 126).

DEFINITION 7: For any relation $R$, where $I$ is the indifference relation for $R$ similar to that in DEFINITIONS 3 and 6 above (i.e., $aIb$ *iff* *not* $aRb$ and *not* $bRa$):

$$aJ_Rb$$

*iff*

(i) $aRb$, and

(ii) For all $c \in A$, if $aRc$ , then $bRc$ or $bIc$.

In words, "$aJ_Rb$" means that $a$ immediately precedes $b$ with respect to the relation $R$.

DEFINITION 8: A semiorder $\mathsf{S} = \langle A, \prec \rangle$ is said to be *well-behaved* if it has a finite, equal-spaced linear extension and a constant threshold. More precisely, a well-behaved semiorder satisfies the following (for all $a, b, c, d \in A, A_{UL}$, and all $i, j, k, r \in \mathbf{Z}^{+}_{|A|}$):

(i) There exists a finite linear extension $\mathsf{S}_{UL} = \langle A, \prec_{UL} \rangle$ of the semiorder $\mathsf{S} = \langle A, \prec \rangle$.

(ii) If $a_i J \prec_{UL} b_j$ and $c_k J \prec_{UL} d_r$, then $j - i = r - k$.

(iii) For all $a_i, a_j \in A$, $\delta_i = \delta_j$ (i.e., the discrimination threshold defined in DEFINITION 4 is constant for all $a \in A$).

Intuitively, a well-behaved semiorder is a semiordering, with constant threshold, of the elements of an equal-spaced linear order. Thus, well-behaved semiorders may be considered idealizations of a variety of emprical contexts where a linearly ordered set is presented to a human subject who is unable to discriminate every pair in the set with respect to intensity, utility, etc. Consequently, some of the ordered pairs of the underlying linear order may be missing from the data obtained from the subject, and the data will form a semiorder. DEFINITION 6 gives us a

criterion for recovering some of these non-discriminated pairs, and hence for extending the semiorder.

In the second clause of DEFINITION 8, we import the intuitive notion of "equal spacing" to semiorders, but this requires some explanation. Equal-spaced structures play an important role in measurement theory, where they are sometimes called "equal-difference" structures. These are discussed in Suppes (1957, 1972), Scott and Suppes (1956), and Suppes and Zinnes (1963). Suppes (1972, p. 45) pointed out that "Finiteness and equal spacing are characteristic properties of many standard scales, for example, the ordinary ruler, the set of standard weights used with an equal-arm balance in the laboratory or shop, or almost any of the familiar gauges for measuring pressure, temperature, or volume." Of course, there is nothing qualitatively inherent in pressure, length, etc. that brings about the equal spacing. This is something that the scientist imposes on the measurement scale in order to simplify measurement. In these cases, the precision in measurement is limited by the constant differences in millimeters, the differences in weight between the standard blocks, the differences between marks on a gauge, etc. Nevertheless, the equal-spacing assumption is made without loss of generality, since we may either select stimuli so that they are equally spaced with respect to the relevant relation, or we may take arbitrary stimuli that are not in the equal-spaced set and place them in intervals bounded by adjacent stimuli in the set. Thus, by decreasing the spacing between a standard equally-spaced set, any arbitrary stimulus may be measured within any desired degree of accuracy. This is essentially the idea behind building "standard sequences" of equally-spaced elements, which is discussed in Davidson et al. (1957), Luce (1967), and Krantz et al. (1971).

It is known that equal-difference structures are uniquely representable and interval scalable (see Suppes (1972) for proofs). However, it only makes sense to define equal spacing of

elements with respect to some relation, and for this to be possible over an entire order, all elements in the order must be comparable under the relation. This condition is satisfied by linear orders, but not necessarily by a semiorder. Since some elements of a semiorder may not be comparable with respect to the relation $\prec$, it doesn't make sense to think of them as "equally-spaced" with respect to this relation. The same elements may, however, be equally-spaced with respect to a different relation, and in the above case we specify this through what we refer to as the "underlying" linear ordering by $\prec_{UL}$. This linear order is simply an extension of the semiorder, but we refer to it as "underlying" to emphasize the empirical situations in which the semiorder is actually derived from the linear order in the manner discussed above. As we will see below, well-behaved semiorders have nice properties that make them useful for studying measurement, like their linear equal-difference counterparts.

CONVENTION 3: The indices of any well-behaved semiorder are chosen to agree with the linear extension $\mathsf{S}_{UL} = \langle A, \prec_{UL} \rangle$. That is, if $i$ is the index of $a \in \mathsf{S}_{UL}$, and the element $a$ is also an element of the semiorder $\mathsf{S}$, then $i$ is the index of $a \in \mathsf{S}$. Moreover, for any set $A$ ordered by a relation $R$, $a_i R b_j$ *iff* $i < j$ (for all $a, b \in A$, and all $i, j \in \mathbf{Z}^{+}_{|A|}$). In words, the ordering of the indexing always agrees with the ordering of the relation.

This convention, like the two proceeding it, could be dropped without loss of generality, but its inclusion allows considerable notational simplification in the proofs below.

DEFINITION 9: For any two structures $\mathsf{S} = \langle A, R \rangle$, and $\mathsf{S}' = \langle A', R' \rangle$, where $A$, $A'$ are finite sets and $R$, $R'$ are $k$ - ary relations on $A$, $A'$ respectively, a function $h$: $A \longrightarrow A'$ is a *homomorphism* from $\mathsf{S}$ to $\mathsf{S}'$ *iff* for all elements $a_1, \ldots, a_k \in A$,

$$R(a_1, \ldots, a_k) \quad \Rightarrow \quad R'(h(a_1), \ldots, h(a_k)).$$

Alternatively, we say that $\mathsf{S}$ and $\mathsf{S}'$ are *homomorphic*, or that $h$ is an *embedding* of $\mathsf{S}$ in $\mathsf{S}'$.[8]

DEFINITION 10: For any semiorder $\mathsf{S} = \langle A, \prec \rangle$, an order preserving, real-valued homomorphism $f : A \xrightarrow{1\text{–}1} \mathbf{Re}$ will be called a *closed representation* if there is a nonnegative function $\delta$: $A \longrightarrow \mathbf{Re}$ such that for all $a_i, a_j \in A$,

$$a_i \prec a_j \quad \textit{iff} \quad f(a_i) + \delta(a_i) \leq f(a_j).^{9}$$

We will call any closed representation with $\delta(a_i) = t$ (where $t$ is a constant for all $a_i$) a *representation with constant threshold*. Although we define the mapping into **Re** to enable full generality of the definition, we will only be considering cases of finite equal difference structures, where $\delta$ need only have values in the natural numbers. Since the representations of interest are mappings into ordered sets of numbers, they are sometimes more specifically called *numerical* representations. A *representation theorem* shows the existence of a numerical representation. The first representation theorem for semiorders was proved in Scott and Suppes

[8] In the literature on measurement theory a stronger notion of homomorphism is sometimes used, where a biconditional is supposed rather than a conditional. We need only the weaker form to describe the case of linear extensions (DEFINITION 4). See Chang and Keisler (1973 p. 70) for a discussion of homomorphisms.

[9] Closed representations are usually defined with thresholds in mind, consequently a strict inequality is usually used. Since our theorems are simpler to state if we think in terms of a jnd, we will use a non-strict inequality (c.f. footnote 2).

(1958) for semiorders with finite domains.[10] Additional representation theorems may be found in Fishburn (1970) and Mirkin (1972) for the denumerable case and for interval orders, and in Fishburn (1973a) for sets of arbitrary cardinality. A comprehensive survey of these results may be found in Suppes, et al. (1989).

The following property enables us to separate representations that preserve the underlying linearity of the semiorder from those that don't.

DEFINITION 11: For any semiorder $\mathsf{S} = \langle A, \prec \rangle$, a representation $f$ is said to be *strongly monotonic* iff it satisfies the following (for all $a_i, a_j \in A$):

$$a_i \prec_* a_j \quad \textit{iff} \quad f(a_i) < f(a_j).$$

We call this *strong* monotonicity because the representation is not only preserving $\prec$, but is preserving the induced relation $\prec_*$ as well. The representations that fail to preserve strong monotonicity are precisely those that Swistak (1980) called "paradoxical."

Finally we present a definition of the set of elements discriminated from a given element:

DEFINITION 12: For any $a_i \in A$ that is "left-discriminated" from at least one other element (i.e., where there exists $a_k$ such that $a_i \prec a_k$), we will use a boldface "$\boldsymbol{a}_i^{\prec}$" to designate the set $\{a_k\colon\ a_i \prec a_k\}$ of all elements "right-discriminated" from $a_i$ (for all $k \in \mathbf{Z}^+_{|A|}$).

Notice that all sets defined by DEFINITION 12 are nonempty. In what follows, if $\boldsymbol{a}_i^{\prec}$ is defined by DEFINITION 12 then we will occasionally refer to $a_i$ as "the defining element for $\boldsymbol{a}_i^{\prec}$."

[10] There are many proofs of this result available. See Suppes and Zinnes (1963), Scott (1964), and Rabinovitch (1977) for three different approaches.

## 3. Representation Problems for Semiorders

### *3.1 "Irregularities" and "paradoxes"*

In this section we will illuminate some of the difficulties with obtaining uniqueness for finite semiorders, and provide the motivation for confining our attention to well-behaved semiorders that do not have "exceedingly large" thresholds. We'll consider two unusual situations that arise with representation theorems for semiorders. The first was discovered by Roberts and Franke (1976), who showed that there may be multiple representations with constant threshold for a given semiorder, where the representations are not related by any transformation at all. They called these "irregular" representations, and the following is an example of such a case:

EXAMPLE 1. *An "irregular" representation*: Let $\mathsf{S} = \langle A, \prec \rangle$ be a semiorder where $A = \{a_1, a_2, a_3\}$ (and all elements are distinct), and suppose that $a_1 \prec a_3$ and $a_2 \prec a_3$, but $a_1 \sim a_2$. Let $f$ and $g$ be two representations with constant threshold, where we set $\delta(a_i) = 1$ for all $i \in \mathbf{Z}_3^+$, and let $f(a_1) = f(a_2) = 0$, $f(a_3) = 2$, $g(a_1) = 0$, $g(a_2) = 0.9$, and $g(a_3) = 2$. It is easy to see that both of these representations capture the structure of the semiorder by considering each qualitative pair in turn with its numerical representations (numerical values are listed underneath for ease of reference):

| Qualitative Pair | Numerical Representations |
|---|---|
| $a_1 \prec a_3$ | $f(a_1) + 1 \leq f(a_3)$ |
| | $(0 + 1 \leq 2)$ |
| | $g(a_1) + 1 \leq g(a_3)$ |
| | $(0 + 1 \leq 2)$ |
| $a_2 \prec a_3$ | $f(a_2) + 1 \leq f(a_3)$ |
| | $(0 + 1 \leq 2)$ |
| | $g(a_2) + 1 \leq g(a_3)$ |
| | $(0.9 + 1 \leq 2)$ |
| $a_1 \sim a_2$ | $f(a_1) + 1 \not\leq f(a_2)$ and $f(a_2) + 1 \not\leq f(a_1)$ |
| | $(0 + 1 \not\leq 0$ and $0 + 1 \not\leq 0)$ |
| | $g(a_1) + 1 \not\leq g(a_2)$ and $g(a_2) + 1 \not\leq g(a_1)$ |
| | $(0 + 1 \not\leq 0.9$ and $0.9 + 1 \not\leq 0)$ |

Both $f$ and $g$ capture the fact that $a_3$ is discriminated from both $a_1$ and $a_2$, and that the pair $\{a_1, a_2\}$ is not discriminated. However, $f$ is irregular because there can be no function $\phi$ that transforms $f$ into $g$. Suppose (for contradiction) that there *were* some function $\phi$ such that $g = \phi \circ f$. Then we would have $g(a_1) = \phi(f(a_1)) = \phi(f(a_2)) = g(a_2)$, and at the same time, by definition $g(a_1) \neq g(a_2)$. This is a contradiction, thus (given the implicit assumption that $\phi$ was arbitrary) there can be no such transformation relating $f$ and $g$.

This creates a serious problem for the meaningfulness of the representations. By the definition of meaningfulness given above, scientific statements are only meaningful if their truth value remains invariant under a change of the numerical representation, where the relevant representations are related by a specific algebraic transformation. In the example above, $f$ and $g$ are not related by *any* transformation at all. As Roberts and Franke pointed out (1976 p. 213), the situation may arise where an irregular representation is, say, an interval scale, while another representation of the same structure is not. Roberts and Franke proved that such conflicts of

scale type only occur for irregular representations (1976 p. 215), and of course they can only occur when the representations are not unique.

Unfortunately, these irregular representations will not be eliminated simply by equally spacing the elements of the semiorder. They will also not be eliminated by confining the representation to integers, as the reader can easily see by simply multiplying all values in the above example by 10. To be assured that the numerical representation will not be irregular, the representation itself must be in an equally-spaced set (i.e., a numerical equal-difference structure), such as an arithmetical progression of integers, or any equally-spaced set of rationals. Again, many measurement scales have this characteristic, such as those appearing on rulers, gauges, etc. The existence of such a representation for well-behaved semiorders will be shown in THEOREM 2 below.

An additional problem was pointed out by Swistak (1980), who noted that closed representations with variable thresholds may fail to preserve the underlying linearity of the semiorder that is represented (in our terminology, this is a failure to preserve strong monotonicity). In other words, the representation of $\langle A, \prec \rangle$ may fail to preserve the induced relation $\prec_*$. He rightly considered this to be paradoxical, since in empirical cases where we consider the semiorder $\langle A, \prec \rangle$ to consist of incomplete data about an underlying linear order $\langle A, \prec_{UL} \rangle$, the induced relation $\prec_*$ is giving us additional information about this underlying order. We would certainly want a representation of a semiorder to be consistent with any additional information that may be obtained about the underlying order. But a closed representation of a semiorder may fail to do this, as more ordered pairs from the underlying order are added to the semiorder. This is shown by the following example:

EXAMPLE 2. *A "paradoxical" representation*: Let $\mathsf{S} = \langle A, \prec \rangle$ be a semiorder where $A = \{a_1, a_2, a_3\}$ (and all elements are distinct), and suppose that $a_1 \prec a_3$, but $a_1 \sim a_2$ and $a_2 \sim a_3$. Using DEFINITION 6, it is easy to see that this semiorder induces the ordering $a_1 \prec_* a_2 \prec_* a_3$. Now consider the representation: $f(a_1) = 1$, $f(a_2) = 0$, $f(a_3) = 2$, $\delta(a_1) = 1$, $\delta(a_2) = 3$, and $\delta(a_3) = 1$. One can easily see that this representation correctly captures the structure of the semiorder:

| Qualitative Pair | Numerical Representation |
|---|---|
| $a_1 \prec a_3$ | $f(a_1) + \delta(a_1) \leq f(a_3)$. $(1 + 1 \leq 2)$ |
| $a_1 \sim a_2$ | $f(a_1) + \delta(a_1) \not\leq f(a_2)$ $(1 + 1 \not\leq 0)$ and $f(a_2) + \delta(a_2) \not\leq f(a_1)$. $(0 + 3 \not\leq 1)$ |
| $a_2 \sim a_3$ | $f(a_2) + \delta(a_2) \not\leq f(a_3)$ $(0 + 3 \not\leq 2)$ and $f(a_3) + \delta(a_3) \not\leq f(a_2)$. $(2 + 1 \not\leq 0)$ |

This representation captures the fact that the pair consisting of $a_1$ and $a_3$ is "sufficiently wide" to be discriminated, while the pairs $\{a_1, a_2\}$ and $\{a_2, a_3\}$ are not. However, the representation fails to preserve strong monotonicity, since $a_1 \prec_* a_2$ but $f(a_1) \not< f(a_2)$. Thus, adding the pair $(a_1, a_2)$ to the semiorder would result in a failure of this representation, hence a "paradox."

Both of the above examples arise only because of the great deal of freedom allowed in the choices of $f$ and $\delta$. When this freedom is sufficiently constrained, such cases do not arise. Just as the Roberts and Franke example showed that the choice of $f$ should not be too free, Swistak's paradoxical cases can arise only if $\delta$ is allowed to be variable. These problematic representations give us reason to suppose that representations for semiorders will only achieve empirical

adequacy if suitable constraints are imposed on the representations. Although even paradoxical representations could be considered empirically adequate in the weak sense of reflecting the structure of the *available* data, this is not a sense of empirical adequacy with which any scientist would be happy. When theories apply only to available data, and fail as soon as new data is introduced, we consider them ad hoc and not truly informative about the qualitative structure they represent. We would like representations that work not only for the available data, but also for any new data that may subsequently be introduced. In order to avoid Swistak's paradoxical cases, either i) we are restricted to cases with constant thresholds, or ii) we must specify conditions on representations with variable thresholds which guarantee that the paradoxical cases will be avoided. Unfortunately, it is not at all obvious what kinds of conditions on variable thresholds would be sufficient. Fortunately, it turns out that the assumption of a constant threshold is not only sufficient to avoid the paradox, but it is also descriptively adequate in most applications.[11]

As with the equal-spacing assumption for the underlying order discussed above, the assumption of a constant threshold is not as confining as it may at first appear to be. In real world applications, semiorders are usually applied over such a limited range of utility values that it is a fairly weak idealization to suppose them to have a constant threshold over this limited range. Of course, if one were to consider the entire range of utility values, then it would be implausible to suppose discrimination thresholds to be constant. For example, in cases where we can translate utilities into monetary equivalents, a $10 difference may be decisive when one is deciding between goods in the $20 range, but it would have little to no effect on decisions to

[11] Nevertheless, Swistak (1980) showed that *if* a semiorder has a representation with a variable threshold, then one of its representations must preserve strong monotonicity, and hence avoid the paradox. But this still falls short of a uniqueness theorem for the variable threshold case. I have altered the phrasing of his theorem to agree with our terminology.

purchase in the $10,000 range. This corresponds roughly to the existence of what economists know as decreasing marginal utility. This is important when we are studying the population as a whole, or when studying the dynamic behavior of an individual in a particular market in the long term. But most singular decision problems, including almost all marketing applications, involve such a limited range of utility values that it is innocuous to suppose discrimination thresholds to be constant over the intended range of application.

*3.2 The multiplicity of linear extensions*

An additional problem, which is not directly related to the problems above, but which does bear on the difficulty with obtaining uniqueness, is the existence of multiple linear extensions for partial orders. It is well known from Szpilrajn (1930) that any partial order may be extended to a linear order.[12] This means that for any incomparable pairs $\{a, b\}$ from the domain, we may select either $a \prec b$ or $b \prec a$, and then place the newly ordered pair into the partial order without disturbing the rest of the order, and we may do this for all such pairs. The result will be a linear ordering of the original set. Szpilrajn's theorem simply says that this can always be done – that there *exists* a linear extension for any partial order. But a partial order will generally have many linear extensions, corresponding to combinations of choices in the ordering of the incomparable pairs. This may be seen by consulting the graphs in Figure 1:

[12] This important result has been proved in many different ways. See Los and Ryll-Nardzewski (1951) for a topological proof, Sierpinski (1958, p. 189) for a constructive proof of the denumerable case, and Trotter (1992, p.17) for a brief, elegant proof.

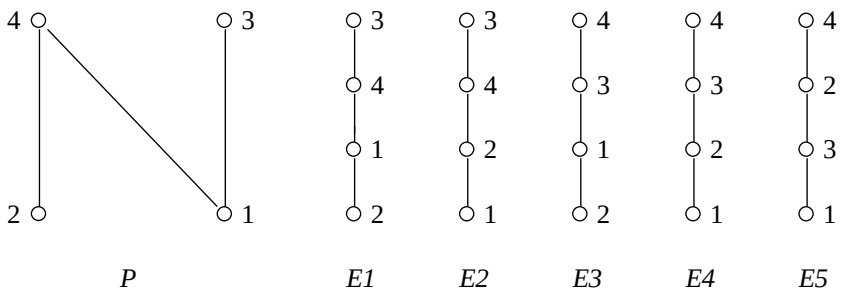


FIGURE 1

*P* contains only the three ordered pairs $(1, 3)$, $(1, 4)$, and $(2, 4)$. Thus *P* is a partial order (also a semiorder), since the pairs $\{1, 2\}$, $\{2, 3\}$ and $\{3, 4\}$ are incomparable. To see how this example bears on empirical situations, we may think of the numbers as originating in their usual ordering (*E4*), and then think of *P* as resulting from a threshold that prevents adjacent numbers from being discriminated. *E1* through *E5* are linear extensions of *P* that result from making different choices in the ordering of the incomparable pairs, without changing the ordering of any of the pairs already in *P*. For example, *E1* results from choosing $2 \prec 1$, $2 \prec 3$, and $4 \prec 3$.[13]

The multiplicity of linear extensions illuminates some of the difficulty with obtaining uniqueness, and it also raises an interesting question that was studied by Fishburn (1973b) and Fishburn and Gehrlein (1974, 1975). In psychophysical experiments one often begins with a linearly ordered stimulus set, from which a human subject, with limited perceptual and cognitive abilities, can identify only a fragment of the order. Although every pair of objects in the stimulus set is originally ordered, the subject "removes" certain pairs by failing to discriminate them, resulting in a truncation of the linear order to a partial order. Consequently, the scientist obtains data from the subject in the form of a partial order. Since in general there will be many

[13] Notice that certain *combinations* of choices will be ruled out, however. For example, we may not choose both $(4, 3)$ and $(3, 2)$, since that combination together with the transitivity of $\prec$ entails $(4, 2)$, but this is incompatible with $(2, 4)$, which is already in *P*. Hence choosing both $(4, 3)$ and $(3, 2)$ is ruled out because choosing both would disturb the rest of the order.

linear extensions of the partially ordered data, it is natural to ask whether it is possible to recapture the *original* linear order, from among the many possible, using only information available in the incomplete data. In terms of Figure 1, we would want to know how we can be sure that we obtain *E4* rather than one of the other extensions, using only the information available in *P*.

Fishburn and Gehrlein (1974, 1975) tackled this question in a slightly weaker form. Just as many *linear* orders will extend a given partial order, even more *weak* orders will extend the same partial order. Fishburn and Gehrlein found several different algorithms for finding the weak order that is "best supported" by a partial order, which is the weak order induced by Definition 6. Also, Fishburn (1973b) showed that for what we are calling "well-behaved" semiorders, if the threshold is no larger than (approximately) half the length of the semiorder, all indifferences can be resolved (i.e., in our terminology, if there is an underlying linear ordering, then it may be fully recaptured). Here "length" means the range of values of the set of objects under consideration with respect to some attribute, which could be utility, or loudness, or brightness, etc. What is of interest is that this result gives rise to both representation and uniqueness theorems, as we will show in the next section. Linear representations in these cases will automatically preserve both regularity and strong monotonicity, and hence avoid the "irregularity" and "paradox" just discussed.

## 4. THEOREMS

In THEOREMS 1 and 2 below we will show that for all well-behaved semiorders with thresholds that are suitably bounded, a linear extension may be found constructively, and this extension may be used to derive a representation of the semiorder in an arithmetical progression of integers. As

a corollary, such well-behaved semiorders are representable in a convex set of integers. We will first present THEOREM 1 (from Fishburn (1973b)), which specifies the condition on the size of the threshold under which the weak order induced by a semiorder $\mathsf{S}$ from DEFINITION 6 is in fact a linear order, (and hence a linear extension of the semiorder). We will give a constructive proof of Fishburn's theorem that gives a procedure for recovering the induced linear order $\mathsf{S}_{UL} = \langle A, \prec_{UL} \rangle$. In THEOREM 2 we show that this linear extension $\mathsf{S}_{UL}$ may be used to construct a numerical representation of the original semiorder $\mathsf{S}$ by an arithmetical progression $\mathsf{Z} = \langle \mathbf{Z}, < \rangle$. Thus, in THEOREMS 1 and 2 we define a series of functions: $h$: $\mathsf{S} \longrightarrow \mathsf{S}_{UL}$, $f$: $\mathsf{S}_{UL} \longrightarrow \mathsf{Z}$, and $f \circ h$: $\mathsf{S} \longrightarrow \mathsf{Z}$. Here $h$ is an embedding of $\mathsf{S}$ into the (unique) linear extension $\mathsf{S}_{UL}$ that recovers its underlying linearity, $f$ is an isomorphism of $\mathsf{S}_{UL}$ into an arithmetical progression of integers, and the representation $f \circ h$ is a homomorphism that preserves both regularity and strong monotonicity. After showing the existence of these mappings, we will then show in THEOREM 3 that the representation $f \circ h$ is unique up to a positive affine transformation. In closing, we will show that this is the strongest kind of uniqueness that may be obtained.

We will begin with a few propositions about semiorders that will be useful in the proofs. The following three propositions hold for all finite well-behaved semiorders $\mathsf{S} = \langle A, \prec \rangle$ where $\mathsf{S}_{UL} = \langle A, \prec_{UL} \rangle$ is an underlying equal-spaced linear extension[14] The reader should recall that we are assuming the indexing of the semiorder follows the indexing of the underlying linear ordering (CONVENTION 3). Proofs are in the appendix.

PROPOSITION 1: *For any* $a_i, a_j \in A$, *if* $a_i \prec a_j$, *then:*
[*for all* $k > j$, $a_i \prec a_k$] *and* [*for all* $m < i$, $a_m \prec a_j$].

[14] The reader may easily check that these propositions also hold more generally for the variable threshold case where the threshold $\delta_i$ of the semiorder is non-decreasing.

PROPOSITION 2: *If* $a_i \prec a_j$, *then* $\left|\boldsymbol{a}_j^{\prec}\right| < \left|\boldsymbol{a}_i^{\prec}\right|$.

PROPOSITION 3: *If* $\left|\boldsymbol{a}_j^{\prec}\right| < \left|\boldsymbol{a}_i^{\prec}\right|$, *then* $a_i \prec_* a_j$.

PROPOSITION 2 follows from PROPOSITION 1 as long as the threshold is non-decreasing (hence also when the threshold is constant). While PROPOSITION 2 states a relation between the pairs in a semiorder and the cardinalities of the sets of discriminated elements, PROPOSITION 3 allows us to determine the induced order even when two elements are *not* discriminated in the semiorder.

One may see the idea behind PROPOSITION 3 by referring back to Figure 1. Suppose that we begin with *E4* as an underlying order (i.e., $1 \prec_{UL} 2 \prec_{UL} 3 \prec_{UL} 4$), and we want to derive this extension from the semiorder *P*. Using PROPOSITION 3, we may deduce that since $\left|\mathbf{2}^{\prec}\right| = \left|\{4\}\right| < \left|\{3, 4\}\right| = \left|\mathbf{1}^{\prec}\right|$, then $1 \prec_* 2$, which is consistent with the desired $1 \prec_{UL} 2$. This is a variant of what Fishburn and Gehrlein (1974, 1975) called the "Cardinal Method" for constructing weak orders.

We will now generalize this procedure. The following theorem first appeared in slightly different form in Fishburn (1973b, p. 470), and states conditions under which the weak order induced by a well-behaved semiorder is in fact a linear order. We refer the reader to that paper for a simple existence proof. We will offer a constructive proof of the theorem, which provides a procedure for recovering this induced linear order.[15]

[15] Fishburn showed that the conditions of THEOREM 1 are *necessary* and sufficient for the result. We present only the sufficient direction, because we are interested only in establishing that the representation that we construct from the linear extension has been arrived at by a constructive procedure.

THEOREM 1: *Let* $\mathsf{S} = \langle A, \prec \rangle$ *be any finite semiorder where* $|A| = n$, *and* $\delta$ *is a positive integer. If*

$$\textit{(i) } n \textit{ is even, and } \delta \leq \frac{n}{2},$$

*or*

$$\textit{(ii) } n \textit{ is odd, and } \delta \leq \frac{n+1}{2},$$

*then the order* $\langle A, \prec_* \rangle$ *induced from* $\prec$ *by* DEFINITION 6 *is the underlying linear extension* $\mathsf{S}_{UL} = \langle A, \prec_{UL} \rangle$. *Moreover, there exists an embedding of* $\mathsf{S}$ *into* $\mathsf{S}_{UL}$.

*Proof*: We will prove the even case, from which the odd case follows with obvious minor adjustments. Since the constructive proof is lengthy we will give an overview. Using CONVENTION 2, we assume that the underlying linear ordering of $A$ is indexed by $\mathbf{Z}_n^+$. We then adopt CONVENTION 3, namely that the semiorder shares the indexing of this underlying linear order. We then use DEFINITION 10 to define sets $\boldsymbol{a}_i^{\prec}$ for all of the elements in $A$ that occur as left elements of some ordered pair in $\mathsf{S}$. Since $\delta$ is constant, these sets may be ordered by cardinality, so we order them thus and call the resulting sequence "*Seq*." Then we will use *Seq* to "reconstruct" the left and right parts of the linear order underlying $\mathsf{S}$. We do this by constructing two sequences: $\langle A^{left} \rangle$, consisting of a linear extension of the "left part" of $A$ in $\mathsf{S}$, and $\langle A^{right} \rangle$, consisting of a linear extension of the "right part" of $A$ in $\mathsf{S}$. We construct $\langle A^{left} \rangle$ by using the ordering of all sets $\boldsymbol{a}_i^{\prec}$ in *Seq*, and then correspondingly ordering the defining elements for those sets. This recaptures the underlying linearity of the left part of $\mathsf{S}$. We construct $\langle A^{right} \rangle$ by *comparing* all sets $\boldsymbol{a}_i^{\prec}$ in *Seq*, extracting a unique element $b^*$ from each $\boldsymbol{a}_i^{\prec}$ (where $b^*$ is not in the successor of $\boldsymbol{a}_i^{\prec}$ in *Seq*), and then ordering these extracted elements parallel to the ordering of the sets they were extracted from. The sequence $\langle A^{right} \rangle$ then

recaptures the underlying linearity of the right part of $\mathsf{S}$. If the semiorder satisfies condition (*i*), then the union of $\langle A^{left} \rangle$ and $\langle A^{right} \rangle$ comprises the desired induced linear extension $\mathsf{S}_{UL}$. It is then possible to merely use the identity *h*: $A \longrightarrow A$ on the elements of $A$ to recapture the underlying order of $\mathsf{S}$ that was only partially explicit in the semiorder itself. *h* is then an embedding of the semiorder $\mathsf{S}$ into the unique linear extension $\mathsf{S}_{UL}$ induced by DEFINITION 6. We will now be more precise.

**STEP 1**: Assume there is an underlying linear ordering of $A$ by the index set $\mathbf{Z}_n^+$, and assume without loss of generality that the semiorder shares this indexing (CONVENTION 3). Since $\mathsf{S}$ is finite, we may inspect all of its ordered pairs. If an element $a_i$ occurs as the left element of any ordered pair, then using DEFINITION 10 define the set $\boldsymbol{a}_i^{\prec}$ of "right elements" discriminated from $a_i$. We now need several lemmas (proofs of all lemmas are in the appendix).

LEMMA 1: *For all distinct i and j,* $|\boldsymbol{a}_i^{\prec}| \neq |\boldsymbol{a}_j^{\prec}|$.

LEMMA 2.1: *Let* $\sup \mathbf{Z}_n^{+\,left}$ *be the greatest* $i \in \mathbf{Z}_n^+$ *for which there exists a* $j \leq n$ *such that* $a_i \prec a_j$. *Then there are exactly* $\sup \mathbf{Z}_n^{+\,left}$ *sets* $\boldsymbol{a}_i^{\prec}$.

LEMMA 2.2: *For all* $a_i, \in A$, $|\boldsymbol{a}_i^{\prec}| = n - (i + \delta - 1)$.

LEMMA 3: *For all* $k$: $1 \leq k \leq n - \delta$, *there exists exactly one* $i \in \mathbf{Z}_n^+$ *such that* $|\boldsymbol{a}_i^{\prec}| = k$.

**STEP 2**: We now want to use the various sets $\boldsymbol{a}_i^{\prec}$ to build sequences, and we want to refer to relative positions in the sequences without specifying the exact value of *i* that occurs in the underlying ordering. Thus we will introduce a subindexing on the *i*'s as follows: "$\boldsymbol{a}_{i_m}^{\prec}$"

($1 \leq m \leq n - \delta_{i_m}$) means "any of $\boldsymbol{a}_i^{\prec}$, $\boldsymbol{a}_j^{\prec}$, etc.," where these may in turn represent $\boldsymbol{a}_1^{\prec}$, $\boldsymbol{a}_{27}^{\prec}$, etc. Since there are only finitely many sets $\boldsymbol{a}_i^{\prec}$, and by LEMMA 1, no two are of the same cardinality, we may effectively arrange the sets into a totally ordered sequence from largest (left) to smallest (right). We may also define an indexing $\zeta$ of the sets with respect to cardinality. By LEMMA 2.1, there are exactly $sup\ \mathbf{Z}_n^{+\ left}$ sets. Thus, we will index the largest set by 1, the next largest by 2, and so on, until the smallest set is indexed by $sup\ \mathbf{Z}_n^{+\ left}$ More precisely, since $sup\ \mathbf{Z}_n^{+\ left} = n - \delta$ (by LEMMAS 2.1 and 3), we may define the indexing $\zeta: \mathbf{Z}_n^{+} \longrightarrow \mathbf{Z}_n^{+}$ from 1 to $n - \delta$ as follows:

DEFINITION 11: Let $|\boldsymbol{a}_{i_m}^{\prec}| = k$. Then define $\zeta(i_m) = n - \delta - (k - 1)$.

Thus, the sequence (*Seq*) of sets $\boldsymbol{a}_{i_m}^{\prec}$ may be defined as follows:

DEFINITION 12: $Seq = ((\boldsymbol{a}_{i_1}^{\prec})_{\zeta(i_1)}, (\boldsymbol{a}_{i_2}^{\prec})_{\zeta(i_2)}, \ldots, (\boldsymbol{a}_{i_{n-\delta}}^{\prec})_{\zeta(i_{n-\delta})})$, or more simply,
$Seq = ((\boldsymbol{a}_{i_1}^{\prec})_1, (\boldsymbol{a}_{i_2}^{\prec})_2, \ldots, (\boldsymbol{a}_{i_{n-\delta}}^{\prec})_{n-\delta})$.

To see that the second expression is merely an abbreviation of the first, it suffices to check that the largest (first) set has cardinality $n - \delta$ (by LEMMA 3), and thus by DEFINITION 11, $\zeta(i_1) = n - \delta - (n - \delta - 1) = 1$. The reader may similarly check that $\zeta(i_m) = m$ for all remaining $i_m$.

**STEP 3**: From *Seq* we may use PROPOSITION 3 to deduce the underlying order of the defining elements for the sets in *Seq*. More precisely, we may define the following sequence which recaptures the underlying order of the left-hand part of S.

DEFINITION 13: Let $\langle A^{left} \rangle = (a_1, a_2, \ldots, a_{n-\delta})$ be a sequence of elements of $A$ where each position in the sequence is defined as follows:

If $\boldsymbol{a}_{i_m}^{\prec}$ is in the $m^{\text{th}}$ position in *Seq*, then $a_{i_m}$ is in the $m^{\text{th}}$ position in $\langle A^{left} \rangle$.

Since each $\boldsymbol{a}_{i_m}^{\prec}$ has a unique defining element $a_{i_m}$, it is immediate from PROPOSITION 3 and the fact that *Seq* is a total ordering by cardinality that $\langle A^{left} \rangle$ is also a total ordering. Even for pairs of elements $a_{i_m}$, $a_{i_{m+1}}$ that were not discriminated in $\mathsf{S}$, the sets $\boldsymbol{a}_{i_m}^{\prec}$, $\boldsymbol{a}_{i_{m+1}}^{\prec}$ defined in terms of them differ in cardinality and were so ordered in *Seq*. $\langle A^{left} \rangle$ is merely the reflection of the ordering of sets in *Seq* to an ordering of the original elements of $A$. That *every* left element (i.e., all $i_m \leq n - \delta$) gets so ordered follows from the fact that for all $i_m \leq n - \delta$, $\boldsymbol{a}_{i_m}^{\prec}$ is well defined. Hence, the set $\boldsymbol{a}_{i_m}^{\prec}$ for each $i_m \leq n - \delta$ appears in *Seq*, and thus by DEFINITION 13 the element $a_{i_m}$ which defines $\boldsymbol{a}_{i_m}^{\prec}$ appears in $\langle A^{left} \rangle$.

**STEP 4**: We will now reconstruct the right-hand part of $\mathsf{S}$. For this we must use a different method, and we need an additional lemma.

LEMMA 4: *For any* $i \in \mathbf{Z}_n^+$ *such that* $\boldsymbol{a}_{i_m}^{\prec}$ *is defined,*

*(i) either some set* $\boldsymbol{a}_{i_{m+1}}^{\prec}$ *is the successor of* $\boldsymbol{a}_{i_m}^{\prec}$ *in Seq, and there exists exactly one k such that* $b_{i_k} \in (\boldsymbol{a}_{i_m}^{\prec})_{\zeta(i_m)}$ *and* $b_{i_k} \notin (\boldsymbol{a}_{i_{m+1}}^{\prec})_{\zeta(i_{m+1})}$, *or there is no set* $\boldsymbol{a}_{i_{m+1}}^{\prec}$, *and* $\boldsymbol{a}_{i_m}^{\prec}$ *is a singleton with element* $b_{i_k}$, *and*

*(ii) for this unique* $k$, $b_{i_k} = a_{i_m+\delta}$ (*i.e.,* $b_{i_k}$ *is the* $\prec$*- least element such that* $a_{i_m} \prec b_{i_k}$).

COROLLARY 1: $(\boldsymbol{a}^{\prec}_{i_{n-\delta}})_{\zeta(i_{n-\delta})} \subset (\boldsymbol{a}^{\prec}_{i_{n-\delta-1}})_{\zeta(i_{n-\delta-1})} \subset \cdots \subset (\boldsymbol{a}^{\prec}_{i_2})_{\zeta(i_2)} \subset (\boldsymbol{a}^{\prec}_{i_1})_{\zeta(i_1)}$.

Since each of the sets $\boldsymbol{a}^{\prec}_{i_m}$ is finite, we may effectively inspect $(\boldsymbol{a}^{\prec}_{i_m})_{\zeta(i_m)}$ and $(\boldsymbol{a}^{\prec}_{i_{m+1}})_{\zeta(i_{m+1})}$ for all pairs $m$ and $m + 1$, and (by LEMMA 4) select the unique element that is in $(\boldsymbol{a}^{\prec}_{i_m})_{\zeta(i_m)}$ but is not in $(\boldsymbol{a}^{\prec}_{i_{m+1}})_{\zeta(i_{m+1})}$. We will call this element "$b^*_{i_k}$." Thus, $b^*_{i_k} \in (\boldsymbol{a}^{\prec}_{i_m})_{\zeta(i_m)}$ and $b^*_{i_k} \notin (\boldsymbol{a}^{\prec}_{i_{m+1}})_{\zeta(i_{m+1})}$.[16] Now we may define the sequence $\langle A^{right} \rangle$ as follows:

DEFINITION 14: Let $\langle A^{right} \rangle = (b_1, b_2, \ldots, b_{n-\delta})$ be a sequence of elements of $A$ where the element $b^*_{i_k}$ appears in the $k^{\text{th}}$ position in the sequence.

The construction of $\langle A^{right} \rangle$ is possible because of the uniqueness of the selected elements, which follows from LEMMA 4. It is also immediate from the uniqueness of each $b^*_{i_k}$ that $\langle A^{right} \rangle$ is a total order.

**STEP 5**: We must now show that $\langle A^{left} \rangle \cup \langle A^{right} \rangle$ is a linear extension, from which it will immediately follow that it is in fact $\mathsf{S}_{UL}$, the underlying linear order. Since we have already shown that $\langle A^{left} \rangle$ and $\langle A^{right} \rangle$ are total orderings of the left and right parts of $\mathsf{S}$, respectively, it suffices to show that there is no element of $A$ left out (i.e., that the sequences meet end to end with no gap, or that they overlap). This means that we must check whether $b_1 \preceq_{UL} a_{n-\delta+1}$. By STEP 4 and DEFINITION 14, $b_1 \in \boldsymbol{a}^{\prec}_{i_1}$ and $b_1 \notin \boldsymbol{a}^{\prec}_{i_2}$. Thus by LEMMA 4, $b_1 = a_{1+\delta}$. Thus we only

[16] For $m = n - \delta$, $(\boldsymbol{a}^{\prec}_{i_{m+1}})_{\zeta(i_{m+1})}$ is not defined, so in this case we select the only element of the singleton $(\boldsymbol{a}^{\prec}_{i_m})_{\zeta(i_m)}$.

need to check that $a_{1+\delta} \preceq_{UL} a_{n-\delta+1}$, and this is clearly true for all $\delta \leq \frac{n}{2}$; that is, whenever condition (*i*) is satisfied. Thus $\langle A^{left} \rangle \cup \langle A^{right} \rangle$ is a linear extension, and it is clear by our use of PROPOSITION 3 that this linear extension is in fact $\mathsf{S}_{UL}$.

**STEP 6**: We must now show that there exists an embedding of $\mathsf{S}$ into $\mathsf{S}_{UL}$. Since we have assumed that the indexing of the semiorder follows the indexing of the underlying order, the identity mapping will suffice for this purpose. More explicitly, we will define *h* as follows:

DEFINITION 15: Let $h: A \longrightarrow A$ be the function defined by: $h(a_i) = a_i$.

*h* preserves the underlying ordering of all elements of *A*, whether they are discriminated in the semiorder $\mathsf{S}$ or not. That is,

$$a_i \prec a_j \Rightarrow a_i \prec_{UL} a_j \Rightarrow h(a_i) \prec_{UL} h(a_j).$$

The first implication follows from PROPOSITIONS 2 and 3, and the second implication is immediate from DEFINITION 15. For any non-discriminated elements $a_i$, $a_j$ where $a_i \prec_{UL} a_j$ but $a_i \sim a_j$, the result follows from the last implication alone. This completes the proof of THEOREM 1. □

In the above proof, the method of recovering the underlying linear structure – the method of "cardinality comparisons" – is elementary, although somewhat tedious. It clearly yields a constructive procedure for recovering the induced, or "underlying" order. Although Fishburn didn't state that the induced linear order is unique, this follows from a result of Roberts (1971). In addition, Suppes (1972) showed that when any linear extension is equal-spaced, it is uniquely representable and interval scalable. These results jointly entail that when the order induced by

DEFINITION 6 (or equivalently, the "underlying" order) is an equal-spaced *linear* order as above, then one may uniquely represent the semiorder in a convex set of integers. We will now prove this in the following representation and uniqueness theorems. First we state (without proof) the following obvious but useful proposition.

PROPOSITION 4: *If* $\mathsf{S} = \langle A, < \rangle$ *is any finite linear order, then* $\mathsf{S}$ *is isomorphic to a convex, totally ordered integer structure* $\mathsf{Z} = \langle Z, < \rangle$.

THEOREM 2: (*Representation Theorem*)

*Let* $\mathsf{S} = \langle A, \prec \rangle$ *be any finite semiorder where* $|A| = n$, $\delta$ *is a positive integer such that either*

$$\textit{(i) } n \textit{ is even, and } \delta \leq \frac{n}{2},$$

*or*

$$\textit{(ii) } n \textit{ is odd, and } \delta \leq \frac{n+1}{2},$$

*and* $\mathsf{S}_{UL} = \langle A, \preceq_{UL} \rangle$ *is the underlying linear order induced by* DEFINITION *6 as in* THEOREM *1. Then there is a homomorphism of* $\mathsf{S}$ *into an arithmetical progression of integers* $\mathsf{Z} = \langle Z, < \rangle$, *and the homomorphism is a closed representation of* $\mathsf{S}$ *with constant threshold*.

*Proof*: Again, we will show only the even case. Let $\mathsf{S}_{UL} = \langle A, \preceq_{UL} \rangle$ be the linear order underlying $\mathsf{S}$ (i.e., induced from $\mathsf{S}$ by DEFINITION 6). By CONVENTION 2, $A$ is indexed by the set $\mathbf{Z}_n^+$, and is such that $i < j$ *iff* $a_i \prec_{UL} a_j$. By PROPOSITION 4, there exists an isomorphism between $\mathsf{S}_{UL}$ and a convex, totally ordered integer structure. If we use the integer structure $\mathsf{Z} = \langle \mathbf{Z}_n^+, < \rangle$, then an isomorphism $f : A \longrightarrow \mathbf{Z}_n^+$ may be simply defined as follows: let

$f(a_i) = i$. $\mathbf{Z}_n^+$ is convex and totally ordered by definition, and it is immediate that $f$ is both 1-1 and onto (*onto* by our having chosen the index set to have the same cardinality as *A*). It is also immediate that $a_i \prec_{UL} a_j$ *iff* $f(a_i) < f(a_i)$. Hence $f$ is the desired isomorphism.

By THEOREM 1, there is a homomorphism *h* from $\mathsf{S}$ into $\mathsf{S}_{UL}$. Thus, since $f$ is an isomorphism from $\mathsf{S}_{UL}$ onto $\mathsf{Z}$, then as long as $f \circ h$ is well defined, it is a homomorphism from $\mathsf{S}$ into $\mathsf{Z}$. That it *is* well defined follows from the fact that dom $f = A$ = ran *h*. The range of $f$ is $\mathbf{Z}_n^+$ (this is the domain of the structure $\mathsf{Z}$), and this is convex and totally ordered by definition.

To see that $f \circ h$ is a closed representation of $\mathsf{S}$ with constant threshold, we need to show that for some constant $c$, $a_i \prec a_j \Leftrightarrow f \circ h(a_i) + c \leq f \circ h(a_j)$. To show the $\Rightarrow$ direction, assume for two *i* and *j* that $a_i \prec a_j$, and assume the hypothesis of the theorem. By DEFINITION 9, $\delta = k - i$, where *k* is the least element of $\mathbf{Z}_n^+$ such that $a_i \prec a_k$. Thus $i + \delta = k$. Since $k$ is *least*, then by the assumption, $i + \delta \leq j$. If we let $c = \delta$, then this is just what the consequent says, since $f \circ h(a_i) = i$ and $f \circ h(a_j) = j$.

For the $\Leftarrow$ direction, assume for two *i* and *j* that $f \circ h(a_i) + \delta \leq f \circ h(a_j)$, and assume the hypothesis of the theorem. By definition of $\delta$, $a_i \prec a_{i+\delta}$. The consequent follows as long as $i + \delta \leq j$. But this is true by the assumption, since $f \circ h(a_i) = i$ and $f \circ h(a_j) = j$. □

LEMMA 5: *If* $\mathsf{S}$ *is a finite semiorder, and* $f$ *and* $f'$ *are two convex, totally ordered representations of* $\mathsf{S}$ *in integers with constant thresholds, then there is a constant* $c \in \mathbf{Z}^+$ *such that* $|f'(a_i) - f(a_i)| = c$ *for all* $a_i \in A$.

THEOREM 3: (*Uniqueness Theorem*)

*Let* $\mathsf{S} = \langle A, \prec \rangle$ *be any finite semiorder. Then any convex, totally ordered integer representation of* $\mathsf{S}$ *with constant threshold is unique up to an affine transformation.*

*Proof*: Let $\mathsf{S}$ be any finite semiorder with underlying linear extension $\mathsf{S}_{UL}$. Let $g$ and $g'$ be any two convex, totally ordered integer representations of $\mathsf{S}$ with constant thresholds ($g, g' : A \longrightarrow \mathbf{Z}$). First we consider the case where $g(a_i) \leq g'(a_i)$ for all $i \in \mathbf{Z}_n^+$. Then let $\phi : \mathbf{Z} \longrightarrow \mathbf{Z}$ be a transformation $\phi \circ g(a_i) = g'(a_i)$ from $g$ to $g'$ defined as follows:

$$\phi \circ g(a_i) = g(a_i) + (g'(a_i) - g(a_i)).$$

This is clearly an affine transformation, since it is of the form $\phi(x) = \alpha(x) + c$ with $\alpha = 1$. We need only check that $\phi$ is an isomorphism. To show that $\phi$ is 1-1, suppose that for some $g(a_i) \neq g(a_j)$, $\phi \circ g(a_i) = \phi \circ g(a_j)$. Then by definition of $\phi$,

$$\phi \circ g(a_i) = g(a_i) + (g'(a_i) - g(a_i)) = g(a_j) + ((g'(a_j) - g(a_j)) = \phi \circ g(a_j).$$

But since by assumption $g(a_i) \neq g(a_j)$, the central equality above entails that $(g'(a_i) - g(a_i)) \neq (g'(a_j) - g(a_j))$. This contradicts LEMMA 5.

Let the image of $A$ under $g$ be $\mathbf{Z}^* \subset \mathbf{Z}$, and let the image of $A$ under $g'$ be $\mathbf{Z}^{**} \subset \mathbf{Z}$. For this case (i.e., $g(a_i) \leq g'(a_i)$), we must show that $\phi$ is onto $\mathbf{Z}^{**}$. Since both $\mathbf{Z}^*$ and $\mathbf{Z}^{**}$ are images of $A$ under isomorphisms of $\mathsf{S}_{UL}$, then $|\mathbf{Z}^*| = |\mathbf{Z}^{**}|$. Then since $\phi$ is 1-1, it is clearly onto $\mathbf{Z}^{**}$.

To show that $\phi$ preserves $\leq$, we must show that

$$g(a_1) - g(a_2) \leq g(a_3) - g(a_4) \Leftrightarrow \phi(g(a_1)) - \phi(g(a_2)) \leq \phi(g(a_3)) - \phi(g(a_4)).$$

To show the $\Rightarrow$ direction, assume $g(a_1) - g(a_2) \leq g(a_3) - g(a_4)$. Then for any constant *c*,

$$[g(a_1) + c] - [g(a_2) + c] \leq [g(a_3) + c] - [g(a_4) + c].$$

Thus in particular, letting $c = (g'(a_i) - g(a_i))$ for all *i* (this constant is the same for all *i* by LEMMA 5),

$$[g(a_1) + (g'(a_1) - g(a_1))] - [g(a_2) + (g'(a_2) - g(a_2))] \leq$$

$$[g(a_3) + (g'(a_3) - g(a_3))] - [g(a_4) + (g'(a_4) - g(a_4))].$$

Then by definition of $\phi$, $\phi(g(a_1)) - \phi(g(a_2)) \leq \phi(g(a_3)) - \phi(g(a_4))$. The $\Leftarrow$ direction merely traces the first direction in reverse.

The case where $g'(a_i) < g(a_i)$ is similar. □

Since only non-unique representations can be irregular, these representations are clearly regular. And since the threshold $\delta$ is constant, they also satisfy strong monotonicity. Thus the problems discussed in SECTION 3 are avoided for well-behaved semiorders. These representations are of a special kind of interval scale called a "difference scale" by Suppes and Zinnes (1963, p. 12), and are unique up to the addition of a constant.[17] It is easy to show that this is the strongest kind of uniqueness that may be obtained. Let's assume (for contradiction) that two representations *g* and $g'$ as above can be related by a (non-trivial) similarity transformation $\psi$ where $g'(a_i) = \psi \circ g(a_i) = \alpha(g(a_i))$, for some positive constant $\alpha \neq 1$. Then

[17] Suppes and Zinnes attribute the coining of this expression to Donald Davidson.

$$\alpha = \frac{g'(a_i)}{g(a_i)}$$

for all $a_i \in A$.[18] But this is impossible by LEMMA 5, since the difference $|g'(a_i) - g(a_i)|$ is constant for all $a_i \in A$, thus the ratio

$$\frac{g'(a_i)}{g(a_i)}$$

cannot be constant for all $a_i \in A$.[19]

## 5. DISCUSSION

Concern about the descriptive inadequacy of utility models has focused mainly on preference reversal phenomena and apparent intransitivities in preference. Less attention has been devoted to the phenomenon that motivated semiorder research; namely, the intransitivity of indifference. Unlike the case of preference, where there is some controversy as to whether the intransitivities apparent in the data reflect genuine intransitivity in judgment, it is not controversial that judgments of indifference can be genuinely intransitive. Thus, for this aspect of choice behavior, semiorders provide a way of obtaining more descriptively adequate models. The theorems above show that when semiorders have constant thresholds that are not too large, then they have numerical representations that are unique up to a positive affine transformation; hence, such semiorders are interval scalable. This permits more descriptively adequate modeling of a wide range of choice behavior by semiorders, and it allows meaningful scientific inferences to be

[18] The case where $g(a_i) = 0$ for some *i* is a trivial case, for in that case it follows that $g'(a_i) = 0$, and since $g, g'$ are convex integer representations, this means that they would in this case be the *same* representation.

[19] Assuming of course that *A* is non-trivial; that is, that $2 < |A|$.

drawn from the representations of these semiorders. But in order to obtain a unique representation, it was necessary to assume that $\delta$ is constant, or equivalently, that $\delta$ is independent of *i*. From the standpoint of the purely mathematical questions about semiorders, this is a somewhat restrictive assumption. It rules out many semiorders of interest, so it does not solve the uniqueness problem for finite semiorders in general. But THEOREM 3 elucidates the difficulty somewhat. Since THEOREM 3 is provable with the assumption of a constant threshold, it is reasonable to infer that the primary source of the difficulty with the *general* uniqueness problem for semiorders lies with the variability of the thresholds.

Similarly, it is the variability of the thresholds that causes problems with possible *empirical* applications of semiorders. Our original concern was that the weak orders and linear orders used for modeling choice behavior have properties that are not universally true of human decision making. But the theorems show us that it is not the mere existence of discrimination thresholds in human behavior that renders linear models descriptively inadequate. To the extent that thresholds play any role in lessening the descriptive adequacy of linear models, they do so because of their variability only. And it appears to be true that subjects do show increasingly large discrimination thresholds as utility increases, analogous to what happens with sensory discrimination thresholds.

Because of the difficulties that arise because of the variability of thresholds, the less variability we have, the better. Fortunately, in most empirical applications we may reduce or eliminate the variability merely by confining our attention to a smaller interval in the range of utility values. In most applications, this is in fact what is done. For example, marketing a product usually involves consideration of possible substitutes, i.e., other goods in the same range of utility values, but it does not usually require consideration of luxury versions of the same type

of product. In fact, it would be fair to say that decision theory is of most interest when utility values are relatively close together. Thus, in any cases where it is possible to confine our attention to a *subset* of the entire range of utility values, we may, without harm, assume constancy of the discrimination threshold for that part of the range. This subset of the range of utility values would be determined by any considerations that enable us to safely assume that the threshold is constant (or near constant) over that interval. For such restricted domains, THEOREMS 1, 2 and 3 tell us that we may then model choice behavior with a semiorder rather than a weak order, obtain a unique numerical representation of the semiorder, and measure utility by an interval scale. This enables us to derive meaningful inferences about processes underlying human behavior from the representations. Since semiorders are preferable to weak orders from the standpoint of descriptive adequacy, this is a possible course for various social sciences to take in improving the descriptive adequacy of models of choice behavior.

APPENDIX

PROPOSITION 1: *Proof*: The Proposition has two parts. First we show that for any $a_i,\ a_j \in A,$ if $a_i \prec a_j,$ then for all $k > j,\ a_i \prec a_k$ (for all $i, j, k \in \mathbf{Z}^{+}_{|A|}$). Let $a_i,\ a_j \in A$ be arbitrary, and suppose $a_i \prec a_j$. Pick any $k > j$ such that $a_k \in A$. Since by CONVENTION 3 the ordering of the indexing follows the ordering of $\prec$, then clearly $a_k$ is $\prec$- right of $a_j$. And since $a_i$ and $a_j$ are separated by at least $\delta$, $a_i$ and $a_k$ are separated by at least $\delta+1$, and thus $a_i \prec a_k$. The proof of the second part: that for all $m < i,\ a_m \prec a_j,$ is clearly similar. □

PROPOSITION 2: *Proof*: Suppose $a_i \prec a_j$, and let $\boldsymbol{a}_i^{\prec}$ and $\boldsymbol{a}_j^{\prec}$ be the sets of elements "right-discriminated" from $a_i$ and $a_j$ respectively, as defined in DEFINITION 12. By CONVENTION 3, the indices $m_j$ of all elements in $\boldsymbol{a}_j^{\prec}$ are $> j$. By PROPOSITION 1, since $a_i \prec a_j$, then for all $m_j > j,\ a_i \prec a_{m_j}$, and thus by DEFINITION 12, all $a_{m_j} \in \boldsymbol{a}_i^{\prec}$. Thus, every element in $\boldsymbol{a}_j^{\prec}$ is also in $\boldsymbol{a}_i^{\prec}$. But in addition, the element $a_j \in \boldsymbol{a}_i^{\prec}$, but $a_j \notin \boldsymbol{a}_j^{\prec}$, hence $\left|\boldsymbol{a}_j^{\prec}\right| + 1 \leq \left|\boldsymbol{a}_i^{\prec}\right|$, and thus $\left|\boldsymbol{a}_j^{\prec}\right| < \left|\boldsymbol{a}_i^{\prec}\right|$. □

PROPOSITION 3: *Proof*: Let $\boldsymbol{a}_i^{\prec}$ and $\boldsymbol{a}_j^{\prec}$ be defined as in DEFINITON 12, and assume the hypothesis that $\left|\boldsymbol{a}_j^{\prec}\right| < \left|\boldsymbol{a}_i^{\prec}\right|$. Then since $\prec$ is trichotomous, it follows from DEFINITION 6 that $\prec_*$ is as well. Hence to show that $a_i \prec_* a_j$, we may suppose (for contradiction) that either $a_j \prec_* a_i$ or $a_j \sim_* a_i$.

Case 1: Suppose $a_j \sim_* a_i$. Then by DEFINITION 6, neither $a_j \prec_* a_i$ nor $a_i \prec_* a_j$, and hence $a_j \not\prec a_i$. Moreover, also by DEFINITION 6, there is no $a_k \in A$ such that both $a_j \prec a_k$ and

$a_i \sim a_k$, and there is no $a_m \in A$ such that both $a_m \prec a_i$ and $a_m \sim a_j$. Thus, $a_i$ and $a_j$ are discriminated from the same number of elements, hence $\left|\boldsymbol{a}_j^{\prec}\right| = \left|\boldsymbol{a}_i^{\prec}\right|$, contradicting the hypothesis.

Case 2: Suppose $a_j \prec_* a_i$. Then in the semiorder either $a_j \prec a_i$ or $a_j \sim a_i$. If $a_j \prec a_i$, then by PROPOSITION 2, $\left|\boldsymbol{a}_i^{\prec}\right| < \left|\boldsymbol{a}_j^{\prec}\right|$, contradicting the hypothesis. If $a_j \sim a_i$, then by CONVENTION 3 and the supposition, $j < i$ and hence $j + \delta < i + \delta$. But in that case, since $a_j \prec a_{j+\delta}$ (by DEFINITION 4), then by PROPOSITION 1, $a_j \prec a_{i+\delta}$, and for all $r > i + \delta$, $a_j \prec a_r$ as well. Hence, by DEFINITION 12, $a_{i+\delta} \in \boldsymbol{a}_j^{\prec}$, and for all $r > i + \delta$, $a_r \in \boldsymbol{a}_j^{\prec}$. But also by DEFINITION 12, the set $\boldsymbol{a}_i^{\prec}$ consists just of the element $a_{i+\delta}$ (the $\prec$- least, by DEFINITION 4) and all elements $a_r$ where $r > i + \delta$ (by PROPOSITION 1). Thus $\boldsymbol{a}_i^{\prec} \subset \boldsymbol{a}_j^{\prec}$. But to see that this is actually a proper inclusion, we need only consider the element $a_{j+\delta}$. By DEFINITION 4, this element is the $\prec$- least such that $a_j \prec a_{j+\delta}$. Yet it is not the case that $a_i \prec a_{j+\delta}$, since if that were true then $i + \delta \leq j + \delta$, in which case $i \leq j$, contradicting $j < i$. Thus, the set $\boldsymbol{a}_i^{\prec}$ is properly included in the set $\boldsymbol{a}_j^{\prec}$, hence $\left|\boldsymbol{a}_i^{\prec}\right| < \left|\boldsymbol{a}_j^{\prec}\right|$, contradicting the hypothesis. Thus $a_i \prec_* a_j$. □

LEMMA 1. *Proof*: If two sets $\boldsymbol{a}_i^{\prec}$, $\boldsymbol{a}_j^{\prec}$ (for $i \neq j$) *were* of the same cardinality, then two different elements $a_i, a_j$ would be discriminated from the same number of elements to their right. But since *A* is finite and all elements of *A* are equally spaced, $a_i$ and $a_j$ are separated by a distance of at least 1. Thus since $\delta$ is constant, either $|\boldsymbol{a}_i^{\prec}| + 1 \leq |\boldsymbol{a}_j^{\prec}|$ or $|\boldsymbol{a}_j^{\prec}| + 1 \leq |\boldsymbol{a}_i^{\prec}|$. □

LEMMA 2.1. *Proof*: This is immediate from the fact that the sets $\boldsymbol{a}_i^{\prec}$ are only defined when there exists at least one element $a_j$ such that $a_i \prec a_j$, and they are defined for every such $a_i$. □

LEMMA 2.2. *Proof*: By DEFINITIONS 9 and 10, $a_{i+\delta}$ is the $\prec$- least element of $\boldsymbol{a}_i^{\prec}$, so $\boldsymbol{a}_i^{\prec} = \{a_{i+\delta}, a_{i+\delta+1}, \ldots, a_n\}$. This set includes all $n$ elements of $A$ *except* the $i$ elements $a_j$ where $a_j \preceq_{UL} a_i$, and the $\delta - 1$ elements between (but not including) $a_i$ and $a_{i+\delta}$. Thus $|\boldsymbol{a}_i^{\prec}| = n - (i + \delta - 1)$. □

LEMMA 3. *Proof*: "Only one" is immediate from LEMMA 1. To see that there is at least one set of cardinality $k$ for each $k$ ($1 \leq k \leq n - \delta$), it suffices to note that for each such $k$ there is a unique element $a_k$ (by the fact that all $k$ in this range are indices), and by DEFINITION 10, a set $\boldsymbol{a}_k^{\prec}$. By LEMMA 2.2, these sets have cardinality $|\boldsymbol{a}_k^{\prec}| = n - (k + \delta - 1)$, which clearly varies from $n - \delta$ to 1 as $k$ varies from 1 to $n - \delta$. □

LEMMA 4. *Proof of (i)*: By LEMMA 1 and by the ordering of *Seq* from largest to smallest, there is at least one such $b_{i_k}$. To see that there is no more than one, we must first notice that for all adjacent $i_m$, $i_{m+1}$, and their corresponding adjacent sets in *Seq*, $(\boldsymbol{a}_{i_{m+1}}^{\prec})_{\zeta(i_{m+1})} \subset (\boldsymbol{a}_{i_m}^{\prec})_{\zeta(i_m)}$. This is an immediate consequence of PROPOSITION 1 and the fact that $\delta$ is constant. Now let's assume (for contradiction) that there are two distinct elements $b$ and $b'$, both of which are elements of $(\boldsymbol{a}_{i_m}^{\prec})_{\zeta(i_m)}$, but neither of which is in $(\boldsymbol{a}_{i_{m+1}}^{\prec})_{\zeta(i_{m+1})}$. Then $\left|(\boldsymbol{a}_{i_{m+1}}^{\prec})_{\zeta(i_{m+1})}\right| + 2 \leq \left|(\boldsymbol{a}_{i_m}^{\prec})_{\zeta(i_m)}\right|$. By LEMMA 3, there is exactly one set in *Seq* of cardinality $k$ for all $k$: $1 \leq k \leq n - \delta$. Thus there is a set, call it "$\boldsymbol{a}_{i_*}^{\prec}$" of cardinality $\left|\boldsymbol{a}_{i_*}^{\prec}\right| = \left|(\boldsymbol{a}_{i_{m+1}}^{\prec})_{\zeta(i_{m+1})}\right| + 1$. But

then $\left|(\boldsymbol{a}^{\prec}_{i_{m+1}})_{\zeta(i_{m+1})}\right| < \left|\boldsymbol{a}^{\prec}_{i_*}\right| < \left|(\boldsymbol{a}^{\prec}_{i_m})_{\zeta(i_m)}\right|$. By the ordering of *Seq* by cardinality, and by the fact that every set $\boldsymbol{a}$ defined by DEFINITION 10 appears in *Seq*, $\boldsymbol{a}^{\prec}_{i_*}$ appears in *Seq* between $(\boldsymbol{a}^{\prec}_{i_m})_{\zeta(i_m)}$ and $(\boldsymbol{a}^{\prec}_{i_{m+1}})_{\zeta(i_{m+1})}$. This contradicts the assumption that $(\boldsymbol{a}^{\prec}_{i_m})_{\zeta(i_m)}$ and $(\boldsymbol{a}^{\prec}_{i_{m+1}})_{\zeta(i_{m+1})}$ are adjacent in *Seq*.

*Proof of (ii)*: The unique $b_{i_k}$ from the first part of this lemma is by definition of $\boldsymbol{a}^{\prec}_{i_m}$ an element of $A$ such that $a_{i_m} \prec b_{i_k}$ and $a_{i_{m+1}} \not\prec b_{i_k}$. We must check that $b_{i_k}$ is the $\prec$- least element of $\boldsymbol{a}^{\prec}_{i_m}$. Suppose (for contradiction ) that there is an element $b_{i_j}$ where $i_j < i_k$ and both $a_{i_m} \prec b_{i_j}$ and $a_{i_{m+1}} \not\prec b_{i_j}$. Then by DEFINITION 9, $i_m + \delta \leq i_j$. But since by assumption $i_j < i_k$, then this entails that $i_m + \delta < i_k$, and thus $i_{m+1} + \delta \leq i_k$. But then $a_{i_{m+1}} \prec b_{i_k}$, which is a contradiction. Thus, $b_{i_k}$ is the $\prec$- least element that is right-discriminated from $a_{i_m}$, which by DEFINITION 9 means that $b_{i_k} = a_{i_m+\delta}$. □

LEMMA 5. *Proof*: Let S be any finite semiorder, and let $f$ and $f'$ be two convex, totally ordered integer representations of S with constant thresholds. We will first consider the case where $f(a_{i_m}) \leq f'(a_{i_m})$ for all $a_{i_m} \in A$. Since these representations are closed representations with constant thresholds, then min (ran $f'$) = $f'(a_{i_1})$, and min (ran $f$) = $f(a_{i_1})$. Let $c = \left|f'(a_{i_1}) - f(a_{i_1})\right|$. Then since both ran $f$ and ran $f'$ are convex and both $f$ and $f'$ satisfy strong monotonicity, the value of $a_{i_{1+k}}$ (for any *k*: $1 \leq k \leq n-1$) may be written $f(a_{i_1}) + k$ and $f'(a_{i_1}) + k$, respectively. Since *k* is a positive constant, then clearly

$$\left|f'(a_{i_1}) - f(a_{i_1})\right| = c \;\Rightarrow\; \left|f'(a_{i_1}) + k - f(a_{i_1}) + k\right| = c.$$

Since all values of $i$ are included in $i_1, i_{1+k}, \ldots, i_{1+(n-1)}$, then this is true for all $a_{i_m} \in A$.

The case where $f'(a_{i_m}) < f(a_{i_m})$ is similar. □

## REFERENCES


Armstrong, W. E.: (1939), The determinateness of the utility function, *Economic Journal* 49, 453-67.

Beeson, M. J.: (1980), *Foundations of Constructive Mathematics*, Springer Verlag, Berlin.

Chang, C. C., and H. J. Keisler: (1973), *Model Theory*, North Holland, Amsterdam.

Davidson, D., and P. Suppes: (1957), *Decision Making: An Experimental Approach*, Stanford University Press, Stanford, CA.

Debreu, G.: (1954), Representation of a preference ordering by a numerical function, *Decision Processes*, edited by R.M. Thrall, C.H. Coombs, and R.L. Davis, Wiley, New York, pp. 159-65.

Fishburn, P.: (1970), Intransitive indifference with unequal indifference intervals, *Journal of Mathematical Psychology* 7, 144-49.

———: (1973a), Interval representations for interval orders and semiorders, *Journal of Mathematical Psychology* 10, 91-105.

———: (1973b), On the construction of weak orders from fragmentary information, *Psychometrika* 38, 459-72.

Fishburn, P., and W. Gehrlein: (1974), Alternative methods of constructing strict weak orders from partial orders, *Psychometrika* 39, 501-16.

———: (1975), A comparative analysis of methods for constructing weak orders from partial orders, *Journal of Mathematical Sociology* 4, 93-102.

Fishburn, P., H. M. Marcus-Roberts, and F. S. Roberts: (1988), Unique finite difference measurement, *SIAM Journal on Discrete Mathematics* 1, 334-54.

Gescheider, G. A.: (1985), *Psychophysics: Method, Theory, and Application, 2nd ed.*, Lawrence Erlbaum Associates, Hillsdale, NJ.

Georgescu-Roegen, N.: (1936), The pure theory of consumer's behavior, *Quarterly Journal of Economics* 50, 545-93.

_____: (1958), Threshold in choice and the theory of demand, *Econometrica* 26, 157-68.

Heyting, A.: (1966), *Intuitionism*, North-Holland, Amsterdam.

Hogarth, R. M. and M. W. Reder: (1986), *Rational Choice: The Contrast between Economics and Psychology*, University of Chicago Press, Chicago.

Jamison, D. T., and L. J. Lau: (1973), Semiorders and the theory of choice, *Econometrica* 41, 901-912.

———: (1977), The nature of equilibrium with semiordered preferences, *Econometrica* 45, 1595-1605.

Kreps, D. M.: (1990), *A Course in Microeconomic Theory*, Princeton University Press, Princeton, NJ.

Kobberling, V., C. Schwieren, and P. Wakker: (2004), Prospect-theory's diminishing sensitivity versus economics' intrinsic utility of money: How the introduction of the Euro can be used to disentangle the two empirically, *Manuscript, Theory and Decision 2007*.***

Los, J., and C. Ryll-Nardzewski: (1951), On the application of Tychonoff's theorem in mathematical proofs, *Fundamenta Mathematica* 38, 233-37.

Luce, R. D.: (1956), Semiorders and a theory of utility discrimination, *Econometrica* 24, 178-91.

_____: (1967), Sufficient conditions for the existence of a finitely additive probability measure, *Annals of Mathematical Statistics* 38, 780-86.

Marschak, J.: (1950), Rational behavior, uncertain prospects, and measurable utility, *Econometrica* 18, 111-41.

Mirkin, B. G.: (1972), Description of some relations on the set of real-line intervals, *Journal of Mathematical Psychology* 9, 243-52.

Narens, L.: (2002), *Theories of Meaningfulness*, Lawrence Erlbaum Associates, Mahwah, NJ.

Rabinovitch, I.: (1977), The Scott-Suppes theorem on semiorders, *Journal of Mathematical Psychology* 15, 209-12.

Roberts, F. S.: (1971), On the compatibility between a graph and a simple order, *Journal of Combinatorial Theory* 11, 28-38.

———: (1989a), Applications of combinatorics and graph theory to the biological and social sciences: Seven fundamental ideas, *Applications of Combinatorics and Graph Theory to the Biological and Social Sciences*, edited by F. S. Roberts, Springer-Verlag, New York, pp. 1-37.

Roberts, F. S., and C. H. Franke: (1976), On the theory of uniqueness in measurement, *Journal of Mathematical Psychology* 14, 211-18.

Savage, L. J.: (1954), *The Foundations of Statistics*, Wiley, New York.

Scott, D.: (1964), Measurement structures and linear inequalities, *Journal of Mathematical Psychology* 1, 233-47.

Scott, D., and P. Suppes: (1958), Foundational aspects of theories of measurement, *Journal of Symbolic Logic* 23, 113-28.

Sierpinski, W.: (1958), *Cardinal and Ordinal Numbers*, Hafner, Warsaw and New York.

Stevens, S. S.: (1946), On the scales of measurement, *Science* 103, 677-680.

Suppes, P. : (1957), *Introduction to Logic*, Van Nostrand, Princeton, NJ.

_____: (1972), Finite equal-interval measurement structures, *Theoria* 38, 45-63.

Suppes, P., and J. L. Zinnes: (1963), Basic measurement theory, *Handbook of Mathematical Psychology (Vol. I)*, edited by R. D. Luce, R. R. Bush, and E. Galanter, Wiley, New York, pp. 1-76.

Suppes, P., D. Krantz, R. D. Luce, and A. Tversky: (1989), *Foundations of Measurement (Vol. II)*, Academic Press, New York.

Swistak, P.: (1980), Some representation problems for semiorders, *Journal of Mathematical Psychology* 21, 124-35.

Szpilrajn, E.: (1930), Sur l'extension de l'ordre partiel, *Fundamenta Mathematica* 16, 386-89.

Trotter, W.: (1992), *Combinatorics and Partially Ordered Sets*, Johns Hopkins University Press, Baltimore.

Vincke, P.: (1980), Linear utility functions on semiordered mixture spaces, *Econometrica* 48, 771-775.

von Neumann, J., and O. Morgenstern: (1944), *Theory of Games and Economic Behavior*, Princeton University Press, Princeton.

# ABSTRACT OF ESSAY 2

TROUBLES WITH CONVENTION T

By

Rolf Henry Johansson

Doctor of Philosophy in Social Science – Mathematical Behavioral Sciences

University of California, Irvine, 2014

Professor Louis Narens, Chair

Tarski's Convention T has been applied successfully to the study of the semantics of formal languages, but there are numerous well-known difficulties with its application to natural languages. These involve falsifications of the equivalence schema generated by substituted sentences involving indeterminate truth values or indexicals, as well as sentences giving rise to the various semantic antinomies. All of these difficulties have arisen from attempts to apply Convention T to sentences that can be seen to be "problematic" (i.e., ambiguous or paradoxical) in isolation. Hence, the problematic cases are generally regarded as involving anomalous sentences, but not necessarily as indications of problems with Conventon T itself. This general attitude toward the problem cases has fostered increased attention to Convention T in recent decades, which is manifest in the central role it plays both in Davidson's semantic program, and more recently in the currently dominant deflationist theory of truth. But an ingenious counterexample to Convention T was discovered by Hintikka (1976a) which differs from all other types of counterexamples in a crucial respect – the sentence substituted into the equivalence schema is neither ambiguous nor paradoxical, yet substitution of the sentence into the equivalence schema yields a false sentence. Hintikka's counterexample received little attention, largely because it was thought to be an isolated case, and hence not necessarily an

obstacle to the wider application of Convention T. I show that Hintikka's counterexample generalizes in quite unexpected ways, and that there are in fact a large number of unambiguous sentences that generate counterexamples of the same general type. I then show that various proposals for dealing with Hintikka's original counterexample are unsatisfactory, and that none of the proposed solutions can resolve all of the counterexamples presented in this essay. The existence of such a large variety of counterexamples corroborates Tarski's and Hintikka's skepticism about the possibility of using Convention T as a foundational criterion for assessing the adequacy of natural-language truth definitions, and hence poses a serious obstacle to Davidson's program in semantics, but even more seriously undermines the deflationist theory of truth.

## 1. Introduction: Convention T and Natural Languages

According to Tarski (1944, 1956), it is unlikely that the project of constructing precise semantic theories for natural languages will be successful. One would certainly expect any such theory to include, at the very least, a precise definition of what it means to be a 'true sentence' in the language under consideration. But with respect to such a definition, Tarski said:

> *The problem of the definition of truth obtains a precise meaning and can be solved in a rigorous way only for those languages whose structure has been exactly specified.* For other languages–thus, for all natural, 'spoken' languages–the meaning of the problem is more or less vague, and its solution can have only an approximate character. (Tarski 1944: 347, italics in original).

It seems that Tarski's reasons for this position stemmed from other more general concerns about the very enterprise of natural language semantics – in particular, concerns about the pervasiveness of ambiguity and context-sensitivity on the one hand, and about the existence of the semantic paradoxes on the other. These two concerns were clearly on his mind when he wrote the following:

> Whoever wishes, in spite of all difficulties, to pursue the semantics of colloquial language with the help of exact methods will be driven first to undertake the thankless task of a reform of this language. He will find it necessary to define its structure, to overcome the ambiguity of the terms which occur in it, and finally to split the language into a series of languages of greater and greater extent, each of which stands in the same relation to the next in which a formalized language stands to its meta-language. It may, however, be doubted whether the language of everyday life, after being 'rationalized' in this way, would still preserve its naturalness and whether it would not rather take on the characteristic features of the formalized languages. (Tarski 1956: 267)

There is an insight underlying both of these passages which, I think, has not received the attention it deserves. It's not the obvious point that natural languages, unlike formal languages,

contain many terms that are ambiguous and context-sensitive. The deeper insight surfaces in Tarski's conjecture that if one attempts a scientific semantics of a natural language, then the very process of precisely defining the structure of the language and disambiguating its terms might thereby *alter* the natural language in such a way that the resulting language would not really be the original object of study, but instead would be something more resembling a formal language.[20] We could perhaps then qualify Tarski's skepticism about natural language semantics as follows: it's not that he thought the application of formal methods to natural languages was in principle flawed, but rather that he was skeptical about the enterprise succeeding *as* a study of natural languages, and suspected that it would end up really being a study of something *else*; namely, the study of some new, semi-formal language that was not the intended object of study.

While one might object to Tarski by pointing out that *all* scientific modeling idealizes to some extent, and thereby distorts the object of study, I think Tarski's position has at least a *prima facie* plausibility on the basis of an additional layer of distortion that happens when one formalizes a natural language. The difference is that in the formalized study of the physical world by the natural sciences, the natural world is an entirely separate entity from the formal languages used to study it, so mathematical idealizations may be thought of as merely simplifying the *process of inquiry*, and not 'thereby' simplifying the *object of study*. In contrast, study of a natural language by means of another (formal) language involves not merely the modeling of some non-linguistic structure by a language, but rather the process of *translation* of

[20] The late Henry Hiż, who was a student of Tarski's in Warsaw, once informed me (in conversation) that Tarski had extensive knowledge of linguistics and languages, and had a much greater appreciation of the complexities of natural languages than one might gather from the bulk of Tarski's contributions to mathematics. This little-known fact about Tarski suggests that his skepticism about the utility of the formalization of natural languages was at least well-informed from the linguistics side, and not merely a bias driven by his daily workings with the precision of mathematics.

one language into another, and this process may indeed result in corruption of the original object of study in a way that mathematical modeling of physical structures does not.

This is clearly a topic worthy of more extended study. For now, however, we needn't settle the question of whether Tarski was right; we need only realize that his skepticism had a reasonable foundation.[21] However, what is relevant for our purposes is that Tarski thought the situation was quite different with respect to a scientific semantics of mathematics and logic. Formal languages do not generate the sorts of problems associated with indexicals, tenses, and other context-sensitive features of natural languages, and truth can be defined in such a way as to avoid the liar paradox. Thus, despite the fact that translation from object to meta-language may still be involved in studying the semantics of formal languages, this process needn't (in principle) result in any distortion of the object language.

Tarski proposed a criterion for defining truth predicates for formal languages in Convention T, which we paraphrase here, omitting some detail:

<u>Convention T</u>*: A truth predicate "is true" is adequate for a language L if and only if it entails all substitution instances of the equivalence schema:*

(ES) *X* is true if and only if *p*,

*where 'X' is a placeholder for a name of a sentence in the object-language, and 'p' is a placeholder for a translation of this sentence into the meta-language*.[22]

---

[21] For an opinion radically opposed to Tarski's, see Richard Montague's comments at the openings of "English as a Formal Language" and "Universal Grammar," (1970a: 188 and 1970b : 222, respectively), where he famously rejected the contention that there is any important theoretical difference between natural and formal languages. Whether Tarski's or Montague's opinion is correct will, I think, not be known until we have formalized at least a large fragment of some natural language, so that we may then be in a position to determine whether the resulting construction can be considered the original natural language in 'formal dress,' or something else entirely.

[22] Of the two occurrences of the biconditional *if and only if* in this paraphrase – one in the initial lines of Convention T, one within the equivalence schema (ES) itself – Tarski was occasionally more relaxed with his phrasing of the first occurrence within Convention T. Thus, in (1956: 187-88) Tarski says "*A formally correct definition of the symbol* '*Tr*' *… will be called an* adequate definition of truth *if it has the following consequences*:" (one of the consequences being the satisfaction of all instances of (ES)). This may mislead one into thinking that Tarski only intended the satisfaction of all instances of (ES) to be a *sufficient* condition for a definition of truth to be adequate. But several other passages in both (1944) and (1956) make it clear that he also intended the satisfaction of all instances of (ES) to be a *necessary* condition for a definition of truth to be adequate. For example, he says "… *if* the definition of truth is to conform to our conception, it *must imply* the following

It will be important in what follows to clearly distinguish the equivalence schema (ES) from Convention T, since the two have sometimes been conflated in the literature.[23] (ES) is merely a *schema*, but Convention T is an *adequacy criterion* which involves reference to (ES) for the purpose of saying something about truth. Whereas (ES) has free variables and individual substitution instances, which (following common practice) we'll hereafter call "T-sentences," Convention T involves *universal quantification over 'X' and 'p,'* where the substitutions into *'X'* and *'p'* are related by translation. I'll return to this distinction shortly.

In spite of the obvious differences between formal and natural languages, many philosophers and linguists have not shared Tarski's skepticism about natural language semantics. Davidson (1967, 1969, 1973) famously argued that Tarski's Convention T provides a template for the construction of a recursive definition of truth for any natural language L, provided that the T-sentences are relativized to speakers and times of utterance. In place of the usual T-sentences, Davidson proposed using substitution instances of the more elaborate:

(D) *X* is true (as English) for speaker *u* at time *t* if and only if *p*.[24]

---

equivalence: The sentence "snow is white" is true if, and only if, snow is white." (1944: 343, emphasis mine), and " … we wish to use the term "true" in such a way that *all equivalences of the form [ES] can be asserted* … " (1944: 344, emphasis mine). In addition, when explaining how Convention T might be phrased for a truth predicate "*Tr*" which is applied to a language that has only finitely many sentences, where one could list all possible substitution instances ($X_1$, $p_1$), ($X_2$, $p_2$), …, ($X_n$, $p_n$) of (ES), Tarski makes use of a *biconditional* phrasing in the initial lines of Convention T, and he translates the biconditional within (ES) into a disjunction of conjunctions. He there says: " … it would suffice to complete the following scheme: $X \in Tr$ *if and only if either* $X = X_1$ *and* $p_1$, *or* $X = X_2$ *and* $p_2$, …, or $X = X_n$ *and* $p_n$, …." (1956: 188; where each $p_i$ is a translation of the sentence $X_i$ into the metalanguage).

[23] For example, Klagge (1977: 377-78), Kirkham (1992: 143-4), and Williams (1999: 549) all refer to (ES) itself as "Convention T." As far as I can tell, this doesn't affect their arguments in any substantive way, but for the purposes of the present paper the distinction is important. Klagge's paper is discussed in section 4 below.

[24] I've modified Davidson's (D) trivially so that it agrees with our notation in (ES). See Davidson (1969: 756). In order to simplify cross-referencing, the equivalence schema and its relatives will be labelled mnemonically with capital letters as above with "ES" and "D," and all counterexamples to the various schemas will be numbered in sequence. Variants of either a schema or counterexample will carry the same letter(s)/number(s) as the original, but will add a prime or small italic letter, respectively.

This schema, perhaps with additional modifications as warranted by the complexity of the substituted sentence, enables one to specify the truth conditions of sentences with overtly context-sensitive indexicals.[25]

In recent decades, Convention T has figured prominently in the literature on truth because the equivalence schema is the centerpiece of currently popular deflationist theories of truth. This central role is emphasized in the following passage from an anthology devoted to deflationism:

> … what constitutes the heart of deflationism – is that deflationists take the instances of [ES] to be *fundamental*, both conceptually and explanatorily. According to the deflationist, neither a conceptual nor a substantive ... analysis of truth is possible, because there is nothing – conceptual or explanatorily – underwriting instances of [ES]. The instances of [ES] are bedrock." (Armour-Garb and Beall 2005: 3).

When Armour-Garb and Beall speak of a 'conceptual' analysis they have in mind specifically Tarski's analysis of truth in terms of the more basic concept of satisfaction, and by a 'substantive' analysis they have in mind non-deflationist attempts to explain why the substitution instances of (ES) hold by 'inflating' (ES) with some additional property, such as corresponding with reality, or cohering with a set of beliefs.

This increased attention to both Convention T and T-sentences has occurred despite Tarski's skepticism, and despite the existence of an interesting counterexample to Convention T presented by Hintikka (1976*a*, 1976*b*) which has received scant attention over the years. The lack of attention to Hintikka's counterexample has been due, I surmise, to a premature assessment of it as merely an isolated case. In what follows I'll show that Hintikka's counterexample was far from isolated, and that what he in fact discovered is a phenomenon which generates a surprising variety of counterexamples. But unlike problematic substitution

[25] By 'additional modifications' I mean that if the object-language sentence contained, say, a demonstrative such as 'this' or 'that,' then the schema would require specification of 'the object demonstrated by speaker *s*,' or some such. See Davidson 1967: 319-20 for details.

instances of the equivalence schema that clearly *are* isolated peculiarities, such as liar-type sentences and sentences with non-denoting expressions or indexicals, in which cases it is the substituted sentence that is the root of the problem – Hintikka's counterexample and others like it employ substitution instances that are *un*ambiguous and *non*-paradoxical, and hence make it clear that it is *the use of the equivalence schema itself* that is problematic. This renders these counterexamples a more serious problem for both the Davidsonian and the deflationist than the more widely-discussed problem cases just mentioned. More troubling is the fact that some of the counterexamples persist even under a propositional formulation of the equivalence schema.

## 2. HINTIKKA'S COUNTEREXAMPLE

Since (ES) employs a material biconditional, one of its logical consequences is of course the conditional schema

(CS) *X* is true if *p*.

Clearly, any theory of truth or meaning that requires the truth of instances of (ES) also requires the truth of instances of (CS). Consequently, any substitution instance that falsifies (CS) will necessarily also falsify (ES). For ease of reference, I'll refer to the substitution instances of (CS) from here on as 'T-conditionals.' Hintikka (1976*a*: 107-8) provided the following counterexample to (CS):

(1) 'Any corporal can become a general' is true if any corporal can become a general

or alternatively,

(1′) If any corporal can become a general, the sentence ‘Any corporal can become a general’ is true.

The reason why (1) is a counterexample to (CS) is that the first ‘any’ in (1) clearly has the force of a universal quantifier, while the second ‘any’ has the force of an existential quantifier. Of course, this is not the *only* way that (1) can be interpreted. One might, for example, try to preserve the truth of (1) by simply interpreting the second ‘any’ also as universal. Alternatively, one might try the more sophisticated approach of keeping the first ‘any’ universal and the second ‘any’ existential, but then consider a ‘non standard’ universe with only one corporal, where that corporal can become a general. In this non-standard universe, the conditional will come out *true* since the existential and the universal quantifiers would then range over the same singleton.

But both of these attempts at a solution miss the point – the mere *existence* of a false reading is a problem for the use of (ES) in Convention T. The very purpose of Convention T is for it to have a foundational role as a *criterion* for the adequacy of truth-definitions (Hintikka 1976*a*: 111). Its purpose is to provide a standard which enables us to see whether a predicate we are using to capture our intuitive concept of truth actually does capture that intuitive concept. Hence, it is especially intended to cover cases where the natural language terms are given their usual interpretations in the actual world. Attempts to preserve the truth of instances of (ES) by giving the schema an *alternate* reading undermine the foundational status of Convention T as a semantic criterion, especially when they can only preserve truth by resorting to a reading that seems forced and feels unnatural.[26]

[26] The readings are unnatural whether they result from the relatively minor ploy of giving an infrequently-used interpretation to any of the terms, or as a result of using a non-standard model. In Hintikka’s original articles he speaks of his counterexample as ‘false,’ rather than ‘easily falsifiable,’ or ‘false according to the dominant reading.’

This fact led Hintikka to consider his counterexample a serious problem for semantic theories, such as Davidson's, that demand the entailment of *all* substitution instances of the equivalence schema. The problem is this: for Davidson, Convention T and T-sentences jointly form *the sole criterion of adequacy for a definition of truth*, as is made clear in the following passages:

> Convention T and T-sentences provide the first and best link between familiar truths about truth and formal semantics; they alone constitute an unmistakable test that a theory has captured a concept of truth we are interested in. (Davidson 1973: 77)
>
> … although T-sentences do not define truth, they can be used to define truth-predicatehood: any predicate is a truth-predicate that makes all T-sentences true. (Davidson 1973: 76)

By requiring a definition of truth to make *all* T-sentences true (in order to capture our intended concept of truth), Davidson set the bar rather high. As long as there is even *one* acceptable reading of an instance of (ES) which falsifies it, one has a counterexample on one's hands. A false instance indicates that the employed definition fails to capture our intended concept of truth; i.e., it fails to capture our intuitive notion of what "is true" *means*.

But there is a deeper problem: if Convention T is to provide the cornerstone for a compositional semantic theory, as it is in Davidson's program, then the instances of (ES) are specifying how the *truth conditions* of complex sentences are built compositionally out of the satisfaction of their parts. In order for (ES) to correctly state these conditions for any given sentence, it is of course essential that *i*) the T-sentence in question is actually *true*, and *ii*) the meaning of the substituted sentence is *not altered* after substitution into (ES). These are fairly

---

I have considered the alternate readings more sympathetically only for completeness' sake, but the above comments should make it clear that Hintikka's phrasing was well-justified.

minimal requirements, but neither of them is satisfied in the case of (1). Because of this, Hintikka concluded that:

> ... there is no reason to expect the schema [ES] not to have false instances in other cases as well. ... Hence we just cannot trust the schema [ES] to yield only true substitution-instances. This already puts an entirely new complexion on attempts to base one's semantics on Convention T. (Hintikka 1976*b*: 63)

Notice that Hintikka's counterexample is quite different from the sorts of problem cases that have received much more attention in the literature on truth. The sentence 'Any corporal can become a general' has a determinate truth value, and hence it is not subject to problems associated with non-denoting expressions, such as those discussed in Dummett's classic paper on truth (1959). Dummett's concern was with a type of sentence *S* that is neither true nor false because it contains a non-denoting expression. In such cases, the sentence '*S* is true' will be *false*, and consequently the equivalence expressed by (ES) clearly will not hold. In addition, Hintikka's sentence contains no indexicals, and thus it is not the kind of example that motivated Davidson to provide an utterance-based version of the equivalence schema. Moreover, it contains no self-reference, so it will not lead to any of the known semantic antinomies. These features separate Hintikka's counterexample and others to be presented below from the better-known indeterminate cases, essentially context-sensitive cases, and liar-type cases.

To see the importance of Hintikka's discovery, we will next show that there are many additional counterexamples of the same general type. These additional cases are also unambiguous and non-paradoxical, but more importantly, they involve the reinterpretation of a variety of syntactic categories besides quantifiers. Hence they corroborate the claim, first made by Hintikka, that the problem arises not because of 'peculiarities' with any particular terms in the

lexicon, but rather because of the form of the equivalence schema itself. We will hereafter call all such cases 'Hintikka-type counterexamples' to (ES).

## 3. HINTIKKA-TYPE COUNTEREXAMPLES

Suppose we begin with the following sentence:

(2) Otto is ever grateful.

This sentence can be paraphrased roughly as 'Otto is *always* grateful.' Now suppose we wish to specify the truth conditions of (2) by employing Convention T, substituting a quotation of (2) for *X* and (2) itself for *p* into (ES). We then obtain as a logical consequence the following substitution instance of (CS):

(2*a*) 'Otto is ever grateful' is true if Otto is ever grateful.

This sentence is problematic in a way analogous to Hintikka's (1). In (2*a*), the right hand occurrence of (2) has the dominant reading:

(2*b*) $\exists t$ (*t* is a time & Otto is grateful at *t*).

But clearly, the relatively weak condition expressed in (2*b*) – that is, the condition of Otto being grateful *at some time* – is not a sufficient condition for the truth of (2), i.e., Otto *always* being

grateful. Consequently (2*a*) is *false*, since it in effect says that (2*b*) *is* a sufficient condition for the truth of (2).[27]

It is important to notice that (2) has a clear, unambiguous meaning, where 'ever' means *always* and cannot mean *at some time*. However, the substitution of (2) into a T-conditional results in two occurrences of the same string of words, where the object-language occurrence of 'ever' is interpreted as *always*, but this interpretation is altered by context in the T-conditional so that the dominant interpretation of the meta-language occurrence of 'ever' in (2*a*) is *at some time*. While this is not the only way that (2*a*) can be interpreted, it is clearly the most natural reading.

Example (2*a*) and Hintikka's (1) above hinge on the reinterpretation of an adverb ('ever') and a quantifier ('any'), respectively. Now consider the following verb case:

(3) 'Obama should win the election' is true if Obama should win the election.

This is also a counterexample to (CS) because the first 'should' is clearly interpreted as synonymous with *ought to*, while the second 'should' has the dominant interpretation *happens to*. In a case like (3) the falsification of the T-Conditional may be clearer if it is rephrased in the following way:

(3*a*) If Obama should win the election, then 'Obama should win the election' is true.

It is of course true that when the verb 'should' means *ought*, this *ought* itself may be understood in either a moral or a probabilistic sense (i.e., as either *deserves* to win or *is likely* to win,

[27] As with Hintikka's (1), there may of course be cases where (2*a*) is *vacuously* true, e.g., whenever (2*b*) is false. But as noted in the previous section, what's at issue is merely whether there are *any* false readings of an instance of (ES). With this caveat stated, I'll hereafter ignore this point as it pertains to examples below.

respectively). But this does not affect the status of (3) as a counterexample to (CS). To see that (3) is a counterexample to (CS), it suffices to notice that the sentence

(3*b*) Obama should win the election

cannot mean that Obama *happens* to win the election. In (3*b*), 'should' must be interpreted as *ought*, in either of the two senses just mentioned. Thus, while 'Obama *happens* to win the election' may well be true, it is clear that the truth of this sentence is neither a necessary nor a sufficient condition for the truth of the sentence 'Obama *ought* to win the election.' But (3) is the result of substituting (3*b*) and its quotation into (CS), and (3) in effect says that *happens to* is a sufficient condition for *ought to*, which is clearly wrong. Consequently, (3) is false.

The following counterexamples show that this phenomenon extends, surprisingly, to noun phrases in addition to quantifiers, adverbs, and verbs. Dominant interpretations are given below in parentheses:

(4) 'Mary comes home for Thanksgiving' is true if Mary comes home for Thanksgiving.
(*every Thanksgiving*) (*this Thanksgiving*)

(5) 'Bob cleans the gutters in the fall' is true if Bob cleans the gutters in the fall.
(*every fall*) (*this fall*)

When examples (2*a*), (3), (4), and (5) are considered alongside Hintikka's original example (1), it becomes clear that the general-then-specific reading of most T-conditionals seems to be driven by the conditional itself, and not by the particular word or phrase whose meaning is changed.

Before closing this section, we will consider two substitution instances of (CS) where the falsehood of the resulting T-conditional is less certain, yet it is nonetheless clear that the

resulting conditional does not unambiguously specify the truth conditions of the substituted sentence. Consequently, they provide further evidence of the inadequacy of the use of Convention T as a semantic criterion. Consider the following verb and adverb cases (again, the dominant interpretations of the relevant terms appear below in parentheses):

(6) 'I may bring a guest' is true if I may bring a guest.
(*might*) (*am allowed to*)

(7) 'Pierre can just finish in time' is true if Pierre can just finish in time.
(*barely*) (*only*)

In both (6) and (7), as with the other examples above, it is possible to force readings of the right-hand sides of the conditionals so that the meta-language interpretations of 'may' and 'just' are the same as their object-language interpretations. But the changing interpretations indicated above appear to be dominant for most speakers, and as with the clearly false cases discussed earlier, this is more than enough to cause problems for Convention T.[28] In spite of the fact that (6) does involve the indexical 'I,' and hence is not unambiguous in the way that other Hintikka-type cases are, this example illustrates Hintikka's main point even more forcefully. When reading (6), it is natural to keep the reference of the *overtly* ambiguous term 'I' the same in both of its occurrences, yet the seemingly unproblematic word 'may' is given two different readings upon substitution into the T-conditional.[29] (6) and (7) are further counterexamples to Convention

[28] That is, the changing interpretations have been dominant for most of the colleagues, students, and other audience members to whom this paper has been presented.

[29] Notice: parallel to the earlier counterexamples, when one considers the left hand side of (7) in isolation, 'just' must be interpreted as meaning *barely*, and cannot be interpreted as meaning *only*. With (6), however, although it is possible to read the left-hand occurrence of 'may' as *is allowed to*, this is not the dominant interpretation for most speakers.

T because in both cases, the truth of the quoted object language sentence is independent of the truth of the respective disquoted occurrence of the same sentence in the meta-language.

All of these counterexamples manifest a key property of natural languages *not* shared by formal languages: the context-sensitivity of many not-overtly-ambiguous words in the lexicon. Both Tarski and Hintikka believed that this feature of natural languages limited the application of Convention T beyond formal languages. I'll now show that despite the recent, more widespread application of Convention T to natural languages, the more conservative position shared by Tarski and Hintikka still withstands the objections that have been leveled against it.

## 4. PROPOSED SOLUTIONS

The additional counterexamples to Convention T show, at the very least, that the phenomenon Hintikka discovered is far more pervasive than his original example might lead one to believe. The phenomenon pertains not only to quantifiers, but also (at least) to adverbs, verbs, nouns, and definite descriptions. By performing straightforward substitutions of other ordinary, unambiguous phrases into T-conditionals, it is not difficult to find additional Hintikka-type counterexamples for each of the relevant syntactic categories.[30] Consequently, the phenomenon affects *many* Tarski-biconditionals expressed in English, and therefore clearly poses a problem for theories of semantics or truth for natural languages that require the truth of all substitution instances of (ES).

[30] I was able to find several dozen Hintikka-type counterexamples in English – far too many to discuss adequately in a single paper – including cases where the shifting semantics occurred for indefinite descriptions and plural nouns. Consider, for example, the sentence: "'A caterpillar becomes a butterfly' is true if a caterpillar becomes a butterfly." Or consider: "'Raccoons are found near the garbage' is true if raccoons are found near the garbage." Both of these instances of (CS) are parallel to the examples (2*a*), (3), (4), and (5) in that the object language occurrences of 'a caterpillar' and 'raccoons' are general, while the metalanguage occurrences are easily given a particular reading.

To my knowledge, Hintikka's counterexample inspired only one paper-length discussion – by Klagge (1977), but also some more recent, albeit relatively brief comments by Kirkham (1992), Peregrin (1999b), and Lepore & Ludwig (2005). There does not appear to be any agreement on exactly how the counterexample should be dealt with, and this, I think, counts as at least *prima facie* evidence that there is much more to Hintikka's counterexample than meets the eye. We'll now look at the various proposed solutions, and see that they all fail to provide a satisfactory resolution of the problem.[31]

### *4.1 On ambiguity versus context-sensitivity*

In Hintikka's original article (1976a: 108-09), he considered, but rejected the potential allegation that his counterexample is simply due to the fact that 'any' is ambiguous in English. Although none of his later commentators has explicitly made this allegation, some have pointed to context-sensitivity, and at least Peregrin suggested that the issue was a 'peculiarity' of 'any' (1999: xvii). Before discussing the relationship between ambiguity and context-sensitivity, it will be instructive briefly to consider Hintikka's own explanation of his 'any' counterexample, in order to get a clearer understanding of the underlying phenomenon generating the counterexamples.

Hintikka argued that ambiguity is *not* the root of the problem in (1) and (1´), although he defended this view by offering a rather ambitious hypothesis about the meaning of 'any' in English. He argued that 'any' is always universal in English (and hence not ambiguous), although its *scope* may change as it interacts with its grammatical environment in a T-conditional

[31] I should note that Kirkham has so far been alone (apart from the present author) siding with Hintikka in this debate. He has acknowledged the seriousness of Hintikka's original counterexample for Davidsonians, and sees no way out for them. See Kirkham (1992: 244).

(so one could still say that 'any' is context-sensitive).[32] One of his reasons for believing this was that formalizations of 'any' using an existential quantifier can be given paraphrases using a universal quantifier.[33] He offered the following formalizations of (1′):

(1∃) $\exists x$ ($x$ is a corporal & $x$ can become a general) → 'any corporal can become a general' is true

which, under a standard prenex conversion, is easily seen to be logically equivalent to the formalization

(1∀) $\forall x$ [($x$ is a corporal & $x$ can become a general) → 'any corporal can become a general' is true].

Given the logical equivalence of (1∃) and (1∀), it becomes possible to argue that 'any' has the 'deep' interpretation of a universal quantifier in English, and hence that the 'underlying' form of (1′) is (1∀), but that the logical scope of the quantifier is altered by its grammatical context in the T-conditional – in particular, by its interaction with the word 'if.' The surface grammatical form is then rendered by (1∃).[34] The result is a substitution instance of (CS) (and hence of (ES)) that is false under its standard interpretation.

---

[32] See Hintikka (1976*b*: 64) for the claim that 'any' is always universal, and see Hintikka (1997) for a discussion of the difference between what he calls 'binding scope,' which is indicated by the placement of parentheses, and 'priority scope,' which is determined by the underlying logical form of the sentence. Priority scope is what is relevant for the 'any' case under consideration here.

The reader should also be careful not to confuse Hintikka's claim that 'any' is always universal with a related but *different* claim, which has become known as his 'any-thesis.' The 'any-thesis' says, roughly, that a use of 'any' is grammatical if and only if substitution of 'every' for 'any' results in a grammatical sentence not identical in meaning with the original 'any' sentence. See Hintikka (1977, 1980) for a defense of his 'any-thesis.'

[33] The reverse is *not* generally true, though, as can be seen by considering a sentence with an 'any' expression in the lead position (e.g. 'Anyone can get the job done'), where one will be unable to give the expression an existential reading.

[34] I wish to emphasize that I am not claiming that 'underlying' form or 'surface' grammatical form necessarily corresponds, or doesn't correspond, to any particular level of representation postulated by different syntactic

Although the logical equivalence (1∃) and (1∀) may lend some plausibility to the hypothesis that 'any' is always universal in English, this hypothesis was challenged by Carlson (1980) and Higginbotham (1982), both of whom provided examples of sentences involving 'any' where a universal paraphrase does not appear to be available. For example, Higginbotham argued that in the following examples, 'any' has only existential force:

(8) That teacher rarely fails anybody.
(9) John will know if anybody left.

Higginbotham (1982: 267-68) showed that a logical paraphrase analogous to the above paraphrase of (1∃) by (1∀) will not work in these cases, and hence that a universal reading of 'any' in (8) and (9) fails to give the correct interpretation. Even though Hintikka's universality claim concerned a 'deep' level of representation rather than the surface form or logical form, the existential force of 'any' in (8) and (9) is so strong that it does not seem likely that either could be given a plausible 'deep' representation where 'any' is universal.

I surmise that part of the reason for the lack of attention to Hintikka's counterexample over the years is that some may have believed – *wrongly*, I think – that Hintikka's main criticism of Convention T rested entirely on his hypothesis about the universality of 'any,' and the examples of Carlson and Higginbotham support a persuasive counterargument that Hintikka's universality hypothesis does not appear to be correct as stated.[35] But in light of the additional counterexamples presented in the last section, none of which involve the word 'any' at all, it

---

theories, especially to what linguists have called 'LF.' Hintikka seems to have had in mind that (1∀) is a 'deep' structure representation and that (1∃) is either a 'surface' structure representation, or what today would be considered LF. The shudder quotes indicate that what Hintikka said in the 70's may well be different from what he might say today, given the changes in the understanding of levels of representation in linguistics over the last few decades.

[35] However, it may still be true that some weaker form of the hypothesis is correct – e.g., that 'any' is universal in all but a few isolated contexts.

should be clear that Hintikka's criticism of natural language applications of Convention T is actually independent of his hypothesis about the universality of 'any.' The counterexamples must have some other source.

One may then wonder why Hintikka went to such lengths to defend his hypothesis that 'any' is always universal, if indeed this hypothesis is independent of his main criticism of Convention T. I can think of at least one good reason: If 'any' *were* always universal, it would then be clearer that the source of the problem with (1) is not the *ambiguity* of an isolated word, but rather *context-sensitivity* as a general phenomenon. This is an essential distinction, because by pointing to context-sensitivity rather than ambiguity, it becomes clearer that there is a genuine problem with using Convention T as a semantic criterion, since (ES) involves the substitution of words into a *new* context. I think that was the main point Hintikka meant to emphasize, and as I'll now argue, he was indeed right to point to context-sensitivity and not to ambiguity as the main culprit.

To see this, we should first clearly separate the concepts of ambiguity and context-sensitivity. A term that appears in a given context and can receive more than one interpretation in *that* context may be called 'ambiguous.' On the other hand, a term which receives one interpretation in one context and a different interpretation in other contexts – but can only receive a *single* interpretation in some of these contexts – would properly be called 'context-sensitive,' but not necessarily 'ambiguous.' Hence, context-sensitivity is a less extreme form of meaning variance than ambiguity. If we fix a term but vary the context, and find that the meaning of the term can vary across the different contexts, then the term is merely context sensitive. But if we fix *both* a term and a context, and find that the meaning of the term can still vary within that given context, then the term is ambiguous. Of course, these properties do not form disjoint sets; e.g., some terms we call 'context-sensitive' may, in some of their contexts, also have more than one

interpretation, and hence may also be ambiguous. But for the most part these properties are distinct. There are many terms in the lexicon that would be classified as context-sensitive because they receive different interpretations in different contexts, but most of these terms will be unambiguous in any given context.

Since this distinction is central to understanding Hintikka's criticism of Convention T, it may help to consider some simple examples. In the following cases:

(10) John listed his accomplishments
(11) The boat listed in the wind

it would be wrong to say that 'list' is ambiguous in either (10) or (11), (except in the relatively minor respect in which (10) does not specify whether the listing is oral or written). Clearly we are here dealing not with a single ambiguous word, but rather with *two* context-sensitive homonyms, which we might call '$list_1$'(*to itemize in a series*) in (10), and '$list_2$' (*to lean to one side*) in (11). Although it may well be possible to construct example sentences where a term which has a homonym is itself ambiguous, in the vast majority of cases the meaning of such a term will be made clear by context, as it is in (10) and (11).[36] In fact, the usual effect of embedding a context-sensitive term within a sentence is specification of a *unique* intended word or meaning.

To return to the Hintikka-type counterexamples, of course a necessary condition for generating such a counterexample is the presence of at least one context-sensitive term in the substituted object-language sentence (i.e., one term that is capable of having different interpretations in different contexts). But it should now be clear that this is very different from

[36] I'm using 'term' for any sequence of letters or sounds that is interpretable as one or more distinct words.

claiming that the relevant term or the object-language sentence it occurs in is ambiguous in isolation. The object-language sentences in all of the Hintikka-type counterexamples considered above contain *context-sensitive* terms, but the object-language sentences containing them are *unambiguous* in isolation.[37]

One could say, however, that ambiguity is relevant to Hintikka-type counterexamples only in the following sense: since it is possible to give the entire T-conditional two readings – one generating a false Hintikka-type counterexample, and one preserving truth – then *the T-conditional itself* is ambiguous, and its two interpretations are made possible by giving two interpretations to the meta-language occurrence of one of the terms in the substituted sentence. But three points have now been made concerning the multiple readings of such T-conditionals: *i*) the mere *existence* of a false reading is sufficient to generate a counterexample to Convention T, *ii*) in Hintikka-type counterexamples, the false reading of the T-conditional is the dominant one – indeed, it appears that a true reading can in some cases only be 'seen' by forcing the interpretation in a way that betrays the most natural (false) reading of the conditional, and perhaps the most troubling, *iii*) if it is possible to take an *unambiguous* natural-language sentence, substitute it into the equivalence schema, and end up with an *ambiguous* sentence (i.e., an ambiguous resulting T-sentence), then so much the worse for the use of Convention T as a semantic *criterion*!

It is worth noting that a very large number of terms in any natural language are context-sensitive.[38] It seems to have been this that concerned Tarski and Hintikka the most. Although

[37] That is, they are not ambiguous in any sense *relevant* to their status as counterexamples. Ambiguity may still be present in a way that does not undermine the status of a counterexample *as* a counterexample; cf. the above discussion of two senses of *ought* in connection with example (3).

[38] Although there are different possible ways of counting "words," by taking a sample of over 1,000 *entries* in the *American Heritage Dictionary*, a very conservative estimate reveals that over 10% of the entries list multiple meanings. Since this dictionary contains over 100,000 entries, and these form only a subset of all terms of English, one can estimate that there are over 10,000 terms in American English that can receive different interpretations in

we may find ways to deal with the *obviously* context-sensitive features of natural language – pronouns, demonstratives, and tenses – natural languages contain *many* other expressions that are context-sensitive, even though their context-sensitivity may not be as readily apparent. The sheer quantity of such context-sensitive terms in the lexicon of any natural language creates an even greater problem for the use of Convention T, because in many cases we may not know – *prior* to the substitution of a relevant object-language sentence into (ES) – whether a given term is context-sensitive. Hence there does not appear to be any way to deal uniformly with such a large number and variety of context-sensitive terms.

### *4.2 Formalizing the meta-language*

Klagge suggested that the source of Hintikka's counterexample was an interplay of 'any' and 'if,' and that it could be avoided by finding some context-independent (with respect to 'any') part of our metalanguage (1977: 379-80). His proposal for doing this was, oddly, not exactly to 'find' a context-independent part of our metalanguage, but rather to *create* one by adding formalized connectives to English. He suggested that one could avoid Hintikka's counterexample by simply replacing the natural language 'if and only if' with one its symbolic translations. But Hintikka had already entertained this possibility in his original articles, and argued that:

> This attempt will lead to expressions whose truth-value has not been determined, for the original sentence *p* will of course have to figure in a substitution-instance of the schema [CS]. Hence these substitution-instances of [CS] will then be mixed expressions containing both formalized connectives and English words like 'any' and 'can'. There simply are no grounds for deciding how 'any' behaves vis-à-vis such foreign elements as formalized connectives, and consequently no satisfactory solution is available in this

different contexts (and hence are context-sensitive). Nevertheless, much more often than not, sentences containing these terms are *not* ambiguous.

way. (The explanatory value of the whole condition formulated in terms of [ES] would be destroyed by this indeterminacy) (Hintikka 1976: 110).

Perhaps Hintikka had not made this point emphatically enough. Klagge seems to have assumed that schemes phrased with a mixture of formal and natural-language expressions will be clearly interpretable. But as Hintikka was quick to point out, this is just not so. Consider Klagge's comment:

> Convention T, as Tarski and Davidson intend it, is:
>
> (T) $X$ is true $\leftrightarrow$ $p$.[39]

We can ignore the conflation here of Convention T with (ES) itself, because Klagge's claim has a more substantive problem. His claim is that both Tarski and Davidson "intended" to render Convention T in a meta-language that included formalized logical connectives, but I think it is quite unlikely that (at least) Tarski intended this. In Tarski's most formal monograph devoted to this topic (1956), in *every* potential instance where a more careless writer might make use of a formal connective in the meta-language, Tarski explicitly *avoids* it. This was clearly deliberate, so it is somewhat tendentious to suggest that Tarski "intended" otherwise.

Moreover, it is well-known that formal connectives only capture part of the meanings of their usual natural language paraphrases, so one cannot assume that any translation from natural language to formal language (or vice versa) will preserve the original interpretation. To give just one example, consider the following English sentence:

(12) If the machine starts whenever I press its 'ON' switch, then the machine starts whenever I simultaneously press its 'ON' switch and pull its plug.

[39] See Klagge (1977: 378). I have only made the trivial adjustments of changing Klagge's $\pi$ to $X$ and $\equiv$ to $\leftrightarrow$ in order to agree with our notation.

This sentence, though obviously false under a standard interpretation, has a standard translation into a theorem of the propositional calculus (using obvious substitutions):

$$(P \rightarrow R) \rightarrow (P \wedge Q \rightarrow R).$$

This translation might well be provided in an introductory logic class that translates every instance of English "if … then" as "→," but it is well-known that the conditional in English has several other interpretations besides that of the material conditional. One may well explain the lack of translatability in the above example in terms of non-monotonic reasoning, or in terms of non-material conditionals in English, or some such. But examples like this only corroborate Hintikka's claim about the impossibility of "explaining away" his counterexample by using a semi-formal translation. Any translation along the lines of what Klagge suggests will involve some combination of formal connectives and English expressions, but without precise rules for interpreting *mixed* expressions, we simply have no grounds for saying that any one interpretation is the correct one. We may agree with Klagge that the intended meaning of the biconditional in (ES) is of a material biconditional, but this does not license the substitution of a formal connective into (ES) where Tarski took pains to use natural language. Doing so alters the metalanguage and renders it something that itself is in need of interpretation, when the metalanguage was intended to be used as a vehicle, the interpretation of which is not in question, so that it could be *used* to study the language that is the object of study (the object language). Given the foundational and explanatory importance attached to Convention T and to the equivalence schema, any indeterminacy in their interpretations is clearly unacceptable.

It appears that Klagge may have had in mind the fact that we occasionally *read* expressions containing formal connectives by employing standard natural-language counterpart expressions,

and that there should then be no special problem reversing this process; i.e., in formalizing natural-language expressions. But this is far from true. There are many complexities of natural languages that have yet to find suitable means of formalization (consider the enormous complexity of pragmatics, just for starters). Moreover, when one reads a formal symbol in some way with the help of a natural language, the use of a natural language as meta-language again becomes vulnerable to the formulation of any of the counterexamples. Thus, this attempt at a solution either introduces uncertainty into the very interpretation of Convention T, or it merely masks the still-present counterexample in the clothing of a formalization, allowing the counterexample to persist whenever anyone reads it in the meta-language.

### *4.3 Paraphrasing the conditional*

As another possible way around the counterexample, Klagge suggested that one need only give the conditional one of its many other natural language paraphrases, and merely find one for which the counterexample does not persist. But this attempt at a solution has at least two flaws: *i*) even if a paraphrase is considered successful for one counterexample, there does not appear to be a paraphrase that will eliminate all of the counterexamples, and *ii*) the best candidates among the possible paraphrases seem to considerably distort and even undermine the explanatory role of Convention T.

To see this, suppose that instead of (CS) we used its counterpart conditional:

(CS´) *X* is true only if *p.*

Then it is still easy to construct counterexamples paralleling those above. Consider:

(13) ‘Otto is ever grateful’ is true only if Otto is ever grateful.

This counterexample to (CS´) is subject to problems somewhat different from those involved in the counterexamples to (CS). Here the meta-language occurrence of ‘ever’ (right) is still read as *at some particular time*, while the object language occurrence (left) is still read as *always*. Hence, the right side may be true while the left is false, which means that the conditional may still be (vacuously) true. However, the right side clearly does not state the proper truth conditions for the object-language sentence on the left.

One would think that the natural-language paraphrase with the most promise of eliminating the counterexamples is one that avoids the use of ‘if’ altogether, as in the following paraphrase of (CS):

(CS´´) It is not the case that *p*, or *X* is true.[40]

While this does appear to weaken Hintikka’s counterexample (see Klagge 1977: 379 for details), this approach will not work as a general solution. Consider:

(14) It is not the case that Otto is ever grateful, or ‘Otto is ever grateful’ is true.

and

(15) It is not the case that I may bring a guest, or ‘I may bring a guest’ is true.

[40] Kirkham (1992, p.244), for example, was of the opinion that Hintikka’s counterexample arose because “When ‘any’ follows ‘if’ it is an *existential* quantifier.” Nevertheless, after consideration of Hintikka’s counterexample he agreed that “… the truth conditions of a compound sentence are not always a function of the truth conditions of its component clauses …. I see no obvious way for Davidson to escape this objection.” But as I’ll show in the examples immediately below, the real source of the problem *cannot* be merely that ‘if’ generates different interpretations from object-language to meta-language, since counterexamples persist even with paraphrases that avoid use of ‘if’ altogether. The additional paraphrases suggest that the real source of the counterexamples is context-sensitivity more generally.

In the case of (14), the left hand reading of 'ever' in the meta-language is still predominantly *at any particular time*, while the object language occurrence on the right is still *always*. Thus, this reading of (14) is a disjunction which has the form:

(14´) $\sim \exists t$ (*t* is a time and Otto is grateful at *t*) $\vee$ 'Otto is ever grateful' is true.

This disjunction could be paraphrased: 'Either Otto is never grateful, or the sentence 'Otto is always grateful' is true.' But this clearly fails to give the intended truth conditions of the sentence 'Otto is ever grateful.' A similar problem is faced by (15), where the meta-language occurrence of 'may' still has the dominant reading *am allowed to*, while the object-language occurrence still has the dominant reading *might*.

Klagge's use of (CS´´) was only applied to Hintikka's 'any' case, and even if his paraphrase is regarded as successful at weakening that one counterexample, it is not successful at resolving (14) or (15), and the prospects for resolving the remaining counterexamples uniformly do not look promising. The counterexamples are sufficiently different from one another that the existence of a universal paraphrase that resolves all of them seems quite unlikely.

But there is a deeper problem with this approach to resolving the counterexamples. By taking each component conditional in (ES) and paraphrasing it in any of the above-mentioned ways (or in any of the standard alternative phrasings), one no longer has a version of (ES) that is worth including in one's semantic theory for its intended purpose. To make this clearer, since it would require too much unnecessary detail to consider every possible paraphrase, we'll look at the strongest contender – the paraphrase (CS´´), which uses a disjunction that avoids any use of the word "if." Consider the following sentence, which is analogous to (14):

(16) It is not the case that snow is white, or 'Snow is white' is true.

If one replaces (ES) with a conjunction of (CS´´) and its counterpart disjunction (for the other direction of the biconditional), then any theory that is required to entail all substitution instances of (ES) would have to entail (16). But this requirement has bizarre consequences – (16) is *trivially* true, and not for the reasons that a deflationist might give. At the risk of abusing terminology, (16) is a version of the law of excluded middle, expressed partly in the object-language and partly in the meta-language. If we wish to adhere to some version of Convention T where each component conditional in the equivalence schema is replaced with a disjunction of the form of (CS´´) or its counterpart disjunction, we are then forced to say that our truth predicate is adequate so long as (16) and its equally trivial counterpart are satisfied. But since (16) will be satisfied *no matter what color snow is*, it hardly merits a place in any discussion of the truth conditions of 'Snow is white.' Hence, the truth of substitution instances of (CS´´) really tells us nothing about the truth conditions of the substituted sentences. This 'solution' to Hintikka-type counterexamples strips Convention T of whatever explanatory power it had to begin with.

### *4.4 Dispensing with disquotation*

Lepore and Ludwig (2005) acknowledged that context-sensitive sentences *are* a problem for disquotational versions of Convention T, but they intimate that other versions of Convention T, presumably propositional versions, will not be subject to counterexamples generated by context-sensitivity. In a footnote following their presentation of a variant of the 'any' counterexample, they say:

> The criticism of using truth theories in Tarski's style in application to natural languages fails once we move away from the disquotational paradigm, which is required in any case to accommodate context sensitive sentences.[41].

While this is the only proposal that fully (and correctly) acknowledges that the underlying issue is context sensitivity, I think it nevertheless overestimates the possible sweep of a propositional solution to the counterexamples, possibly because it also underestimates the variety and sheer number of the counterexamples.

Let's take, for example, Horwich's propositional statement of (ES) (1990: 7):

(P) It is true *that p* if and only if *p.*

While this is only one way of formulating a propositional version of (ES), it is typical of its class, and it is sufficient to illustrate the problem with this attempt at a solution. It must be noticed that even though quotation marks are absent in (P), the schema *is still formulated mostly in a natural language*. Consequently, use of this schema (or any of its variants) will not entirely eliminate interaction of the object language *p* with the syntax of (P) itself. To make this clearer, consider that (P) still gives rise to the following Hintikka-type counterexamples:

(17*a*) It is true that Otto is ever grateful if Otto is ever grateful.

or

(17*b*) It is true that Obama should win the election if Obama should win the election.

[41] Lepore and Ludwig use the example: 'Anyone can do it' as the object-language sentence. See Lepore and Ludwig (2006: 364, ftn. 280).

The interpretations of ‘ever’ and ‘should’ in these cases are parallel to the disquotational versions. In (17*a*) the left-hand (propositional) ‘ever’ is still read as *always*, while the right-hand occurrence still has the dominant interpretation *at any particular time*. Similarly, in (17*b*) the left-hand ‘should’ still reads as *ought to*, while the right-hand occurrence still reads as *happens to*.

Thus, even if it should turn out that moving to a propositional paradigm ends up resolving some problematic cases, since all variants of (P) will involve embedding one natural-language sentence within a mostly-natural-language schema, it is extremely unlikely that even any variant of (P) will resolve all Hintikka-type counterexamples. Contrary to Lepore and Ludwig’s claim, it does not appear that resorting to a propositional version of the equivalence schema rather than a disquotational version will rid us of all Hintikka-type cases.

## 5. Discussion

The difficulty of finding any uniform resolution of the counterexamples should give us pause for thought. Although Tarski seemed to think that attempting a precise semantic theory for natural languages was a thankless enterprise, Hintikka has not been quite as pessimistic. In fact, he said:

> Of course the truth-conditions of a complex sentence must hang together in some specifiable way with its structure and with its component expressions. However, there is no reason to expect that this dependence be so simple as to make a recursive truth-characterization possible or to make possible the kind of use of Convention T Davidson envisages. (Hintikka 1976b: 66)

Thus, it was not natural language semantics in general that was Hintikka’s target, but rather the attempt to transfer to natural languages the precise methods Tarski had applied so successfully to

formal languages. By considering the additional counterexamples provided in (2a) through (7) above, it should now be clear that Hintikka's criticism is even more forceful than he may have thought. For many syntactic categories of English, there is an unknown but probably large number of lexical items *within* each of those syntactic categories that interact with their grammatical environments. Since we can't be sure *a priori* that a given sentence will not produce a false instance of the equivalence schema, Hintikka argued that Convention T should not be the centerpiece of a compositional semantics for natural languages:

> ...the very intended use of [ES] as a systematic tool in the semantics of natural languages was to employ it as a means of spelling out what the truth-conditions of complex sentences are, i.e., how their truth-value depends on the truth-values of their parts. The behavior of 'any' which my counterexample illustrates shows that no such systematic dependence can obtain in general, for no definite truth-value can be assigned to a subordinate any-clause independently of its (verbal) context. For it is part and parcel of the meaning of 'any' that it interacts with its context. Hence in a modified (and deeper) sense my counterexample would survive even the formalization of the Tarski condition ... (Hintikka 1976*a*: 111).

We can now say that similar remarks would apply to many other lexical items in English, and many additional counterexamples such as those presented in Section 3 above.

While Convention T and the equivalence schema have been successfully applied to formalized languages, where the interpretations of terms can be specified precisely, and the truth of complex expressions can be defined in terms of the more basic property of satisfaction, attempts to transfer this success to natural languages have come up against serious limitations. Some of these are now well-known: *i*) the existence of the semantic paradoxes, *ii*) the existence of propositions with indeterminate truth values, and *iii*) the existence of sentences with indexicals. I have argued that in contrast to these well-known problem cases, Hintikka-type counterexamples are even *more* pernicious problems for the would-be Davidsonian or Deflationist. What separates Hintikka-type counterexamples from the more widely-discussed

types of counterexamples is precisely that the object-language sentence, when read in isolation, is neither ambiguous (as cases with indexicals can be) nor 'problematic,' (as are the cases with indeterminate truth values and those generating liar-type paradoxes). Rather, as Hintikka claimed about his original example, it seems clear that the problem in all of the counterexamples is that by substituting an otherwise *un*ambiguous and *un*problematic English sentence into a T-Conditional, one gets either a false sentence, or a shift in meaning that subverts the use of the resulting Tarski biconditional to provide truth conditions for the substituted sentence. Either way, the proposal to use the equivalence schema and Convention T as the *foundation* for a semantics of natural languages is seriously undermined.

More clearly than any of the other known difficulties for Convention T, Hintikka-type counterexamples show that trouble arises not only from particular features of substituted object-language sentences, but *from the use of Convention T and from the form of the equivalence schema themselves*. This is clearly a difficulty for the Davidsonian, but it is an even greater difficulty for the Deflationist. A common claim of deflationists is that "uncontroversial instances" of the equivalence schema capture essentially all that there is to say about truth. But in light of the many Hintikka-type counterexamples, in addition to the liar cases, indeterminate cases, and cases with indexicals, it begins to appear that the expression "uncontroversial instances" is just a convenient disclaimer that enables deflationists to avoid having to explain what are actually a very large number of anomalies for their theory. Regardless of what criteria one emphasizes when choosing among competing theories, when a theory has many exceptions, of a wide variety of types, that's a fairly good indication that the theory is just *wrong*.

Of course, deflationists are right to think that there is a sense in which (ES) is "simple," but it in no way follows from that observation that truth itself must be simple. In fact, the variety of

sentence types that generate falsifications of such a seemingly innocuous schema as (ES) tells us, I think, quite the opposite – that truth is indeed something much more mysterious than the "uncontroversial instances" of the equivalence schema alone might lead us to believe. After all, virtually any concept can be made to *look* simple by limiting the relevant data to the simplest cases at the outset. But the hallmark of a good theory is that it explains the *difficult* cases. It seems, though, that the deceptive simplicity of the equivalence schema, and the fact that it does yield true instances in a large majority of cases, have obscured the very substantial differences between formal and natural languages, and have misled many into believing that truth itself must also be simple. But the variety of false substitution instances of (ES) tells us that nothing could be further from the truth.

## REFERENCES


Armour-Garb, B. P. and Beall, J.C. (2005) 'Deflationism: the basics,' in B. P. Armour-Garb and J.C. Beall (Eds.), *Deflationary Truth* (pp. 1-29). Chicago and La Salle, IL: Open Court.

Carlson, G. (1980) 'Polarity *any* is existential,' *Linguistic Inquiry*, 11, 799-804.

Davidson, D. (1967) 'Truth and meaning,' *Synthese*, 17, 304-23.

_____ (1969) 'True to the facts,' *Journal of Philosophy*, 66, 748-64.

_____ (1973) 'In Defense of Convention T,' in H. Leblanc (Ed.), *Truth, Syntax and Modality* (pp. 76-86). Amsterdam: North-Holland.

_____ (1999) 'The centrality of truth,' in J. Peregrin (Ed.), *Truth and its Nature (if any)* (pp. 105-15) Dordrecht: Kluwer.

Dummett, M. (1959) 'Truth,' *Proceedings of the Aristotelian Society*, n.s., LIX, 141-62.

Feferman, A.B. and Feferman, S. (2004). *Alfred Tarski: Life and Logic*. Cambridge: Cambridge University Press.

Field, H. (1972) 'Tarski's theory of truth,' *Journal of Philosophy* 69: 347-75.

_____ (1986) 'The deflationary concept of truth,' in G. MacDonald and C. Wright (Eds.), *Fact, Science, and Morality*, XXX-XXX. Oxford: Basil Blackwell.

Higginbotham, J. (1982) 'Comments on Hintikka's paper,' *Notre Dame Journal of Formal Logic*, 23, 263-71.

Hintikka, J. (1976*a*) 'A counterexample to Tarski-type truth-definitions as applied to natural languages,' in A. Kasher (Ed.), *Language in Focus* (pp. 107-12). Dordrecht: Reidel.

_____ (1976*b*) 'The prospects for Convention T,' *Dialectica*, 30, 61-66.

_____ (1977) 'Quantifiers in natural language: Some logical problems II,' *Linguistics and Philosophy*, 1, 153-72.

_____ (1980) 'On the *any*-thesis and the methodology of linguistics,' *Linguistics and Philosophy*, 4, 101-22.

_____ (1991) 'Defining truth, the whole truth, and nothing but the truth,' in J. Hintikka (Ed.), *Lingua Universalis vs. Calculus Ratiocinator (Selected Papers 2)* (pp. XXXX). Dordrecht: Kluwer.

_____ (1997) 'No scope for scope?,' *Linguistics and Philosophy*, 20, 515-44.

Horwich, P. (1990). *Truth*. Oxford: Basil Blackwell.

_____ (1999) 'Deflationary truth, aboutness and meaning,' in J. Peregrin (Ed.), *Truth and its Nature (if any)* (pp.163-71). Dordrecht: Kluwer.

Kirkham, R.L. (1992) *Theories of Truth*. Cambridge, MA: MIT Press.

Klagge, J. C. (1977) 'Convention T regained,' *Philosophical Studies*, 32, 377-81.

Lepore, E. and Ludwig, K. (2005) *Donald Davidson: Meaning, Truth, Language, and Reality*. New York: Oxford University Press.

Montague, R. (1970a) 'English as a formal language,' in R. Thomason (Ed.) (1974), *Formal Philosophy: Selected Papers of Richard Montague*. New Haven, CT: Yale University Press, pp. 188-221.

_____ (1970b) 'Universal Grammar,' in R. Thomason (Ed.), pp. 222-46.

Peregrin, J. (Ed.) (1999a) *Truth and its Nature (if any)*. Dordrecht: Reidel.

_____ (1999b) 'Tarski's legacy,' in J. Peregrin (Ed.), (pp. vii-xvii).

Tarski, A. (1956) 'The concept of truth in formalized languages,' trans. from the German (1935) by J. H. Woodger. In J. Corcoran (Ed.) *Logic, Semantics, Metamathematics*, 2nd edition, (pp. 152-278). Indianapolis, IN: Hackett. Original in Polish (1933).

_____ (1944) 'The semantic conception of truth,' *Philosophy and Phenomenological Research*, 4, 341-76.

Williams, Michael (1999) 'Meaning and Deflationary Truth,' *Journal of Philosophy*, 96, 545-64.

# ABSTRACT OF ESSAY 3

ELEMENTARY FORMULAS FOR THE *n*TH PRIME AND FOR THE NUMBER OF PRIMES UP TO A GIVEN LIMIT

By

Rolf Henry Johansson

Doctor of Philosophy in Social Science – Mathematical Behavioral Sciences

University of California, Irvine, 2014

Professor Louis Narens, Chair

The existence of a formula for the $n^{th}$ prime $p_n$ was for a long time considered impossible (that is, a formula by which one could calculate the $n^{th}$ prime using $n$ as input). Prior to the twentieth century, the closest thing to such a formula was Euler's well-known polynomial that generates a somewhat long sequence of primes for its early inputs. By the mid-twentieth century, several formulas had been discovered that did represent only primes, but none of them enabled one to generate all of the primes.

The first formula that enables one (in principle) to generate all and only primes was discovered by Srinivasan (1961), and a related formula was presented by Ghandi (1971), but these formulas are disappointing because they rely on an inclusion-exclusion process which requires the computation of $2^{n-1}$ terms. Other formulas by Willans (1964), Jones (1975), and Hardy and Wright (1979) made essential use of Wilson's famous theorem which effectively provides a "definition" of primality: *p is prime iff* $(p-1)! \equiv -1 (mod\ p)$. These formulas are also impractical for computation because they require the computation of factorials of $n$.

In this paper I present elementary formulas for the $n^{th}$ prime and for the number of primes up to a given limit which improve upon existing formulas by avoiding both the computation of

factorials and the exponential growth of terms. The formulas are based on the idea of "embedding" characteristic functions – a characteristic function for non-divisibility is used to construct a characteristic function for primality, and no use is made of either Wilson's theorem or the inclusion-exclusion process. The resulting formulas, though not in principle superior to the sieve of Eratosthenes as tools for generating primes, nevertheless provide an elementary, compact expression for the primes that is not computationally intractable.

## 1. INTRODUCTION

The prime numbers have the interesting property of being precisely definable, while appearing to be distributed without any discernible pattern. On closer inspection, however, their distribution does reveal structure. The best known theorem concerning their distribution is the prime number theorem, which states that the number of primes less than or equal to $n$ (or "$\pi(n)$") is asymptotically equal to $n / \log n$. Although asymptotic results are certainly of interest, and can be useful for various applications, they are not generally as informative as constructive procedures that enable us to find the $n^{th}$ member of a sequence. Ideally, what we would like for any sequence $S$ is a generating function expressed by a formula: $f(n) = y$, where $y$ is the $n^{th}$ member of the sequence $S$. The existence of such a formula expressing a function that generates the primes was for a long time thought unlikely, if not impossible. In the first edition of their well-known book on number theory, Hardy and Wright (1938) put the problem as follows:

> *Is there a simple general formula for the $n^{th}$ prime $p_n$* (a formula, that is to say, by which we can calculate the value of $p_n$ for any given $n$ with less labour than by the use of the sieve of Eratosthenes)? No such formula is known and it is unlikely that such a formula is possible.

The belief that such a function is not possible was more or less the consensus at least through the early part of the 20th century. One reason for the skepticism was no doubt that unlike, e.g., the sequence of multiples of a given number, which can be simply expressed by either their multiplicative or divisibility properties with respect to that given number, the primes are defined by their *non*-divisibility by *any* number *other than* themselves and 1, and there does not appear to be any simple function expressing the more complex property "is *not* divisible by *any* number *other than* …." Nevertheless, there was some progress in the 1940s, and it is of some interest to

see how Hardy and Wright changed their position in subsequent editions of their text. In the fifth edition, the following passage was inserted immediately after the above quoted passage:

> On the other hand, it is possible to devise a number of 'formulae' for $p_n$. Of these, some are no more than curiosities since they define $p_n$ in terms of itself, and no previously unknown $p_n$ can be calculated from them. ... Others would in theory enable us to calculate $p_n$, but only at the cost of substantially more labour than does the sieve of Eratosthenes. Others still are essentially equivalent to that sieve (1979: 5-6).

This new paragraph, appearing for the first time in the fifth edition of 1979, is a reminder of how recently we have come to know of the existence of formulas for primes. Most formulas for primes that have been discovered thus far fall into Hardy and Wright's first category: formulas that require knowledge of the $n^{th}$ prime in order to compute the $n^{th}$ prime. These are disappointing for obvious reasons. The earliest such formulas had an exponential form, and although none of them yield *all* of the primes, they do yield *only* primes, and for that reason are appropriately called "prime representing functions." We will give examples of such formulas in the next section.

The second largest category of formulas consists of those that would "in theory" enable us to calculate $p_n$. Here Hardy and Wright probably had in mind several formulas based on Wilson's theorem that require the computation of factorials. These formulas are far superior to the first group in that several of them *do* yield all and only the prime numbers. But this comes at a great price – in all such formulas, the computation of $p_n$ requires the computation of $n!$, so these formulas are useless for computing any but the very smallest primes. Nevertheless, these formulas are of interest because they do represent all and only primes. In the third section we will discuss Willans' formula, the first such formula to be discovered.

The last group of formulas considered by Hardy and Wright were those that are "essentially equivalent" to the sieve of Eratosthenes, although Hardy and Wright's use of "essentially equivalent" was rather liberal; in practice, these formulas are still computationally intractable for large primes. The best-known example of this kind is Gandhi's (1971) formula, even though a similar, simpler formula had been discovered prior to Gandhi by Srinivasan in (1961). These formulas are inferior to Willans' formula from an epistemological standpoint, because the computation of the $n^{th}$ prime requires knowledge of all prior primes, which Willans' formula does not. In addition, the procedure for eliminating multiples of primes (i.e., what makes them "essentially equivalent" to Eratosthenes' sieve) employs an inclusion-exclusion computation, and this requires the computation $2^{n-1}$ terms. We will present these formulas in Section 4.

The last two categories of formulas suffer from computation problems – either the computation of factorials of *n* or the computation of an exponentially growing number of terms. For any but the very smallest primes, these are prohibitive computational difficulties. All three categories of formulas are disappointing if we are hoping to find a function that generates primes in a simple and efficient manner, and captures the *definition* of the primes in some intuitive sense.

In this paper, I will present formulas for $\pi(n)$ and for the $n^{th}$ prime that improve upon the existing formulas both in computational efficiency *and* in capturing a definition of primality. They still do not fully answer Hardy and Wright's original question, because they rely on a sifting method "essentially" equivalent to Eratosthenes' sieve (although in this case, an equivalence much closer to Eratosthenes). But unlike earlier formulas, they do not involve the computation of factorials or the exponential growth of terms. Hence, it is conceivable that they could compete with algorithms used to generate large primes, but an empirical test of this

conjecture is beyond the scope of the present paper. We will discuss these formulas after giving a historical survey.

## 2. A Brief History of Prime Representing Functions

In 1772, long before any formulas for primes had been discovered, Euler discovered a now-famous polynomial with integral coefficients that yields a long string of primes for its early inputs: $f(n) = n^2 + n + 41$. This yields primes for all values where $n = 0,1,\ldots,39$, but then $f(40) = 1681 = 41^2$. Others have produced variants of this polynomial that yield the same outputs, each output produced twice or more times, or in reverse order. Still others have found other quadratics that yield distinct primes for even more initial values of *x*, the best so far being $f(n) = 36n^2 - 810n + 2753$, which yields primes for $n = 0,1,\ldots,44$. The interested reader may consult Dickson's book for references (1952:420), and a more recent survey of these results by Boston and Greenwood (1995).

In spite of the existence of quadratics of this sort, it is well known that no polynomial $f(n)$ with integral coefficients can yield *only* primes unless it is constant, and therefore improper. Since this fact considerably constrains the possible forms that any prime-generating function can take, it will be instructive to go through its proof. We'll show that any nonconstant polynomial that yields as a value even a single prime number will also yield infinitely many composite numbers as values. Let $f(n) = a_0 n^k + a_1 n^{k-1} + \cdots + a_k$ be any nonconstant polynomial (i.e., $k \geq 1$), and suppose that for some natural number *r*, $f(r) = p$, where *p* is a prime. Then obviously $p \mid f(r)$, but from this it is easy to show that *p* must also divide many additional

values of $f(n)$ (which will all thereby be composite), in particular, $p \mid f(r+mp)$ for all $m = 1, 2, \ldots$. This can be seen by looking at the binomial expansion of the terms:

$$f(r+mp) = a_0(r+mp)^k + a_1(r+mp)^{k-1} + \cdots + a_k$$

$$= a_0 r^k + a_0\left[\binom{k}{1} r^{k-1}(mp) + \binom{k}{2} r^{k-2}(mp)^2 + \cdots + \binom{k}{k-1} r(mp)^{k-1} + (mp)^k\right] +$$

$$a_1 r^{k-1} + a_1\left[\binom{k}{1} r^{k-2}(mp) + \binom{k}{2} r^{k-3}(mp)^2 + \cdots + \binom{k}{k-2} r(mp)^{k-2} + (mp)^{k-1}\right] + \cdots + a_k .$$

Remembering that $p \mid f(r)$ (i.e., $p \mid (a_0 r^k + a_1 r^{k-1} + \cdots + a_k)$), if we rearrange terms in the binomial expansion by placing all of the terms *not* involving *p* first, then since *p* appears as a factor in all remaining terms, *p* also divides each of those terms, and thus $p \mid f(r+mp)$. Since $m \geq 1$, then although there may be one *m* where $f(r+mp) = 0$, $f(r+mp)$ will be composite for all other values of *m*. Thus, any non-constant polynomial that yields even one prime will also yield infinitely many composite numbers.

In 1943, Reiner showed a slightly more general result. Defining a *prime-representing function* to be any function $f(x)$ that yields a prime for every positive integral value of *x*, he showed the following:

THEOREM (REINER): *If* $f_i(x), g_i(x) \quad (i = 1, \ldots, n)$ *are polynomials with integral coefficients and positive leading coefficients, the following is not a prime-representing function:*

$$f(x) = \sum_{i=1}^{n} f_i(x)^{g_i(x)} .$$

Buck (1946) generalized this negative result further, and showed that no nonconstant rational function can be a prime-representing function. These negative results dashed any hopes that multiplication or division of polynomials might produce a prime-representing function.

Buck further conjectured that "no simple function, finitely expressible" can be a prime-representing function. Similar claims had been made by other mathematicians, but they were all shown to be wrong when Mills (1947) presented the first function, expressible as a simple formula, that always represents primes:

THEOREM (MILLS): *There is a real number* $\theta$ *such that* $\left\lfloor \theta^{3^x} \right\rfloor$ *is a prime-representing function.*

Here "$\lfloor x \rfloor$" denotes the greatest *integer* $\leq x$, and $\theta$ is a number roughly equal to 1.3064.... This result depends crucially on a prime gap result of Ingham (1937). Mills' paper inspired many others to generalize the result in different ways (see Kuipers (1950), Wright (1951), and Niven (1951) for the first extensions, Ore (1952) and Wright (1954) for the most general results, and both Dudley (1969) and Ribenboim (1989) for a history of Mills-type functions).

Unfortunately, all prime-representing functions of this type suffer from several defects. On the one hand, although they do yield *only* primes, none of them yields *all* primes. In fact, they

all fall considerably short of this goal. Consider Wright's (1951) exponential prime-representing function:

$g(n) = \left\lfloor 2^{2^{2^{\cdot^{\cdot^{\cdot^{2^{\alpha}}}}}}} \right\rfloor$ is prime for every $n \geq 1$ (where there is a string of *n* exponents, the first of which is $\alpha$).

The square brackets again denote the greatest integer function, and all square brackets in all subsequent formulas in this paper are also to be interpreted as this function. Here $\alpha$ may have many possible values (see Wright (1959) for the details). To get an idea of the size of the gaps between outputs of this function, consider that if $\alpha = 1.9287800\ldots$, then $g(1) = 3$, $g(2) = 13$, $g(3) = 16381$, and $g(4)$ has approximately 5000 digits! Of course, the number of exponents here quickly expands the gaps between the output primes, but all exponential prime-representing functions generate a similar-looking sequence. Very early on, the functions generate enormous numbers. A further problem is that in Mills' and Wright's formulas, the computations of $\theta$ and $\alpha$, respectively, are complicated and require knowledge of the $n^{th}$ prime to compute the $n^{th}$ prime. This is a defect in all Mills-type functions. Moreover, these functions clearly don't capture the definition of primality in any intuitive sense.

There have been many more papers written on quadratics that have prime-rich intervals than there have been on prime-representing functions. Part of the reason for this, of course, is that prior to Mills' paper of 1947, there simply were no known prime-representing functions, so until then Euler-type quadratics were the next best thing. But even though Mills' function and its extensions succeeded where some thought it was not possible, they are still not of any use in

computing the $n^{th}$ prime, and they don't come anywhere close to defining the primes in any interesting sense. These early results can be considered, as Hardy and Wright put it, mere "curiosities."

## 3. Formulas Based on Wilson's Theorem

Before discussing the first true formulas for primes, we should consider what form we might reasonably expect any formula for primes to have, given the negative results concerning the use of polynomials and rational functions. What property of primality should appear as most salient in any such formula? A reasonable place to begin would be with Wilson's theorem, which gives necessary and sufficient conditions for primality: [42]

Theorem (Wilson): *p is prime iff* $(p-1)! \equiv -1 \pmod{p}$.

This theorem has been used by several authors to express formulas for primes. The first such formula was presented by Willans in 1964. His formula employs the following characteristic function for primes:

$$f(j) = \left[ \cos^2 \pi \frac{(j-1)! + 1}{j} \right] = \begin{cases} 1 \text{ if } j \text{ is prime or } j = 1 \\ 0 \text{ if } j \text{ is composite} \end{cases}$$

To see how this function works as a characteristic function for primality, we need only consider immediate implications of Wilson's theorem. When $j$ is prime or $j = 1$, then by Wilson's Theorem $\frac{(j-1)! + 1}{j}$ is an integer, and thus $\cos^2 \pi \frac{(j-1)! + 1}{j} = 1$ and $[1] = 1$ (remember,

[42] "Wilson's" theorem was actually first conjectured by Leibniz, and first proved by Lagrange. See Dudley (1978: 43) for the historical details, and p. 46-47 of the same book for an elementary proof.

"[$x$]" is the greatest *integer* $\leq x$). When $j$ is not prime, then by Wilson's Theorem $\frac{(j-1)!+1}{j}$ is not an integer, and thus $\cos^2 \pi \frac{(j-1)!+1}{j} = b$, where $0 < b < 1$, and thus $[b] = 0$. By means of a characteristic function for primes, one may easily construct a formula for the number of primes up to and including $m$ (i.e., "$\pi(m)$") by simply adding successive outputs of the characteristic function. For example, Willans provided the following formula for $\pi(m)$:

$$\pi(m) = -1 + \sum_{j=1}^{m} f(j).$$

By exploiting in addition some elementary properties of the $n^{\text{th}}$ root function, he was able to construct a formula for primes (again, "[$k$]" is the greatest *integer* $\leq k$):

$$p_n = 1 + \sum_{m=1}^{2^n} \left[ \left( \frac{n}{1 + \pi(m)} \right)^{\frac{1}{n}} \right].$$

This formula generates all and only the primes by exploiting the following idea: for each positive integer input $n$, it finds the $n^{\text{th}}$ prime $p_n$ by simply adding 1 for every positive integer $m$ up to $m = p_n - 1$, at which point the 1 at the beginning of the formula brings the sum to $p_n$, the desired $n^{\text{th}}$ prime. For all values of $m$ greater than $m = p_n - 1$, the greatest integer function simply yields an extraneous zero, hence the sum will remain $p_n$ up through $m = 2^n$.

To see how the greatest integer function yields 1 only up to $m = p_n - 1$, and 0 thereafter, one need only consider the following two properties of the $n^{\text{th}}$ root function, which are easily proved (for all positive integers $n$ and all positive real $k$):

(1) $\left[k^{\frac{1}{n}}\right] = 1$ when $1 \le k \le n$

(2) $\left[k^{\frac{1}{n}}\right] = 0$ when $0 \le k < 1$

Letting $k = \dfrac{n}{1+\pi(m)}$, consider the computation of the $n^{\text{th}}$ prime $p_n$ using Willans' formula. To generate this prime, one takes the sum of the values of the greatest integer function for all integers $m$ from 1 to $2^n$ (The reason for this upper limit is discussed below). For $m = 1$, $\pi(m) = 0$, so $k = n$ and thus by property (1), $\left[k^{\frac{1}{n}}\right] = 1$. As $m$ increases, $\pi(m)$ either increases or remains constant at each successive $m$ (depending on whether $m$ is prime or composite, respectively), and since $n$ is fixed, the fraction $k$ either decreases or remains constant. By property (1), for all values of $m$ where the fraction $k$ decreases from $n$ down to and including 1, the output of $\left[k^{\frac{1}{n}}\right]$ is 1. The last $m$ for which $\left[k^{\frac{1}{n}}\right]$ is 1 will be $m = p_n - 1$, because that will be the last $m$ such that $\pi(m) = n - 1$, and thus it will be the last $m$ where the fraction $k = 1$. Then, for all $m \ge p_n$, $\pi(m) \ge n$, and hence the fraction $k$ is such that $0 < k < 1$, and thus by property (2), $\left[k^{\frac{1}{n}}\right] = 0$.

By disregarding the extraneous zeroes in the sum, we can more clearly see what Willans' formula does by writing it as follows:

$$p_n = 1 + \sum_{m=1}^{p_n - 1} 1.$$

Notice that though Willans' formula does require using the $n^{\text{th}}$ prime as an input, it does not require knowledge that the $n^{\text{th}}$ prime is prime. The reason for using the upper limit of $2^n$ is that

we want to be assured of yielding the $n^{th}$ prime as output, which in turn requires that we take the sum at least to $m = p_n - 1$. The exact limit is derived from a well-known prime gap result which states that $p_n \leq 2^n$ (for all $n > 0$).[43]

The reader may notice that when written explicitly (i.e., without abbreviating $\pi(m)$ and $f(j)$), Willans' formula is rather cumbersome. Nevertheless, it accomplishes much more than the earlier Mills-type prime-representing functions, because it does yield all and only the primes in their usual order. From a computational standpoint, the main difficulty is the computation of $(j-1)!$, which is prohibitive except for very small $j$. For example, to compute the 20th prime (i.e., 71) using Willans' formula, one must compute $(2^{20})!$ . Other formulas for primes have also been based on Wilson's theorem, for example, those by Jones (1975), Papadimitriou (1975), and Hardy and Wright (1979: 414), and all of these require a similar factorial computation.

## 4. Formulas Based on the Sieve of Eratosthenes

The last class of formulas mentioned by Hardy and Wright consists of those that are "essentially equivalent" to Eratosthenes' sieve. Formulas of this type have been formulated by Srinivasan (1961) and Gandhi (1971).[44] Unfortunately, these formulas also suffer from a computational problem: they depend on an inclusion-exclusion process, and hence require the computation of an exponentially growing number of terms. In addition, unlike Willans' formula, which does not require knowledge of any particular primes, these formulas are recursive and require knowledge of all primes $p_1 \ldots p_{n-1}$ in order to compute the $n^{th}$ prime $p_n$. The

[43] See Hardy and Wright (1979) pgs. 17 and 414 for details. As Hardy and Wright point out, this upper limit follows from Bertrand's Postulate that there is always at least one prime between $n$ and $2n$. (more precisely: for all positive integers $n$, there is a prime $p$ such that $n < p \leq 2n$). This upper limit can be substantially reduced for large $n$ by using better prime gap results, but this is unimportant for our purposes.

[44] It is of some historical interest to note that it was Srinivasan who discovered the first true formula for primes, even though Willans' and Gandhi's formulas have received much more attention. Srinivasan's formula is also a bit simpler than Gandhi's in that it does not involve the use of logarithms.

recursiveness, however, does not itself lengthen the computations beyond what is required in the use of Willans' formula, and for some applications can even shorten the computation (if one knows the $n^{th}$ prime and merely wishes to find the $(n+1)^{st}$ prime, for example).

The core idea in these formulas is the sieve of Eratosthenes, which isolates primes by eliminating all multiples of primes other than the primes themselves (i.e., by eliminating all composite numbers). The remaining numbers will be the primes as well as the number 1 itself, and those eliminated will be those divisible by some prime $p$ or some product of primes $p_i \cdots p_k$. Many composite numbers in Eratosthenes' sieve are eliminated more than once, since they are multiples of more than one prime, so in order to get an exact count of the total number of primes less than or equal to $x$ ("$\pi(x)$"), one uses an "inclusion-exclusion" process. This process works as follows: one begins with $x$, and then counts the total number of multiples of each prime less than or equal to $x$ (i.e., all multiples of 2, *plus* all multiples of 3, etc.). Then one subtracts (i.e., "excludes") the sum of all multiples of all primes, as well as the number 1 itself. Some of the numbers in this total count of multiples were counted more than once, since they were multiples of more than one prime (e.g., 6 was counted twice, as a multiple of 2 and as a multiple of 3; 30 was counted 3 times, as a multiple of 2, 3, and 5, etc.) Thus, to assure that each composite number is counted exactly once, we compensate for this overcount by adding back (i.e., "including") the total from the previous count of products of *two* distinct primes, as well as the total number of primes less than or equal to $\sqrt{x}$, (since all primes divide themselves, and hence that total was also included in the last count). At this point, the number of composites divisible by either one or two distinct primes has been counted exactly, but those divisible by more than two primes have not been counted only once. Thus, one compensates again by subtracting the number of numbers divisible by a product of *three* distinct primes (since those

numbers were counted three times in each of the last two steps), etc. This process can be expressed in a formula as follows (where $p_1 < p_2 < \cdots < p_m$ are all primes less than or equal to $\sqrt{x}$):

$$\pi(x) = [x] + \pi(\sqrt{x}) - 1 - \sum_i \left[\frac{x}{p_i}\right] + \sum_{i<j} \left[\frac{x}{p_i \cdot p_j}\right] - \sum_{i<j<k} \left[\frac{x}{p_i \cdot p_j \cdot p_k}\right] + \cdots.$$

(where $p_i$ , $p_j$ , $p_k$ , run over the primes $p_1$ , $p_2$ , … , $p_m$). This formula, first published in a more general form by Legendre, is now usually expressed more compactly using the möbius function[45] as follows:

$$\pi(x) = -1 + \pi(\sqrt{x}) + \sum_{d \mid p_1 \cdot p_2 \cdots p_m} \mu(d) \left[\frac{x}{d}\right]$$

(where $d$ runs over all divisors of the product $p_1 \cdot p_2 \cdots p_m$ of all primes less than or equal to $\sqrt{x}$).[46] We'll hereafter refer to the latter formula as "Legendre's formula."

Unlike the prime number theorem, which gives an asymptotic value for $\pi(x)$, Legendre's formula gives a precise count of the number of primes less than or equal to $x$. However, this precision comes at a great cost. Although the idea behind Eratosthenes' sieve is quite simple, the computation requires taking a sum over all divisors $d$ of the product $p_1 \cdot p_2 \cdots p_m$ (where $m = \pi(\sqrt{x})$), hence there will be $2^m$ terms in this sum. This follows from the basic combinatorial identity:

---

[45] The möbius function is defined as follows: $\mu(1) = 1$, $\mu(n) = (-1)^r$ if $n$ is the product of $r$ distinct primes, and $\mu(n) = 0$ if $n$ is divisible by any prime to a power higher than 1. For example, $\mu(p) = -1$ for every prime $p$; $\mu(p_1 \cdot p_2) = 1$ for any product of two distinct primes, and $\mu(p^2) = \mu(p^3) = \mu(p^4) = \cdots = 0$ for all higher powers of primes. What's relevant here are mainly the alternating 1's and –1's, as the divisors $d$ are products of an even or odd number of distinct prime factors, respectively.

[46] The möbius function and the inclusion-exclusion process are discussed in many number theory texts, but an especially clear discussion relevant to the problem of counting primes can be found in Brauer (1946), as well as in the monographs by Halberstam & Roth (1966) and Halberstam & Richert (1974).

$$\sum_{r=0}^{m}\binom{m}{r}=2^m$$

which is easily proved by induction. In Legendre's formula, we are "choosing" divisors as "combinations" of primes.[47]

In spite of this computational difficulty, the use of the möbius function to count primes does make possible various compact expressions for the the $n^{\text{th}}$ prime $p_n$. The first such formula using this idea; indeed, the very first true formula for the $n^{\text{th}}$ prime was presented by Srinivasan (1961):

$$p_n=\left\lfloor\frac{\displaystyle\sum_{d\mid p_1\cdots p_{n-1}}\frac{\mu(d)\,d\,2^d}{(2^d-1)^2}-\frac{1}{2}}{\displaystyle\sum_{d\mid p_1\cdots p_{n-1}}\frac{\mu(d)}{2^d-1}-\frac{1}{2}}\right\rfloor$$

A decade later, Gandhi (1971) provided a similar formula using logarithms:

$$p_n=\left\lfloor 1-\log_2\left(-\frac{1}{2}+\sum_{d\mid p_1\cdots p_{n-1}}\frac{\mu(d)}{2^d-1}\right)\right\rfloor$$

Proofs of Gandhi's formula have been given by Vanden Eynden (1972) and Golomb (1974), and a sample computation can be found in Grosswald (1984). Both Srinivasan's and Gandhi's formulas were generalized by Namboodiripad (1971).

---

[47] For example, for $m=\pi(\sqrt{x})=3$,

$$\sum_{d\mid p_1\cdot p_2\cdot p_3}\mu(d)\left\lfloor\frac{x}{d}\right\rfloor=\left\lfloor\frac{x}{1}\right\rfloor-\left\lfloor\frac{x}{p_1}\right\rfloor-\left\lfloor\frac{x}{p_2}\right\rfloor-\left\lfloor\frac{x}{p_3}\right\rfloor+\left\lfloor\frac{x}{p_1\cdot p_2}\right\rfloor+\left\lfloor\frac{x}{p_1\cdot p_3}\right\rfloor+\left\lfloor\frac{x}{p_2\cdot p_3}\right\rfloor-\left\lfloor\frac{x}{p_1\cdot p_2\cdot p_3}\right\rfloor,$$

and there are $2^3$ terms in this sum.

The most difficult computations in both of these formulas are the summations over values of the möbius function $\mu(d)$, which have $2^{n-1}$ terms. The number of these terms is significantly larger than the number of terms in computations using the Legendre formula, which have $2^{\pi(\sqrt{n})}$ terms in the sum. It is this exponential growth of terms that renders these formulas impractical for computing any but the very smallest primes. In addition, in both Srinivasan's and Gandhi's formulas, computation of the $n^{\text{th}}$ prime $p_n$ requires knowledge of all primes less than or equal to $p_{n-1}$.

## 5. Elementary Formulas for Primes

I will now present a method for deriving formulas for primes that does not suffer from the computational difficulties of computing factorials or the exponential growth of terms. This approach utilizes two main ideas: *i*) the standard definition of the primes as those numbers not divisible by any numbers other than themselves and 1, and *ii*) the construction of a characteristic function for primality by taking a product of values of a characteristic function for non-divisibility. Using the latter function, we are able to construct considerably simplified formulas for primes that do not have the computational difficulties of the formulas discussed above, and have the added merit of more intuitively capturing the "definition" of the primes.

The first step is the construction of a characteristic function for the set of all numbers *m* that do *not* divide a given number *a*: (where *a*, *m* are any positive integers, and we continue to use "[*x*]" to indicate the greatest *integer* $\leq x$):

PROPOSITION 1: $C_m(a) = 1 + \left\lfloor \frac{a-1}{m} \right\rfloor - \left\lfloor \frac{a}{m} \right\rfloor = \begin{cases} 1 \ \textit{if} \ \ m \nmid a \\ 0 \ \textit{if} \ \ m \mid a \end{cases}$

*Proof*: If $m > a$, then clearly $m \nmid a$ and $\left\lfloor \frac{a-1}{m} \right\rfloor = \left\lfloor \frac{a}{m} \right\rfloor = 0$, hence $C_m(a) = 1$. If $m \le a$, we must consider two cases: *i*) *m* divides *a*, and *ii*) *m* does not divide *a*. If $m \mid a$, then for some positive *q*, $m \cdot q = a$. Then $\left\lfloor \frac{a-1}{m} \right\rfloor = \left\lfloor \frac{m \cdot q - 1}{m} \right\rfloor = \left\lfloor q - \frac{1}{m} \right\rfloor = q - 1$, and $\left\lfloor \frac{a}{m} \right\rfloor = \left\lfloor \frac{m \cdot q}{m} \right\rfloor = q$, hence $C_m(a) = 1 + (q-1) - q = 0$. If $m \nmid a$, then for some positive $q$ and remainder $r$ (where $1 \le r < m$), $m \cdot q + r = a$. It suffices to show that in this case $\left\lfloor \frac{a-1}{m} \right\rfloor$ and $\left\lfloor \frac{a}{m} \right\rfloor$ have the same values. First, $\left\lfloor \frac{a-1}{m} \right\rfloor = \left\lfloor \frac{m \cdot q + r - 1}{m} \right\rfloor = \left\lfloor q + \frac{r}{m} - \frac{1}{m} \right\rfloor$, and since $1 \le r < m$, the difference $\frac{r}{m} - \frac{1}{m}$ is either 0 (if $r = 1$), or some fraction *k* where $0 < k < 1$ (if $r > 1$). In either case, $\left\lfloor q + \frac{r}{m} - \frac{1}{m} \right\rfloor = q$, and thus $\left\lfloor \frac{a-1}{m} \right\rfloor = q$. In addition, $\left\lfloor \frac{a}{m} \right\rfloor = \left\lfloor \frac{m \cdot q + r}{m} \right\rfloor = \left\lfloor q + \frac{r}{m} \right\rfloor$, and since $0 < \frac{r}{m} < 1$, $\left\lfloor q + \frac{r}{m} \right\rfloor = q$ and thus $\left\lfloor \frac{a}{m} \right\rfloor = q$ as well. Hence, when $m \nmid a$, $C_m(a) = 1 + q - q = 1$. □

One can think of this as a characteristic function for "non-divisibility" of *a* by *m*. Since the primes are those numbers not divisible by any number other than themselves and 1, it is possible to use the above formula to check for the primality of *a* by simply taking the product of all

values of $C_p(a)$ for all primes $p \le \sqrt{a}$. Since there are no primes $\le \sqrt{2}$ or $\le \sqrt{3}$, we begin checking primality at $a = 4$, and present the following characteristic function for primes:

PROPOSITION 2: For all $a \ge 4$, $s(a) = \prod_{p \le \sqrt{a}} C_p(a) = \begin{cases} 1 & \text{if } a \text{ is prime} \\ 0 & \text{if } a \text{ is composite} \end{cases}$

*Proof*: Case 1: *a* is prime. Then for all primes $p \le \sqrt{a}$, $p \nmid a$. Thus by PROPOSITION 1, $C_p(a) = 1$ for all such $p \le \sqrt{a}$, hence the product $\prod_{p \le \sqrt{a}} C_p(a) = 1$. Case 2: *a* is composite. Then for some prime $p' \le \sqrt{a}$, $p' \mid a$. Let $p_1, p_2, \ldots, p', \ldots, p_k$ be all of the primes $\le \sqrt{a}$ ($p'$ of course needn't be unique). Then by PROPOSITION 1, $C_{p'}(a) = 0$, and thus $s(a) = \prod_{p \le \sqrt{a}} C_p(a) = C_{p_1}(a) \cdot C_{p_2}(a) \cdots 0 \cdots C_{p_k}(a) = 0$. Hence, as long as any prime divides *a*, the product $s(a) = 0$. □

The function *s* enables us to construct a very elementary formula for $\pi(x)$, which we formulate in the following:

THEOREM 1: For all $x \ge 4$, $\pi(x) = 2 + \sum_{a=4}^{x} s(a)$

*Proof*: Using this identity for an arbitrary $x \ge 4$, $\pi(x) = 2 + s(4) + s(5) + s(6) + \cdots + s(x)$, and since by PROPOSITION 2, $s(a) = 1$ if *a* is prime, and $s(a) = 0$ if *a* is composite, this function will clearly count 1 for all and only the primes $\le x$, and hence yields an exact count of $\pi(x)$ as desired. □

Written more explicitly, our formula for $\pi(x)$ is:

$$\pi(x) = 2 + \sum_{a=4}^{x} \prod_{p \le \sqrt{a}} \left( 1 + \left[ \frac{a-1}{p} \right] - \left[ \frac{a}{p} \right] \right) \qquad \text{(for all } x \ge 4\text{)}.$$

It is of interest to compare this formula to the Legendre formula of the last section. Calculations using the Legendre formula require $2^{\pi(\sqrt{x})}$ terms in the main sum, while our formula for $\pi(x)$ requires at most (# of terms in each calculation of $C_p(a)$) × (# of terms in the product for the largest *a*) × (# of terms in the sum). There are *i*) 3 terms in each calculation of $C_p(a)$, *ii*) the largest *a* in the calculation will be *x* itself, and for this integer there will be one calculation of $C_p(x)$ in the product for each of the primes $\le \sqrt{x}$, and there are exactly $\pi(\sqrt{x})$ such primes, and finally *iii*) there are clearly $x-3$ terms in the main sum (since we begin at $a = 4$). Hence the number of terms using our formula is bounded above by $3 \cdot \pi(\sqrt{x}) \cdot (x-3)$. This number does not grow anywhere near as rapidly as $2^{\pi(\sqrt{x})}$. To get an idea of the difference, while the number of primes less than or equal to the first few hundred *x* will be calculated more easily using Legendre's formula, the calculation of $\pi(5041)$ requires 1,048,576 terms (since $\sqrt{5041} = 71$, 71 is the 20th prime, and $2^{20} = 1,048,576$). By comparison, using our formula, the calculation of $\pi(5041)$ will be bounded above by at most $3 \cdot 20 \cdot (5038) = 302,280$ terms. In the calculation of $\pi(10,000)$, the Legendre formula requires more than 16,000,000 terms, while our formula in this case is still bounded above by fewer than 720,000 terms. This is still impractical for an individual working by hand, but the differences will be substantial when comparing computing times for very large *x* using a computer.

Since we now have a formula for $\pi(x)$ that is much simpler than Willans' formula for

$\pi(x)$, which was based on Wilson's theorem and required computing factorials, it is clearly possible to simplify Willans' formula for the $n^{\text{th}}$ prime $p_n$ by substituting our formula for $\pi(x)$ from THEOREM 1 into his formula for the $n^{\text{th}}$ prime, making minor adjustments for the early inputs. However, this would still have the undesirable large upper limit in his function for the $n^{\text{th}}$ prime. We can instead construct an even more elementary function for the the $n^{\text{th}}$ prime, however, by deriving a formula that enables us to calculate the prime $p_n$ in terms of the previous prime $p_{n-1}$, as was done by Srinivasan and Gandhi. Using such a recursive procedure of course requires knowledge of previous primes in a way that Willans' formula does not, but the result nevertheless improves upon the formulas of Srinivasan and Gandhi by avoiding the exponential growth of terms.

By PROPOSITION 2, for $a \geq 4$, $s(a) = 1$ if $a$ is prime and $s(a) = 0$ if $a$ is composite. So for all such $a$, $1 - s(a) = 0$ if $a$ is prime and $1 - s(a) = 1$ if $a$ is composite. We can thus use $1 - s(a)$ to add 1 for each composite $a$ between any two primes ($p \geq 4$), and this enables us to construct an elementary formula for the $n^{\text{th}}$ prime $p_n$ in terms of the previous prime $p_{n-1}$. We must assure in addition that the formula *stops* adding 1's once it has reached the $n^{\text{th}}$ prime $p_n$, and it will be necessary to add an additional 1 for the $n^{\text{th}}$ prime itself. This is captured in the following formula (where the upper limit $\theta = \min a$ such that $s(a) = 1$):

THEOREM 2: For $n > 2$, $p_n = p_{n-1} + 1 + \sum_{p_{n-1}+1 \leq a \leq \theta} (1 - s(a))$

*Proof*: The function $s$ is defined only for all $a \geq 4$, because it is checking for divisibility by primes $p \leq \sqrt{a}$, and 4 is the first number to have a prime less than or equal to its square root.

Hence we correspondingly define the formula above only for primes greater than 3, so the first such prime that can be calculated is $p_3 = 5$. It is clear that since $1 - s(a) = 1$ when *a* is composite and $1 - s(a) = 0$ when *a* is prime, then it suffices to check that the bounds on the sum are set so that the formula will add the correct number of ones. That is, to the initial prime $p_{n-1}$ the formula should add *i*) 1 for each composite number greater than $p_{n-1}$ but less than the next prime, and *ii*) 1 for the $n^{\text{th}}$ prime $p_n$ itself. When this last 1 is added, i.e., when a total of $p_n - p_{n-1}$ 1's have been added to $p_{n-1}$, the upper limit should halt the computation, and the total will clearly be equal to $p_n$. This is accomplished by setting the lower limit at one greater than the initial prime $p_{n-1}$, which is clearly the first composite number for which the formula should add 1. The upper limit $\theta$ is then defined so that the addition of terms halts at the first *a* such that $1 - s(a) = 0$; i.e., $\theta$ is the *minimum a* such that $s(a) = 1$. By definition of *s*, this number will be prime, and hence it will be the first prime after the initial prime $p_{n-1}$, or $p_n$ itself. But since $1 - s(p_n) = 0$, it is necessary to add an extra 1 before the main sum to correspond to this final prime, in order to add exactly $p_n - p_{n-1}$ 1's to the initial prime $p_{n-1}$, as desired. □

Written explicitly, our formula for $p_n$ in terms of $p_{n-1}$ is:

$$p_n = p_{n-1} + 1 + \sum_{p_{n-1}+1 \le a \le \theta} \left( 1 - \prod_{p \le \sqrt{a}} \left( 1 + \left\lfloor \frac{a-1}{p} \right\rfloor - \left\lfloor \frac{a}{p} \right\rfloor \right) \right)$$

This formula has several advantages over the recursion formulas of Srinivasan and Gandhi. The primary advantage is the avoidance of an exponential growth of terms, which arose in their formulas because of the use of the möbius function over all divisors of each of the inputs. But there is an additional simplification, in that the above formula only requires checking prime

divisors less than or equal to the *square root* of each *a*, rather than *all* divisors less than or equal to *a*, as in Srinivasan's and Gandhi's formulas. This should enable the construction of a recursive function that could be tested on a computer, but that is beyond the scope of this paper.

## REFERENCES


Boston, N. and Greenwood M.L. (1995) 'Quadratics representing primes,' *American Mathematical Monthly*, 102, 595-599.

Brauer, A. (1946) 'On the exact number of primes below a given limit,' *American Mathematical Monthly*, 53, 521-523.

Buck, R.C. (1946) 'Prime-representing functions,' *American Mathematical Monthly*, 53, 265.

Dickson, L.E. (1919) *History of the Theory of Numbers*, Volume 1. New York: Chelsea, 1952 (reprint of 1919 edition).

Dudley, U. (1969) 'History of formula for primes,' *American Mathematical Monthly*, 76, 23-28.

_____, (1978) *Elementary Number Theory*, 2nd Ed. New York: W.H. Freeman and Co.

Gandhi, J.M. (1971) 'Formulae for the nth prime,' *Proceedings of the Washington State University Conference on Number Theory*, Washington State University Press, Pullman, WA, 1971, pp. 96-106.

_____, (1975) *Mathematical Reviews*, 50, 963.

Golomb, S.W. (1974) 'A direct interpretation of Gandhi's formula,' *American Mathematical Monthly*, 81, 752-754.

Goodstein, R.L. and Wormell, C.P. (1967) 'Formulae for primes,' *The Mathematical Gazette*, 51, 35-38.

Grosswald, E., (1984) *Topics from the Theory of Numbers*, 2nd Ed., Boston: Birkhäuser.

Halberstam, H. and Richert, H.-E. (1974) *Sieve Methods*. London: Academic Press.

Halberstam, H. and Roth, K.F. (1966) *Sequences*. Oxford: Clarendon Press.

Hardy, G.H. and Wright, E.M. (1979) *An Introduction to the Theory of Numbers*, Fifth Edition, Oxford: Clarendon Press.

Ingham, A.E. (1937) 'On the difference between consecutive primes,' *The Quarterly Journal of Mathematics, Oxford Series*, 8, 254-256.

Jones, J.P. (1975) 'Formula for the *N*th prime number,' *Canadian Mathematical Bulletin*, 18, 433-434.

Jones, J.P., Sato, D., Wada, H. and Weins, D. (1976) 'Diophantine representation of the set of prime numbers,' *American Mathematical Monthly*, 83, 449-464.

Kuipers, L. (1950) 'Prime-representing functions,' *Indagationes Mathematicae*, 12, 57-58.

Lagarias, J.C., Miller, V.S. and Odlyzko, A.M. (1985) 'Computing $\pi(x)$: The Meissel-Lehmer Method,' *Mathematics of Computation*, 44, 537-560.

Lehmer, D.H. (1959) 'On the exact number of primes less than a given limit,' *Illinois Journal of Mathematics*, 3, 381-388.

Mapes, D. (1963) 'Fast method for computing the number of primes less than a given limit,' *Mathematics of Computation*, 17, 179-183.

Mills, W.H. (1947) 'A prime-representing function,' *Bulletin of the American Mathematical Society*, 53, 604.

Mollin, R.A (1997) 'Prime-producing quadratics,' *American Mathematical Monthly*, 104, 529-544.

Namboodiripad, K.S. (1971) 'A note on formulae for the nth prime,' *Monatshefte fur Mathematik*, 75, 256-262.

Neill T.B.M. and Singer, M. (1965) 'Letter to the editor,' *The Mathematical Gazette*, 49, 303.

Niven, I. (1951) 'Functions which represent prime numbers,' *American Mathematical Society Proceedings*, 2, 753-755.
Ore, O. (1952) 'On the selection of subsequences,' *American Mathematical Society Proceedings*, 3, 706-712.
Papadimitriou, M. (1975) 'A recursion formula for the sequence of odd primes,' *American Mathematical Monthly*, 82, 289.
Reiner, I. (1943) 'Functions not formulas for primes,' *American Mathematical Monthly*, 50, 619-621.
Ribenboim, P. (1989) *The Book of Prime Number Records*, Second Edition. New York: Springer Verlag.
Riesel, H. (1994) *Prime Numbers and Computer Methods for Factorization*, Second Edition. Boston: Birkhauser.
Sato D. and Strauss, E.G. '*P*-adic proof of non-existence of proper prime representing algebraic functions and related problems,' *Journal of the London Mathematical Society*, (2) 2, 45-48.
Srinivasan, B.R. (1961) 'Formulae for the *nth* prime,' *Journal of the Indian Mathematical Society* (2), 25, 33-39.
Vanden Eynden, C. (1972) 'A proof of Gandhi's formula for the *n*th prime,' *American Mathematical Monthly*, 79, 625.
Venugopalan, A. (1983) 'Formula for primes, twinprimes, number of primes and number of twinprimes,' *Proceedings of the Indian Academy of Science (Mathematical Science)*, 92, 49-52.
Willans, C.P. (1964) 'On formulae for the *n*th prime number,' *The Mathematical Gazette*, 48, 413-415.
Wright, E.M. (1951) 'A prime-representing function,' *American Mathematical Monthly*, 58, 616-618.
_____, (1954) 'A class of representing functions,' *The Journal of the London Mathematical Society*, 29, 63-71.
Uspensky J.V. and Heaslet, M.A. (1939) *Elementary Number Theory*. New York: McGraw Hill.