NHH

Norges
Handelshøyskole

*Norwegian School of Economics
and Business Administration*

FIVE ESSAYS ON BOUNDED RATIONALITY AND GAME THEORY

by

Hans Krogh Hvide

A dissertation submitted for the degree of dr. oecon.

Acknowledgments

Many people have contributed to this thesis and I would like to express my gratitude to a few of them. In particular, thanks to my supervisor Terje Lensberg. Without his encouragement and criticism, this thesis would have been of inferior quality, and - not the least - without him it would have been less fun to work on it. Next thanks to Ken Binmore - who has been in the supervising committee - for his availability and for his detailed comments and suggestions to drafts. For stimulating talks and suggestions, I thank Geir Asheim, Jerker Denrell, Kåre P. Hagen - who also has been in the supervising committee, Eirik G. Kristiansen, Jim March, Philippe Mongin, and Bertil Tungodden. A special thanks to Jim March at Scancor who provided excellent research opportunities while visiting Stanford University in January-December 1996. Finally, thanks to my wife Siri Brekke for patience and support.

For financial support, thanks to the SIS program, to Sturla Amundsens legacy, and to Nils Treschows legacy.

Bergen, December 10., 1997.

Hans Krogh Hvide

Contents

Chapter 1

Introduction

If we draw an imaginary line between the subject «bounded rationality» and the subject «game theory», the five works of this thesis can be seen as five points on that line. While chapter 2 and chapter 3 are close to the «bounded rationality» end of the line, chapter 4 is somewhere in between, and chapter 5 and chapter 6 are squeezed into the «game theory» end of the line, where there is no bounded rationality left. Since the main motivation for the thesis, and also the most interesting part of it, are the papers with «bounded rationality» as an ingredient (chapters 2-4), I will spend most of the introduction outlining that part of the dissertation, and treat the background for chapter 5 and chapter 6 quite cursorily at the end.

The concept of limited or bounded rationality can be traced back to the seminal works of Herbert Simon in the 50's. Simon's well-known critique of economics was that economics modeled human beings not as they appear to us - with cognitive defects, inconsistent choices etc., but as superhuman beings with grossly unrealistic cognitive skills. Simon's project became to model agents that are «intentionally rational but only limited so», or in another phrasing, «rational choice that takes into account the cognitive limitations of the decision-maker - limitations of both knowledge and computational capacity». Central to Simon's project became to model the procedural aspects of decision making.

Even though Simon's critique of perfect rationality was quite immediately well received by the profession, Simon and his associates only partially succeeded in their attempts to model bounded rationality. With some notable exceptions, the state of the art today is surprisingly similar to that forty years ago. Even though many economists view bounded rationality as more realistic and a more appropriate assumption than perfect rationality, there are surprisingly few papers that explore the implications of bounded rationality, in the sense of modeling the implications of limited cognitive abilities of the decision-maker. Moreover, it is not clear how well those papers that do model bounded rationality succeed in capturing the essence of the tricky concept.

Before we turn to speculating over why the success of bounded rationality models in the sense above has been moderate, let me emphasize that there is plenty of work on «bounded rationality» that falls outside the scope of this introduction, because their understanding of «bounded rationality» is different. Let me just mention one important direction, the evolutionary minded works that in the last decades starts out with Nelson & Winter (1982), and has roots at least back to Cyert & March (1963) and various psychological approaches to learning in the fifties and in the sixties. This literature is «behavioristic» or «inductive» in motivation, and takes as theoretical models some version of learning less rational than Bayesian learning. Some recent work in this direction can be found in Fudenberg & Levine (1996) and Young (1997), and to a certain extent in Weibull (1995), but perhaps the most radical and interesting works are done in the connection to other disciplines, for example to the genetic algorithm program of Holland (1974), (1991). In spite of this direction's promise we will ignore it in the remainder of this introduction.

The bounded rationality direction we consider in this thesis focuses on the cognition of individual agents, and is more rationalistic in flavor than the evolutionary minded works cited above. It emphasizes the procedural aspects of decision making by focusing on the role of information and information processing, and, in short, views agents as having costs to - or limited ability in - processing information correctly. Furthermore, given this limited information processing ability, agents are assumed to act in some sense optimally. This gives rise to the idea that bounded rationality refers to choice that is imperfect in the sense that it is often not the "correct" one, but is sensible in that it can be understood as an attempt by the agent to do reasonably well given his cognitive limitations (Lipman 1995). Examples of papers from this literature are Abreu & Rubinstein (1988); Dow (1991); Fershtman & Kalai (1993); Lipman (1991), (1995);

Piccione & Rubinstein (1997); and Rubinstein (1986), (1993). For a recent book on the topic, see Rubinstein (1997). In the rest of this introduction we will refer to this literature and tools it contains as MBR (Models of Bounded Rationality).

In spite of its intrinsic interest, the tools of MBR has to a little extent been incorporated into mainstream game theory and economics. There are at least three reasons for that. In the remainder of the introduction I will briefly explain these three reasons, and moreover try to explain how chapter 2-4 of the thesis can be related to them. The first criticism is a philosophical point. Even though MBR captures some aspects of bounded rationality, it has still not solved a basic problem, namely where the dividing line between models of bounded rationality and models of irrationality should be set. Up to now, it seems that every fact seemingly inconsistent with the perfect rationality paradigm can be «explained» by a suitably defined notion of bounded rationality. In short, there seems to be few bounds to our concept of bounded rationality. In chapter 2, which is forthcoming in *Theory and Decision*, I attempt to contribute to this problem. Instead of trying to define bounded rationality positively - by for example proposing a specific way of boundedly rational information processing, I try to define bounded rationality negatively - by defining some bounds wherein a theory of bounded rationality must evolve. The bounds I look for are logical. The specific setting is one with two points in time, time 1 and time 2. The agent receives some information - in the form of «sentences» - at time 1, and deduces some knowledge on the basis of this information and his information processing ability. The formal language is the epistemic logic that originates in the beautiful Hintikka (1962). Between time 1 and time 2 the agent may forget some of his knowledge. The question I pose is whether we can put any restrictions on the following: knowledge that cannot be forgotten, truths that cannot be known by the agent, and knowledge that must be forgotten. The method of proof is reductio ad absurdum; for example I assume that a certain piece of knowledge p is forgotten and if this leads to inconsistency then I interpret it to imply that p cannot be forgotten by a consistent decision maker; if p is forgotten then his knowledge is inconsistent at time 2.

The second criticism is more specific. I have argued that the motivation behind MBR is to model agents that have limited ability in processing information correctly. Given this motivation, the following feature of MBR is ironical; agents have «reasonable» cognitive dysfunctions like absent mindedness or limited attention span, but at the same time they are able to do sophisticated optimizing exercises taking these cognitive constraints as given. For example, in Dow (1991) an agent is absent minded and is aware of this fact. From receiving a continuous one-dimensional signal on day 1, say a price in a market, he is on day 2 only able to remember whether the signal he received on day 1 was low or high (e.g., below or above $10). The pretty complex problem the agent faces is to - before he receives the price signal - construct a language that determines what he should mean by a «low» price and by a «high» price. This approach to the semantics of natural language is neat but suffers from at least two problems. The first problem is the obvious one that if agents are boundedly rational in the first place it seems unrealistic to assume full ability in solving the complex optimization problem of constructing an optimal language. The second problem is probably less fundamental but is of considerable practical interest; how have the agents become perfectly aware of their cognitive constraints? Is perfect awareness a reasonable assumption? It is the second problem we approach in chapter 3. Let us mention in passing that Dow's language instead of being the outcome of some deliberate cognitive process might be seen as having evolved from some trial and error process. This is an argument to explore for future work.

The starting point for chapter 3 is the intuition that if MBR, which assumes perfect self-awareness, should be taken literally, one should try to come up with some plausible learning argument that supports perfect awareness. To be able to specify a learning process towards perfect awareness we should first ask the basic question of what we mean by an agent being uncertain about properties of himself. Surprisingly, at least to me, this question has barely been posed in the decision-theoretic minded literature (an exception is Binmore, 1987). I propose a heuristic framework to deal with this problem. Without going into details, an agent is viewed as a two-layer information processing unit. Level 1 does the «dirty» work of processing information of the external world and transmitting its conclusions to level 2, which again makes its decisions on basis of beliefs about the quality of level 1.

Roughly speaking, this part of the paper concludes that with the exception of a disturbing circularity aspect when modeling boundedly rational agents, uncertainty about the world and uncertainty about oneself can be modeled in pretty much the same fashion. From this it does not follow that perfect awareness is a plausible assumption; it just says that models of learning about the world seem to be a good first approximation when modeling learning about oneself.

The second part of the paper is more applied. The basic question is whether we should care whether some kind of imperfect awareness seems more reasonable than perfect awareness. I think the answer is yes, and list some reasons why I think so. The most prominent of these reasons is that imperfect awareness may be important to our understanding of some social phenomena. On basis of Asubel (1991), I consider a specific social phenomenon: The seemingly non-competitive prices in the credit card market. The novelty of this section is a speculation over the dynamic forces in an overconfident market.

Chapter 4 relates to the third, and probably most important, criticism of MBR; that even though the models cast light on the decision making process, there has been a lack of good applications. In chapter 4 the topic is also self-awareness, but the chapter is different in spirit from chapter 3. While chapter 3 attempted to raise and partly answer questions of «foundational» character, chapter 4 applies some ideas on self-awareness to an education setting. It is more «game theory» - multi-agent, perfectly rational decision-making, than «bounded rationality» in spirit. I therefore have labeled an agent's opinions about himself «self-knowledge» instead of «self-awareness». The starting point is the following puzzle (Blaug, 1992, Weiss 1995). Say that we are in a «Spencian» world, where education does not enhance worker productivity. Moreover assume that firms can observe worker output without considerable cost. Then why is there any need for education? Why do not firms replace education by a cheaper screening mechanism like performance wage?

I propose a hypothesis, «The Self-Knowledge Hypothesis», to solve the education puzzle. The Self-Knowledge Hypothesis, basically an old educators' argument, amounts to saying that there is nothing inconsistent in Spencian education and performance wages living peacefully side by side provided that one motive for taking education is that education gives agents a more accurate estimate of their own abilities. The main result is stronger; in a simple model I show that an institutional setting with both education and performance wages generates at least as much social surplus as an institutional setting with performance wages alone. It turns out that a sufficient condition for this result is that agents' prior beliefs about themselves satisfy a certain condition (C). Intuitively speaking, condition (C) says that for every overconfident agent in the population there should be one underconfident agent and vice versa. Condition (C) is a considerably weaker assumption on beliefs than that made so far in literature on agents that lack self-knowledge (e.g., see Jovanovic 1979, Weiss 1983), but may - in light of experimental evidence - be too strong to be realistic in a strict sense. At any rate, I think (C) serves as an intuitive and useful benchmark assumption on beliefs at population level.

The remaining two chapters of the dissertation, on implementation theory, are in the game theory end of the imagined line between bounded rationality and game theory. While perhaps the best known part of non-cooperative game theory, the equilibrium selection literature, explores «solutions» and convergence to solutions for given games, the task of implementation theory is the inverse; to design games that implement certain normatively appealing solutions. Classical examples of this modeling technique are Vickrey's (1961) second price auction, and Groves' (1973) mechanism for implementing truth-telling in the valuation of public goods.

In chapter 5 and chapter 6 I study the implementation of efficient provision of «effort» in a class of simple partnership games, where a partnership game just means budget-balance; wages of the workers must equal the income of the partnership. The chapters are two comments on the mechanism proposed by Legros & Matthews (1993) and Vislie (1994). In chapter 5 I show with a very simple model that the implementation result obtained in Legros & Matthews (1993) and Vislie (1994) is sensitive to the agents being uncertain about the exact relationship between effort and output of the other agents. In chapter 6 I show that their mechanism is also sensitive to the tightness of the participation constraint; if one or more partners have an outside option that is more attractive than the equilibrium outcome their sharing rule breaks down. I construct a sharing rule that implements the efficient outcome in Nash equilibrium regardless of size of the outside options.

References

Asubel, L. M. (1991). The Failure of Competition in the Credit Card Market. *American Economic Review*, **81**, 50-81.

Axelrod, R. (1984). *The Evolution of Co-operation*. Basic Books Inc.

Binmore, K. (1987). Modeling Rational Players: Parts I and II. *Economics and Philosophy* **3**; **4**.

Cyert, R. & March, J. (1992). *A Behavioral Theory of the Firm*, 2nd edition. Blackwell.

Fudenberg, D. & Levine, D. (1997). *Learning in Games*. The MIT press.

Groves, T. (1973). Incentives in Teams. *Econometrica*, **41**, 617-31.

Hintikka, J. (1962). *Knowledge and Belief.* Cornell University Press.

Holland, J. (1975). *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor.

Holland, J. & Miller, J. (1991). Artificial Adaptive Agents in Economic Theory. *American Economic Review Papers and Proceedings* **81**, 365-70.

Lipman, B. (1995). Information Processing and Bounded Rationality: a Survey. *Canadian Journal of Economics* **1**, 42-67.

Maynard Smith, J. (1982). *Evolution and the Theory of Games*. Cambridge University Press.

Neyman, A. (1981). Bounded Rationality Justifies Cooperation in the Finitely Repeated Prisoners' Dilemma game. *Economics Letters* **19**, 227-29.

Radner, R. (1980) & (1986). Can Bounded Rationality Resolve the Prisoner's Dilemma?. In: *Essays in Honor of Gerard Debreu.* Mas-Colell, A. & Hildebrand, W. (eds.).

Rubinstein, A. (1997). Modeling Bounded Rationality. *MIT Press*, forthcoming.

Sidney, R. & Winter, N. (1982). *An Evolutionary Theory of Economic Change.* Harvard University Press.

Vickrey, W. (1961). Counterspeculation, Auctions, and Competitive Sealed Tenders. *Journal of Finance*, **16**, 1-17.

Weiss, A. (1995). Human Capital vs. Signalling Explanation of Wages. *Journal of Economic Literature* **9**, 133-54.

Young, H. P. (1997). .... . Forthcoming, MIT Press.

Chapter 2

Bounds to Memory Loss[2]

Abstract

If we express our knowledge in sentences, we will find that these sentences are linked in complex patterns governed by our observations and our inferences from these observations. These inferences are to a large extent driven by logical rules. We ask whether the structure logic imposes on our knowledge restricts what we forget and what we remember. The model is a two period S5 logic. In this logic, we propose a memory loss operator: the agent forgets a sentence p if and only if he knows p at time 1 and he does not know p at time 2. Equipped with the operator, we prove theorems on the relation between knowledge and memory loss. The main results point to classes of formulas that an agent cannot forget, and classes of formulas he must forget. A desirable feature is that most results hold in the S4 logic. The results illustrate bounds to memory loss, and thus to bounded rationality. We apply the model to single-agent conventions: conventions made between an agent and himself.

Keywords: Bounded Rationality, Imperfect Recall, Memory Loss Operator, Redundancy, Reasoning Through Time, S5 Logic, Single-Agent Conventions.

[2] Thanks to Ken Binmore, Joseph Halpern, Kaare Johnsen, Mamoru Kaneko, Terje Lensberg, Jim March, Arild Waaler, three referees, and particularly to an editor of this journal for valuable comments and suggestions. Thanks also to participants at SITE 1996, Stony Brook 1996 and IAREP 1995, and to seminar participants at University College London, Norwegian School of Economics and Business, and University of Oslo. Remaining errors are, of course, my own. Financial support from Sturla Amundsens Legacy and SIS V is gratefully acknowledged.

1. Introduction

One possible interpretation of «bounded rationality» is «bounded recall». In this paper we propose a formal language describing the change in the epistemic state of an absent-minded agent when he forgets some knowledge (where an epistemic state is, informally, just a list of formulas describing an agent's knowledge). Although the language we propose clarifies what is meant by an agent being absent-minded, it would be of narrow interest if it did not produce non-trivial theorems. We think the most interesting theorems of the language concern the "bounds to memory loss", restrictions on what an absent-minded agent cannot forget and restrictions on what he cannot remember. For example, say that I know that p is true, and furthermore I know that p implies q. Suppose I am sufficiently rational to be able to deduce q. My knowledge then has the structure of a modus ponens argument; I know p, I know that p implies q, and I know q. Now say that I forget q (and only q!). Thus I do not know q anymore. Is my new epistemic state consistent? Clearly not; since I per assumption still know p, and that p implies q, I can re-deduce q, contradicting the claim that I do not know q.

The contradiction can be interpreted in the following way: An agent with an epistemic state described by p, p implies q, and q, and which furthermore is familiar with *modus ponens*, cannot forget q without also having forgotten either p or that p implies q (or both). Thus, we have obtained a restriction (bound) to memory loss for an agent who knows the use of *modus ponen*s. The bound to memory loss in the example above stems from the logical structure we assign to knowledge. Therefore, to understand the results on memory loss we obtain later it is critical to understand the notion of knowledge that we refer to. Briefly, we use the epistemic language with sentences as primitives that originates from the classic "Knowledge and Belief," by J. Hintikka (1962). The strongest logic of that language, S5, whose descriptive value Hintikka himself disapproved of, assumes that the agent has exceptionally strong logical powers; he deduces all formulas that follow logically from a given set of observations. This property of S5 - usually referred to as «logical omniscience» - corresponds in a precise sense to the reasoning of a perfectly rational agent depicted in informational economics.[3] To arrive at logical omniscience, the S5 logic includes unrealistic axioms, such as "if the agent does not

[3] This relation is made clear in Fagin et al. (1995), Proposition 2.5.2.

know p, then he knows that he does not know p." Fortunately, this negative introspection axiom is not - as shown later - necessary to obtain most of our results.

The advantage of a formal language becomes evident when considering how to assess the existence of possible restrictions to memory loss. A frequently occurring method of proving that a restriction exists is *reductio ad absurdum.* To briefly explain, a restriction exists, given acceptable axioms on knowledge and a definition of memory loss, if supposing its negation leads to inconsistency; that is for a given sentence p, both the sentences "the agent knows p" and "the agent does not know p" are true (at the same point in time). For example, say that a candidate restriction specifies that forgetting p implies forgetting q, where p and q are members of some domain. We would then investigate whether forgetting p is inconsistent with not forgetting q. If so, the restriction exists.

Memory loss implies a time structure. The model we construct in this paper covers two points in time, put simply, before and after memory loss. At time 1, some epistemic state prevails. An agent is provided some facts by "nature", the "basic facts", taken as exogenous, to deduce other pieces of knowledge. An agent's knowledge consists of the conjunction of the basic facts and all deductions from these facts. An agent forgets a fact ϕ (between time 1 and time 2) if ϕ is known at time 1 and ϕ is not known at time 2. Thus, the epistemic state at time 2 stems from the basic facts, the logical capacity of the agent, and his forgetting. Even though there is no conceptual problem in extending the language to cover an arbitrary number of periods (and agents), and moreover to incorporate learning, we abstract from it.[4]

The paper is structured as follows. First we fix a language and propose a formal definition of memory loss in that language. Then we present the S5 logic, and in Proposition 1 we prove theorems on the relation between forgetting and knowledge. In Proposition 2 we prove properties of two extreme forms of self-insight: that the agent knows that he is going to forget (*ex-ante* awareness), and that he knows that he has forgotten (*ex-post* awareness). Some philosophical points related to Proposition 2 are then discussed. Proposition 3 regards the

[4] With more than two points in time, we would need to put a time label on the F operator as well. In the extension, let K_i and F_i range over n points in time, where $n > 2$. $F_1\phi$ means "forget that ϕ between time 1 and time 2", and $F_2\phi$ "forget that ϕ between time 2 and time 3". The definition of memory loss between t and t+1 then becomes: $F_t\phi \equiv K_t\phi \wedge \neg K_{t+1}\phi$. Notice that there is also not a problem with incorporating beliefs in the language. In that case, the language would be identical to that in Battigalli & Bonanno (1996).

"chain of forgetting", i.e. the logical link between forgotten formulas. Proposition 4 shows properties of temporal knowledge; knowledge of future memory loss and knowledge. Proposition 5 proves the impossibility of forgetting theorems when remembering all axioms. The last section is devoted to an application of the language. We discuss a one-person co-ordination problem, and find a sufficient condition for a single-agent convention to be successful.

Doing a conceptual analysis of bounded rationality is by no means a novel idea, the industry of clarifying Herbert Simon's intuitions is at least as old as Simon's work itself. For a very informative survey on recent attempts to model bounded rationality within economics and game theory, see Lipman (1995). For an excellent text-book (with applications to computer science) on the sentence-based language used in the present paper, see Fagin et al. (1995). For applications of the sentence-based language in work related to decision theory, see Modica & Rustichini (1994), Bonanno & Battigalli (1997), and particularly Bacharach & Mongin's (1994) survey on the use of epistemic logic in economics. For the construction of an intricate "game logic," see a sequel of papers by Kaneko & Nagashima (1996). On some interesting decision theoretic aspects of memory loss, see for example Dow (1991) and Piccione & Rubinstein (1997).[5]

2. Memory Loss

This section presents a language containing a notion of knowledge and then defines memory loss in terms of knowledge. Let Θ be a non-empty set of primitive formulas, labeled $p, p', q, q', \ldots$. Throughout, assume that the truth-value of the primitive formulas is fixed; p is true at time 2 iff it is true at time 1.[6] The set of non-epistemic formulas, Γ, is closed under substitution on Θ, under negation $\neg$, and under disjunction $\vee$. Hence if ϕ and ψ are formulas then so are $\neg\phi$, $(\phi \vee \psi)$ and $\neg(\phi \vee \psi)$. The full language, denoted Λ, is closed under Γ and the epistemic operators K_1 and K_2. Thus, if ϕ and ψ are formulas in Λ, then so are $K_1\phi$ and $K_2\psi$.

[5] The latter paper will be published in Games and Economic Behavior along with several other papers on the same topic.

[6] This has a natural interpretation: a primitive formula should be seen as a statement about the world at a specific time, e.g., "it rains in London on December 31, 1997."

The intended interpretation of $K_1\phi$ is "the agent knows ϕ at time 1," and the intended interpretation of $K_2\psi$ is "the agent knows ψ at time 2." Notice that with Λ we may represent "temporal knowledge", i.e., sentences such as "the agent knows that he is going to know ϕ," $K_1K_2\phi$, and "the agent knows that he knew ϕ," $K_2K_1\phi$.[7]

Let the operator F denote forgetting. The intended interpretation of a formula $F\phi$ is "the agent forgets ϕ." We propose the following definition of $F\phi$,[8]

$$F\phi \equiv K_1\phi \wedge \neg K_2\phi,\ \phi\in\Lambda.$$

The definition says that an agent forgetting a formula ϕ is equivalent to that agent knowing ϕ at time 1 and not knowing ϕ at time 2.[9]

3. The S5 Logic

Let ϕ denote an arbitrary (consistent) formula in the language Λ. Let $t = 1,2$, where t indicates time. Then the S5 axioms are as follows:

PC	The set A of all tautologies of propositional calculus	
D	$(K_t\phi \wedge K_t(\phi \Rightarrow \gamma)) \Rightarrow K_t\gamma$	(Distribution axiom)
T	$K_t\phi \Rightarrow \phi$	(Truth axiom)
4	$K_t\phi \Rightarrow K_tK_t\phi$	(Positive introspection)

[7] Surprisingly the philosophical literature is rather sparse on extending modal logics to dynamic settings. The notation for time used here was used by Shoham (1989), and is also used in a parallel paper by Battigalli & Bonanno (1996).

[8] To our knowledge the F operator is novel. An operator used in the distributed computing literature that is somewhat the same spirit is the distributed knowledge operator D, where (interpret the index as persons),

$$D\phi \equiv K_1\phi \wedge K_2\phi,\ \phi\in\Lambda.$$

Thus ϕ is distributed knowledge if both agent 1 or agent 2 knows ϕ. For properties of this operator, see Fagin, et al. (1995).

[9] From an ex-ante point of view (time 1) we can interpret $F\phi$ as "the agent is going to forget ϕ", and from an ex-post point of view (time 2) we interpret $F\phi$ as "the agent has forgotten ϕ". Which interpretation to choose depends on the location in time of the analyst. The identity of the agent and the identity of the analyst may coincide. That depends on the application we have in mind.

5 $\neg K_t\phi \Rightarrow K_t\neg K_t\phi$ (Negative introspection)

D says that if an implication and its antecedent are known by an agent then the precedent is also known. T says that only true formulas can be known. 4 says that if a formula is known, the fact that it is known is also known. Axiom 5, controversially, says that if a formula is not known then the fact that it is unknown is indeed known. We use the following simplified notation: for $n \geq 0$: $K^0\phi \equiv \phi$, $K^n \equiv KK^{n-1}\phi$. Analogously, $(\neg K)^0\phi \equiv \phi$, $(\neg K)^n\phi \equiv \neg K(\neg K)^{n-1}\phi$, and $F^0\phi = \phi$, and $F^n\phi \equiv FF^{n-1}\phi$.

Turning to the inference rules, i.e., how valid formulas are derived,

(MP) Modus Ponens: $\dfrac{\phi \wedge (\phi \Rightarrow \gamma)}{\gamma}$

The set of valid formulas is closed under *modus ponens.*

Now turning to the agent's knowledge, the following rule describes how an S5 rational agent infers knowledge:

(RE) Rule of epistemization: $\dfrac{\phi \Rightarrow \gamma}{K_t\phi \Rightarrow K_t\gamma}$ $t = 1, 2.$

RE says that knowing the antecedent of a valid formula implies knowing the precedent. The main result of S5, which follows from T, D, MP and RE, is that if a formula ϕ is valid then the agent knows ϕ. Formally,[10]

(LO) Logical omniscience: $\dfrac{\phi}{K_t\phi}$ $t = 1, 2.$

[10] For a proof, see Fagin, et al. (1995) or a textbook on modal logic, such as Hughes & Creswell (1968).

To sum up, the set of valid non-epistemic formulas is closed under *modus ponens*, and knowledge is closed under logical omniscience. Notice that there are no particular difficulties in extending the S5 logic to two points in time: standard completeness and soundness results hold, analogous to the two person case.[11]

In the next section, simple theorems on memory loss are deduced, followed by some comments on the language. In the remainder of the paper, parenthesis will be place behind the heading of each proposition to indicate the epistemic logic sufficient to derive the proposition. S4 emerges from subtracting axiom 5 from S5, T is identical to S4 without axiom 4, and D is identical to T without the truth axiom.

4. Theorems on Memory Loss

Proposition 1. Properties of F. (S5)

a) $F\phi \Rightarrow \phi$

b) $\neg K_1\phi \Rightarrow K_1\neg F\phi$

c) $K_2\neg K_1\phi \Rightarrow K_2\neg F\phi$

d) $K_2\phi \Rightarrow K_2\neg F\phi$

e) $\neg F\phi \Leftrightarrow \neg K_1\phi \vee (K_1\phi \wedge K_2\phi)$

f) $K_1F\phi \Rightarrow K_1K_1\phi \wedge K_1\neg K_2\phi$

Proof.

a)$F\phi \Rightarrow K_1\phi$ by definition, and $K_1\phi \Rightarrow \phi$ by T. b)$\neg K_1\phi \Rightarrow K_1\neg K_1\phi$ by axiom 5. Apply *modus tollens* on the definition of $F\phi$ to obtain $\neg K_1\phi \Rightarrow \neg F\phi$. By RE, $K_1\neg K_1\phi \Rightarrow K_1\neg F\phi$. c)follows from applying RE on b) to get $K_2\neg K_1\phi \Rightarrow K_2K_1\neg F\phi$ which implies $K_2\neg F\phi$ by applying T and LO. d)$K_2\phi \Rightarrow \neg F\phi$ by definition of $F\phi$. $K_2K_2\phi \Rightarrow K_2\neg F\phi$ by RE and $K_2K_2\phi$ is equivalent to $K_2\phi$ by axiom 4 and T. e)by negating the definition of memory loss and by PC arguments, $\neg F\phi \Leftrightarrow (\neg K_1\phi \wedge K_2\phi) \vee (\neg K_1\phi \wedge \neg K_2\phi) \vee (K_1\phi \wedge K_2\phi)$.

[11] See Fagin, et al. (1995), Theorem 3.3.1, for the case of n = 2. Simply interpret the person index used by Fagin, et al. on the modal operators as a time index. Analogously, a model with n points of time is also sound and complete.

Clearly $(\neg K_1\phi \wedge K_2\phi) \vee (\neg K_1\phi \wedge \neg K_2\phi) \Leftrightarrow \neg K_1\phi$. Thus $(\neg K_1\phi \wedge K_2\phi) \vee (\neg K_1\phi \wedge \neg K_2\phi) \vee (K_1\phi \wedge K_2\phi) \Leftrightarrow \neg K_1\phi \vee (K_1\phi \wedge K_2\phi)$. f)$K_1F\phi \Leftrightarrow K_1(K_1\phi\wedge\neg K_2\phi)$ follows from the definition of the memory loss operator. $K_1(K_1\phi \wedge \neg K_2\phi) \Rightarrow K_1K_1\phi \wedge K_1\neg K_2\phi$ is proved on page 51 of Fagin et al. (1995).

a)is trivial but reassuring. Note that it implies $FF\phi \Rightarrow F\phi$: If I have forgotten that I was going to forget ϕ then it also must be true that I have forgotten ϕ. b)says that if I do not know ϕ then I know that I am not going to forget ϕ. c)says that if I know that I did not know a fact then I know that I have not forgotten the fact. d)says that if I know something then I know that I have not forgotten it. e)is just a restatement of the definition of memory loss. f)$K_1F\phi$ means that the agent knows that he will forget ϕ. We denote such clairvoyance as *ex-ante awareness* (of memory loss). For example, that Anne knows that she is going to forget Beth's telephone number (α) is expressed as $K_1F\alpha$.

Let us make two comments on the language. First, when applying the language, we formalize statements like "the agent knows that he is going to forget ϕ", $K_1F\phi$, and "the agent know that he is going to know ϕ", $K_1K_2\phi$. What is the meaning of such sentences involving knowledge about the future? Is the future already known? To simplify matters we have assumed that the truth-value of the primitive formulas is constant (implying that the truth-value of more complex non-epistemic formulas is constant). Therefore, if something is known to be true today, it will also be known to be true tomorrow; and the agent's uncertainty about whether he will know ϕ tomorrow reflects uncertainty about his own absent-mindedness, not uncertainty about the world. Consequently, if we abstract from an agent's learning, his knowledge about future knowledge boils down to knowledge today combined with knowledge about future memory loss. In such a setting it seems plausible that the agent can have knowledge about future knowledge, as in the statement "I know that I am going to forget that Helen's telephone number is y at time 1".

Second, no restrictions have been put on the nature of time. In fact, we do not have to interpret the subscript as a time operator. If the subscripts 1 and 2 are interpreted as persons, rather than time, the F operator is given another interpretation: person 1 has informational advantage ϕ over person 2 if and only if $F\phi$ holds.

5. Awareness

Proposition 2. Awareness of memory loss. (S4)

a) $K_2F\phi$ *is inconsistent*

b) For all m= 2,3, ..., $K_1F\phi \Leftrightarrow F^m\phi$.

Proof.

a)$K_2F\phi \Rightarrow K_2\neg K_2\phi$ by definition and RE, which then implies $\neg K_2\phi$ by T. $K_2F\phi \Rightarrow K_2K_1\phi$ which implies $K_2\phi$ by T and RE. This is inconsistent. b)First we show the implication from left to right by showing that the formula $K_1F\phi \wedge \neg F^m\phi$ is inconsistent for any $m \geq 2$. $K_1F\phi \wedge \neg FF\phi$ implies $K_2F\phi$ by definition, which is inconsistent by 2a). Therefore $K_1F\phi \Rightarrow FF\phi$. By RE, $K_1K_1F\phi \Rightarrow K_1FF\phi$, and by axiom 4, $K_1F\phi \Rightarrow K_1K_1F\phi$. But $K_1FF\phi \wedge \neg FFF\phi$ implies $K_2FF\phi$ by definition. By Proposition 1a), $K_2FF\phi$ implies $K_2F\phi$. This is inconsistent by a). Thus it has been shown that $K_1F\phi \Rightarrow F^m\phi$ is valid for $m = 2$ and $m = 3$. The rest of the proof goes through by induction on m. Now the implication from right to left. For $m \geq 2$, $F^m\phi \Rightarrow FF\phi$ by 1a), and $FF\phi \Rightarrow K_1F\phi$ by definition.

a)says that knowing that I have forgotten is inconsistent. The result is important because it shows that there are some true formulas (about memory loss) that cannot be known by the agent.[12] Note that even if I cannot know the exact content of what I have forgotten, I may know something about it. For example, suppose that I have forgotten that Helen's telephone number is y. Then K_2Fy is inconsistent, i.e., the statement "I know that I have forgotten that Helen's telephone number is y" is inconsistent. However, there is nothing inconsistent in knowing that I have forgotten that Helen's telephone number is y or, say, y',

[12] That is of course given that the agent forgets something in the first place. To see that forgetting is at all possible, consider the simple model where there is only one fact, p. Then there is clearly nothing inconsistent in $K_1p \wedge \neg K_2p$.

where y' is different from y.[13] In that case $K_2(Fy \vee Fy')$. b)Look at the statement for $m = 2$. Then we get $K_1F\phi \Leftrightarrow FF\phi$. *Ex-ante* awareness is equivalent to forgetting that I was going to forget. The implication from left to right says that *ex-ante* awareness of memory loss must be forgotten, while the implication from right to left follows from the definition of memory loss. Let us dwell a minute on the plausibility of the former. Say that I know that Helen's telephone number is y (she just told me). Thus K_1y is valid. Moreover, I know that I am very absent-minded with numbers, so I know that I am going to forget y. Thus K_1Fy is valid. Then b) says that I cannot forget y and at the same time remember K_1Fy, simply because if I remember K_1Fy this implies K_2K_1Fy, which is inconsistent. Returning to the problem of placing bounds on memory loss, Proposition 2b) does so by pointing to a type of knowledge, *ex-ante* awareness, that must be forgotten within the language.

Moreover, *ex-ante* awareness of memory loss implies forgetting infinitely many formulas. I forget the fact itself, I forget that I know that I will forget; I forget that I know that I am going to forget that I know that I am going to forget, and so on for all *m*. An amusing way of stating this result goes as follows: If rational folk are those who know what they are going to forget, then rational folk forget more than the less rational.

Even if the agent can have knowledge about his own forgetting, there is nothing in the language that forces the agent to have any such knowledge.[14] To be specific: That the agent forgets ϕ neither implies that he knows at time 1 that he is going to forget ϕ (thus $F\phi \Rightarrow K_1F\phi$ is not a theorem of the logic), nor that he knows that he has forgotten ϕ at time 2 (thus $F\phi \Rightarrow K_2F\phi$ is not a theorem of the logic). For obvious reasons $F\phi \Rightarrow K_2F\phi$ should not be an axiom of the language, but should the «awareness axiom», $F\phi \Rightarrow K_1F\phi$, be? We think the awareness axiom is implausible in the abstract setting considered here, simply because it is implausible that an agent has perfect awareness of his own (bounded) cognitive abilities.[15] Such awareness seems to require too much from introspection, particularly for an agent whose rationality is already bounded. However, for an agent to be able to derive the optimal decision rules in Dow

[13] It would not be hard to construct a first-order language to express sentences such as "I know that I have forgotten a number with the property that it is Helen's telephone number".

[14] For some applications one may want to model the analyst's knowledge as different from the agent's knowledge. There is nothing inconsistent in the analyst having more knowledge about the agent's forgetting than the agent has himself.

[15] For more on this problem see Binmore (1987) and Hvide (1997).

(1991) and Piccione & Rubinstein (1996) it seems that he must be endowed with the awareness axiom.

6. Bounds to Memory Loss

We have seen that *ex-ante* awareness of memory loss must be forgotten by an agent. In this section we continue finding bounds to memory loss. First let us make it somewhat more precise what we mean by "an agent".

An agent makes deductions from basic facts. The basic facts may be interpreted as the agent's perception of the world or facts provided him by an external source. Since the agent is conscious of all formulas in Λ, we can think of the basic facts as taking the form of truth assignment to some or all elements in Λ. By making deductions on the basic facts the agent deduces new knowledge. For example by the distribution axiom he deduces q from knowing p and $p \Rightarrow q$. Obviously, an agent whose reasoning satisfies the S5 logic will be able to deduce more from a given set of basic facts than an agent whose reasoning satisfies the S4 logic; for a given set of basic facts, the deductions of an S4 rational agent will be a subset of the deductions of an S5 rational agent. Let K denote the set of known elements at time 1. For $\phi \in \Lambda$, we say that $\phi \in K$ iff $K_1\phi$ holds. Denote the set of basic facts as K_B and the set of deduced facts as K_D. Then $K_D = K \setminus K_B$.[16]

Say that we (the analysts) know that an agent has forgotten a formula ϕ. Can we say anything about other forgotten formulas? The following proposition gives a result on the "chain of forgetting":

Proposition 3. The chain of forgetting. (D)
$(\phi \Leftrightarrow \gamma) \Rightarrow (F\phi \Leftrightarrow F\gamma)$, $\phi, \gamma \in \Lambda$.

[16] Denoting the conjunction of basic facts β, $\phi \in K_D$ iff $\beta \Rightarrow \phi$ is provable in the S5 logic. For the concept of provability, see Fagin et al. (1995).

Proof.

It suffices to show that given $\phi \Leftrightarrow \gamma$, assuming that $\neg F\phi$ and $F\gamma$ leads to inconsistency. $F\gamma \Rightarrow (K_1\gamma \wedge \neg K_2\gamma)$ by definition. $K_1\gamma \Rightarrow K_1\phi$ by RE. $(K_1\phi \wedge \neg F\phi) \Rightarrow K_2\phi$ holds by Proposition 1e), which by RE implies $K_2\gamma$. Inconsistent.

Proposition 3 states that forgetting a formula ϕ implies forgetting all formulas equivalent to ϕ. The intuition behind is that having forgotten a formula ϕ is not consistent with being able to derive ϕ at time 2, which would have been the case if a formula equivalent to ϕ were remembered.[17]

It is of interest to check if we can obtain stronger results on the chain of memory loss than Proposition 3. First notice that $(F\phi \Leftrightarrow F\gamma) \Rightarrow (\phi \Leftrightarrow \gamma)$ clearly does not hold, as there is nothing inconsistent with $(\neg F\phi \wedge \neg F\gamma) \wedge \neg(\phi \Leftrightarrow \gamma)$. I can remember both ϕ and γ, without ϕ and γ having to be equivalent. The rule $(\phi \Rightarrow \gamma) \Rightarrow (F\phi \Rightarrow F\gamma)$ does not (and should not either) follow from the definition of memory loss. I may forget the axioms of a theory (ϕ) without forgetting its conclusions (γ). The rule $(\phi \Rightarrow \gamma) \Rightarrow (F\gamma \Rightarrow F\phi)$ looks plausible; if I forget conclusions of a theory I must forget the axioms; if not I could simply re-deduce the forgotten conclusions. To see that the intuition is false, simply consider the case when I do not know the axioms at either point in time (making $F\phi$ untrue even if $F\gamma$ and $\phi \Rightarrow \gamma$ hold).

Notice that from observing that an agent must forget all formulas equivalent to a given forgotten formula, it is simple to prove that an agent cannot forget a finite number of formulas.[18] Now consider knowledge about memory loss and knowledge in the future; *temporal knowledge*, K_T, where $K_T \subset K$. $\tau \in K_T$ iff $\tau = K_2\phi$, $\tau = \neg K_2\phi$, or $\tau = \neg F\phi$, where $\tau \in \Lambda$;

[17] Notice that because $\phi \Leftrightarrow K_1\phi$ is not a valid formula, Proposition 3 does not exclude the possibility of forgetting introspective knowledge of a fact, say $K_1K_1\phi$, without forgetting the fact itself. This is how it should be; it seems perfectly plausible to remember a fact without knowing that one knew it before.

[18] First, define let A_ϕ be the set of all statements that are equivalent to ϕ. Thus $\gamma \in A_\phi$ iff $\gamma \Leftrightarrow \phi$, $\gamma \in K$. Label the elements in A_ϕ as ϕ_1, ϕ_2, ... , in any order. To see that the sequence$\{\phi_j\}_{j=1,2,...}$ is (countable) infinite, simply observe that the number of tautologies is infinite. An agent must forget either none or infinitely many formulas. (D). To prove this, observe that $F\phi \Rightarrow K_1\phi$ by definition, and that $K_1\phi \Rightarrow K_1\phi_j$, $\forall j$, by RE. Since the sequence $\{\phi_j\}_{j=1,2,...}$ is infinite, it follows from Proposition 3 that an agent forgets either none or infinitely many formulas. The reasoning behind this is simply that an S5 rational agent always knows infinitely many formulas that are equivalent to a given formula. Notice that logical omniscience is not necessary for that result. Briefly, for that result to hold any logical system containing all tautologies and where $K_1\phi \Leftrightarrow K_1\phi_i$ is valid would be sufficient: the result does not depend on an unrealistic introspection assumption in the S5 logic.

I know that I am going to know a formula ϕ, know that I will not know a formula ϕ, I know that I am going to forget a formula ϕ, and I know that I am not going to forget a formula ϕ. The reasoning involved in making deductions about temporal knowledge we label as *reasoning through time.*

Proposition 4. Temporal knowledge cannot be forgotten. (S5)
$FK_2\phi$ is inconsistent, $F\neg K_2\phi$ is inconsistent, and $F\neg F\phi$ is inconsistent if $K_1\phi$ is valid.

Proof.
We start by considering $FK_2\phi$. $FK_2\phi \Rightarrow (K_1K_2\phi \wedge \neg K_2K_2\phi)$ by definition, and $K_1K_2\phi$ implies $K_2\phi$ by T. $\neg K_2K_2\phi \Rightarrow \neg K_2\phi$ by applying *modus tollens* on axiom 4. Inconsistent. Now consider $F\neg K_2\phi$. By definition, $F\neg K_2\phi \Rightarrow (K_1\neg K_2\phi \wedge \neg K_2\neg K_2\phi)$. By T, $K_1\neg K_2\phi \Rightarrow \neg K_2\phi$, while $\neg K_2\neg K_2\phi \Rightarrow K_2\phi$ by definition of $F\phi$ and axiom 5. Inconsistent. Now we show that $F\neg F\phi$ leads to inconsistency if $K_1\phi$ holds. By definition, $F\neg F\phi \Rightarrow (K_1\neg F\phi \wedge \neg K_2\neg F\phi)$. By Proposition 1e) and RE, $K_1\neg F\phi \Rightarrow K_1(\neg K_1\phi \vee (K_1\phi \wedge K_2\phi))$. Recall that we have assumed that $K_1\phi$ holds. $[K_1\phi \wedge K_1(\neg K_1\phi \vee (K_1\phi \wedge K_2\phi))] \Rightarrow [K_1\phi \wedge K_1K_1\phi \wedge K_1K_2\phi]$ since $K_1\phi \wedge K_1\neg K_1\phi$ is inconsistent by T. Since $[K_1\phi \wedge K_1K_1\phi \wedge K_1K_2\phi]$ and $\neg K_1K_2\phi$ is inconsistent we have that $[K_1\phi \wedge K_1(\neg K_1\phi \vee (K_1\phi \wedge K_2\phi))] \Rightarrow K_1K_2\phi$. Furthermore, $K_1K_2\phi \Rightarrow K_2\phi$ by T, and $K_2\phi \Rightarrow K_2\neg F\phi$ by Proposition 1d). This again is inconsistent.

Proposition 4 shows that temporal knowledge cannot be forgotten. Thus, forgetting either of the formulas $K_2\phi$, $\neg K_2\phi$, or $\neg F\phi$ (the last one if $K_1\phi$ holds) is inconsistent in the S4 logic.

The final proposition of the section presents a result that severely limits the formulas that an agent may forget. The idea is simple (and was touched upon under the discussion of Proposition 3). Say that an agent has an identical informational basis at time 1 and time 2. Then no formula can be forgotten. Since an agent's logical ability is identical at both points in time, the set of formulas deduced at time 1 and the set of formulas deduced at time 2 are identical, and thus knowledge at time 1 and knowledge at time 2 are identical. Other cases follow a similar intuition. For example, if the basic facts at time 2 is a subset of the basic facts at time 1, then agent cannot forget set of facts deducible from the subset. We will prove this point in the particularly simple case where the agent remembers all of the basic facts.

Proposition 5. Impossible to forget theorems when axioms are remembered. (D)
Assume that $\neg F\phi$ *holds for all* $\phi \in K_B$. *Then* $F\gamma$ *is inconsistent for any* $\gamma \in K_D$.

Proof.
We show that assuming $\neg F\phi$ for $\phi \in K_B$ is inconsistent with $F\gamma$, $\gamma \in K_D$. If $\neg F\phi$ holds for all $\phi \in K_B$, then by Proposition 1e) $K_2\phi$ holds for all $\phi \in K_B$. Since the agent is S5 rational at both points in time, $K_2\gamma$ must hold for all $\gamma \in K_D$. Inconsistent.

An important question is whether the propositions go through if the logical ability of the agent is weaker than S5. Let us review the propositions and comment on that issue. Proposition 2 shows that *ex-post* awareness of memory loss is inconsistent. It also shows that if an agent were *ex-ante* aware of forgetting ϕ in the future, the agent must not only forget ϕ but also the fact of the *ex-ante* awareness. Both awareness results derived in Proposition 2 hold in the T logic (S5 without axioms 4 and 5), which is a quite weak logic since it does not presuppose any introspectional ability on the part of the agent. Proposition 3, the chain of forgetting (if ϕ is forgotten then all formulas in ϕ's equivalence class is forgotten) goes through in the D logic (T without axiom T). Of course, which formulas are equivalent differs from logic to logic. For example, in contrast to the S4 logic, not knowing in S5 is equivalent to knowing that not knowing. Proposition 4, which states that forgetting some formulas and remembering others may consistently be modeled within the language, holds in S4 (axiom 5 is only needed in proving one of the statements). Proposition 5, the impossibility of forgetting deduced facts without forgetting some basic fact, goes through in the D logic. However, the deduced facts in the D logic will be a subset of the deduced facts in the T logic, the deduced facts of the T logic will be a subset of the deduced facts in the S4 logic, and so forth. For example in S5, lack of information about p implies knowledge about lack of knowledge about p. Thus $\neg Kp$ will be a deduced fact and $K\neg Kp$ will hold at both points in time, while in S4 such negative introspective knowledge cannot be derived.

In sum, we have characterized knowledge that must be forgotten and knowledge that cannot be forgotten by an agent satisfying S4 or a stronger logic. Consider first the formulas that must be forgotten. If I know that I am going to forget a formula ϕ, then I must not only forget ϕ, but also forget the fact that I knew that I was going to forget ($K_1F\phi$). Moreover, forgetting a formula ϕ implies forgetting all formulas equivalent to ϕ. For example, forgetting ϕ implies that an agent forgets all levels of introspective knowledge of ϕ ($K_1{}^n\phi$, n = 1, 2, ...). Shifting over to formulas that cannot be forgotten it was demonstrated that neither $K_2\phi$, $\neg K_2\phi$ nor $\neg F\phi$ (the last when $K_1\phi$ holds) can be forgotten by an S5 rational agent. Thus temporal knowledge cannot be forgotten.

As with many theoretical insights from the bounded rationality literature, the practical value of the results is not evident. Consider the following attempt.[19] In criminal cases, three alternative circumstances may lead to conviction: Confession, compelling evidence, or inconsistency in interrogations. Interrogations may last for days and weeks; suspects in custody seldom admit their crime (nor are let free) the first day. As time goes by between interrogations, suspects may forget details, which in itself is not enough to lead to conviction. That interrogators know that suspects may be absent minded can be used strategically by suspects. They may (falsely) claim to have forgotten sensitive details. Of course, the interrogator knows that, and in lack of direct psychological tests, epistemic tests may be needed to expose lying about memory loss. The results provide an interrogator with such a test; it tells which pieces of knowledge a suspect may forget and may remember without the suspect being inconsistent. For example, suppose the interrogator asks the subject at day 1, "Are you sure you will maintain tomorrow (day 2) that you were located at x at the time of the murder?" If the suspect answers "yes", this may be interpreted as K_1K_2x. Say that when tomorrow comes, the suspect claims that he has forgotten where he was at the time of the murder. This may be translated as $\neg K_2x$. The suspect has failed an epistemic test.

[19] This application was suggested by Sjur Flåm.

7. Application: Single-Agent Conventions

It is an everyday event that we act according to rules of behavior. Some of those rules regulate our interaction with other people, others are solely a means to regulate our interaction with ourselves. Let us consider an example of the latter, storage of keys to the car, and derive epistemic conditions for such a convention to be successful.

After driving home the agent decides where to store his car keys. When he needs the car again he attempts to guess where he placed the car keys the previous day. Of course, if he had perfect recall this would be a simple coordination problem. But since he is absent-minded, and furthermore is aware that he is absent-minded, he tends to stick to the following convention: place the car keys on the shelf in the living room. The location of this particular storing place does not give him any intrinsic pleasure, nor is it initially better than any other storing places. The reason why he sticks to using the shelf is that, since he is more accustomed to using the shelf, he more easily remember where he put the keys than if he were to use another storing place. Thus, what starts out as being strictly conventional behavior becomes optimal behavior.[20]

Denote a storing convention *c*. Under what conditions can an absent-minded agent be certain of getting a high payoff (find the keys at once) given that he has adopted the convention *c*? To answer this question we need to do some reasoning related to what the agent knows on day 1 and 2. Evidently, on both days the agent must know the convention; K_1c and K_2c must hold. But that is not enough. If he believes that he may forget *c*, he cannot at t_1 be certain whether he will know c at or not at t_2. If he forgets *c* he might believe at t_2 that the convention was *d* (e.g., "put keys in right pocket", and act accordingly. Thus K_1K_2c must hold. But again, if the agent at time 2 is uncertain whether he knew that he was going to remember *c* at t_1, he might believe that he acted according to *d*. Thus $K_2K_1K_2c$ must also hold. This also holds for $K_1K_2K_1K_2c$. The same argument can be done for time 2.

[20] At some point a better storing place may be available to me (better in the sense that it has higher intrinsic value). Then whether I should switch to this storing place or not is a trade-off between the long term gains of having the keys a better place and the short-term loss of forgetting more often where the keys are.

The above argument motivates the observation below. First a definition. We say that a single agent convention c is *Idiosyncratic Knowledge* for an agent if K_1c, K_2c, K_1K_2c, K_2K_1c, $K_1K_2K_1c$, $K_2K_1K_2c$ etc., holds. Then we can make the following observation: *A sufficient condition for the agent to solve the co-ordination problem by designing a convention is that the convention is Idiosyncratically Known to the agent.*

The argument above shows that the solution to a single agent coordination problem is similar to the solution of a multi-agent co-ordination problem: the convention is common knowledge among the agents.[21] The similarity exists for precisely the same reason that multi-agent coordination problems can be modeled as single agent decision problems with imperfect recall and vice versa.[22] One difference, however, is that while communication among agents in space - at least potentially - works in both directions, memory works only from past to present. The similarity between the multi-agent case and the intertemporal case is that memory in single agent problems plays the same role as communication in a multi-agent setting. Acts of speech are communication through space between distinct agents. Memory is communication through time between different selves of the same agent. In coordination games free communication establishes a coordinating mechanism for the agents. Perfect memory in the same way coordinates an agent to choose to look for the keys and to store them in the same location.

Although making conventions for ourselves is common, reasoning through time of the type described above is - analogous to multi-agent conventions - not common. Thus, there seems to be a gap between what our intuition and the logic tell us. A paradoxical feature of the solution is that it requires a lot of recall from the agent; not only must he know that he knew c yesterday, he must also know that he knew that he would know c at time 2, he must know that he knew that he will know etc. In sum, stating the conditions seems to come close to stating that the agent has perfect recall and realizes it. It then falls naturally to ask why we choose to make conventions with ourselves in spite of the fact that they require a very high degree of recall to be completely successful. A pragmatic answer is that even if we can never be certain

[21] Not only must the convention be known, but it must also be known that it is known, known that it is known that it is known and so forth.

[22] Take, for example, the game of bridge. It can be modeled either as a four-person game with incomplete information where North-South and East-West have identical preferences, or as a two-person incomplete information game with imperfect recall.

that our conventions will be successful, expected payoff may be higher from using a convention than by directing our recall problem in other ways (like employing someone to remind us), or not directing it at all. An avenue for future research could be to model the adoption of conventions of the type discussed above. An exciting task in this project we think is to model the agent's beliefs about his own absent-mindedness.

8. Conclusion

We have put bounds to memory loss by pointing to formulas that cannot be forgotten and formulas that must be forgotten. The results have theoretical value in two ways. They show that models assuming that the probability of forgetting is distributed uniformly over knowledge, or anything like that, are too simplistic when knowledge is conceived to have some linked structure governed by logic. Moreover, and we believe most importantly, the result answers one criticism used against modeling agents as boundedly rational: the concept of bounded rationality is too vague to be taken seriously by economic practitioners. The results illustrate that there are bounds to at least on one interpretation of bounded rationality, namely the hypothesis that the agent is absent-minded.

The results are based on defining a memory loss operator in an essentially static version of the S5 logic. One weakness of this adoption is that the logic does not formalize the notion of time, as in the logic presented by Fagin, et al.(1995). It is for example not evident how to separate the notion of memory loss from the notion of asynchrony, the concept that an agent does not know what time it is. The asynchrony concept is prevalent in the puzzling absent-minded driver problem of Piccione and Rubinstein (1996). A strength, compared to Fagin, et al. (1995), is that the number of axioms in the language presented in this paper is fewer, and therefore the F operator is open to more than one interpretation; instead of denoting memory loss it may formalize a notion of asymmetric information.

The framework in this paper may also be seen as a conceptual analysis clearing the ground for experimental work on absent-mindedness. First, empirical results showing that very "rational" agents (those who are highly *ex-ante* aware of their absent-mindedness) forget more than those who are less aware are not necessarily paradoxical. In fact this is what we should expect, simply because being *ex-ante* aware implies forgetting a host of formulas describing one's awareness. Second, why do we often forget premises of a theory (basic facts) even though we remember the conclusions (deduced facts)? For example, few economists forget the content of the fundamental theorems of welfare economics, but how many are able to state the exact underpinnings of the theorems? Although psychological factors may be important, this paper shows there is a logical explanation to the fact that conclusions are less likely to be forgotten than premises. Agents with a moderate level of logical sophistication cannot know the premises of a theory without knowing its conclusions, thus forgetting conclusions alone is not possible.

We have not discussed which elements of knowledge we believe are most likely forgotten. The problem should partly be left to empirical work, and partly to a theory of decisions under imperfect recall, which is not yet developed. An anticipation we get from working on this paper is that a theory of decision under imperfect recall should carefully two things. First it should consider the mechanisms an agent can use to reduce the impact of his absent-mindedness, "internal" or "external".[23] In the last part of the paper we discussed an internal mechanism, storing place conventions. The idea behind this internal mechanism is that an agent, being aware of his absent-mindedness, adopts a convention that gradually reduces the complexity of recall. It was shown that epistemic conditions for such a single agent convention to be successful is that the convention is «idiosyncratically known» by the agent; not only is the convention known, it is also known that it is known at the other point in time, and known that it is known that it is known etc. Second, the behavioral implications of an agent's absent-mindedness depends on his level of absent-mindedness, but - at least as important, also on the agent's beliefs about his absent-mindedness. For example, the behavioral implications of a given level of absent-mindedness are quite different for a person that believes he has perfect memory than for a person that believes he is virtually without memory. Therefore, we think is important for a theory of decision under imperfect recall to properly model the evolution of an agent's beliefs about his own absent-mindedness.

[23] Bergson (1919), a classic on memory and memory loss, emphasizes Humean association in explaining what clusters of knowledge that are remembered, and the use of mental techniques, rather than logical inference, in describing how humans retrieve knowledge from memory.

9. References

Bacharach, M. & Mongin, P. (1994). Epistemic Logic in Game Theory, *Theory and Decision*, **37**, 1-24.

Battigalli, P. & Bonanno, G. (1997). The logic of belief persistency. *Draft.*

Bergson, H.: 1919, *The Spiritual Energy*, PUF-Quadrige 1985 edition.

Dow, J. (1991). Search Decisions with Limited Memory, *Review of Economic Studies*, **58**, 15-41.

Fagin, R., Halpern, J. Y., Moses, Y. & Vardi, M. Y. (1995). *Reasoning about Knowledge*, The MIT press.

Gärdenfors, P. (1988). *Knowledge in Flux***,** The MIT press.

Haack, S. (1978). *Philosophy of Logics*, Cambridge University Press.

Halpern, J. Y. (1986). Reasoning about Knowledge. An Overview. In: *Theoretical Aspects of Reasoning about Knowledge, vol. I.* Morgan Kaufman Publishers.

Hintikka, J. (1962). *Knowledge and Belief.* Cornell University Press.

Hughes, G. E. & Creswell, M. J. (1968). *An Introduction to Modal Logic.* Methuen Press.

Hvide, H. K. (1997). Self-Awareness, Uncertainty, and Markets with Overconfidence. *NHH discussion paper 9/97.*

Kaneko, M. & Nagashima, T. (1996). Game Logic and its Applications I + II; I: *Studia Logica*, **57**, 325-54, and II: forthcoming in 1997 in *Studia Logica.*

Lemmon, E. J. (1959). Is there Only one Correct System of Modal Logic? *Proceedings of the Aristotelian Society XXXIII*, 23-44.

Lipman, B. L. (1995). Information processing and bounded rationality: a survey. *Canadian Journal of Economics*, **38**, 42-67.

Modica, S. & Rustichini, A. (1994). Awareness and Partitional Information Structures. *Theory and Decision*, **37**, 107-124.

Piccione, M. & Rubinstein, A. (1997). On the Interpretation of Decision Problems with Imperfect Recall. *Forthcoming in Games and Economic Behaviour.*

Shoham, Y. (1989). Time for action, *Proceedings, 11th International Joint Conference on Artificial Intelligence*, 954-959.

Chapter 3

Self-Awareness, Uncertainty, and Markets with Overconfidence[1]

Abstract

Standard decision theoretic models take as given that agents have perfect self-awareness; they have complete knowledge of their own abilities. In the first part of the paper we combine philosophical and empirical arguments to attack the perfect awareness assumption. In the second part we ask whether uncertainty about oneself needs to be modeled differently than uncertainty about the world, and argue that with the exception of a disturbing circularity aspect, the answer is no. In the last part of the paper we speculate over the implications for market behavior of a certain form of lack of self-awareness; overconfidence. The originality we claim is in the projects we suggest - and do not properly undertake - along the way.

Keywords: Bounded Rationality, Introspection, Learning, Overconfidence, Self-Awareness, Self-Knowledge, Uncertainty.

[1] Thanks to Ken Binmore, Alexander Cappelen, Jerker Denrell, Terje Lensberg, James March, Heine Rasmussen, Fred Schroyen, Bertil Tungodden and participants at the FUR VIII meeting for valuable comments and discussions.

> *I confess that in 1901, I said to my brother Orville that man would not fly for fifty years ... Ever since, I have distrusted myself and avoided all predictions.*
>
> Wilbur Wright in 1908.

1. Introduction

Let us begin with an example.

Example 1. In a classroom, the teacher informs the students of the axioms, the inference rules, and the definitions of Euclidean geometry. He continues by instructing the students to deduce five theorems of that geometry. A student who accomplishes it, receives $10. A student who makes a try, but only manages to deduce four theorems or less, receives nothing. A student who leaves class without trying, receives $5.

What would an agent choose? The common sense suggestion - which seems healthy - is that an agent's choice depends on his beliefs about his (logical) ability. If he believes that he is weak in proving theorems he takes the $5 and runs, and if he believes that his logical ability is strong he tries to deduce the five theorems. Notice that common sense views the decision as one under uncertainty; an agent may be uncertain about his own ability and thus have formed beliefs about it. Common sense stands in contrast with established theories. A theory of perfect rationality is plainly not of much help in analyzing the problem since a perfectly rational agent would simply deduce five theorems on the spot and grab the $10. Supposing that an agent cannot deduce five theorems on the spot, which seems reasonable, we seem to be in the realm of "bounded rationality". What can recent models of bounded rationality say about choice in the theorem-proving problem? Also very little, we argue. To explain why, let us do a quick survey of the bounded rationality literature we alluded to.

The basic idea of the literature is that bounded rationality refers to choice that is imperfect in the sense that it is often not the "correct" one, but is sensible in that it can be understood as an attempt by the agent to do reasonably well given his cognitive limitations (Lipman 1995). To be precise; boundedly rational agents maximize some objective function, just like perfectly rational agents, but with the difference that boundedly rational agents optimize taking into account their own cognitive constraints.[2] This sounds plausible but has an implausible corollary; that agents are perfectly aware of their own cognitive constraints.[3] We denote this assumption the *perfect awareness assumption*. Under perfect awareness, whether or not to participate in the theorem-proving gamble is a trivial choice, just as it was for a perfectly rational agent. If the agent has weak logical ability he knows it, and chooses the $5. If his logical ability is strong he knows that too, and prefers theorem proving to the $5.

Thus there do not seem to exist theories on decision making that properly capture the choices of agents that are not perfectly aware of their own abilities. Should we care? We try to answer whether imperfect awareness is important by asking and tentatively answering two questions. First, does uncertainty about oneself need to be modeled differently than uncertainty about the world? Second, can imperfect awareness shed new light on behavior in markets? Before discussing these two questions we propose some arguments in favor of imperfect awareness; in part 1 we combine philosophical and empirical arguments to attack the perfect awareness assumption. We propose a heuristic framework to define self-awareness; an agent is modeled as two layers, where the lower one does the «dirty work» of observing the world and calculating beliefs about the world, and the upper level receives these beliefs from the lower level and chooses an action for the agent as a whole. Within this tentative framework we discuss properties of agents that are imperfectly aware of their abilities; we model imperfect awareness as the upper level being uncertain about the functioning of the lower level.

Part 2 is mainly motivating the third and the fourth part. We list three reasons for why we think imperfect awareness is important. First, it seems that imperfect awareness can shed light on learning theories; second, imperfect awareness may mean that we have to do some

[2] For example, an agent knows that he is going to forget certain facts, and given this knowledge constructs an optimal decision rule (Dow 1991, Piccione & Rubinstein 1997); an agent knows that he has limited attention span and therefore concentrates effort on a small amount of markets (Fershtman & Kalai 1993, Rubinstein 1993); or an agent knows his cost to processing information and therefore takes care not to assemble too much information (Conlisk 1988, 1996). For a recent book dealing with optimizing boundedly rational agents, see Rubinstein (1997).

[3] Even though the bounded rationality models have been interpreted in terms of deliberate optimization it is not obvious this is the only tenable interpretation. As with models of perfect rationality, an "as-if" defense is an interesting alternative. See Hvide (1998).

rethinking on what models of bounded rationality should look like, and third, it seems that imperfect awareness can explain economic phenomena that otherwise are not easily explained.

In part 3 we elaborate on whether it is reasonable to assume that agents will become perfectly aware of their own cognitive constraints. To answer this question we first ask in what sense we need new models to model imperfect awareness. Our tentative answer is that there is - with the exception of a certain circularity issue involved when modeling boundedly rational learning - not a big difference between being uncertain about the world and being uncertain about oneself.

In part 4 we elaborate on why we think imperfect self-awareness is important to the functioning of certain markets. Our starting point is a finding from the psychology of judgment literature: It seems that agents not only are imperfectly aware of their abilities, but also they seem to be consistently overconfident about them. We speculate over what overconfidence may imply for market analysis in general, and the credit card market in particular. This part is based on empirical findings from Asubel (1991).

There is a range of related work on self-awareness (not necessarily using this term) within at least three traditions; the philosophy of mind literature, the decision under uncertainty literature, and finally the psychology of judgment literature. With the exception of some work by Daniel Dennett, the philosophy of mind literature tends to focus on ontological and epistemological aspects, and ignore decision making. The decision under uncertainty literature has with some exceptions (some of them to be addressed later) not yet been involved with making models of self-awareness. The psychology of mind literature tends to focus on cognitive biases and hypothetical choices while ignoring incentives and modeling of decisions. We emphasize that the paper - with the exception of part four - to a large extent is a convex combination of works within the above three traditions.[4]

[4] Let us list three general references. Our basic view on self-awareness corresponds well with the much-quoted Binmore (1987b), which offers a more satisfying model of self-awareness than we do. Lipman (1995) gives an overview of some recent work on bounded rationality. Some of the references to the psychology of judgment literature is from chapter 19 of Plous (1993).

2. Self-Awareness

We understand an agent's self-awareness as the beliefs he holds about his own cognitive abilities. With cognitive abilities we mean abilities in information processing and in problem solving.[5] We begin by proposing a simple framework - that takes the viewpoint of an outside observer - for defining self-awareness. Sometimes we shall use just «awareness» instead of «self-awareness».

Implicit in the notion of self-awareness is a hierarchical model of the mind. Ours looks like this: A certain part of the brain receives information about the world and transforms the information into beliefs. These beliefs are in turn, with or without deliberation, delivered to other parts of the brain, which then acts upon the beliefs transmitted. The sender of these beliefs we denote by level 1. The receiver of the beliefs we denote level 2. Level 2 is imagined functioning in pretty much the same way as level 1, but the spirit of it is that level 1 has specialized in «computational» problems while level 2 functions in a more heuristic way. A useful analogue is that level 1 is the personal computer, and level 2 is the personal computer user.

A central intuition is that level 2 may be suspicious to the quality of the output from level 1, and thus «corrects» it. But in that case we can imagine a level 3 that wants to correct the correction of level 2, a level 4 that wants to correct the correction of level 3 and so on. For example, say that a person assesses the length between two points A and B. Level 1 computes for 15 seconds and comes up with an answer, "The distance between A and B is 50 yards". Now, the person may have a history associated with assessing distances which have taught him, i.e., level 2, that he is bad in assessing distances. Specifically he may know, for example, that on his first hunch he tends to overestimate the distance. Thus he comes up with a revised belief, "I believe the distance between A and B is 40 yards". Of course, the agent may have beliefs about how level 2 is functioning as well. He may reason, "I often believe that my ability in assessing distances is worse than it is. In fact my immediate hunch often makes my best guess". This makes him revise his belief again, "I believe that the distance between A and

[5] In some examples we will also understand memory capacity as a cognitive ability.

B is 50 yards". This way of forming beliefs about the information of lower levels obviously poses a regress problem.

Our view on the regress should be stated right away. Theoretically there is an infinite regress but surely there must be a cut-off point where the agent stops reasoning. This cut-off level *n* should be endogenously determined through (expected) cost-benefit considerations. At some point the cost of continued reasoning about lower-level functioning exceeds the expected gain of continuing.[6] Say that the reasoning stops at level *n*. In that case we have the following procedure. The agent does some reasoning about level 1 and is led upward the «ladder of doubt» until he reaches level *n*, which is the highest level he finds it worth considering (of course *n* may vary from problem to problem). From level *n* he descends the ladder again to arrive at a conclusion regarding the output from level 1. Given this conclusion the agent chooses an action. It follows that «self-awareness» is not only level 2's conjectures of level 1 but the conclusion the sequence of levels from level 2 and upwards reaches on the functioning of level 1. To ease exposition we will refer to this hierarchy of reasoning about oneself starting from level 2 simply as level 2.

There is some controversy whether the finite-layer approach to decision making we sketch here is appropriate. As noted by among others Mongin & Walliser (1988) and Lipman (1991), modeling a person's decision making process may (from the perspective of an outside observer) at advantage be modeled as an infinite regress converging to a fixed point rather than a finite regress with an «artificial» cutoff. There are subtle issues concerned here, but one reason to prefer the fixed point model is its tractability; various results from mathematics can be applied. In spite of its tractability it is not obvious that it also comes closer in realism. We proceed taking the finite layer model as given. [7]

[6] This may sound simple but the problem of finding an optimal *n* is in general a very complex problem. As Lipman (1991) and Conlisk (1989), (1996) point out in a similar setting, this problem may indeed not have a solution.

[7] A different model of the mind could be a circular arrangement where the different parts, say two, take turns in deliberating each others output. A problem with such a model may be how to incorporate the fact that some part of the mind must make the final decision, without that bringing in an implicit hierarchy.

Perfect Self-Awareness

A benchmark case occurs when agents are perfectly aware of their information processing. By perfect self-awareness we do not necessarily mean that an agent's mind is «transparent to itself», but rather that the outcomes of cognitive processes are known to an agent. For example an agent may know from experience that he is able to deduce five theorems of Euclidean geometry without having a clear hunch on how he really does it. That kind of knowledge is clearly empirical.

An a priori defense of perfect awareness, on the other hand, could go like this. Perfect awareness follows from the Cartesian «fact» that the mind is transparent to itself. Through introspection the mind can reveal every feature about its own functioning and thus perfect awareness is probable, if not obvious. In a strict sense this statement is clearly false in view of Gödel's theorem, which briefly states that any moderately complex logical system cannot be complete without being inconsistent; there are propositions about the system that are valid but still cannot be proved within the system.[8] The grain of truth in the Cartesian position lies in the fact that we are probably better at predicting the functioning of our own cognition than predicting the functioning of other people's cognition. We have what philosophers of mind call «privileged access» to our own mind; in an obvious sense a person can look into his own mind in a way that another person cannot, but it does not follow that he can dispassionately assess what he observes.

With the model outlined above we defy the transparency defense; perfect awareness is impossible to obtain through introspection. Introspection takes the form of level 2 «scanning» level 1. Trivially, to have perfect awareness we must be able to scan the scanner, scan the scanner of the scanner and so forth. This leads to a vicious regress; the scanning operation may itself be scanned, and so on, but we must in the end reach an unscanned scanner (level n+1). Of course, the unscanned scanner is not a logically unscannable scanner, for it is always

[8] See Binmore (1987b) for a more thorough discussion on the implications of Gödel's theorem for self-awareness.

possible to imagine a further scanning operation; although the series must end somewhere of economic reasons, it need not have ended at the particular place it did end.[9]

It is not difficult to find support in the philosophy of mind literature for the view that degree of self-awareness is an empirical question, not an a priori one. For example Churchland & Sejnowski (1989) states that, «Inner knowledge, like outer knowledge, is conceptually and theoretically mediated - it is the result of complex information processing. Whether our intuitive understanding of the nature of our inner world is at all adequate is an empirical question, not an a priori one». Armstrong (1968), p. 115 with a similar point; «I do not think that we can overestimate the importance for the philosophy of mind of a completely ungrudging acceptance of the possibility of introspective error and of unconscious mental states. Again and again, the Cartesian picture of our own mind as something perfectly transparent to us stands in the way of philosophical progress. We must see our cognitive relations to our own mind as like our cognitive relation to anything else in nature. We know in part, guess in part, in part we are mistaken and in a large part we are simply ignorant.»

Also work done by numerous psychologists suggest that perfect self-awareness is an implausible assumption. In fact, a seemingly robust findings in the psychology of judgment is that people tend to be overconfident in assessing their abilities (see e.g., De Bondt & Thaler 1994, Plous 1993, Vallone et al. 1990, and Liechtenstein et al. 1982). [10] Let us return to the overconfidence issue in part 4.

3. Imperfect Self-Awareness

In the previous section we concluded that perfect self-awareness seems dubious for both philosophical and for empirical reasons. Should we care? Is imperfect self-awareness

[9] As expressed in a later section, instead of viewing introspection as self-scanning one may view it as a simulation exercise. By saying that level 2 introspects level 1 we then mean that level 2 takes the information it has about level 1 and simulates the functioning of level 1. The outcome of this exercise is level 2's estimate of the functioning of level 1. Of course, we can imagine a level 3 that simulates the simulation of level 2 and so forth.

[10] Even if the evidence in favor of overconfidence seems strong, there are situations where humans seem to be underconfident in their assessment of themselves, for example subjects tend to be underconfident of their ability to choose the larger of two irregular areas (Dawes, 1997).

important to understanding social phenomena? To motivate, consider three reasons for why we think the answer is yes. In part 4 and part 5 we elaborate on the second and third reason sketched below.

Learning about the world

First, imperfect awareness is indirectly important to learning about the world;[11] seemingly innocent information about oneself may have much stronger effect on beliefs about the world than information about the world itself. For example, my firm ranking of Mozart ahead of Beethoven is more likely to be upset by a finding of my own lack of musicality than the finding of a flaw in one of Mozart's main symphonies. Another example. In a tricky case, even by Sherlock Holmes' standards, Holmes is able to deduce from the fact that the dog did not bark that the burglar had not visited the house that night. Holmes way of inferring this is through the familiar *modus tollens*. Say that p = «the burglar went into the house at time x» implies q = "the dog barks at time x". Then by modus tollens, $\neg q \Rightarrow \neg p$. This ingenious way of reasoning shocked Dr. Watson to exclaim: «Holmes, you are incredible! Not only do you infer facts from what did happen, but also from what did not happen.» An interpretation of the story is that not only did Dr. Watson learn about Holmes' ingenuity through this experience, but also he learned about his own level 1's lack of reasoning power; it was not able to use modus tollens. From now on his level 2 could - when receiving beliefs from level 1 - take into account level 1's weakness in logic, and thereby for example put wider confidence intervals to estimates obtained from level 1.

[11] For example, learning about the world in a Bayesian framework implies receiving a signal which leads to revisions of posterior beliefs (about the world). Bayesian learning would in our framework mean that level 1 improving the beliefs (about the world) it transmits to level 2. To use a Bayesian framework for modeling self-awareness is not innocuous, however. As pointed out by Ken Binmore, Bayesian decision theory applies only in small worlds, but a world that includes oneself is necessarily large.

Bounded Rationality

Second, it seems that imperfect awareness is important to models of bounded rationality.[12] To the point, degree of self-awareness seems crucial to the behavioral implications of an agent having certain «cognitive limits». For example, whether or not an agent would accept the bet in example 1 would not only depend on his true theorem-proving ability, but also on his beliefs about his theorem-proving ability. Without going into details, the results obtained in the literature on bounded rationality referred to in the introduction relies heavily on the perfect awareness assumption. In that case, it becomes important to investigate whether learning processes would tend to converge to perfect awareness.[13]

To be able to model learning of cognitive limits we should first ask the basic question of in what sense being uncertain about properties of oneself is different from being uncertain about properties of the world. That is what we do in part 4.

Overconfidence

Third, it seems that a wide range of social phenomena can be better understood by applying an explanation with imperfect self-awareness as an ingredient. There has already been done some work in this direction. Let us give three examples. In a classic matching model, Jovanovic (1979), an agent may accept a low paid job if this job gives him more information about his abilities. Orphinades & Zervos (1995) discusses the optimal behavior for agents that enjoy some activity but worry about being «hooked». The papers that have used imperfect self-awareness has been rather sloppy in their assumptions on what kind of deviation from correct beliefs about oneself that can be accepted. Hvide (1997) proposes a consistency condition, condition (C), which briefly says that for each overconfident agent in the population there is one underconfident agent. Hvide (1997) goes on to discuss how imperfect self-awareness may

[12] Lipman (1995) offers some interesting comments on the role of self-awareness for boundedly rational agents. Among others he points out that S4 epistemic models (Hintikka type of epistemic logic without the negative introspection axiom.) of reasoning (see Geanakoplos 1989 for an application) seems to rely on lack of self-awareness from the agents.

[13] It seems important but perhaps too difficult in the short run to model how these cognitive limits may change as one learns about them. Some comments to this problem are offered in the next section.

explain why we have «Spencian» unproductive, education in spite of individual performance being contractible by firms. In light of experimental evidence showing that real world agents tend to be overconfident, (C) seems to be unrealistic. Therefore, an interesting project would be to construct models where (C) is violated in the direction of overconfidence. We do some preliminary speculations on such a project in part 5.

4. Imperfect Self-Awareness Compared to Uncertainty about the World

What is the difference between uncertainty about oneself and uncertainty about the world? Is there any difference? We have found four candidate properties. The fourth property points to a difficult circularity problem when modeling boundedly rational agents who learns about themselves, while the first three seem rather inconsequential.

First, the regress issue makes perfect awareness in a trivial sense impossible. Not only can level 2 have doubts about the problem solving abilities of level 1, level 2 can also have doubts about the sense data it receives from level 1. For example, level 2 can doubt that level 1 tells the truth when level 1 informs level 2 that the sun shines outside. For our purposes this argument presses the skepticism a bit too far.

Second, when being uncertain about oneself one may - in contrast to when being uncertain about the world - try to resolve this uncertainty with introspection. What we mean by introspection is that level 2 reasons to answer questions of the following type (which may or may not be counterfactual). Say that level 1 were to perform task x. Then how would its performance be? It seems clear that for many tasks, for example in the theorem-proving of example 1, introspection at least potentially improves an agent's beliefs. However, introspection seems to be pretty much equivalent to simulation, and introspection of introspection equivalent to simulation of simulation and so forth. Thus introspection does not seem to constitute a fundamental difference between uncertainty about the world and uncertainty about oneself.

Third, obtaining information about oneself may change properties of oneself, in contrast to assumptions in standard decision theoretic models. a) as mentioned before, if level 2 obtains information about level 1's abilities through experimentation of some sort, such experimentation may lead to a change in level 1's ability. For example, say that level 2 wonders about level 1's theorem proving ability, and three days in a row tests level 1 by telling it to deduce some theorems of Euclidean geometry. The simple point is that such testing may, in addition to giving level 2 data on the ability of level 1, improve the theorem-proving ability of level 1. Thus collecting data about ability may change ability. This seems to be an awkward problem; the parameters change as we learn about them. However, that does not pose a particularly difficult estimation problem, where we estimate the change in ability as a function of trials.[14] b) to become aware of some cognitive constraint may in itself have an altering effect. This is a well-known lesson from psychotherapy; when a client becomes aware of some traumatic experience the effect of this experience may gradually fade away. It is obscure to us how and when exactly a mechanism like this works. For example, it seems unreasonable to claim that knowing about one's absent mindedness reduces absent mindedness in any significant way. However, for some "irrational" processes, like failing to deduce that √9 = +-3, it may well be that this cognitive constraint vanishes when one becomes aware of it. c) awareness may give level 2 an incentive to change level 1 by for example taking a math course. We may see this as level 2 reprogramming level 1.

There is a fourth difference that may be important. In Bayesian models of learning, practically all uncertainty about the world can be resolved with sufficient information. With learning about oneself it may be different; all uncertainty cannot be resolved, there may be bounds to the degree of self-awareness that is possible. To be specific, if the process of learning about oneself involves using the same properties of oneself as one is learning about, this circularity may put bounds to the degree of awareness that is possible.

Consider two examples. First, the theorem proving example; if making judgments about one's theorem-proving ability makes use of the same kind of ability as theorem proving does, then we may expect a bad theorem-prover to also be bad in making judgments about his theorem-

[14] A simple method is to use logistic regression, common in the literature on epidemiology. With this method we can for example estimate a probability p for succeeding in doing a certain task, when p is a function of number of trials. Of there is a problem in guessing the right functional form to estimate, as there is in «normal» regressions.

proving. Conversely, a good theorem prover can be expected to be better in making judgments about his theorem proving abilities than a bad theorem prover.[15]

Of course there does not have to be the positive correlation between abilities that the examples indicate. There are two other interesting possibilities. First, there is nothing inconsistent in level 1 being screamingly «irrational», and level 2 being perfectly rational. For example, level 1 may be a useless theorem-prover, and level 2 can be sophisticatedly aware of this fact. Second, one could also think of cases where it is the other way around; level 1 has a high ability in theorem-proving but level 2 is unaware of this fact.

Our intuition is that a high ability level 2 and a low ability level 1 is more likely than the opposite; a low ability level 2 and a high ability level 1. That is quite obvious if we look at the most salient cause of high ability, practice. As discussed in the previous section, practice has two effects, increasing ability and giving information about ability. Thus more practice implies both higher ability and lower variance on estimates about ability.[16] This implies a certain asymmetry; agents that are good at theorem-proving have a more realistic opinion of themselves as theorem provers than bad theorem provers.[17] Furthermore, the argument suggests that in a population of agents we can expect a positive correlation between abilities of levels 1 and 2.

Both the practice argument and the related circularity argument suggest a positive correlation between ability and quality of conjectures about ability. Experimental data suggest otherwise. To be specific, experimenters have investigated the closely related question of degree of correlation between accuracy and confidence in estimation.[18] In a number of experiments, investigators have first asked a group of subjects their estimate of certain parameters, and then their degree of confidence in their estimate. To a large extent, these studies suggest that confidence to estimates is virtually uncorrelated to how accurate the estimate actually is. For example, a famous study, Goldberg (1959), assessed the correlation between correlation and

[15] A second example: An absent-minded person needs some memory to become aware of his absent-mindedness; to some extent he must be able to record in which situations he tends to forget and in which situations he tends to remember. At the extreme, a person without memory can in a certain sense not know that he is without memory.

[16] A similar point is made in March & Shapira (1987).

[17] There are surely other reasons, e.g., evolutionary, for why we would expect a positive correlation, but for brevity we skip them here.

[18] Dubbed «calibration» in the psychology of judgment literature.

confidence in clinical diagnoses. Goldberg found two surprising results. First, all three groups of judges - experienced clinicians, trainees, and non-psychiatrists - correctly classified 65 to 70 percent of the patients. There were no differences based on clinical experience; secretaries performed as well as psychologists with four to ten years of clinical experience. Second, there was no significant relationship between individual diagnostic accuracy and degree of confidence.[19]

While in example 1 we consider self-awareness from an ex-ante point of view («making predictions about oneself») the Goldberg study takes an ex-post point of view. It asks subjects of an estimate of some uncertain quantity and then asks the subjects to assess their confidence in their estimate. Should we expect different results on ex-ante and ex-post confidence? Since ex-post confidence is built on some concrete estimation experience, ex-ante confidence is based on even less information. This could have two effects. One that the subjects become more cautious, and two that their confidence becomes even more biased.[20]

To sum up, we have considered arguments for why learning about oneself needs to be modeled differently than learning about the world. We found three candidate properties that we viewed as inconsequential, and one property - the circularity aspect - that could potentially make a difference.

The circularity aspect is a potential difficulty when modeling learning by boundedly rational agents. It seems clear that the question of whether perfect awareness is obtainable for bounded rational agents, and under which conditions, needs careful modeling. However, it is not obvious that imperfect awareness implies bounded rationality. For example, the signaling model of Weiss (1983) and the job matching model of Jovanovic (1979) include agents who are uncertain of how well they will perform in certain jobs. This may have the interpretation that the agents have some uncertainty about the nature of the job, but it may also be consistent with imperfect self-awareness. Even if we were convinced that the latter interpretation is the right one, it is not clear that we would prefer to model the agent as boundedly rational. We

[19] Perhaps psychiatry, with its lack of secure knowledge, is not the best field to find examples from. The Goldberg study is illustrative, and at any extent, later literature on calibration has shown similar tendencies of their subjects.

[20] Investigations performed by Valloner et al. (1990) suggest that subjects are just as overconfident in ex-ante.

may simply choose to model the agent as perfectly rational but with some lack of information, just as Weiss and Jovanovic do.

Generally, to ignore bounded rationality seems to be a wise strategy when considering borderline cases between boundedly and unboundedly rational agents; particularly considering the state of bounded rationality models. In the next section we will implicitly choose exactly this strategy when discussing a market with overconfident agents. To explain how overconfidence comes about, one would probably need an explanation based on bounded rationality, but, we think, overconfidence can be a very interesting phenomenon also from a rational, lack of perfect information, point of view.

4. Markets with Overconfidence

In this last part of the paper, we explore the implication of overconfidence for market settings. Let us emphasize that the material below should be viewed as preliminary speculations.

In the single agent case it is simple to define overconfidence; an agent is overconfident if his beliefs about his ability are higher than his actual ability (suppose that ability is measured along one dimension). When defining a measure of confidence for a population it is not obvious how to weigh the underconfident against the overconfident. We propose a simple measure; a population is (under-) overconfident if the average belief about ability is (lower) higher than actual average ability.[21] Let us formalize this definition in a simple model. Suppose there are two types of agents in the population, those with low ability and those with high ability. Let the population share of the low ability type be θ_L and the population share of the high ability type be θ_H, and let the two types be indistinguishable in physical appearance. Each agent holds a subjective belief b on his ability type. The interpretation of a certain belief, say 3/4, is that a person believes that he is a high type with probability 3/4 and a low type with probability 1/4. Let $f_L(b)$ denote the density of beliefs for the low type, and $f_H(b)$ the density of beliefs for the high type.

[21] An alternative measure of overconfidence could be that the distribution of beliefs first order stochastically dominates the distribution of abilities.

Suppose nobody is underconfident or overconfident. Then,

$\int_0^1 zf_L(z)dz = 0$, and, $\int_0^1 zf_H(z)dz = 1$. Hence, $\theta_L \int_0^1 zf_L(z)dz + \theta_H \int_0^1 zf_H(z)dz = \theta_H$.

We therefore define overconfidence by the criterion,

$$\theta_L \int_0^1 zf_L(z)dz + \theta_H \int_0^1 zf_H(z)dz > \theta_H.$$

Correspondingly, the population is underconfident if and only if the expression on the left side is less than θ_H.

Overconfidence in the Credit Card Market

In considering overconfidence in a market setting we look at the credit card market.[22] Other markets that could be analyzed in roughly the same fashion is the market for education and certain betting markets.[23]

When considering which credit card to go for, consumers should compare the fixed fees, the transaction costs for ordinary purchases, and the interest rate they pay on overdrawn accounts. The relevance of the latter cost depends on the probability an agent assesses for him coming into a situation where his liquidity indicates that it is rational for him to borrow. For the sake of argument, suppose that his assessment of how probable it is for him to borrow on this high rate depends on his conjectured ability in liquidity engineering. Suppose further that there are two types of agents, the low type and the high type. The low type is bad in liquidity engineering (and thus pays a large fine) while the high type is good in liquidity engineering.

[22] Underconfidence in a population could be used in much the same way as overconfidence is here to explain why we have certain insurance markets.

[23] Golec & Tamarkin (1995) test empirically whether bettors prefer long shots because they are risk-lovers or because they are overconfident. They find support for the overconfidence hypothesis.

In case of overconfidence, credit card companies could make a profit by offering a credit card contract that would be good for the «above average» ability in liquidity engineering person to accept, but bad for the «below average» ability in liquidity engineering person to accept. Such a contract would typically have a small fixed fee, a small transaction fee, and a large penalty for overdrawn accounts. Too many agent would self-select to buying credit cards and firms would make a profit.[24] In the long run, profits are eliminated by free enter of firms, but a rationale for the credit card industry would still be to «fool» the overconfident.[25]

But should not beliefs change along the way? Intuitively, we would expect «market experience» to adjust beliefs to a state where there are no profits to be made by firms. Overconfident agents would (after consistently paying larger fines than expected) gradually realize that they were overconfident and adjust their beliefs about themselves downwards.[26] If we imagine a process where beliefs are gradually modified with experience, what restriction on the distribution of beliefs must hold for there to be no profit opportunities? Let us propose such a no-profit condition, condition (C).[27]

$$\text{(C)}\ \frac{\theta_H f_H(b)}{\theta_L f_L(b)+\theta_H f_H(b)} = b, \text{ for all } b\in[0,1].$$

[24] We are not assuming that firms know more than individuals about ability versus perceived ability of the population. Even if an agent knows that his socioeconomic group is overconfident in aggregate, it is not clear that he would or should adjust his beliefs downward anyway (even if he should, whether people actually do is an empirical question). This view needs to be explored but seems consistent with Golec & Tamarkin (1995): «Overconfidence might be eliminated if bettors could clearly reject the hypothesis that their subjective error variances are smaller than that of the market. Noisy condition and small samples, however, will often thwart such rejection. Hence, overconfidence is probably not obvious to many bettors.»

[25] In fact, Asubel (1991) reports that due to agents' overconfidence in liquidity engineering (Asubel does not use that term), it will be of little point for firms to compete along the penalty for overdrawn accounts dimension; a lower penalty for overdrawn accounts will only attract those few that are bad at liquidity engineering and knows it. Instead it seems that credit card companies compete along the transaction fee dimension, to such extent that the transaction fees are lower than their marginal costs!

[26] Two comments. First, it is not easy to come up with specific advise as to the degree of ex-ante overconfidence of a population it is reasonable to assume. The discussion in the previous section indicates that the degree of overconfidence should be lower for «high ability» agents than for «low ability» agents. Second, as also argued in the previous part of the paper, repeated car driving may not only change beliefs but also actual car driving ability. We abstract from these considerations here.

[27] For an application of condition (C) to a sorting context, see Hvide (1997). Notice that (C) is a strict weakening of the assumptions made in the literature hitherto (Weiss 1983, Jovanovic 1979): An agent's beliefs about his type should equal the average productivity of the socioeconomic group he belongs to. In a setting where there is only one socioeconomic group, like here, all agents should have exactly the same beliefs about themselves, and furthermore these beliefs should be identical to the population average.

The interpretation is that for any belief *b*, the fraction of high ability agents among those with belief *b* is equal to exactly *b*. Notice that (C) implies that beliefs are correct on average.

Under (C), even if they constructed a mechanism where agents revealed their true belief about themselves, firms could not make profit on those that had incorrect beliefs about themselves; a person with belief *b* would be of high ability with probability exactly equal to *b*.

We have implicitly considered two different equilibrium conditions for the credit card market. First, that firms entered to make profit opportunities disappear, and second that - from market experience - beliefs tend to converge to condition (C), in which case no firms could make a profit on overconfidence.[28]

An intuition - closely related to the intuition behind (C) - is that a situation with overconfidence and risk-neutrality would not be stable in an evolutionary sense. Genes that carry systematic information processing errors will be wiped out in the long run simply because «bad» information processing will be reflected in «bad» action choices.[29] Whether a situation where genes are overconfident can be evolutionary stable was asked in a thought-provoking article by Mike Waldman (1994). One of his points is that a gene that produces overconfident assessments may be evolutionary stable if the gene also carries a predisposition for having a utility function that eliminates the cognitive bias. If a gene is too overoptimistic in its assessments of its own judgments, then it may still survive if it has a sufficiently «cautious» utility function.[30]

An implication of this argument is that an overconfident economy may in fact be in equilibrium (i.e., no profit opportunities) if risk aversion exactly offsets the effect of

[28] We are uncertain of which equilibrium condition is the most plausible one: Perhaps the economy first reaches an equilibrium where firms compete away the profits made from overconfidence and then gradually converges to condition (C). This question needs careful modeling.

[29] A less dramatic interpretation is that agents could be more fit by adjusting their beliefs about themselves downwards. Compare to the discussion above.

[30] Waldman's main point is that in a world with sexual inheritance (at least two parents) these genes may survive even if they do not have «evolutionary optimal» behavior. Waldman's argument is arguably a very abstract one; it seems that his model just as well applies to explain why a population of genes that are underconfident is evolutionary stable. Therefore, some auxiliary assumption is necessary to make an overconfident population a plausible outcome of an evolutionary process.

overconfidence. Whether such an offset is likely or not seems to be a very interesting problem for experimental work.[31]

Notice that the argument also points to a fundamental flaw in the psychological analysis of overconfidence. Even if the motivation of this literature seems to be whether real life actions can be expected to be overoptimistic,[32] overconfidence alone does not imply anything on the quality of actions (compared to some objective standard); what is interesting is how overconfidence and risk preferences interact in determining behavior.

5. Conclusion

The interest of the present paper mainly lies in generating new questions. Consider three possible research projects.

The first possible project is to construct a model of bounded rationality where the circularity aspect is treated. The argument went as follows: First we established that imperfect awareness seems plausible both for philosophical and for empirical reasons. Moreover, imperfect awareness seemed to make a difference for bounded rationality models; for example the implications for behavior of absent-mindedness seem widely different depending on whether we assume perfect or imperfect awareness. The question then becomes how to model imperfect awareness in a decision theoretic setting. We argued that it seemed dubious to model imperfect awareness of boundedly rational agents as Bayesian uncertainty because of a certain circularity aspect: We expect agents with high ability level for a certain task, say theorem proving, to be better at assessing their theorem proving ability than agents that were weak theorem provers. One interesting research question seems to be what theoretical results

[31] There is also an interesting theoretical problem here; what are the conditions for such an offset to be the case? Without going into detail, it seems that in a partial model like here (only one market) there do exist utility functions with the property that they offset practically all degree of overconfidence. If there are several markets, however, such an utility function would - in a Savage setting - have to offset the agent's beliefs in several markets (his ability at car driving, cookery, mathematics, poker play, and so forth). What restrictions that has to be put on beliefs in other markets for there to exist such an overall offsetting utility function seems to be an open question.

[32] This motivation is quite obvious, and is clearly spelled out in Plous (1993).

can be obtained on the degree of self-awareness that can be obtained by a boundedly rational agent.

The second project is to undertake experiments where agents are faced with problems like in the theorem proving example. Even if psychologists have found that overconfidence is prominent among humans, overconfidence alone gives no criterion to judge whether a set of actions were «bad» or over-ambitious in some objective sense; one has to take risk preferences into consideration as well. Surprisingly, there seems to have been constructed few experiments similar to example 1, where beliefs about oneself are linked to actions in a setting with «proper» incentives.[33]

The third project, as discussed at some length in the previous section, is a variety of questions concerned with markets where agents are overconfident. For example, can a market with overconfident beliefs be in equilibrium? If the market is not in equilibrium (there are profits to be made by firms), then what force is stronger, the entry of profit-making firms or adjustment of beliefs to a situation with less overconfidence?

6. References

Armstrong, D. A. (1968). *A Materialist Theory of the Mind.* Routledge International Library of Philosophy.

Asubel, L. M. (1991). The Failure of Competition in the Credit Card Market. *American Economic Review*, **81**, 50-81.

Binmore, K. (1987a). Modeling rational players: Part I. *Ec. & Phil.*, **3**, 179-214.

Binmore, K. (1987b). Modeling rational players: Part II. *Economics and Philosophy*, **4**, 9-55.

Churchland, Pat. & Sejnowski, T. J. (1989). Neural Representation and Neural Computation. In : Nadel, L. et al. *Neural Connections, Mental Computations*, MIT Press.

Conlisk, J. (1996). Why Bounded Rationality? *Journal of Economic Literature*, **34**, 669-700.

De Bondt, W.F.M., Thaler, R. H. (1994). Financial Decision Making in Markets and Firms: A Behavioral Perspective. *NBER Working Paper Series 4777.*

[33] At the completion of this paper, I learned that Lovallo & Camerer (1996) have taken the first steps in checking the implications of overconfidence for decision-making.

Dennett, D. C. (1981). *Brainstorms*. The MIT Press.

Dow, J. (1991). Search Decisions with Limited Memory. *RES*, **58**, 15-41.

Elster, J, (ed.). (1986). *Multiple Selves*. Cambridge University Press.

Fershtman, C. & Kalai, E. (1993). Complexity Considerations and Market Behavior. *Rand Journal of Economics*, **24**, 224-235.

Flåm, S. D. & Risa, A. E. (1996). Search and Self-Confidence. *Working paper 10/96*, Department of Economics, University of Bergen, Norway.

Geanakoplos, J. (1989). Game Theory without Partitions. *Cowles Foundation Working Paper.*

Golec, J. & Tamarkin, M. (1995). Do Bettors Prefer Long Shots Because They are Risk-Lovers or are They just Overconfident? *Journal of Risk and Uncertainty*, 11, 51-64.

Heath, C. & Tversky, A. (1991). Preference and Belief: Ambiguity and Competence in Choice under Uncertainty. *Journal of Risk and Uncertainty*, **4**, 5-28.

Hogarth, et al. (1992). *Rational Choice: the Distinction between Economics and Psychology.* The University of Chicago Press.

Hvide, H. K. (1997). Self-Awareness, Spencian Education and Performance Wages. *NHH Discussion paper 10/97.*

Hvide, H. K. (1998). An As-If Interpretation of Some Recent Models of Bounded Rationality. *Draft.*

Kahneman, D. & Slovic, P, & Tversky, A. (1982). *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press.

Lipman, B. L. (1991). How to Decide How to Decide How to ...: Modeling Limited Rationality. *Econometrica*, **50**, 1105-25.

Lipman, B. L. (1995). Information Processing and Bounded Rationality: a Survey. *Canadian Journal of Economics XXVIII*, **1**, 42-67.

Lovallo, D. & Camerer, C. (1996). Overconfidence and Excess Entry: An Experimental Approach. *CalTech Social Science Working Paper 975.*

March, J. G. & Shapira, Z. (1987). Managerial Perspectives on Risk Taking. *Managment Science*, **33**, 1404-17.

Orphinades, A. & Zervos, D. (1995). Rational Addiction with Learning and Regret. *Journal of Political Economy*, **103**, 740-58.

Piccione, M. & Rubinstein, A. (1997). On the Interpretation of Decision Problems with Imperfect Recall. Forthcoming in: *Games and Economic Behaviour.*

Plous, S. (1993). *The Psychology of Judgment and Decision Making.* McGraw-Hill Inc.

Rubinstein, A. (1993). On Price Recognition and Computational Complexity in a Monopolistic Model. *Journal of Political Economy*, **101**, 22-38.

Rubinstein, A. (1997). *Modeling Bounded Rationality.* MIT Press.

Selten, R. (1991). Bounded Rationality. *Journal of Institutional and Theoretical Economics*,

Vallone, R. P. et al. (1990). Overconfident Prediction of Future Actions and Outcomes by Self and Others. *Journal of Personality and Social Psychology* **58**, 582-92.

Waldman, M. (1994). Systematic Errors and the Theory of Natural Selection. *American Economic Review*, **84**(3), 482-497.

Chapter 4

Self-Knowledge, Spencian Education and Performance Wages[1]

Abstract

If workers are uncertain which sector of the economy they are best fitted for, and education makes them learn about their abilities, then agents may choose education even if it does not increase their productivity. That simple argument, which we label the «Self-knowledge hypothesis», suggests how «Spencian», unproductive education can survive even if firms easily can observe and contract upon future worker performance. It follows from the Self-knowledge hypothesis that education may improve the allocation of workers and thus be of social gain. Our main result is stronger; under a certain condition (C) on agents' prior beliefs about themselves, social surplus in an institutional setting with both education and performance wages is larger than the social surplus in an institutional setting with performance wages alone.

Keywords: Education, Job Matching, Overconfidence, Performance Wages, Self-Knowledge, Sorting.

[1] Thanks to Ken Binmore, Chaim Fershtman, Kåre P. Hagen, Eirik G. Kristiansen, Terje Lensberg, Per E. Manne, participants at seminars in Bergen, and participants at the Topics in Game Theory and Microeconomics conference in Copenhagen in June 1997 for valuable comments.

1. Introduction

A central question in the economics of education is whether the prime function of education is to increase the productivity of workers or whether it is just sorting. While human capital theory (Becker 1964, Mincer 1974) focuses on education as a productivity augmenting investment, the screening hypothesis (Arrow 1973, Spence 1973) views education chiefly as a means for able individuals to sort themselves from less able individuals. A large amount of empirical work aside,[2] the following argument (e.g., Blaug, 1992, p. xiii, and Weiss 1995, p. 145) is perhaps the strongest objection to the screening hypothesis.[3] Say that we are in a «Spencian» world, where education does not enhance worker productivity. Moreover assume that firms easily can observe and contract upon future worker performance. *Then why is there any need for education? Why do not firms replace education by performance wages?*

The main contribution of the present paper is to give a simple and intuitive resolution to the «Education puzzle»; how unproductive education can survive even if other screening mechanisms are cheaper. Say that the economy consists of several sectors, where each sector requires a certain type of ability from their workers to function effectively. In such an economy, individuals obtain the highest wage rate in the sector they are best fit for. But individuals, it is assumed, are uncertain about their ability type. Workers' problem then becomes to choose an optimal sector to work in given their beliefs about their future performance. With plausible restrictions on beliefs, some agents will choose a «wrong» sector to work in. Now, our central assumption is that education makes agents learn about their abilities. In that case, some agents may choose education to learn about what their optimal sector is. That argument, which we label the «Self-knowledge hypothesis», suggests a resolution to the Education puzzle.

[2] The main difficulty of the empirically minded literature seems to be that the main testable implication of both the human capital theory and the screening hypothesis is that wage rates increase with education level. For an early contribution to the empirical literature see Riley (1979), and for a recent one, see Altonji (1995).

[3] Particularly Weiss (1995) is clear on this point; «The most strongly voiced objection to the sorting approach [to education] is: 'There must be cheaper ways to learn about workers!'. The implicit complaint is that if unobserved differences were important, firms would test for them directly, or workers would test themselves.»

A natural question is what the welfare implications of the Self-knowledge hypothesis are. On one hand, education will cause fewer agents to self-select to the «wrong» sector of the economy. On the other hand, the costs of education might exceed the benefit of improved sorting. What can we say about the net effect? As we demonstrate, the net effect depends on the prior beliefs of the population. For example, if the population is wildly overconfident ex-ante about their abilities - which some experimental evidence seem to suggest,[4] then the private incentives for education do not coincide with the social incentives. Overconfident agents expect the additional self-knowledge obtained from education to be of little value, in spite of additional knowledge having high value from society's viewpoint (by leading to an optimal place to work in).[5] When analyzing the welfare implications of the Self-knowledge hypothesis, we are thus naturally led to ask what are proper restrictions on the prior beliefs of the population. Up to now, papers with agents that lack full self-knowledge have used the following restriction on beliefs (e.g., Jovanovic 1979, Weiss 1983): Agents' beliefs about their own ability type should equal the average ability of the socioeconomic group they belong to.[6] With only one socioeconomic group, as in our setup, that maxim implies that all agents have exactly the same beliefs about themselves, and moreover that those beliefs are identical to the average ability of the socioeconomic group. We strongly feel that this common prior type of restriction on beliefs is too strong to be plausible,[7] and therefore propose a novel «consistency» condition on beliefs, denoted condition (C). Roughly speaking, condition (C) allows overconfident agents in the population, but overconfident agents should be counterbalanced by underconfident agents, and vice versa. Without going into technical details, condition (C) has the desirable property of being a weakening of the assumption on prior beliefs made in Jovanovic (1979) and Weiss (1983).

[4] See e.g., Plous (1993) for an overview of the psychological literature on overconfidence.

[5] As we shall see, the converse - that an overconfident agent overestimates the value of education - may also be the case.

[6] A socioeconomic group is for our purposes just a group of people whose members firms do not see any point in discriminating between. Of course, in a more complex setup, such discrimination could also be endogenous.

[7] Not only casual empiricism but also psychologists work on self-beliefs (see footnote 4) suggest a diversity of beliefs within socioeconomic groups.

Let us line out the paper. In the first part, we present the basics of our setup; there are two types of workers and two sectors, with several firms in each sector. In the second part, we make a formal restatement of the Education puzzle. To be specific, we first construct a pure education model, and solve for a separating equilibrium in education choice. The role of education here is just signaling. We then construct a pure performance wage model, and solve for a separating equilibrium in sector choice. To focus on the critical case where alternative sorting mechanisms to education are cheap, we set firms' cost of monitoring workers perfectly to zero. In Proposition 4 we show that the social surplus in an institutional setting with education alone is smaller than the social surplus in an institutional setting with performance wages alone.

In the third part of the paper we put together the two models from the second part by constructing a model where both performance wages and education are present. We assume, as before, that education is unproductive in the sense that it does not change an agent's productivity. However, we do assume that education is productive in the sense of making agents learn about their own ability type.[8] We then ask what the welfare properties of this institutional setting are. In Theorem 1 we prove our main result: An institutional setting that has both performance wages and education yields larger social surplus than an institutional setting with performance wages alone. Theorem 1 is quite remarkable; not only does the combination of education and performance wages make sense in an equilibrium argument, it is efficient! A natural question is whether there are weaker assumptions on beliefs than condition (C) that also gives the efficiency result of Theorem 1. We answer that question in Theorem 2.

In the last part of the paper we consider the effects of a minimum wage on the allocation of workers. Without a minimum wage, those with intermediate beliefs in themselves choose to educate, while those with high beliefs in themselves choose performance wages. The reason for that is simply that the value of information is larger for the agents

[8] Thus although consistent with the assumptions underlying the screening hypothesis, the Self-knowledge hypothesis strictly speaking points to a different rationale for education than both human capital theory and the screening hypothesis. It keeps the intrinsic non-productivity of education of the screening hypothesis, but has the flavor of human capital theory in that education increases a worker's expected productivity, since education improves his occupational choice.

with intermediate beliefs in themselves than for those with high beliefs in themselves. In Theorem 3 we prove that, surprisingly, a minimum wage may lead to a reversal of that separation; a minimum wage may lead those with intermediate beliefs to choose performance wages, while those with high beliefs choose education. The intuition is that the workers with the highest beliefs in themselves choose education to avoid paying «insurance fee» (the difference between minimum wage and productivity) for workers that choose performance wages but turn out to be a low-productivity type.

Notice that the Self-knowledge hypothesis is a formal treatment of the old argument of educators that education may improve students' knowledge about themselves. Thus in an obvious sense the Self-knowledge hypothesis is not a novelty of the present paper. But to our knowledge the only work in economics of education literature where the Self-knowledge hypothesis has been discussed at some length (not under that name) is in Stiglitz (1975).[9] The contrast to our work is that Stiglitz does not support the Self-knowledge hypothesis with a formal argument, nor does he discuss the role of prior beliefs, which we show is essential to understand the welfare implications of the hypothesis.[10] For recent overviews of the economics of education literature see Blaug (1992) and Weiss (1995), and for a collection of prevalent economics of education papers, see Blaug (ed.) (1992).

[9] Stiglitz (1975) states among other (page 292): «Part of the social marginal product of educational institutions is finding each individual's comparative advantage (as educators are wont to say, 'helping the individual find out about himself')».

[10] There are papers in the job matching literature that contain a similar intuition to ours. For example, in Jovanovic (1979) an agent may accept a low paid job for some time if it makes him learn about his productivity in other jobs. On a more technical level, Weiss (1983) proves existence of a separating equilibrium in a model with a continuum of types and a continuum of education levels. A feature of his model is that agents are uncertain of their productivity. The contrasts to our paper are several, for example Weiss offers no comparison of institutional arrangements or clues to how lack of full self-knowledge may resolve the Education puzzle.

2. Preliminaries

There is a continuum of risk-neutral workers on the unit interval. The workers are of two types, Low and High, with population shares θ_L and θ_H, respectively. The following table gives the marginal productivities of the two types in the two sectors of the economy, N («unskilled») and S («skilled»).

Table 1. Marginal productivities.

	Sector N	Sector S
Low	π_o	π_L
High	π_o	π_H

Where by assumption, $\pi_L < \pi_0 < \pi_H$. In words, the two types have equal marginal products in the low sector, while the high type has higher marginal product than the low type in the skilled sector. Moreover, the high type has higher productivity in the skilled sector than in the unskilled sector, while the low type has lower marginal product in the skilled sector than in the unskilled sector.

There are two or more risk-neutral firms engaged in Bertrand competition over workers. Each worker has a belief about his ability, expressed by $b \in B := [0,1]$. The interpretation of an agent having a certain belief, say 3/4, is that the agent believes that he is the high type with probability 3/4 and that he is the low type with probability 1/4. Let $f_L : B \rightarrow \Re$ denote the frequency distribution of beliefs for the low type, and let $f_H : B \rightarrow \Re$ denote the frequency distribution of beliefs for the high type. Correspondingly, denote the cumulative frequency function of beliefs for the low type $F_L(b)$, and the cumulative frequency function of the high type $F_H(b)$. Thus a share $F_i(b)$ of type *i* have belief less or equal to *b*. The functions $F_L(b)$ and $F_H(b)$ are assumed to be differentiable,[11] and moreover to be known by the firms. Now to the key assumption of the paper. To ensure consistency of beliefs at the population level we assume that,

$$\text{(C)}\ \frac{\theta_H f_H(b)}{\theta_L f_L(b)+\theta_H f_H(b)} = b, \text{ for all } b \in [0,1].$$

Condition (C) says that for a given belief *b*, the share of high ability workers having this belief is exactly *b*.[12] The justification of (C) is that firms do not have an informational advantage over workers; if a worker reveals his true belief *b* about himself to a firm, then a firm, by revising this belief using the belief distribution functions, should not have an incentive to believe anything else than *b* about the worker's type. Thus firms cannot make a profit on agents' misjudgments of themselves.[13]

Condition (C) has some important properties. Let P(*b*) denote the fraction of high ability agents among those agents with beliefs on the interval [*b*, 1]. Then,

$$\text{(1) } \mathrm{P}(b) := \frac{\theta_H[1-F_H(b)]}{\theta_H[1-F_H(b)]+\theta_L[1-F_L(b)]}$$

Lemma 1.

Condition (C) implies that,

i) P(*b*) is monotonically increasing in *b*.

ii) P(*b*) > *b*, for all $b \in [0, 1)$,

iii) Beliefs are correct on average.

[11] As can easily be checked, the results we obtain also hold for weaker requirements on the distribution functions, e.g., that they are continuous from the left.

[12] Notice that (C) implies that there is positive mass of beliefs for both the low type and the high type in all points on the interval (0, 1). Consequently, condition (C) is not in a strict sense a weakening of Jovanovic (1979) and Weiss (1983), as suggested in the introduction. To make (C) a weakening in a strict sense, we would have to admit zero mass in points or on intervals of (0, 1). To rewrite (C) to accommodate this possibility would make the paper slightly more technically demanding but would not alter its conclusions.

[13] Hvide (1997) speculates over what an economy with beliefs violating (C) could look like. An interesting task that lies outside the scope of both papers is to use tools of evolutionary theory to model the evolution of beliefs about oneself and to see whether anything like (C) is a plausible outcome of such a process.

Proof.

See Appendix A.

The intuition behind i) is the following. As the cutoff *b* increases, there is a fraction of agents that previously were above *b* that moves *b*. The share of low ability agents in this group of people is greater than the share of low ability people in the group of people that remains above *b*. Therefore the share of high ability agents in the group above *b* increases as *b* increases. While (ii) has a similar intuition, (iii) is obvious.

We define social surplus in a certain institutional setting *i*, SS_i, as the social surplus in the separating equilibrium of that institutional setting. Social surplus we define to be the sum of profits and wages subtracted the cost of education.[14]

3. A Formal Restatement of the Education puzzle

In this part we provide a formal restatement of the Education puzzle. In model I, education is the only sorting mechanism. Here the role of education is just signaling. To get separating equilibria in educational choice we assume that the Low type has higher cost of education than the High type.[15] In model II, performance wage is the only sorting mechanism. As argued for in the introduction, we assume that performance is perfectly observable and contractible without any cost. Moreover we assume that workers are employed for one period only.

[14] We make the innocuous assumption that workers have no cost of effort at work.

[15] We could, as Weiss (1983) does, question this assumption (cost of education, measured in alternative cost should, if anything, be larger for the high type than for the low type) and rather assume that even though the cost of education for the two types are equal, the probability of passing a final test is higher for the high type. Thus the cost of grade points is larger for the low cost type. We expect such a model to give qualitatively the same results as ours.

Model I: Education

Firms offer a wage contract conditional on education level, and workers choose to educate or not given the contract offer. There is only one education level. Cost of education for the High type agent is γ_H, and cost of education for the Low type agent is γ_L, where $\gamma_L > \gamma_H$.[16] We look for a separating equilibrium where only educated workers are employed in the skilled sector (those without education are offered a wage less than π_o in the skilled sector). This separating equilibrium has a *b* such that all agents with belief smaller than *b* choose to not educate, and all agents with belief at least *b* choose to educate. Denote this cutoff belief for $\hat{b}$. Let *w* denote the wage for educated in the skilled sector. Consider an agent's decision. He chooses to educate iff $w - b\gamma_H - (1-b)\gamma_L > \pi_o$, which implies that,

$$(2)\ \hat{b} = \frac{\pi_o + \gamma_L - w}{\gamma_L - \gamma_H}$$

$$(3)\ w = \pi_H P(\hat{b}) + \pi_L(1-P(\hat{b})),$$

where P(*b*) is given by equation (1). A *separating equilibrium* is a solution to (2) and (3) with $\hat{b} \in (0,1)$. We find a simple sufficient condition to ensure existence and uniqueness of separating equilibrium. Define condition (*) as,

$$(*)\ \gamma_H < \pi_H - \pi_o < \gamma_L.$$

Proposition 1. (Existence & Uniqueness).

If (*) holds then there exists a separating equilibrium. Furthermore this equilibrium is a unique separating equilibrium.

[16] We could, as Weiss (1983) does, question this assumption (cost of education, measured in alternative cost should, if anything, be larger for the high type than for the low type) and rather assume that even though the cost of education for the two types are equal, the probability of passing a final test is higher for the high type. Thus the cost of grade points is larger for the low cost type. We expect such a model to give qualitatively the same results as ours.

Proof.

See Appendix A.

Intuitively, the first inequality of (*) ensures that the expected cost of education for a high-confident agent is sufficiently low to make education profitable, and the second inequality of (*) ensures that the expected cost of education for a low-confident agent is sufficiently high to make education unprofitable.

Proposition 2 is an Akerlof (1970) type of result on the efficiency properties of the separating equilibrium of model I.

Proposition 2.

From a social point of view, too many workers educate in a separating equilibrium.

Proof.

Social efficiency implies that the social product of the marginal worker taking education equals his wage, i.e., that $\hat{b}\pi_H + (1-\hat{b})\pi_L = P(\hat{b})\pi_H + (1-P(\hat{b}))\pi_L$. But since $P(b) > b$ from lemma 1.ii), the left side is smaller than the right side. Thus from a social point of view, too many educate in a separating equilibrium.

Separation of agents in model I.

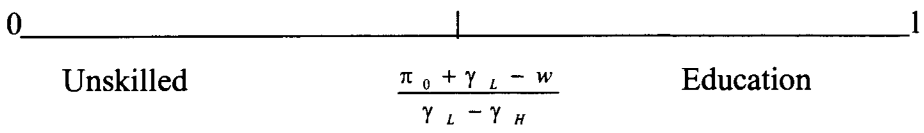


Figure 1

From a social point of view, too many workers are allocated to the S sector because wage for the educated is based on average productivity, which attracts workers with lower beliefs than socially optimal to education. Proposition 2 holds as long as there are finitely many education levels.

In contrast to in Spence's education model, education is not a waste of resources in model I since education improves the allocation of workers. Thus a pooling equilibrium is not necessarily more socially efficient than the separating equilibrium; if θ_H is large, a pooling equilibrium is more efficient than a separating equilibrium, and if θ_H is small a pooling equilibrium is less efficient. The following proposition gives a simple sufficient condition for the separating equilibrium to be more efficient than a pooling equilibrium.

<u>Proposition 3.</u>

If $\theta_H \leq \frac{\pi_0 - \theta_L \pi_L}{\pi_H}$, then the separating equilibrium yields a higher social surplus than a pooling equilibrium.

<u>Proof.</u>

See Appendix A.

Model II:. Performance Wage

We now present a pure performance wage model. First firms offer a wage contract, and then agents choose sector.

In an equilibrium, firms offer,

(4) $w(\mathrm{H}) = \pi_\mathrm{H}$

(5) $w(\mathrm{L}) = \pi_\mathrm{L}$

An agent chooses the performance contract iff $b\pi_\mathrm{H} + (1\text{-}b)\pi_\mathrm{L} \geq \pi_\mathrm{o}$, which implies that the cutoff belief is given by,

(6) $\hat{b} = \frac{\pi_0 - \pi_L}{\pi_H - \pi_L} > 0.$

Separation of agents under performance wages.

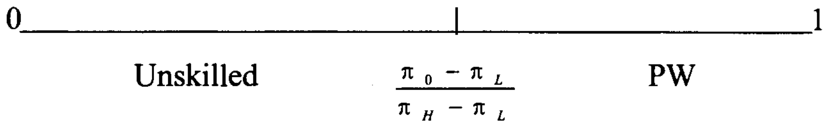


Figure 2

Those with low self-confidence choose to work in the unskilled sector, and those with high self-confidence choose to work in the skilled sector.

Denote SS_E for the social surplus in the separating equilibrium of the pure education model, and SS_P for the social surplus in the separating equilibrium of the pure performance wages model.

Proposition 4.

$SS_E < SS_P$

Proof.

Straightforward, and hence omitted.

There are two reasons why the institutional setting with education alone is less efficient than the institutional setting with performance wages alone. The first reason is the obvious one that PW does not have an intrinsic cost, while education does. The second reason, as stated in Proposition 2, is that too many educate in model I since the wage rate for the educated equals the average product of the educated. In model II, on the other hand, the allocation of workers is efficient. The reason is that from a social planner's point of view, the expected ability for an agent with belief *b* is equal to *b*, and thus a (risk-neutral) social planner agrees on an agent's valuation of the two alternatives.

If the institutional setting with PW alone outperforms an institutional setting with Spencian education alone, then why do we have Spencian education at all? A possible explanation is monitoring costs. If the cost of screening through performance monitoring is larger than the cost of screening through education, then it is not surprising if we should observe education, simply because education may be a cheaper screening institution than performance wages. For completeness, we consider a simple model with monitoring cost in Appendix B. In the next part we focus on the Self-knowledge hypothesis as a resolution to the Education puzzle.

4. A Resolution of the Education puzzle

In this part we formalize the intuition underlying the Self-knowledge hypothesis, and show how the Self-knowledge hypothesis can resolve the Education puzzle. We make two simplifying assumptions. First, that those that educate prior to working have their ability fully revealed, and second that workers can only work for one period, regardless of whether they choose to educate or not.[17]

Model III: Education + Performance Wages

We now consider a model with both education and performance wages. When speaking of «three subgroups» below, we refer to those that choose the unskilled sector directly, those that choose the skilled sector directly, and those that educate.[18]

[17] While the first assumption is innocuous, the second is not; it seems reasonable that self-knowledge may also be obtained by working in one of the sectors (apprenticeships), see e.g., Jovanovic 1979. That mechanism is only be a point in a model where agents are allowed to switch sectors after some time.

[18] We could distinguish between educated that choose the unskilled sector and educated that choose the skilled sector but that is insignificant.

Proposition 5.

i) Given that all three subgroups are present in a separating equilibrium, then those with a low belief choose the unskilled sector, those with an intermediate belief choose education, and those with a high belief choose performance wages.

ii) There exists an equilibrium where all three subgroups are present if and only if

(**) $\gamma_L < (\pi_0 - \pi_L)(\pi_H-\pi_o-\gamma_H)/(\pi_H-\pi_o)$. Furthermore this equilibrium is unique.

Proof.

See Appendix A.

Separation of agents in model III.

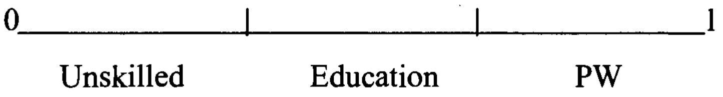


Figure 2

Those with low beliefs in themselves choose the unskilled sector, those with intermediate beliefs in themselves choose to educate to find out about their own type, while those with the highest belief in themselves skip education and take PW directly. Notice that an agent with belief equal to zero will choose the unskilled sector, and an agent with belief equal to one will choose PW. Therefore it is sufficient for all subgroups to be present that education is present. That is exactly what (**) ensures.

Denote SS_{PE} for the social surplus in the separating equilibrium.

Proposition 6.

$SS_{PE} \geq SS_P$, with strict inequality provided that education is present in equilibrium.

Proof.

See Appendix A.

Proposition 5 implied that those most uncertain about their ability choose to pay the cost to find the truth about it, while those that are pretty certain about their ability skip education. Again, from condition (C), the private and social value of information coincide, and therefore social surplus increases from model II to model III. To sum up, we have the following conclusion.

Theorem 1.

$SS_E < SS_P \leq SS_{PE}$

Proof.

The first inequality follows from Proposition 4, and the second inequality follows from Proposition 6.

Distribution of Beliefs and the No Improvement Property

We have seen that condition (C) ensures that the private and the social incentives for the various alternatives coincide (in model II and model III). Thus condition (C) is sufficient for social surplus to be larger in model III than in model II. It is of some interest to check whether we can find weaker requirements on beliefs than (C) that makes Theorem 1 hold. Say that a distribution of beliefs satisfies the *No Improvement Property* (NIP) if a social planner - independently of parameter values - cannot increase social surplus by altering agents' sorting decisions in a separating equilibrium. Then we have the following result.

Theorem 2.

A distribution of beliefs satisfies NIP if and only if it satisfies (C).

Proof.

See Appendix A.

The intuition behind Theorem 2 is simple and can be illustrated by an example: Say that the cutoff between education and performance wages in model III is 3/4; agents with belief lower and equal to 3/4 choose education, and the agents with belief higher than 3/4 choose performance wages. Furthermore assume that the share of high ability agents among those with belief 3/4 is larger than 3/4. In that case a social planner can increase social surplus by forcing the agents with beliefs in a small neighborhood of 3/4 to drop education and rather choose performance wages directly.

The Self-knowledge hypothesis may provide an explanation of other seemingly paradoxical phenomena in the labor market than the Education puzzle. As pointed out by several authors (e.g., Blaug 1992), if the screening hypothesis is true and human capital theory is false, then why do some people turn self-employed after educating? Consider a simple example. Mr. Brown has just completed a BA in Business Administration at a prominent business school, but instead of joining a firm he starts his own (producing good/ service x). Why did Mr. Brown undertake education in the first place, if he (supposedly) did not plan to enter the regular labor market? Of course, a reason may be that he wished to signal ability to potential buyers of x. However, it seems at least as plausible that Mr. Brown realized through educating that he (or x) was so good that to start working at the bottom of the ladder in some corporate firm would be an unsound gambit.

5. Allocation of the Educated

In model I we showed that in a separating equilibrium some agents educated to signal a high confidence in themselves. Firms in sector S, taking into account condition (C), took high confidence as an indicator of high ability and thus offered educated agents a high wage. Thus in the separating equilibrium all educated are employed in sector S. In model III there were both performance wages and education, and the main bulk of the educated did not necessarily go to work in sector S. The reason is that those that educate in model III are those with an intermediate belief in themselves, and the model could not tell us whether most of these agents choose the S sector after education, in other words whether most of them were High ability agents or not.

In the next section we provide an alternative hypothesis to the screening hypothesis of model I to why most educated agents go to the S sector rather than to the N sector. The starting point is the intuition that firms in sector S value self-knowledge of their employees, and therefore offer a more attractive contract to agents with education than to agents without education.

Effects on allocation of a minimum wage in sector S

For firms to value self-knowledge of its employees, it has to be costly for firms to employ agents with lack of self-knowledge (there is no such cost in model III since workers are paid their marginal product). We consider an example of such a cost: There is a minimum wage w_M in sector S, where $\pi_L < w_M < \pi_o$.[19] By imposing w_M we mean that a firm in the S sector can choose whether to employ a certain agent or not, but if it does employ him it has to pay him wage at least w_M. The main result of the section is that a minimum wage in the skilled sector can explain why a majority of the educated are allocated to the skilled sector. For simplicity we consider equilibria where all three subgroups are present. In that case there are two possible equilibrium separations.

The first possibility is that a minimum wage in sector S only changes the cutoff-line between E and PW; the ordering of subgroups along the unit line stays the same. Of course social surplus decreases compared to in model III since the cutoff-line in model III induces efficiency. The second possibility, and a more interesting one, is that a minimum wage turns around who chooses to educate and who chooses PW; those with an intermediate belief in themselves choose PW and those with high beliefs in

[19] An alternative explanation to why firms value self-knowledge could be risk considerations, but of course that requires a different model.

themselves choose E. The point with this equilibrium is that those with high beliefs in themselves educate primarily to screen off those with lower beliefs in themselves.

Let us explain why. If there is a minimum wage in the skilled sector, then there is a cost w_M - π_L for a firm to engage a Low ability worker.[20] Consequently a firm employing workers without education must cover the deficit it runs on employing unaware low ability agents by giving a wage less than π_H to unaware high ability agents. Thus the equilibrium wage for uneducated high ability agents, w_H, is less than π_H.

For the educated High ability agents it is different. Since education makes an agent learn his type, and since the minimum wage in the skilled sector is lower than the wage in the unskilled sector, all educated agents of the Low type will self-select to the N sector. Thus those educated working in the S sector will be solely of the High type, and receive wage π_H in a Bertrand equilibrium. Then, given that the cost of education for the High type is sufficiently low (less than π_H - w_H), agents with a high belief in themselves will educate and receive the high wage π_H. Thus, in contrast to in model III, *education is undertaken by agents with higher confidence than those that choose performance wages directly*. Then, by condition (C), a large share of those people that educate will turn out to be High ability agents, and thus choose sector S rather than sector N. That completes the argument. Let us present a formal statement of the finding. Let $w_M \equiv \pi_L + m$.

Theorem 3.

With a minimum wage in sector S, there are two possible equilibrium separations where all subgroups are present. For i)$m < \pi_o - \gamma_L - \pi_L$ (provided the right side is positive), the separation is in the same order as in model III. However, if ii)$m > \pi_o - \gamma_L - \pi_L$, agents with an intermediate belief choose the PW, and agents with high beliefs choose to educate.

[20] Notice that a firm would never offer more than w_M for a Low performance in a Bertrand equilibrium.

Proof.

See Appendix A.

Corollary 3.1

In case ii), the share of the educated that employed in the S sector is at least θ_H.

Proof.

Since the educated in case ii) belong to those with the highest beliefs, then by the monotonicity of P(b), the share of high ability agents among those educated will be at least θ_H. Only in the limit (when all agents choose to educate) the share of educated agents that choose the S sector will approach θ_H.

Notice that the efficiency properties of an equilibrium where the agents with high confidence educate just to screen off the less confident, resembles the efficiency properties of Spence's education model; education is practically speaking of no social gain. Agents with high beliefs in themselves (and thus high expected ability from a social planner's viewpoint) take education not to improve their sorting but to screen off the less able. Moreover, the mediocre (in expected terms) take PW because they know that if they turn out to be the low type they are to some extent «insured» against receiving a low pay. The «insurance fee» is paid by those with higher (expected) ability.

Usually the rationale for a minimum wage is not efficiency arguments, but distribution arguments. Let us consider the distribution effects of imposing a small m. A minimum wage is surely for the worse for those with high beliefs in themselves (for simplicity consider agents with belief equal to 1), since they pay the penalty of π_H - w_H. On the other hand, a minimum wage does not affect the agents with the lowest beliefs, since they choose the unskilled sector anyway. The possible positive distribution effect are for those with intermediate beliefs. With a minimum wage they are «insured» against a bad outcome and may be better off by choosing PW instead of E (as shown in the proof of theorem 3, only the payoff from choosing the PW alternative changes when introducing a minimum wage). However, it is not necessary that there is anyone that gains from the minimum wage (except for the education institution); we can imagine scenarios where a minimum wage makes PW *less* attractive for those with intermediate beliefs in themselves.

6. Conclusion

We have considered a Spencian world where education does not promote productivity, but where firms may contract wage on performance, and where monitoring is perfect and without cost. First we saw in Proposition 4 that in such a world education is inferior to performance wages as a sorting mechanism. To explain why there can be any education at all (and not only performance wages), we suggested that education, though not leading to a productivity increase, gives agents a better estimate of their own ability. Thus education makes agents better able to sort themselves to their optimal sector of the economy. We then showed that the social surplus of an institutional setting with both education and performance wages was at least as large as the social surplus with performance wages only. The main reason for that result is that beliefs satisfied condition (C), which makes the private and social incentives for education coincide. Although condition (C) may be too strong to be realistic in a strict sense, it seems extremely useful in bringing the literature on agents that lack full self-knowledge a useful step towards realism in that beliefs do not have to be identical within a socioeconomic group.

We then turned to assuming an exogenous fixed minimum wage in the skilled sector. Surprisingly that may lead to a reversal in the separation of agents: While in model III it was the mediocre (in expected terms) that educated and the best that turned «entrepreneurs», in model IV it is the opposite: The best chooses to educated while the mediocre chooses PW. The intuition for the result is that the workers with the highest beliefs in themselves choose education to avoid paying «insurance fee» for workers that choose performance wages but turn out to be a low-productivity type.

Our setup has the advantage of being simple and intuitive and at the same time yielding rich results. However the setup does not capture a number of seemingly important factors like on the job learning and cost of monitoring. A natural extension of the present work would therefore be to consider a prolonged version of model III, where workers may either learn about their abilities through education or through work experience. If a worker turns out to have lower performance than expected in a sector, he may choose to switch to another sector. In combination with such an extension would be required to weaken the assumption of full revelation of ability through education.

The empirical relevance of the Self-knowledge hypothesis is an open and seemingly important question. With which educations would we expect the forces of the self-knowledge hypothesis to be at work? In skill-oriented educations like law and engineering degrees it seems that students learn much about their abilities along few dimensions, but at the sacrifice of generality. On the other hand, in educations like BA's in Business Administration or in certain humanistic disciplines, a student encounters such a variety of different problems that he potentially can learn about his abilities not along one, but along many dimensions. That argument suggests that the main empirical relevance of the Self-knowledge hypothesis is with the latter type of educations.

7. References

Akerlof, G. (1970). The Market for 'Lemons': Quality Uncertainty and the Market Mechanism. *Quarterly Journal of Economics*, **84**, 488-500.

Antonji, J. (1995). The Effects of High School Curriculum on Education and Labor Market Outcomes, *Journal of Human Resources*.

Arrow, K. (1973). Higher Education as a Filter. *Journal of Public Economics*, **2**, 193-216.

Becker, G. (1964). *Human Capital*. New York: Colombia University Press.

Blaug, M. ed. (1992). *The Economic Value of Education: Studies in the Economics of Education*. International Library of Criticial Writings in Economics. University Press, Cambridge.

Blaug, M. (1992). Introduction in: *The Economic Value of Education: Studies in the Economics of Education.* International Library of Criticial Writings in Economics. University Press, Cambridge.

Greenberg, J. (1984). Job Matching Through the Signaling Mechanism. *CORE Discussion Paper* 8405.

Hvide, H. K. (1997). Self-Awareness, Uncertainty and Markets with Overconfidence. *NHH Discussion paper 9/97.*

Jovanovic, B. (1979). Job Matching and the Theory of Turnover. *Journal of Political Economy*, **87**, 972-990.

Miller, R. A. (1984). Job Matching and Occupational Choice. *Journal of Political Economy*, **92**, 1086-1117.

Mincer, J. (1974). Schooling, Experience and Earnings. New York: NBER & Colombia University Press.

Pissarides, C. (1990). *Equilibrium Unemployment Theory.* Oxford: Blackwell.

Riley, J. (1979). Testing the Educational Screening Hypothesis. *Journal of Political Economy*, **87**, 227-252.

Spence, A. M. (1974). *Market Signaling: Information Transfer in Hiring and Related Screening Processes.* Cambridge, Mass.: Harvard University Press.

Stigler, G. (1962). The Economics of Information, *Journal of Political Economy*, **69**.

Stiglitz, J. E. (1975). The Theory of 'Screening', Education, and the Distribution of Income. *American Economic Review*, **71**, 393-410.

Weiss, A. (1983). A Sorting-cum-Learning Model of Education. *Journal of Political Economy*, **91**, 420-442.

Weiss, A. (1995). Human Capital vs. Signalling Explanation of Wages. *Journal of Economic Perspectives*, **9**, 133-54.

8. APPENDIX A

Proof of lemma 1.

i). Rearrange (C) and insert into the definition of P(b). Then, for $z \in$ B,

$$\text{(A1) } \mathrm{P}(b) = \frac{\int_b^1 f_H(z)dz}{\int_b^1 \frac{1}{z} f_H(z)dz}.$$

Differentiate to get,

$$\text{(A2) } \mathrm{P}'(b) = \frac{-f_H(b)(\int_b^1 \frac{1}{z} f_H(z)dz) - (-\frac{f_H(b)}{b}\int_b^1 f_H(z)dz)}{(\int_b^1 \frac{1}{z} f_H(z)dz)^2} = \frac{f_H(b)\int_b^1 \frac{z-b}{bz} f_H(z)dz}{(\int_b^1 \frac{1}{z} f_H(z)dz)^2} > 0.$$

ii) Use the expression from (A1) to get,

$$\text{(A3) } \mathrm{P}(b) - b = \frac{\int_b^1 f_H(z)dz - b\int_b^1 \frac{1}{z} f_H(z)dz}{\int_b^1 \frac{1}{z} f_H(z)dz} = \frac{\int_b^1 \frac{z-b}{z} f_H(z)dz}{\int_b^1 \frac{1}{z} f_H(z)dz} > 0.$$

iii). Suppose nobody is underconfident or overconfident. Then $\int_0^1 zf_L(z)dz = 0$, and, $\int_0^1 zf_H(z)dz = 1$. Hence the correct average belief equals, $\theta_L \int_0^1 zf_L(z)dz + \theta_H \int_0^1 zf_H(z)dz = \theta_H$. By rearranging (C) and integrating we find the average belief for a distribution of beliefs that satisfy (C) as, $\theta_L \int_0^1 z f_L(z)dz + \theta_H \int_0^1 z f_H(z)dz = \theta_H \int_0^1 f_H(z)dz = \theta_H[F_H(1)- F_H(0)] = \theta_H$, as stated.

Proof of Proposition 1.

We wish to show that (*) $\gamma_H < \pi_H - \pi_o < \gamma_L$ implies that there exists a separating equilibrium, and moreover that the separating equilibrium is unique.

Define $H(\hat{b}) \equiv \alpha - \delta P(\hat{b})$, where $\alpha \equiv \frac{\pi_0 + \gamma_L - \pi_L}{\gamma_L - \gamma_H}$, $\delta \equiv \frac{\pi_H - \pi_L}{\gamma_L - \gamma_H}$.

Notice that from lemma 1 we have $H'(\hat{b}) = - \delta P'(\hat{b}) < 0$.

Insert (4) into (3) to obtain,

(A4) $\hat{b} = H(\hat{b})$.

Thus separating equilibria are fixed points to $H(\hat{b})$ on (0,1). Existence. Since $H(\hat{b})$ is downward sloping, stating existence and uniqueness is equivalent to stating i)H(1) < 1 and ii)H(0) > 0. [In words, the curve defined by the function $H(\hat{b})$ intersects the 45 degree line exactly once iff i) and ii) holds. If either i) or ii) does not hold then the curve defined by $H(\hat{b})$ does not cross the 45 degree line at all.] Since P(1) = 1, i) is equivalent to $\alpha - \delta < 1$, which is equivalent to $\gamma_H < \pi_H - \pi_0$. Thus the first inequality of (*) is equivalent to i). Now the second inequality of (*). Since $P(0) = \theta_H$, ii) is equivalent to $\theta_H < \alpha/\delta$, which by definition is equivalent to $\theta_H < \frac{\pi_0 + \gamma_L - \pi_L}{\pi_H - \pi_L}$. It is trivial to show that the second inequality of (*), $\gamma_L > \pi_H - \pi_o$, implies that the numerator is larger than the denominator, which implies that ii) is satisfied. Thus we have established that the first inequality of (*) implies i), and that the second inequality of (*) implies ii).

Proof of Proposition 2.

A pooling equilibrium is a stable situation where all agents go to the same sector (mixed strategies excluded). Denote PN the situation where all agents go to the N sector, and PS the situation where all agents go to the S sector. In PN the productivity (and wage) is simply π_o, and in PS the wage is $\theta_L\pi_L + \theta_H\pi_H$. Thus PN is an equilibrium if $\pi_o > \theta_L\pi_L + \theta_H\pi_H$, and PS is an equilibrium if the opposite inequality holds. It is simple to show that a separating equilibrium is always more efficient than a PN pooling equilibrium. It suffices to show that the average social surplus generated by those educated exceeds π_o. The social surplus of those educated is simply wage minus cost of education for that group, i.e., w - $P(\hat{b})\gamma_H$ - $[1- P(\hat{b})]\gamma_L$. But since the marginal worker in an education equilibrium equates wage to cost of education plus π_o, i.e., $w - \hat{b}\gamma_H - (1- \hat{b})\gamma_L = \pi_o$, we have, by lemma 1 ii) [which states that $P(b) > b$], that, $w - P(\hat{b})\gamma_H - [1- P(\hat{b})]\gamma_L > \pi_o$ Thus an education equilibrium generates larger social surplus than PA. We also know that PA generates larger social surplus than PB for $\theta_H \leq \frac{\pi_0 - \theta_L\pi_L}{\pi_H}$. That completes the proof; a sufficient condition for an education equilibrium to be more efficient than a pooling equilibrium is that $\theta_H \leq \frac{\pi_0 - \theta_L\pi_L}{\pi_H}$.

Notice that this condition is not necessary. Since social surplus in PB is π_H in the limit when θ_H approaches 1, while social surplus in an education equilibrium goes towards $\pi_H - \gamma_H$ when θ_H approaches 1, by continuity there exists a θ_H' such that social surplus is larger in PB than in an education equilibrium. A θ_H larger than θ_H' is both necessary and sufficient for a pooling equilibrium to be more efficient than an education equilibrium. To avoid tedious details, we skip the construction of θ_H'.

Proof of Proposition 3.

We show that if all three subgroups are present in a separating equilibrium, then they are present in the order (N, E, PW); those with low beliefs choose N (unskilled sector), those with intermediate beliefs choose E, and those with high beliefs choose PW. By the same token we derive the necessary and sufficient condition for existence (and uniqueness).

The unskilled sector yields payoff,

(A5) $U(N) = \pi_o$.

Education yields,

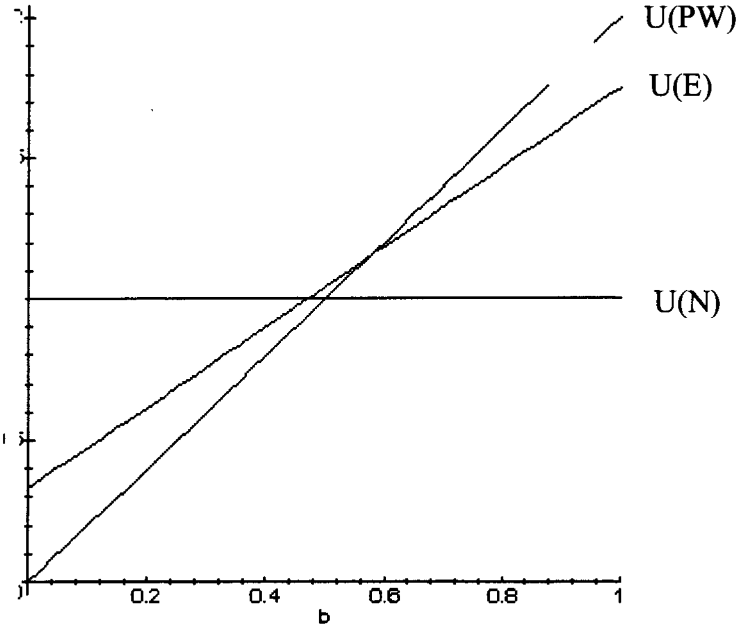
(A6) $U(E) = b(\pi_H - \gamma_H) + (1\text{-}b)(\pi_o - \gamma_L) = (\pi_0 - \gamma_L) + b(\pi_H - \pi_o + \gamma_L - \gamma_H)$

PW yields,

(A7) $U(PW) = b\pi_H + (1\text{-}b)\pi_L = \pi_L + b(\pi_H\text{-}\pi_L)$

Notice that the payoffs are linear in *b*. We get the following figure:

Separation in model III. Figure A1.



Since $U(PW|b{=}1) = \pi_H > U(E|b{=}1) = \pi_H - \gamma_H$, and since payoffs are linear in *b*, the only possible equilibrium ordering where all three subgroups are present is (N, E, PW). To ensure existence of an equilibrium where all subgroups are present it is necessary and sufficient that U(E) and U(PW) intersect above the π_o line. By straightforward calculations that can be shown to be the case if and only if,

(**) $\gamma_L < (\pi_0 - \pi_L)(\pi_H-\pi_o-\gamma_H)/(\pi_H-\pi_o)$.

By linearity, (**) also ensures that there is a unique equilibrium where all three subgroups are present.

Proof of Proposition 4.

Proposition 4 follows from the fact that by condition (C), the private and social incentives of the various alternatives coincide; a social planner has the same valuation to the different alternatives for an agent with belief *b*' as the agent with belief *b*' does himself. If there is anyone taking education in a separating equilibrium, they do so because this alternative is better for them, and by condition (C), also society is better off.

Proof of Theorem 2.

That (C) ⇒ (NIP) follows from the discussion in the text. For the reverse implication we must show that if (C) does not hold then there exist parameter values such that a social planner could increase social surplus in a separating equilibrium by altering some agents' decisions. If (C) does not hold, then there exists at least one *b* such that G(*b*) ≠ *b*, where G(*b*) ≡ $\frac{\theta_H f_H(b)}{\theta_L f_L(b)+\theta_H f_H(b)}$. Assume *b*'∈ (0,1) is such a point, and without loss of generality assume that G(*b*') < *b*'; agents with belief *b*' are on average overconfident. But, since f_i is continuous, G(*b*) is also continuous, and consequently there exists an ε > 0 such that G(*b*') < *b*' for *b* ∈ [*b*'- ε, *b*']. In words, if agents with belief *b*' are overconfident, then there exists an interval around *b*' with the property that for all *b* in this interval, agents are overconfident on average. It is trivial to show, and is hence omitted, that there exists parameter values {γ_L, γ_H, π_H, ...} such that in an equilibrium where both N and E are present (it is not necessary for PW to be present), *b*' is within ε distance to the right of $\hat{b}_1$, the cutoff between N and E. In words, suppose that in equilibrium agents with a belief in the interval $[\hat{b}_1, b']$ choose to educate but are «close» to choosing N. Then a social planner could increase social surplus by forcing the agents with beliefs in the interval $[\hat{b}_1, b']$ to choose N rather than E.

<ins>Proof of Theorem 3.</ins>

Define a type I equilibrium as one with the following separation of agents.

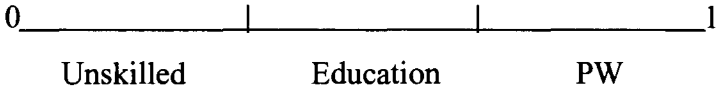


Figure A2

A type I equilibrium is an equilibrium with the same ordering of the subgroups as in model III.

Define a type II equilibrium as one with the following separation of agents.

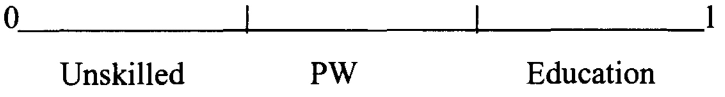


Figure A3

Notice that in a type II equilibrium, the subgroup PW and the subgroup E have changed place. Now to the expected payoffs, which again are linear in *b*. The payoff for entering the unskilled sector directly is the same as before. Also the expected payoff from taking education is unchanged. Thus,

(A8) $U(N) = \pi_o$,

and,

(A9) $U(E) = b(\pi_H - \gamma_H) + (1-b)(\pi_o - \gamma_L) = (\pi_O - \gamma_L) + b(\pi_H - \pi_o + \gamma_L - \gamma_H)$

However, PW yields,

(A10) $U(PW) = bw_H + (1\text{-}b)w_M = w_M + b(w_H - w_M)$,

where w_H is endogenously determined. Let ρ be the fraction of high ability agents among those that undertake PW. Then the profit of a firm offering contracts to those without education is:

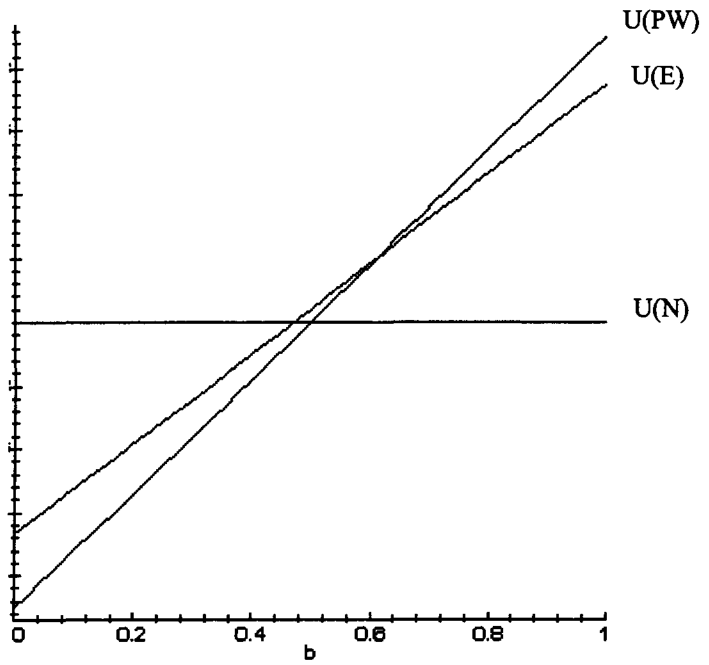
(A11) $\Pi = \rho(\pi_H - w_H) - (1\text{-}\rho)(w_M - \pi_L)$.

In an equilibrium firms obtain zero profit, so (recall that $m \equiv w_M - \pi_L$),

(A12) $w_H = \pi_H - \pi_L + w_M - (w_M - \pi_L)/\rho = \pi_H - m(1 - \rho)/\rho$

Consider first the case when the intercept of the U(E) function is larger than the intercept of the U(PW) function, i.e., $\pi_o - \gamma_L > w_M$, which is equivalent to i)$m < \pi_o - \gamma_L - \pi_L$. In that case the only possible equilibrium where all subgroups are present is a type I equilibrium.

Separation when $m < \pi_o - \gamma_L - \pi_L$. Figure A4.

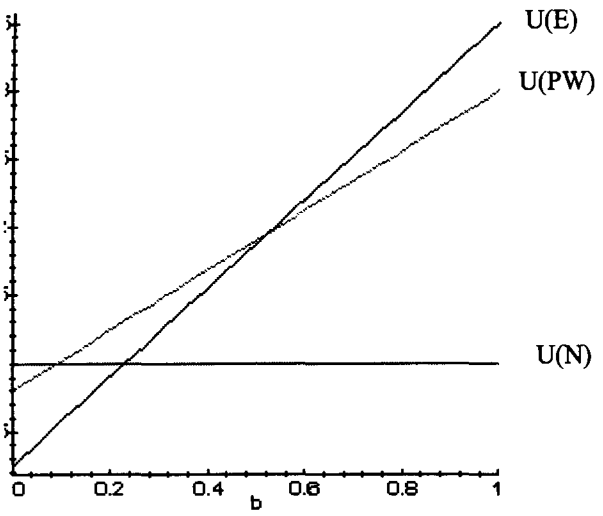


A low m ensures that the intercept of the U(E) lies above the intercept of the U(PW) line, and, in that case, a type I equilibrium is the only possible equilibrium where all subgroups are present.

When increasing m, then w_M increases (by definition), while w_H decreases. The latter needs a small argument. Suppose that w_H had increased as well. In that case, PW would have become relatively better for all agents, and there would be some agents that before chose E that now choose PW. But, since we are in a type I equilibrium, that causes ρ (the share of high ability agents that choose PW) to decrease. Thus, by (A12), w_H decreases, which is a contradiction. Thus in a type II equilibrium, $\frac{\partial w_H}{\partial m} > 0$. From this argument it follows that an increase in m causes the U(PW) line to rotate clockwise. Recall that a change in m does not change the position of the U(N) line and the U(E) line.

There are two quite apparent necessary conditions for a type II equilibrium: a)the intercept of the U(PW) line exceeds the intercept of the U(E) line [i.e., that m exceeds $\pi_o - \gamma_L - \pi_L$], and b)the U(E) line and the U(PW) line intersect above π_o. If both a) and b) hold, we get the following figure.

Separation when $m > \pi_o - \gamma_L - \pi_L$. Figure A5.



The figure illustrates the underlying payoffs in a type II equilibrium, i.e., an equilibrium where the high belief agents undertake education and the intermediate belief agents undertake PW, in contrast to in model II. For a type II equilibrium to exist, the intercept of the education line must be below the intercept of the PW line, and moreover the two lines must intersect above π_o.

To see that a type II equilibrium is indeed possible, i.e., that there exists combinations of *m* and w_H that gives the separation in the figure above, consider the following example. We start out with $m = 0$ [which gives a type I equilibrium], and then construct a type II equilibrium by increasing *m*.

Let $\theta_L = \theta_H = 1/2$, and let the distribution of beliefs be $F_H = b^2$, and $F_L = b(2-b)$. It is simple to check that these distributions satisfy (C). By straightforward calculations we get $P(b) = (1 + b)/2$, and the share of high ability agents on an interval (b_i, b_j) becomes $(b_i + b_j)/2$. Let productivities be given by $\pi_H = 2$, $\pi_o = 1$, and $\pi_L = 0$, and costs by $\gamma_L = 2/3$, and $\gamma_H = 1/4$. Notice that (*) is satisfied, i.e., $3/4 < (1-0)(2-1-0)/(2-1) = 1$, so there exists a type I equilibrium for $m = 0$. Let us solve for that equilibrium.

Let b_1 be the belief that gives intersection between U(N) and U(E), and let b_2 be the cutoff between U(E) and U(PW), i.e.,

$b_1 := \{b\text{:U(E)} = \pi_o\}$, and $b_2 := \{b\text{:U(E)} = \text{U(PW)}\}$.

First find b_1, i.e., b that solves $(1 - 2/3) + b(5/3 - 1/4) = 1$, which implies that $b_1 = 8/17 \approx .47$. Then find b_2, i.e., b that solves $(1\text{-}2/3) + b(5/3 - 1/4) = 2b$, to get $b_2 = 4/7 \approx .57$. Thus we have the following type I equilibrium for $m = 0$.

0_____________|______________|______________1 Figure A6

Unskilled .47 Education .57 PW

We now construct a type II equilibrium. Suppose $m = 1/2$. Let $b_3 := \{b\text{:U(PW)} = \pi_o)$. In that case, we get the following type II equilibrium:

$b_3 = 32/(67 + \sqrt{137}) \approx .41.$

$b_2 = 16/(3\sqrt{137} - 71) \approx .89,$

$w_H = (99 + \sqrt{137})/64 \approx 1.73$

Check that $\rho = (b_2 + b_3)/2 = [16/(3\sqrt{137} - 71 + 32/(67 + \sqrt{137}] \approx .65$, which inserted into (18) yields, $w_H = 5/2 - (2 \text{ x } 0.65)^{-1} \approx 1.73$.

0_____________|______________|______________1 Figure A7

Unskilled .41 PW .89 Education

With the parameter values given in this example, it can easily be checked that there exists a type II equilibrium for $m \in (1/3, 5/8)$.

9. APPENDIX B: COST OF MONITORING

For illustration, we let firms have two alternatives: Either to skip monitoring completely, or to monitor perfectly but at an exogenous cost *r* per worker. Notice that there is no point in monitoring the educated on efficient performance wages, since only the high type accepts such contracts in equilibrium. Therefore, only the uneducated are monitored if any. We assume that monitoring of the uneducated is an equilibrium choice by the firms and investigate the properties of such an equilibrium. In a monitoring equilibrium, the firms propose the following wage schedule to the uneducated:

(B1) $w(\text{H}|\text{uneducated}) = \pi_H - r$

(B2) $w(\text{L}|\text{uneducated}) = \pi_L - r$,

and the following to the educated,

(B3) $w(\text{H}|\text{educated}) = \pi_H$

(B4) $w(\text{L}|\text{educated}) = \pi_L$

In that case, education is better than PW if,

$b(\pi_H - \gamma_H) + (1 - b)(\pi_0 - \gamma_L) \geq b(\pi_H - r)+(1 - b)(\pi_L - r)$, which implies,

(B5) $b(\pi_L - \pi_0 + \gamma_L - \gamma_H) \geq \pi_L - \pi_0 + \gamma_L - r$.

Case (i). Assume γ_L is «low», i.e., $\gamma_L < \pi_0 - \pi_L + \gamma_H$. Then $b \leq \frac{\pi_L - \pi_0 + \gamma_L - r}{\pi_L - \pi_0 + \gamma_L - \gamma_H}$ implies that education is better than PW. If $m > \gamma_H$, then education is better than PW for all agents, and if $r < \gamma_H$ then PW is preferred by the agents with the highest beliefs, and education is preferred by the agents with intermediate beliefs. The cutoff point is of course $\frac{\gamma_L + w_M - \pi_0}{\pi_H - w_H + \gamma_L - \gamma_H + w_M - \pi_0}$.

Case (ii). Assume $\gamma_L > \pi_0 - \pi_L + \gamma_H$. Then $b > \frac{\pi_L - \pi_0 + \gamma_L - r}{\pi_L - \pi_0 + \lambda_L - \gamma_H}$ implies that education is preferred to PW. If $r > \gamma_H$ then only the agents with the highest beliefs prefer education to PW, and if $r < \gamma_H$ then all agents prefer PW to education.

To sum up, if $r > \gamma_H$ then the agents with the highest beliefs take education, simply because net wage is higher. Whether agents with intermediate beliefs take education or not depends on γ_L. If γ_L is «low» then they do take education, and if γ_L is «high» they prefer PW to education. Of course if γ_L is very high, there will only be a very small share of agents that choose to educate. If, on the other hand, $r < \gamma_H$ then the agents with the highest beliefs take PW. If cost of education for low type agents is low, then agents with intermediate beliefs choose education rather than PW, and if cost of education for low type is high, then agents with intermediate beliefs choose PW rather than education.

Chapter 5

Complementary Teams, Linear Sharing Rules and Uncertainty: a Note[1]

Abstract

Two recent articles, Legros & Matthews (1993) and Vislie (1994), show that in a non-cooperative production game with strictly complementary (non-observable) inputs, interpreted as effort levels, there exists a linear budget-balancing sharing rule that implements the efficient effort vector in Nash strategies. This is an important insight because it shows that theorem 1 in Holmström's "Moral Hazard in Teams" (1982) does not generalize to the case when the inputs are strict complements. In this note we test the linear implementability result for robustness. First we note that under certainty the implementability result can be generalized to decreasing returns to scale Leontief technologies. Second, and more importantly, we show that with uncertainty in the individual outputs, linear implementation breaks down. Intuitively, the reason is that uncertainty smoothens the kink in the Leontief production function, making the social efficient effort choice inconsistent with Nash equilibrium and budget-balance

Keywords: Budget-balance, Leontief technology, Linear Implementation, Teams, Uncertainty.

[1] Thanks to Geir Asheim and Terje Lensberg for useful comments.

1. Introduction

Two recent articles, Legros & Matthews (1993) and Vislie (1994), show that in a non-cooperative production game with strictly complementary (non-observable) inputs, interpreted as effort levels, there exists a linear budget-balancing sharing rule that implements the efficient effort-vector in Nash strategies.

In this note we test the linear implementability result for robustness. First we observe that under certainty the implementability result can be generalized to hold for all non-increasing returns to scale Leontief technologies, not only constant return to scale technologies as in Legros & Matthews (1993), and Vislie (1994). Second, and more importantly, adding uncertainty to the model (with a multiplicative random term to individual output) makes linear implementation break down. Intuitively, the reason is that uncertainty smoothens the kink in the Leontief production function, and, analogously to in Holmström (1982), the social efficient effort vector cannot be a Nash equilibrium unless we relax budget-balance. In a Nash equilibrium the partners do not take into the account the positive externalities when increasing their effort, and therefore the efficient effort vector cannot be a Nash equilibrium.

2. The Model under Certainty

Output x is determined by a Leontief technology, where $x = f(\min [b_1e_1,b_2e_2])$; e_i is agent i's choice of effort, where $e_i \in [0,E_i]$, and E_i is finite. For simplicity we let $i \in \{1, 2\}$. $f(..)$ is a differentiable, increasing and concave function with $f(0) = 0$. Cost of effort is given by $v_i(e_i)$, where $v_i(e_i)$ is increasing and convex. The utility of agent i is $\beta_i x - v_i(e_i)$, where $\beta_i > 0$ and $\Sigma_i\beta_i = 1$.

Define the efficient effort-vector, e^*, as the vector maximizing social surplus:

$$e^* \equiv \arg\max_{e} [f(\min [b_1e_1,b_2e_2]) - \Sigma_i v_i(e_i)], \qquad (1)$$

Now we restate the result from Legros & Matthews (1993) and Vislie (1994).[2]

Observation 1.

There exists a linear sharing rule β_i* that implements $e*$ in Nash strategies.

Proof.

We prove by construction. Pareto optimality implies that $e*$ is symmetric, i.e., that $e_i* = e_j*b_j/b_i$. For all symmetric e, we can write social surplus as,

$$f(b_1e_1) - v_1(e_1) - v_2(b_1e_1/b_2) \qquad (2)$$

Since (2) is differentiable, e_1* is the unique solution to,

$$f'(b_1e_1)b_1 - v_1'(e_1) - v_2'(b_1e_1/b_2)b_1/b_2 = 0, \qquad (3)$$

which implies that,

$$\frac{1}{f'(b_ie_i*)}\Sigma_i\frac{v_i'(e_i*)}{b_i} = 1. \qquad (4)$$

Now define $\beta_i* \equiv \frac{v_i'(e_i*)}{f'(b_ie_i*)b_i}$. Given that agent j chooses e_j*, agent i maximizes,

$$\frac{v_i'(e_i*)}{f'(e_i*)b_i}f[\min(b_ie_i, b_je_j*)] - v_i(e_i) \qquad (5)$$

It follows that agent i maximizes his payoff by choosing effort level e_i*, and thus $e*$ is a Nash equilibrium under $\beta*$.

[2] Observation 1 is a slight generalization of the result from Legros & Matthews (1993) and Vislie (1994), since we allow decreasing returns to scale in the team output function.

3. The Model under Uncertainty

Let output be given by $f(A)$, where $A = \min[A_1, A_2]$, and $A_i = b_i e_i \varepsilon_i$; ε_i is a stochastic term assumed to be distributed independently of ε_j. ε_i has full (non-negative) support. Let $G_i(z)$ be ε_i's distribution function, assumed to be twice differentiable. Denote ε_i's density function by $g_i(z)$.

Assuming that the team is risk-neutral, the efficient effort vector maximizes expected surplus. Realized surplus, $H(e,\varepsilon)$, equals,

$$H(e,\varepsilon) \equiv f(A) - \Sigma_i v_i(e_i), \tag{6}$$

and the ex-ante efficient vector, e^*, equals,

$$e^* \equiv \underset{e \in E}{\arg\max}\, \mathrm{E}[H(e,\varepsilon)] \tag{7}$$

Proposition 1.

Given the above specification of the model, there does not exist a linear sharing rule that implements the efficient effort vector.

Proof.

We show that $E[f(A)]$ is differentiable with respect to e_i. It is then straightforward to show that e^* is not implementable with a linear sharing rule. Let $M_i(a)$ be A_i's distribution function. Thus,

$$M_i(a) \equiv \mathrm{Prob}(A_i \le a) = \mathrm{Prob}(b_i e_i \varepsilon_i \le a) = \mathrm{Prob}(\varepsilon_i \le a/b_i e_i) = \int_0^{a/b_i e_i} g_i(z)dz = G_i(a/b_i e_i) \tag{8}$$

Thus for $e_i > 0$, A_i's density function, $m_i(a)$, equals,

$$m_i(a) = \frac{g_i(a / b_i e_i)}{b_i e_i} \tag{9}$$

Let M(a) be the distribution function of A, i.e., the distribution function of min(A_1, A_2). By independence, M(a) equals,

$$\mathrm{M}(a) \equiv \mathrm{Prob}(\mathrm{A}_1 \leq a \text{ or } \mathrm{A}_2 \leq a) = 1 - \mathrm{Prob}(\mathrm{A}_1 > a \text{ and } \mathrm{A}_2 > a) =$$
$$1 - [1\text{-}\mathrm{M}_1(a)][1\text{-}\mathrm{M}_2(a)] = \mathrm{M}_1(a) + \mathrm{M}_2(a) - \mathrm{M}_1(a)\mathrm{M}_2(a), \tag{10}$$

and A's density function consequently becomes,

$$m(a) = m_1(a)[1 - \mathrm{M}_2(a)] + m_2(a)[1 - \mathrm{M}_1(a)] =$$

$$\frac{g_1(a / b_1 e_1)}{b_1 e_1}[1 - \mathrm{G}_2(a/b_2e_2)] + \frac{g_2(a / b_2 e_2)}{b_2 e_2}[1 - \mathrm{G}_1(a/b_1e_1)] \tag{11}$$

Observe that $m(a)$ is differentiable with respect to a. We then have that,

$$E[f(A)] = \int_0^\infty f(a)m(a)da, \tag{12}$$

which is differentiable with respect to e_i since $m(..)$ is differentiable with respect to e_i. The rest of the proof is straightforward. From (6) and budget-balance, efficiency implies:

$$\frac{\partial E[H(e^*,\varepsilon)]}{\partial e_i} = \beta_i \frac{\partial E[f(A)]}{\partial e_i} + \beta_j \frac{\partial E[f(A)]}{\partial e_i} - v_i'(e_i) = 0. \quad \forall i,\ i{\neq}j. \tag{13}$$

However, in a Nash equilibrium,

$$\beta_i \frac{\partial E[f(A)]}{\partial e_i} - v_i'(e_i) = 0. \qquad \forall i \tag{14}$$

Consistency of (8) and (9) requires,

$$\beta_j \frac{\partial E[f(A)]}{\partial e_i} = 0. \qquad i \neq j \tag{15}$$

But since $\beta_i \frac{\partial E[f(A)]}{\partial e_i} = (1 - \beta_j) \frac{\partial E[f(A)]}{\partial e_i}$, consistency of (8) and (9) is impossible except in the trivial case $e^* = (0,0)$. □

Thus under uncertainty the efficient vector, e^*, is not implementable with a linear sharing rule: Nash equilibrium implies that the agents do not take into account the positive externality when inducing effort, in contrast to behavior in social optimum.

Proposition 1 is easy to derive but is quite interesting to interpret. For the uncertainty model to be more appropriate than the certainty model it does not have to be the case that individual output is deterministic per se; $G_i(z)$ may be interpreted as reflecting agent j's uncertainty about agent i's output function. In the extension of that it would be interesting to check whether, in the case where individual output is deterministic, one needs common knowledge of the individual output functions for e^* to be implementable, or whether any higher order uncertainty makes implementability break down.

4. An Example

To get intuition on how proposition 1 works, consider the following simple example with uniformly distributed stochastic terms. Let $b_1 = b_2 = 1$ and $f(A) = A$. Furthermore, let $\varepsilon_i \sim_{\text{iid}} U[0,2]$. First assume $e_1 < e_2$. In that case,

$$E[f(A)] = \int_0^{2e_1} a g_A(a) da =$$

$$\int_0^{2e_2} \frac{a}{2e_1} da + \int_0^{2e_2} \frac{a}{2e_2} da - 2\int_0^{2e_2} \frac{a}{4e_1e_2} da = [\frac{a^2}{4e_1} + \frac{a^2}{4e_2} - \frac{a^2}{4e_1e_2}]_0^{2e1} = \frac{e_1}{e_2}(e_1 + e_2 - 1).$$

By using the same procedure for $e_1 \geq e_2$ we get,

$$E[f(A)] = \begin{cases} \frac{e_1}{e_2}(e_1 + e_2 - 1) & \text{for } e_1 < e_2 \\ \frac{e_2}{e_1}(e_1 + e_2 - 1) & \text{for } e_1 \geq e_2, \end{cases} \qquad (16)$$

which is continuous. As can easily be checked, $E[f(A)]$ is differentiable since the derivative from the left equals the derivative from the right in the point $e_1 = e_2$. Thus e^* is not implementable with a linear sharing rule.

5. References

Holmström, B. (1982). Moral Hazard in Teams. *Bell Journal of Economics*, **13**, 324-340.

Legros, P. & Matthews, S. (1993). Efficient and Nearly-Efficient Partnerships. *Review of Economic Studies*, **68**, 599-611.

Vislie, J. (1994), Efficiency and Equilibria in Complementary Teams. *Journal of Economic Behavior and Organization* **23**, 83-91.

Chapter 6

Leontief Partnerships with Outside Options[1]

Abstract

A weakness of the sharing rule proposed by Legros & Matthews (1993) and Vislie (1994) is that under broad conditions it does not satisfy individual rationality. We construct a sharing rule s^* that satisfies both incentive compatibility and individual rationality.

Keywords: Budget-balance, Leontief technology, Linear Implementation, Outside Options, Teams.

[1] Thanks to Terje Lensberg for useful comments. This chapter is a revised and shortened version of the NHH 11/ 96 discussion paper «Delay in Joint Projects».

1. Introduction

Legros & Matthews (1993) and Vislie (1994) construct an incentive compatible linear sharing rule for a simultaneous action (that are non-observable or non-contractible) partnership game with strictly complementary inputs. The fraction agent *i* receives of joint output, β_i*, depends only on his marginal productivity and marginal cost of effort in optimum.

Even though the model where inputs are strict complements is formally speaking quite restricted, it gives a natural starting point to for example an equilibrium analysis of the completion time of partnership projects, where - as a first approximation - the agents precommitt their effort level,[2] and where the completion time of a project is when the last agent in the partnership finishes his subtask.[3] In view of that application, we think the following weakness of the rule β* proposed in the above papers is important; if an agent has an attractive outside option (but not sufficiently attractive to make the partnership inefficient) the payoff associated with β* may induce him to choose the outside option rather than to participate in the partnership. In other words, β* may not satisfy individual rationality.

We solve this problem by constructing a (non-linear) sharing rule *s** that makes the socially efficient action both incentive compatible and individually rational. With *s**, which in fact is a continuum of incentive compatible sharing rules, we can divide net surplus in any way we like between the agents, and still satisfy individual rationality. Thus agents with high participation constraints can be given a larger share of surplus than what is possible with β*.[4]

[2] See Abreu, Milgrom & Pearce (1990) for a discussion of precommitment and change of strategies underway in a timing game.

[3] A similar point occurs in the R&D literature, see e.g., Dasgupta & Maskin (1987).

[4] Notice the nice connection to bargaining theory.

2. The model

Joint payoff x is determined by a Leontief technology, where $x = f(\min\,[b_1e_1,b_2e_2])$; e_i is agent i's choice of effort, where $e_i \in [0,E_i]$, and E_i is finite. For simplicity we let $i \in \{1, 2\}$. $f(..)$ is a differentiable, strictly increasing and concave function with $f(0) = 0$. Notice that $f(..)$ is invertible. Cost of effort is given by $v_i(e_i)$, where v_i is increasing and convex. Let s be a *sharing rule*, i.e., a function $s{:}\Re \rightarrow \Re^n$ that distributes the joint payoff x between the n agents. The utility of agent i is $s_ix - v_i(e_i)$, where $s_i > 0$ and $\Sigma_i s_i = 1$.

Define the efficient effort-vector, e^*, as:

$$e^* = \arg\max_{e}\,[\,f(\min\,[b_1e_1,b_2e_2]) - \Sigma_i v_i(e_i)\,], \tag{1}$$

The vector of fractions β^*, where $\beta_i^* \equiv \dfrac{v_i'(e_i^*)}{x'(e^*)b_i}$, satisfies incentive compatibility as shown by Legros & Matthews (1993).[5] Thus, if participation constraints are sufficiently low, β^* implements e^* in Nash strategies.

It is simple to show that β^* may not implement e^* if participation constraints are raised. Consider the following example. Adam and Betty consider writing a book together. Abstracting from quality considerations, their joint payoff is a function of when they finish the book. The time of completion is given by the maximum of individual completion time. The time agent i finishes his subtask is the inverse of e_i; the more effort, the earlier completion time. Their joint payoff is determined by the function $x = 2\min(e_1, e_2)$, where e_i is agent i's (precommitted) effort level. Both Adam and Betty has cost of effort equal to $\frac{1}{2}e_i^2$. It is simple to verify that the efficient effort vector is $e_1 = e_2 = 1$, which if chosen by the participants gives joint (net) payoff equal to 1. The sharing rule β^* dictates Adam and Betty to share the output in equal shares, which gives them

[5] See chapter 5 for a proof.

both utility equal to 1/2. Clearly e^* is a Nash equilibrium for both participation constraints equal to zero.

Suppose Betty may choose to either write a book with Adam or to work on her own project. This latter project excludes working with Adam but gives her utility ¾ (any number between ½ and 1 will make our point). Adam, on the other hand, has outside option zero. Trivially, the equilibrium outcome under β^* gives Betty less than doing her own project, and thus the efficient action (writing the book with Adam) is not an equilibrium action for her. We construct a (non-linear) sharing rule that solves this problem, but let us first note that there do not exist other incentive compatible linear sharing rules than β^*.

Lemma 1.

β^* is the unique incentive compatible linear compatible sharing rule.

The proof is straightforward and hence omitted.

Now to the main result. Let $g_i : \Re \rightarrow E_i$ be a function that *given an outcome x, computes agent i's effort given that he did not waste effort.*[6] Furthermore, let the *quasi-surplus*, *Q*, be defined as, $Q(x) \equiv [x - \sum_{i=1}^{n} v_i(g_i(x))]$. For a given x, Q measures the total surplus if no agent wasted effort in realizing x. Observe that for a symmetric effort vector actual surplus and quasi-surplus are equal, while for an asymmetric effort-vector the quasi-surplus is larger than the total surplus since the total cost component in the quasi-surplus is smaller.

Define $s^*(x)$ as,

$$s_i^*(x) \equiv v_i(g_i(x)) + m_i\mathrm{Q}(x), \text{ with } m_i > 0 \text{ and } \sum_{i=1}^{n} m_i = 1. \qquad (2)$$

[6] Formally, $g_i(x) \equiv (f^1(x))/b_i$.

Proposition 1.

By an appropriate choice of m_i, the sharing rule $s^*(x)$ satisfies both incentive compatibility and individual rationality.

Proof.

Clearly $s_i^*(x)$ is budget-balancing since,

$$\sum_{i=1}^{n} s_i{}^*(x) = \sum_{i=1}^{n} v_i(g_i(x)) + \sum_{i=1}^{n} m_i Q(x) = \sum_{i=1}^{n} v_i(g_i(x)) + \sum_{i=1}^{n} m_i [x - \sum_{i=1}^{n} v_i(g_i(x))] = \sum_{i=1}^{n} v_i(g_i(x)) + x - \sum_{i=1}^{n} v_i(g_i(x)) = x. \tag{3}$$

Consider a unilateral deviation from e^* by agent i. Denote this deviation e_i', where $e_i' < e_i^*$. Define $x' \equiv x(e_i', e_{-i}^*)$, and $x^* \equiv x(e_i^*, e_{-i}^*)$. Then,

$$u_i(e_i^*, e_{-i}^*, s^*) - u_i(e_i', e_{-i}^*,s^*) =$$

$$\{v_i(g_i(x^*)) + m_i\ [x^* - \sum_{i=1}^{n} v_i(g_i(x^*))] - v_i(e_i^*)\} - \{v_i(g_i(x') + m_i\ [x' - \sum_{i=1}^{n} v_i(g_i(x'))\ ] - v_i(e_i')\}$$

Since agent i wastes effort neither in e^* nor in (e_i', e_{-i}^*), we have that $v_i(g_i(x')) = v_i(e_i')$ and that $v_i(g_i(x^*)) = v_i(e_i^*)$. Thus the difference transforms into,

$$m_i\{[x^* - \sum_{i=1}^{n} v_i(e_i^*)] - [x' - \sum_{i=1}^{n} v_i(e')]\}, \tag{4}$$

which is non-negative by the definition of e^*. Thus e^* is incentive compatible under s^*. That s^* may be rigged to satisfy any participation constraint follows from the fact that by adjusting m_i we can distribute net surplus in any way we want between the agents, and thus the agents with higher participation constraints can be given stronger incentives to participate in the partnership than under β^*.

The example revisited.

Consider s^* with $m_1 = 1/5$ and m_2 equal to 4/5. Then Betty chooses between utility 3/4 if she does her own project, and utility 4/5 in equilibrium payoff in the partnership game. The participation constraint is satisfied, and e^* is a Nash equilibrium.

4. References

Abreu, D., Milgrom, P. & Pearce, D. (1991). Information and Timing in Repeated Partnerships. *Econometrica* **59**, 1713-33.

Dasgupta, P. & Maskin, E. (1987). The Simple Economics of Research Portfolios. *Economic Journal* **97**, 581-95.

Holmström, B. (1982). Moral Hazard in Teams. *Bell Journal of Economics* **13**, 324-340.

Hvide, H. K. (1996), Delay in Joint Projects. *NHH discussion paper 11/96.*

Legros, P. & S.A. Matthews (1993) Efficient and Nearly-Efficient Partnerships. *Review of Economic Studies* **68**, 599-611.

Vislie, J, 1994, Efficiency and Equilibria in Complementary teams. *Journal of Economic Behavior and Organization* **23**, 83-91.