Journal of Mechanism and Institution Design

CONTENTS

Volume 6, Issue 1, December 2021
ISSN: 2399-844X (Print), 2399-8458 (Online)

Editorial board

Published by
The Society for the Promotion of Mechanism and Institution Design
Editorial office, Centre for Mechanism and Institution Design
University of York, Heslington, York YO10 5DD
United Kingdom
http://www.mechanism-design.org

ISSN: 2399-844X (Print), 2399-8458 (Online), DOI: 10.22574/jmid

The founding institutional members are
University of York, UK
Alpen-Adria-Universität Klagenfurt, Austria
Southwestern University of Economics and Finance, China.

Cover & Logo Artwork @ Jasmine Yang
LaTeX Editor & Journal Design @ Paul Schweinzer (using 'confproc')
LaTeX Editorial Assistants @ Theresa Marchetti & Daniel Rehsmann

Printed in York and Klagenfurt by the Centre for Mechanism and Institution Design

A Letter from the Editor

The year of 2021 has turned out to be yet another difficult year. On the one hand, the Covid pandemic is still running at large and has raised the number of deaths up to 5,274,373 as of December 6, according to Worldometer. On the other hand, many countries have experienced unprecedented natural disasters such as severe droughts, torrential rain, and flooding. So many people lost their lives! So many people have been badly hurt! These apocalyptic tragedies and warnings have made us wonder what has gone wrong with humanity and whether we can do better for our future. For instance, modern transportation makes travelling convenient but also makes diseases spread easily and the environment more vulnerable. The internet makes communication fast and working from home possible but also makes unhealthy information spread quickly. How to manage and regulate the use of modern technologies for the betterment of society is clearly an important subject of study and is well within the field of this Journal.

Over the last two years, many university courses and seminars have been conducted online. Most people have had no chance to attend conferences and workshops in person. As the situation is getting better because of the development of vaccination and the weakening of the pandemic, there is a hopeful sign that we will be able to travel and meet people in a near future. As the new year is coming, after long isolation and impersonal virtual contact, we are really looking forward to meeting colleagues and friends face to face and reviving personal relationships at the Conference on Mechanism and Institution Design, to be hosted by the National University of Singapore, Singapore, July 11-15, 2022, and organized by Jingfeng Lu. There will be four keynote talks to be given by distinguished economic theorists: Fuhito Kojima of University of Tokyo & Stanford University, Dan Kovenock of Chapman University, Alessandro Pavan of Northwestern University, and Rakesh Vohra of University of Pennsylvania.

Our Society for the Promotion of Mechanism and Institution Design is an independent learned society, a recognized UK charity body, managing the bi-annual Conference on Mechanism and Institution Design and its flagship Journal of Mechanism and Institution Design. Its objective is to advance education and research for the public benefit in the subject of mechanism and institution design. The Journal is not associated with any commercial publisher and its sole and ultimate goal is to publish high quality articles in the stated field and its closely related ones. It is totally free and open-access to everyone and does not impose any charge on the authors. We are experimenting a BRAND NEW MODEL of open-access journal, going beyond any existing scientific journals by managing it by ourselves from refereeing, editing, designing, to production. But we rely on voluntary contributions and donations. We have received generous support over the past years and we hope that more and more individuals and academic organizations will share our vision and support us in the future. We highly appreciate your support.

We are deeply grateful to our distinguished colleagues Randall Calvert, Paul Healy, and James Walker, who will step down from our editorial board by the end of this year, for their valuable service, advice, and support. At the same time, we are pleased to announce that two outstanding colleagues, Shigehiro Serizawa and Jun Zhang, will join our editorial board from January 2022. We warmly welcome them and look forward to working with them.

Zaifu Yang, York, December 6th, 2021.

Journal of Mechanism and Institution Design
ISSN: 2399-844X (Print), 2399-8458 (Online)
DOI:10.22574/jmid.2021.12.001

OBJECT-BASED UNAWARENESS: AXIOMS

Oliver J. Board

Paul | Weiss, USA
ojboard@paulweiss.com

Kim-Sau Chung

Hong Kong Baptist University, China
kschung@hkbu.edu.hk

ABSTRACT

This paper provides foundations for a model of unawareness, called object-based unawareness (OBU) structures, that can be used to distinguish between what an agent is unaware of and what she simply does not know. At an informal level, this distinction plays a key role in a number of papers such as Tirole (2009) and Chung & Fortnow (2016). In this paper, we give the model-theoretic description of OBU structures by showing how they assign truth conditions to every sentence of the formal language used. We then prove a model-theoretic sound and completeness theorem, which characterizes OBU structures in terms of a system of axioms. We then verify that agents in OBU structures do not violate any of the introspection axioms that are generally considered to be necessary conditions for a plausible notion of unawareness. Applications are provided in our companion paper.

Keywords: Awareness, object-based unawareness, modal logic.

JEL Classification Numbers: D83, D91.

This paper was first circulated in 2007. The literature has since grown much bigger than is reflected in our references. We apologize for not being able to do justice to this subsequent literature. We thank Eddie Dekel, Lance Fortnow, Joseph Halpern, Jing Li, Ming Li, and seminar participants at various universities for very helpful comments. We also thank the editor of the Journal for inviting us to submit this paper. All errors are ours.

1. INTRODUCTION

There are two literatures on unawareness, and it is often not clear that the authors in each group are aware of each other's contributions.

The first unawareness literature (let's call it the *applied* literature) consists of applied models, such as Tirole (2009) and Chung & Fortnow (2016), where agents are uncertain whether they are aware of everything that their opponents are aware of, and have to strategically interact under these uncertainties. For example, in Tirole (2009), a buyer and a seller negotiate a contract as in the standard holdup problem. At the time of negotiation, there may or may not exist a better design for the product. Even if a better design exists, however, the contracting parties may not be aware of it. If a party is aware of it, he can choose whether or not to point it out to the other party. But even if he is not aware of it, he *is* aware that a better design may exist and his opponent may be aware of this better design. In Tirole's words, "parties are unaware, but aware that they are unaware"; and they have to negotiate under this uncertainty. Chung & Fortnow (2016) consider the plight of an American founding father drafting a Bill of Rights that will be interpreted by a judge 200 years later. The founding father is aware of some human rights, but is uncertain whether or not there are other human rights that he is unaware of. Here, as in Tirole (2009), the founding father is unaware, but aware that he may be unaware; and he has to choose how he should write the Bill of Rights in the face of this uncertainty.

The second unawareness literature (let's call it the *foundational* literature) attempts to provide a more rigorous account of the properties of unawareness: see e.g. Fagin & Halpern (1988), Modica & Rustichini (1994), Modica & Rustichini (1999), Dekel et al. (2016), Halpern (2001), Li (2006), Halpern & Rego (2006), Sillari (2006), and Heifetz et al. (2006b), Heifetz et al. (2006a). These authors are motivated by the concern that *ad hoc* applied models, if not set up carefully enough, may go awry in the sense that agents in those models may violate rationality in some way, as captured by various introspection axioms (which we shall refer to as the DLR axioms hereafter).[1] This concern is articulated in Modica & Rustichini (1994), and Dekel et al. (2016). The rest of this literature proposes various models that are set up carefully enough to

[1] In particular, two of the key axioms that lie behind Dekel et al.'s (2016) impossibility result are *KU-introspection* ("the agent cannot know that he is unaware of a specific event") and *AU-introspection* ("if an agent is unaware of an event E, then he must be unaware of being unaware of E").

take these concerns into account.

These two literatures are somewhat disconnected. For example, Tirole makes no reference to any work in the foundational literature, nor does he explain whether or not his agents satisfy the introspection axioms that are one of the main concerns of that literature. Similarly, none of the papers in the foundational literature explain whether Tirole's model may fit in their framework, and if not, whether Tirole's agents violate some or all of the introspection axioms. This paper and its companion paper, Board & Chung (2008), attempt to connect these two literatures.

There is a reason why it is difficult to directly compare Tirole's model with the majority of the models proposed in the foundational literature. To propose a model, and to provide foundations for it, an author needs to explain how her model should be interpreted. In several of the papers discussed above (e.g. Fagin & Halpern (1988)), this is done by showing how a formal structure assigns truth conditions to each sentence of some formal language; i.e., by the procedure of systematically giving yes/no answers to a laundry list of questions such as: "At state w, does agent i know that it is sunny in New York?" In many of the papers more familiar to economists (e.g. Li (2006)), although this procedure is not performed explicitly, there is typically a clear way to assign truth conditions to an appropriately-specified language according to the author's description of her model. But the procedure of assigning truth-conditions is well-defined only if the set of questions (to be given yes/no answers) is defined clearly. This set of questions is determined by the language associated with the proposed model, and is chosen (either explicitly or implicitly) by the author. But this also means that we can understand a proposed model only up to its associated language. If we ask a question that does not belong to the associated language, we cannot expect to find an answer.

Unfortunately, questions such as "At state w, does agent i know that he is not aware of everything?" do *not* belong to the language of many of the studies in foundational literature (notable exceptions include Halpern & Rego (2006) and Sillari (2006), which we return to in the next paragraph). More generally, the languages underlying many of the studies cited above do not contain quantifiers; while sentences such as "agent i is aware of everything" (implicitly) do. This provides some explanation as to why it is difficult to compare the two literatures. In other words, while in Tirole's model, "parties are unaware, but aware that they are unaware", it is difficult to figure out when or if this would be true of the agents in most of the models proposed in the

foundational literature. Those models do not address such questions, and hence our understanding of them is somewhat limited.

Several contributions by logicians and computer scientists, however, present models that *do* address these questions (e.g., Halpern & Rego (2006) and Sillari (2006)). These papers explicitly present and analyze formal languages that contain quantifiers, and are thus richer than the languages underlying the models discussed above. Their models, however, are very different from the applied models used by Tirole (2009) and Chung & Fortnow (2016). For example, Halpern & Rego (2006) decribe an agent's awareness by means of a *syntactic* awareness function mapping states to sets of sentences of the formal language, to be interpreted as the set of sentences the agent is aware of. Certain restrictions are then imposed on the form of this function to capture a plausible notion of awareness. This "list of sentences" approach is more general, but the cost of this additional generality is less structure. This may explain why this approach, while not uncommon in the formal logic literature, is rarely seen in economics.[2]

In the specific case of Halpern & Rego (2006), there is a more specific reason why it could not be used to provide foundations for applied models such as Tirole (2009) and Chung & Fortnow (2016). In both of these papers, although agents know what they are aware of, they may be uncertain whether or not they are aware of everything. Such uncertainty cannot arise in Halpern & Rego (2006), however.[3] To capture this kind of uncertainty, they would have

[2] To provide an analogy that may help elucidate this comparison, consider the difference between Aumann's information partition model, where a partition of the state space is used to derive an agent's knowledge of events, and a "list of sentences" approach where knowledge is instead modeled by a list of sentences describing exactly what that agent knows.

[3] For readers who are familiar with Halpern & Rego (2006), this can be proved formally as follows. Recall the following definition in Halpern & Rego (2006): "*Agents know what they are aware of* if, for all agents i and all states s, t such that $(s,t) \in \mathcal{K}_i$ we have that $\mathcal{A}_i(s) = \mathcal{A}_i(t)$." So it suffices to prove that, in any instance of Halpern & Rego (2006) structure, if there is a state t such that agent i is uncertain whether or not there is something he is unaware of, then there must be another state s such that $(s,t) \in \mathcal{K}_i$ but $\mathcal{A}_i(s) \neq \mathcal{A}_i(t)$. Let $\alpha = \exists x \neg A_i x$ represent "there is something that agent i is unaware of". Therefore, $\neg\alpha$ means "there is nothing that agent i is unaware of". Let $\beta = A_i\alpha \wedge A_i\neg\alpha \wedge \neg X_i\alpha \wedge \neg X_i\neg\alpha$ represent "agent i is aware of both α and $\neg\alpha$ but he does not know whether α or $\neg\alpha$ is true (recall that X_i is Halpern & Rego (2006) explicit knowledge operator). In short, β means "agent i is uncertain whether or not there is something he is unaware of". Let M be any instance of Halpern & Rego (2006) structure, and t is a state such that $(M,t) \models \beta$. Then we have $(M,t) \models \neg K_i\alpha \wedge \neg K_i\neg\alpha$ (recall that K_i is Halpern & Rego (2006) implicit knowledge operator). Therefore, there exists

to consider a framework in which the formal language is allowed to vary across states: an agent who is unable to distinguish between two states with different languages could thus be uncertain about how many sentences there are, and hence uncertain about how many she is unaware of.

To summarize, while the assumption that "agents are unaware, but are aware that they are unaware" plays a key role in much of the applied literature on unawareness, the foundations of these models remain unclear. We do not know whether agents in these models violate some or all the introspection axioms that are one of the main concerns of the foundational literature. This paper and its companion paper, Board & Chung (2008), attempt to provide this missing foundation.

In these two papers, we describe a model, or more precisely a class of models, called *object-based unawareness structures* (OBU structures). Readers will find that these structures encompass models used in the applied literature. In comparison with the applied literature, however, we provide complete and rigorous foundations for these structures. The underlying language we use is rich, and in particular contains quantifiers, enabling us to describe explicitly whether or not agents are aware that they are unaware. We will provide an axiomatization for these structures, verifying that all of the appropriate introspection axioms are satisfied. The value of thinking about agents who exhibit this kind of uncertainty has already been demonstrated by the existing applied literature; we demonstrate the tractability of our framework by considering further applications.

A key feature of our structures is that unawareness is object-based:[4] a seller may be unaware of a better design, or a founding father may be unaware of a particular human right. In contrast, in models of unforeseen contingencies, agents are unaware of contingencies, or states. This raises the question of whether the agents in our model are aware of every state. We do not have answer to this question. As we argued above, our understanding of any given model is constrained by the language we choose to work with. Although our language is one of the richest in the foundational literature, there are questions that fall outside of it. We do not have answers to these questions, simply because we do not speak that language.

a state s such that $(t, s) \in \mathcal{K}_i$ and $(M, s) \models \neg\alpha$, and another state s' such that $(t, s') \in \mathcal{K}_i$, and $(M, s') \models \alpha$. Since $\alpha = \exists x \neg A_i x$, there exists ϕ such that $\phi \in \mathcal{A}_i(s)$ and $\phi \notin \mathcal{A}_i(s')$. But that means at least of one $\mathcal{A}_i(s)$ and $\mathcal{A}_i(s')$ is different from $\mathcal{A}_i(t)$.

[4] We discuss other possible sources of unawareness in the conclusion.

The division of labor between this paper and its companion paper, Board & Chung (2008), is as follows. In this paper, we give the model-theoretic description of OBU structures by showing how they assign truth conditions to every sentence of the formal language. We then prove a model-theoretic sound and completeness theorem, which characterizes OBU structures in terms of a system of axioms. We then verify that agents in OBU structures do not violate any of the introspection axioms that are generally considered to be necessary conditions for a plausible notion of unawareness. This paper also contains a more complete literature review, as well as a discussion of several variants of OBU structures.

In our companion paper, Board & Chung (2008), we give a set-theoretic description of the OBU structures. Although less formal than the model-theoretic treatment, we hope this will be more accessible to the general audience. In parallel to the model-theoretic sound and completeness theorem in this paper, we prove set-theoretic completeness results in Board & Chung (2008).

The second half of Board & Chung (2008) considers two applications. First, we use the model to provide a justification for the *contra proferentem* doctrine of contract interpretation, commonly used to adjudicate ambiguities in insurance contracts. Under *contra proferentem*, ambiguous terms in a contract are construed against the drafter. Our main result is that when drafter (the insurer) has greater awareness than the other party (the insured), *and when the insured is aware of this asymmetry*, *contra proferentem* minimizes the chances that the insured forgoes gain of trade for fear of being exploited. On the other hand, when there is no asymmetric awareness, efficiency considerations suggest no reason to prefer *contra proferentem* over an alternative interpretive doctrine that resolves ambiguity in favor of the drafter.

From the perspective of our framework, an argument common among legal scholars as far back as Francis Bacon, that *contra proferentem* encourages the insurer to write clearer contracts, misses the point. If a more precise contract increases the surplus to be shared between the insurer and the insured, market forces provide incentives to draft such a contract regardless of the interpretive doctrine employed by the court. The advantage of *contra proferentem* is rather that it enables the insurer to draft more acceptable contracts, by expanding the set of events that he can credibly insure.

Our second application examines speculative trade. We first generalize the classical No Trade Theorem to situations where agents are delusional

but nevertheless satisfy a weaker condition called terminal partitionality. We then introduce the concepts of *living in denial* (i.e., agents believe, perhaps incorrectly, that there is nothing that they are unaware of) and *living in paranoia* (i.e., agents believe, perhaps incorrectly, that there is something that they are unaware of). We show that both living in denial and living in paranoid, in the absence of other forms of delusion, imply terminal partitionality, and hence the no trade theorem result obtains.

The structure of this paper is as follows: Section 2 contains our main result. Section 2.1 first defines the language, or equivalently, the set of sentences that we are to assign truth values to. Section 2.2 presents our axioms. Section 2.3 then presents the class of structures, which we shall call the object-based unawareness (OBU) structures, that is axiomatized by exactly the axioms presented in Section 2.2. Section 2.4 formally states our characterization theorem.

Section 3 gives an example of our framework in use, showing how one can use an OBU structure to model those American founding fathers who were opposed to including the Bill of Rights in the constitution.

In section 4, we verify that our structures satisfy the DLR axioms. Then, in Section 5, we discuss how to incorporate other axioms of interest.

Section 6 reviews the previous unawareness literature, and Section 7 concludes.

2. OBJECT-BASED UNAWARENESS

We start by introducing our language. Formally, a *language* is a set of sentences; it should be rich enough to express everything we might want to say about the agents in our model, such as "the founding father knows that there are some human rights he is not aware of". As we explain below, this requires working with a version of first-order modal logic rather than propositional modal logic, which (implicitly or explicitly) forms the basis of the majority of the models developed in the preceding literature on unawareness.

Although we do not describe object-based unawareness (OBU) structures until Section 2.3, let's use $\mathcal{M}$ to denote the class of such structures that characterizes the axiom system we are about to introduce, with M being a typical structure within that class. Within the set of all sentences, there is a subset that is of particular interest, namely those sentences that are *valid* in $\mathcal{M}$. The definition of a valid sentence will also have to wait until Section 2.3,

but roughly speaking valid sentences are those sentences that are always true in every structure M in $\mathcal{M}$. In fact, the notion of validity (if not the word itself) appears in other contexts that will be familiar to many economists. For example, the sentence "it cannot be common knowledge that agents disagree on the probability of the same fact" is a valid sentence in the class of partitional information structures with a common prior, and is a result proved by Aumann (1976) using those structures. Axiomatization of a class of structures such as $\mathcal{M}$ is in effect axiomatization of its set of valid sentences.

Axiomatization of the set of valid sentences takes the form of a procedure to generate a set of *provable formulas* that coincides with the set of valid sentences. The procedure has two parts. First, some sentences are declared provable directly. These sentences are called *axioms*. Second, sentences other than axioms can also be qualified as provable indirectly by association with existing provable formulas. The association rules are called *inference rules*. One way to interpret this exercise is to think of axioms and inference rules as two different forms of "hidden assumptions" of the class of structures being axiomatized. Every provable formula one may prove using a given class of structures must have its root in some of these "hidden assumptions."

One last remark before we start: although we have been using the word "sentences" casually so far, we shall stop doing so below. The reason is that logicians typically reserve the word "sentences" for something else, and use the word *formulas* to refer to what we have been calling "sentences." We shall follow this convention below.

2.1. The Language

Our language (to be formally defined shortly) is a version of first order modal logic. Roughly speaking, first order modal logic is first order logic augmented by modal operators, and first order logic is an extension of propositional logic. Examples of formulas in propositional logic include $\neg\alpha$ (read "it is not the case that α"), $\alpha \wedge \beta$ (read "α and β"), $\alpha \rightarrow \beta$ (read "whenever α is true, β will be true as well"), etc. First order logic extends propositional logic by including formulas such as $\forall x Tall\,(x)$ ("every x is tall"). Modal operators are represented by letters such as K_i and A_i that can be affixed to simpler formulas and result in longer ones: $K_i\alpha$ (read "agent i knows that α") and $A_i\alpha$ (read "agent i is aware that α"). In this way, modal operators allow us to construct formulas that describe the mind of an agent. Their meaning will be governed

by axioms, which are the subject of the next subsection.[5]

In addition to K_i and A_i, we will have a third modal operator denoted L_i in our language. Although we will not present axioms that govern the meaning of L_i until the next subsection, it is useful give an informal interpretation right now. We would like to use L_i to represent an alternative kind of knowledge that differs slightly from K_i. In particular, L_i stands for the kind of "know" that appears in the following English sentence: "If Madison had been aware of the right to universal suffrage, he would have known that it was important, and would have included it in the Bill of Rights". Here, "know" refers to knowledge in the benchmark case—a hypothetical case where Madison were not plagued by his unawareness of the right to universal suffrage. It is not the same as actual knowledge, because Madison *was* plagued by unawareness. In the previous literature, what we call "benchmark knowledge" (L_i) and "actual knowledge" (K_i) have been called "implicit knowledge" and "explicit knowledge", respectively. Although we do not think these names are ideal, they have entered standard usage and we shall follow henceforth this convention.[6]

We now formally describe the language. Fix a set N of agents. Fix an infinite set X of *variables*, with typical elements $x, y, z, x_1, x_2, x_3, \ldots$, etc. Fix some vocabulary consisting of a set of relation symbols (*predicates*); e.g. $Tall(x)$ ("x is tall"), $Taller(x, y)$ ("x is taller than y"), etc. This generates a set Φ of *atomic formulas*, $P(x_1, \ldots, x_k)$, where P is a k-ary predicate and $x_1, \ldots, x_k \in X$ are variables. Our language $\mathcal{L}$ is the smallest set of *formulas* that satisfies the following conditions:

- if $\phi \in \Phi$, then $\phi \in \mathcal{L}$;
- if $\phi, \psi \in \mathcal{L}$, then $\neg\phi \in \mathcal{L}$ and $\phi \wedge \psi \in \mathcal{L}$;
- if $\phi \in \mathcal{L}$ and $x \in X$, then $\forall x\phi \in \mathcal{L}$;
- if $\phi \in \mathcal{L}$ and $i \in N$, then $L_i\phi \in \mathcal{L}$ and $A_i\alpha \in \mathcal{L}$ and $K_i\alpha \in \mathcal{L}$.

[5] Augmenting propositional logic (rather than first-order logic) with modal operators generates *propositional modal logic*: this is the language used in most of the previous literature on unawareness. It should be clear that propositional modal logic does not include formulas such as "I am not sure whether or not you are aware of something that I am not".

[6] We believe the names "implicit knowledge" and "explicit knowledge" obscure the fact that L_i is merely an conceptual tool, and is used only as an intermediate step to define K_i, which in turn is our ultimate interest.

We use the following standard abbreviations:

- $\alpha \vee \beta$ for $\neg(\neg\alpha \wedge \neg\beta)$;
- $\alpha \rightarrow \beta$ for $\neg\alpha \vee \beta$;
- $\alpha \leftrightarrow \beta$ for $(\alpha \rightarrow \beta) \wedge (\beta \rightarrow \alpha)$;
- $\exists x\alpha$ for $\neg\forall x\neg\alpha$;
- $U_i\alpha$ for $\neg A_i\alpha$.

Finally, we require that there is a special unary predicate called E. The intended interpretation of $E(x)$ is "x is real". The meaning of "real" will depend on the specific application. For example, a founding father writing the Bill of Rights may come up with the following list of human rights: freedom of speech, freedom to bear arms, freedom of choosing one's own fate. . . . Upon further reflection, he may realize that only the first two are "real" rights, while the third one is merely an artificially-created concept. For an agent trying to enumerate animals, on the other hand, horses and cows are "real", while unicorns are not.

2.2. The Axiom System

Before we state our axioms, we need one more definition. We say that a variable x is *free* in the formula α if it does not fall under the scope of a quantifier $\forall x$.[7] For example, x is free in $\forall y\, Taller(x, y)$ but not in $\forall x\forall y\, Taller(x, y)$.

We are now ready to present our axiom system, called *AWARE*, for the language $\mathcal{L}$. As we explained earlier, an axiom system contains both axioms and inference rules. We shall group the axioms and inference rules into several different categories. The first category is borrowed directly from propositional logic, and is common to all axiom systems in this literature:

[7] More formally, we define inductively what it is for a variable to be free in α:

- if ϕ is an atomic formula $P(x_1, \ldots, x_k)$, then each x is free in the formula;
- x is free in $\neg\alpha$, $K_i\alpha$, $A_i\alpha$, and $L_i\alpha$ iff x is free in α;
- x is free in $\alpha \wedge \beta$ iff x is free in α or β;
- x is free in $\forall y\alpha$ iff x is free in α and x is different from y.

PC all propositional tautologies are axioms.

MP from α and $\alpha \to \beta$ infer β.

PC (propositional calculus) is a set of axioms: propositional tautologies include formulas such as $\alpha \vee \neg\alpha$ and $\big(Tall(x) \wedge \exists x\, Taller(x,y)\big) \to \exists x\, Taller(x,y)$. **MP** (*modus ponens*) is an inference rule, and says that if α and $\alpha \to \beta$ are provable formulas, then so is β. Note that any formulas may be substituted for α and β, and any variables for x and y.

The second category governs the universal quantifier $\forall$:

E1 for any variable x, $\forall x\, E(x)$ is an axiom.

E2 for any formula α and variables x and y, $\forall x\alpha \to \big(E(y) \to \alpha[y/x]\big)$ is an axiom.

E3 for any formulas α and β and variable x, $\forall x(\alpha \to \beta) \to (\forall x\alpha \to \forall x\beta)$ is an axiom.

E4 for any formula α and variable x that is not free in α, $\alpha \leftrightarrow \forall x\alpha$ is an axiom.

UG from α infer $\forall x\alpha$.

Axiom **E1** can be rewritten as $\neg\exists x\neg E(x)$, which gives an interpretation to the existential quantifier in terms of the predicate E. Since there are at least two different ways to interpret the existential quantifier, **E1** is an important axiom in the sense that it clarifies which of the two interpretations is adopted in our system. Consider the following sentence:

> "There exist rights that have not been included in the Bill of Rights—think about the freedom to choose one's own fate."

Depending on how we interpret the word "exist", one may or may not agree that "the freedom to choose one's own fate" is an appropriate example in the above sentence. In particular, if we interpret the word "exist" according to **E1**, then we would likely regard that example as inappropriate, because that freedom is not a real right at all. However, one can conceive another interpretation of the word

"exist" that would make that example appropriate. These two interpretations correspond to what logicians call *actualist existence* and *possibilitist existence*, respectively. We consider the possibilitist interpretation as less useful in economics. The reason, roughly speaking, is that most possibilitist axiom systems we are aware of lead to constant-domain structures, which is an especially restrictive property for economic models. We shall return to this point in Section 6, when we compare our current paper with Halpern & Rego (2006) and Sillari (2006).

In axiom **E2**, $\alpha[y/x]$ is the same formula as α with free y replacing every free x.[8] To understand axiom **E2**, one can consider the following conversation:

> FATHER: "One difference between horses and goats is that horses do not have horns."
> SON: "But unicorns have horns."
> FATHER: "Unicorns exist only in fantasy stories. What I meant was: *real* horses do not have horns."

The "E(y)" part of **E2** captures the father's qualification in his second statement, where α is "if x is a horse then x does not have horns". The basic idea is that the quantifier $\forall$ ranges only over "real" things. **E3** is straightforward. In **E4**, if x is not free in α, then adding $\forall x$ at the beginning of α does not change the meaning; for example, "for all things, Aumann is an economist" has the same information content as "Aumann is an economist" (although the former is rather awkward English). To understand **UG**, consider a formula such as "x is either tall or not tall". Suppose we have managed to find a proof for it by means of other axioms and inference rules. We would like to make sure that the formula "for all x, x is either tall or not tall" is also provable; and **UG** is an inference rule that will help make sure of this.

The third category governs the meaning of explicit knowledge:

K for any formula α, $K_i\alpha \leftrightarrow (A_i\alpha \wedge L_i\alpha)$ is an axiom.

We have already alluded to the idea behind **K** in Section 2.1: an agent explicitly knows a fact if and only if he implicitly knows it and he is not plagued by unawareness problems.

[8] So for example $\left(E(y) \wedge \forall x \exists z P(x,y)\right) \rightarrow \exists z P(y,y)$ is an axiom, but $\left(E(y) \wedge \forall x \exists y P(x,y)\right) \rightarrow \exists y P(y,y)$ is not.

The fourth category governs the meaning of awareness:

A1 for any formula α that contains no free variables, $A_i\alpha$ is an axiom.

A2 for any formulas α and β, $(A_i\alpha \wedge A_i\beta) \rightarrow A_i(\alpha \wedge \beta)$ is an axiom.

A3 for any formulas α and β, if every variable x that is free in β is also free in α, then $A_i\alpha \rightarrow A_i\beta$ is an axiom.

To see that these three axioms capture the idea that awareness is object-based, one may heuristically think of a free variable as referring to some specific object, while a variable that is bound by a $\forall$ quantifier refers to generic objects. With this heuristic understanding, our idea that unawareness of a fact must arise from unawareness of specific objects referred to in the fact will have three implications, each correspond to one of the three axioms: if a fact does not refer to any specific objects, an agent will be aware of it (**A1**); if the agent is aware of two facts, then he is aware of a more complicated fact which is the conjunction of the two (**A2**); and if an agent is aware of a fact that refers to a collection of specific objects, then he is also aware of a fact that refers to only a subcollection of them (**A3**). These three implications, combined together, also characterize the idea that unawareness of a fact must arise from unawareness of specific objects referred to in the fact.

The last category governs the meaning of implicit knowledge:

L $L_i(\alpha \rightarrow \beta) \rightarrow (L_i\alpha \rightarrow L_i\beta)$.

LN from α infer $L_i\alpha$.

UGL from $\alpha_1 \rightarrow L_i\big(\alpha_2 \rightarrow \cdots \rightarrow L_i(\alpha_h \rightarrow L_i\beta)\cdots\big)$, where $h \geq 0$, infer $\alpha_1 \rightarrow L_i\big(\alpha_2 \rightarrow \cdots \rightarrow L_i(\alpha_h \rightarrow L_i\forall x\beta)\cdots\big)$, provided that x is not free in $\alpha_1, \ldots, \alpha_h$.

These axioms suggest that our agents are very powerful reasoners indeed, at least implicitly. Both **L** and **LN**, with explicit knowledge replacing implicit knowledge, are present in (the axiom system that underlies) all standard state-space models. In this sense they are standard assumptions in economics. **L**

says that the agent can apply the inference rule *modus ponens* in his head (at least when he is not plagued by unawareness problems). **LN** says that our agents implicitly know all provable formulas of *AWARE*, even formulas that no one has ever written down, let alone found a proof for and published in a journal.

UGL is bit of a mouthful. To understand what it says, it may be useful to put $h = 0$ and simplify it to "from $L_i\beta$ infer $L_i\forall x\beta$," which in turn takes a form similar to **UG**.

From these axioms and inference rules we can define the set of provable formulas. A formula can qualify as provable directly because it is an axiom, or it can qualify indirectly by association. The process of showing that a formula qualifies as a provable formula is called a "proof." Formally, a *proof* is a finite sequence of formulas, each of which is either an axiom or follows from preceding formulas in the sequence by applying an inference rule. A proof of α is such a sequence whose last formula is α. A formula is a provable formula iff it has a proof.

2.3. The Object-Based Unawareness Structures

We now present the class of structures $\mathcal{M}$ (together with a truth-value assignment rule); we shall show (Theorem 1) that this class of structures is axiomatized by the axiom system *AWARE* presented in Section 2.2 above.

An object-based unawareness (OBU) structure is a tuple

$$M = (W, D, \mathcal{E}, \mathcal{P}_1, \ldots, \mathcal{P}_n, \mathcal{A}_1, \ldots, \mathcal{A}_n, \pi),$$

where:

- W is a set of possible worlds, with typical element w;
- D is a set of objects;
- $\mathcal{E} : W \to D$ is an *existence correspondence*: $\mathcal{E}(w)$ is the set of objects which are real at world w;
- $\mathcal{P}_i : W \to 2^W$ is agent i's *possibility correspondence*: $\mathcal{P}_i(w)$ is the set of worlds that agent i considers possible when the true world is w;
- $\mathcal{A}_i : W \to 2^D$ is agent i's *awareness correspondence*: $\mathcal{A}_i(w)$ is the set of objects that agent i is aware of when the true world is w;

- π is an *assignment* at each w of a k-ary relation $\pi(w)(P) \subseteq D^k$ to each k-ary predicate P.

Intuitively, the assignment π describes the properties of each object; these can differ across worlds, so that for example Alice could be taller than Bob in world w_1, while Bob is taller than Alice in world w_2. Let $\mathcal{M}$ be the class of all OBU structures.

The way we intend to use an OBU structure is very standard, and will be formally captured in the truth-value assignment rule. Before we present the rule, let's go through two simple examples first.

As the first example, suppose John is an element in D, and $Tall$ is one of the predicates. To determine whether or not, in world w, agent i knows that John is tall, we first construct the event that John is tall, which is $E := \{w \mid |John \in \pi(w)(Tall)\}$. We then ask two questions: in world w, (1) does i implicitly know that John is tall ($\mathcal{P}(w) \subset E$)? and (2) is i aware of John ($John \in \mathcal{A}_i(w)$)? If both answers are affirmative, then i knows that John is tall in world w.

As another example, suppose we want to determine whether or not, in world w, i knows that everyone is tall. Once again, we first construct the event that everyone is tall, which is $E := \{w \mid D_w \subset \pi(w)(tall)\}$. Note that we only count those people who are "real"—for example, in a world where Jesus has no son, we do not count "Jesus' son" even if he is an element in D. We then ask: in world w, does i implicitly know that everyone is tall? If the answer is affirmative, then i knows that everyone is tall in world w. Note that we do not need to ask the awareness question, because no specific person is referred to in the fact "everyone is tall", and hence by assumption there will be no unawareness problem.

These will all be formally captured by our truth-value assignment rule. Let a *valuation* V on an OBU structure M be a function that assigns a member of D to each variable x. Intuitively, $V(x)$ describes the object referred to by variable x, provided that it appears free in a given formula, just like how a name is associated to an object. The truth value of a formula depends on the valuation, just like whether or not "Bob is tall" depends on which person bears the name "Bob".

We say that the fact represented by the atomic formula $E(x)$ is true at state w of structure M under valuation V, and write

$$(M, w, V) \models E(x),$$

iff $V(x)$ is one of the objects in D_w. For facts represented by more complicated formulas, we use the following rules inductively:

$(M, w, V) \models P(x_1, \ldots, x_k)$ iff $(V(x_1), \ldots, V(x_k)) \in \pi(w)(P)$;

$(M, w, V) \models \neg\alpha$ iff $(M, w, V) \not\models \alpha$;

$(M, w, V) \models \alpha \wedge \beta$ iff $(M, w, V) \models \alpha$ and $(M, w, V) \models \beta$;

$(M, w, V) \models \forall x\alpha$ iff $(M, w, V') \models \alpha$ for every x-alternative V' of V such that $V'(x) \in D_w$;[9]

$(M, w, V) \models A_i\alpha$ iff $V(x) \in \mathcal{A}_i(w)$ for every x that is free in α;

$(M, w, V) \models L_i\alpha$ iff $(M, w', V) \models \alpha$ for all $w' \in \mathcal{P}(w)$;

$(M, w, V) \models K_i\alpha$ iff $(\mathcal{M}, w, V) \models A_i\alpha$ and $(\mathcal{M}, w, V) \models L_i\alpha$.

If α is true at every w in $\mathcal{M}$ under V, we say that α is valid in M under V, and write

$$(M, V) \models \alpha.$$

If α is valid in M under every V, we say that α is valid in M, and write

$$M \models \alpha.$$

If α is valid in every $M \in C \subseteq \mathcal{M}$, we say that α is valid in C, and write

$$C \models \alpha.$$

2.4. The Characterization Theorem

Theorem 1. *$\phi \in \mathcal{L}$ is valid in $\mathcal{M}$ if and only if ϕ is provable in AWARE.*

In logicians' terminology, *AWARE* is a sound and complete axiomatization of $\mathcal{M}$ in $\mathcal{L}$. The proof of Theorem 1 uses standard methodology.[10] We present the complete proof in the appendix.

[9] We say that V' is an x-alternative of V if, for every variable y except possibly x, $V'(y) = V(y)$.
[10] See, for example, the proof of Theorem 16.2 in Hughes & Cresswell (1996).

3. AN APPLICATION

Chung & Fortnow (2016) use a dynamic game with two players (a legislator who is to write the Bill of Rights, and a judge who is to interpret it 200 years later) to formalize the argument of those American founding fathers who opposed the inclusion of the Bill of Rights into the American Constitution. They prove that, in some parameter range, there is a unique equilibrium where the legislator, who *is not sure whether or not there are still other rights that he is unaware of*, optimally chooses to *not* to write the Bill of Rights. That is, he optimally chooses *not* to enumerate even those rights that he is aware of. The reason is that, in equilibrium, how the judge treats those rights not in the Bill depends on how elaborate the Bill is. The more elaborate the Bill is, the more likely the judge will be to rule that it is constitutional for the government to infringe those rights not included into the Bill.

They also prove that, even if the legislator adds the sentence

> "Any other rights not listed in this Bill are equally sacred and the government should not infringe them."

to the Bill, the equilibrium outcome will be the same.

Instead of reproducing the analysis of Chung & Fortnow (2016) here, let's focus on how one can use an OBU structure to model that legislator.

Consider the following object-based unawareness structure. There are two worlds, w_1 and w_2, and two rights, s and f, where s stands for "freedom of speech" and f stands for "freedom to choose one's own fate". The true state is w_2, where only s is "real". However, both s and f are "real" in the other world, w_1. Formally, it means $D_{w_1} = \{s, f\}$, and $D_{w_2} = \{s\}$. Suppose agent i is aware of only s in both worlds (i.e., $\mathcal{A}_i(w) = \{s\}$ for $w = w_1, w_2$). Then, in world w_1, and only in world w_1, there exists some object that the agent is unaware of. If we use P to stand for some arbitrary property that both objects satisfy in both worlds (i.e., $\pi(w)(P) = \{s, f\}$ for $w = w_1, w_2$), then we have:

$$(M, w_1) \models \exists x U_i P(x),$$
$$(M, w_2) \models \neg \exists x U_i P(x).$$

Suppose, in the true state w_2, the agent cannot distinguish w_1 and w_2 (i.e., $\mathcal{P}_i(w_2) = \{w_1, w_2\}$). Then he is not sure whether or not there exists some right

that he is unaware of. I.e., he does not know for sure that there is no such a right:

$$(M, w_2) \models \neg K_i \neg \exists x U_i P(x);$$

and he does not know for sure that there is such a right:

$$(M, w_2) \models \neg K_i \exists x U_i P(x).$$

This lack of explicit knowledge is not due to unawareness, for he can comprehend both facts:

$$\begin{aligned} &(M, w_2) \models A_i\left(\neg \exists x U_i P\left(x\right)\right), \quad \text{and} \\ &(M, w_2) \models A_i\left(\exists x U_i P\left(x\right)\right). \end{aligned}$$

His lack of explicit knowledge is due to his lack of implicit knowledge—he does not (implicitly) know for sure the exact number of rights that really exist.

Note that this is an example with non-constant domains; i.e., D varies across different worlds. Non-constant domains are important in modelling agents who are not sure whether or not there exist things that they are unaware of. Consider the above example again, but suppose D is constant across different worlds. For example, suppose $D_w = \{s\}$ in both worlds. Then $K_i \neg \exists x U_i P(x)$ would have been true in both worlds. Alternatively, suppose $D_w = \{s, f\}$ in both worlds. Then $K_i \exists x U_i P(x)$ would have been true in both worlds. The possibility of non-constant domains in our structures arises from our adoption of the actualist existence axioms. In Halpern & Rego (2006) and Sillari (2006), where they adopt the possibilitist existence axioms, their structures are necessarily characterized by constant domains instead.

4. THE DLR AXIOMS

A first impression of some readers of this paper is that object-based unawareness structures violate Dekel et al.'s (2016) AU Introspection Axiom. AU Introspection is represented by formulas of the form $U_i\phi \rightarrow U_i U_i \phi$. Indeed, every such formula is a provable in *AWARE* (and hence, by Theorem 1, it is valid in $\mathcal{M}$). To see why, first note that $U_i\alpha$ and α have the same free variables, so $A_i U_i \phi \rightarrow A_i \phi$ is a provable formula of *AWARE* (for all $\phi \in \mathcal{L}$ and all i). By simple propositional reasoning, then, $U_i\alpha \rightarrow U_i U_i \alpha$ is also a provable formula of *AWARE*.

Consider however a similar formula: $U_i\alpha \rightarrow U_i\exists x U_i\alpha$. Suppose α stands for $H(x)$, "x is a human right". Then this formula reads "if an agent is unaware that free speech is a human right, then she is unaware that there is any human right that she is not aware of". Clearly we would not want this formula to be a provable formula in *AWARE* (or valid in $\mathcal{M}$): clearly we would like to be able to model agents who are unaware of some things, but aware (or even explicitly know) that there are things they are unaware of. To show that this formula is indeed not valid in $\mathcal{M}$ (and hence not a provable formula in *AWARE*), it suffices to show that its negation is true in some world w of some object-based unawareness structure $M \in \mathcal{M}$ under some valuation V. If α stands for $H(x)$, x is free in α but not in $\exists x U_i\alpha$. Then if $V(x) \notin \mathcal{A}_i(w)$, we have $(M, w, V) \models U_i\alpha$ but $(M, w, V) \not\models U_i\exists x U_i\alpha$, and so $(M, w, V) \models \neg(U_i\alpha \rightarrow U_i\exists x U_i\alpha)$.

Another DLR axiom is Plausibility, which is represented by formulas of the form

$$U_i\alpha \rightarrow (\neg K_i\alpha \wedge \neg K_i\neg K_i\alpha).$$

Again, every such formula is a provable formula in *AWARE*. This follows easily from **A3** and **K**.

Dekel et al. (2016) posit a third axiom of KU Introspection, which is represented by formulas of the form $\neg K_i U_i\alpha$. Such formulas are not provable formulas in *AWARE*. The basic reason is that there are no axioms in *AWARE* to preclude an agent knowing something that is actually false. (In this sense, instead of implicit and explicit knowledge, we should perhaps call L_i and K_i implicit and explicit *belief*.) So an agent may explicitly know/believe that she is unaware of something, even though she is actually aware of it. Adding the Truth Axiom (**T**: $L_i\alpha \rightarrow \alpha$), would make every instance of $\neg K_i U_i\alpha$ a provable formula.[11] In terms of our structures, **T** corresponds to the restriction that the possibility correspondences are reflexive: $w \in \mathcal{P}(w)$. To be more precise, let $\mathcal{M}^r$ be the class of object-based unawareness structures in which each $\mathcal{P}_i$ is reflexive; then the set of formulas that are valid in $\mathcal{M}^r$ is precisely the set of provable formulas of *AWARE*+**T**.

The impossibility result of Dekel et al. (2016) is stated within the confines of standard state-space models, and they argue that Necessitation and Monotonicity are two characterizing features of those models. Both Necessitation and

[11] Since our language has two knowledge operators, there are two ways to write the Truth Axiom. The stronger version is is $L_i\phi \rightarrow \phi$, which we adopted in the text. An alternative, weaker version is $K_i\phi \rightarrow \phi$. Here, even adding the weaker version suffices to make every instance of $\neg K_i U_i\phi$ a provable formula.

Monotonicity are restrictions imposed on their class of state-space models, and can be translated into restrictions on our OBU structures as well. Necessitation corresponds to the restriction that, for any given OBU structure M and valuation V, if the formula α is true (i.e., $(M, w, V) \models \alpha$) in every world w then the formula $K_i\alpha$ is also true (i.e., $(M, w, V) \models K_i\alpha$) in every world w. Monotonicity corresponds to the restriction that, for any given OBU structure M and valuation V, if the formula $\alpha \rightarrow \beta$ is true (i.e., $(M, w, V) \models \alpha \rightarrow \beta$) in every world w then the formula $K_i\alpha \rightarrow K_i\beta$ is also true (i.e., $(M, w, V) \models K_i\alpha \rightarrow K_i\beta$) in every world w. In general, our OBU structures do not satisfy these two restrictions. After all, our OBU structures are not standard state-space models.

5. OTHER AXIOMS OF INTEREST

The axiom system *AWARE* (and, correspondingly, the class $\mathcal{M}$ of all OBU structures) can be thought of as imposing a minimal set of restrictions on the behavior of our language $\mathcal{L}$. Various additional assumptions have been imposed on models of knowledge and unawareness elsewhere in the literature. In this section, we shall discuss several such assumptions. In each case, we offer an axiomatic representation, and explain how it corresponds to a particular subclass of $\mathcal{M}$.

To begin with, the following axioms are standard in the economics literature, and are implicit in the partitional model of knowledge used in the vast majority of economic applications:

T	$L_i\phi \rightarrow \phi$
PI	$L_i\phi \rightarrow L_iL_i\phi$
NI	$\neg L_i\phi \rightarrow L_i\neg L_i\phi$

We have already come across the Truth Axiom **T** in Section 4. **PI** is the Axiom of Positive Introspection, and **NI** is the Axiom of Negative Introspection. Note that all three are stated in terms of implicit knowledge L_i instead of explicit knowledge K_i. These axioms have been interpreted by some as rationality requirements on the agents, but generally they are considered to be unrealistically strong.

As before, we say that an agent's possibility correspondence $\mathcal{P}_i$ is reflexive if $w \in \mathcal{P}(w)$ for all w. We say that it is *transitive* if $x \in \mathcal{P}(w)$ and $y \in \mathcal{P}(x)$ imply $y \in \mathcal{P}(w)$ for all w, x, y; and *Euclidean* if $x \in \mathcal{P}(w)$ and $y \in \mathcal{P}(w)$ imply $y \in \mathcal{P}(x)$ for all w, x, y. Let $\mathcal{M}^r$, $\mathcal{M}^t$, and $\mathcal{M}^e$ denote the subclasses

of $\mathcal{M}$ in which all $\mathcal{P}_i$'s are reflexive, transitive, and Euclidean, respectively. We shall also use, for example, $\mathcal{M}^{re}$ to denote the subclass of $\mathcal{M}$ in which all $\mathcal{P}_i$'s are reflexive *and* Euclidean. The following straightforward extension of Theorem 1 formalizes the notion that reflexivity corresponds to **T**, transitivity to **PI**, and Euclideanness to **NI**:

Theorem 2. *The set of formulas $\alpha \in \mathcal{L}$ that are valid in $\{\mathcal{M}^r, \mathcal{M}^t, \mathcal{M}^e, \mathcal{M}^{rt}, \mathcal{M}^{re}, \mathcal{M}^{re}, \mathcal{M}^{rte}\}$ is exactly the set of provable formulas in AWARE+**T**,**PI**,**NI**, **TPI**,**TNI**,**PINI**,**TPINI**.*

One may also be interested in an axiom that says every agent knows what he is aware of:

KA $A_i\phi \rightarrow K_iA_i\phi$.

A related axiom, A-Introspection ($A_i\alpha \leftrightarrow K_iA_i\alpha$), appears in Heifetz et al. (2006b). Note that **KA** and **A3** imply A-Introspection. *AWARE*+**KA** corresponds to the subclass of $\mathcal{M}$ in which the possibility correspondences satisfy the following restriction: for any w and any $w' \in \mathcal{P}_i(w)$, $\mathcal{A}_i(w) \subseteq \mathcal{A}_i(w')$.

In the presence of **A3** and **K**, it is straightforward to show that **KA** is equivalent to:

LA1 $A_i\phi \rightarrow L_iA_i\phi$ is an axiom.

Inspired by **LA1**, some may be tempted to add its "mirror image" as well:

LA2 $U_i\phi \rightarrow L_iU_i\phi$ is an axiom.

LA2 has appeared in some earlier studies on unawareness.[12] We cannot, however, think of any justification for it,[13] other than the purely aesthetic fact that it looks similar to **LA1**. It is straightforward to show that *AWARE*+**LA1**+**LA2** corresponds to the subclass of $\mathcal{M}$ where, for any w and any $w' \in \mathcal{P}_i(w)$, $\mathcal{A}_i(w) = \mathcal{A}_i(w')$.

[12] It appears, for example, as Axiom **A12** in Halpern (2001).

[13] Or perhaps we should say we are not *aware* of any justification for it.

6. LITERATURE REVIEW

In this section we shall discuss some of the important contributions to the literature on unawareness.[14] All of these papers share a common feature: unawareness is associated with events/facts instead of with objects/things.

In an early paper, Fagin & Halpern (1988) take as their starting point the language of propositional modal logic, and add unawareness modal operators to allow for U-sentences. To construct models that do not preclude U-sentences, they augment the standard Kripke structures[15] with an unawareness function. The unawareness function associates with each state a subset of formulas, listing the facts that the agent is unaware of in that state. They impose no restriction on the unawareness function, so the agent could be aware of a formula but unaware of its negation. They also consider an assumption that awareness is *closed under subformulas*, which rules out this possibility. They provide an axiomatization for their structures analogous to our Theorem 1.

Modica & Rustichini (1999) provide the first treatment of unawareness in the economics literature that avoids the Dekel et al. (2016) critique (and also address concerns raised in an earlier paper of their own, Modica & Rustichini (1994)). Their models, called *generalized standard models*, distinguish between an objective set of possible worlds and a subjective subset, with the latter used to represent facts that the agent is aware of. Halpern (2001) shows that generalized standard models can be viewed as special cases of those in Fagin & Halpern (1988), with appropriate restrictions on the awareness function. Li (2006) uses a similar technique to model multi-agent unawareness; it should be noted that the extension to multiple agents is not trivial in this context.

Heifetz et al. (2006b) deal with the extension to the multi-agent case in a different way. They work with a partially ordered *set of sets* of possible worlds, where the ordering represents the expressive power of each set. For instance, if there are only two primitive propositions of interest, their structure consists of four sets of sets, with the most expressive one describing situations involving both propositions, two less expressive sets describing situations involving the first and the second propositions respectively, and the least expressive set describing only situations that involve neither. These sets are used to represent

[14] A comprehensive bibliography can be found on Burkhard Schipper's website: `http://www.econ.ucdavis.edu/faculty/schipper/unaw.htm`

[15] A *Kripke structure* is, roughly, a more general version of the partitional information structure, together with a function that specifies which primitive propositions hold in which states.

the awareness of agents. In a companion paper, Heifetz et al. (2006a) provide an axiomatization for their structures.

None of the papers discussed so far allows us to model agents' reasoning about their possible lack of awareness. Two recent papers work with languages that include sentences such as "I am not sure whether or not you are aware of something that I am not". Halpern & Rego (2006) use second-order modal logic, augmenting the language of Fagin & Halpern (1988) by including quantifiers *over formulas*. The resulting language includes formulas such as $\forall x K_i A_j x$, to be read as "agent i knows that agent j is aware of every formula". More closely related to our current paper, Sillari (2006) uses a language that is identical to ours.

One difference between these two papers and ours is that they have very different axioms for the existential quantifier. In particular, their axioms correspond to what logicians call the *possibilitist* interpretation of existence, whereas ours correspond to the *actualist* interpretation. Although both kinds of axioms have their proponents, we believe possibilitist existence is less useful when it comes to constructing economic models. The reason, roughly speaking, is that most possibilitist axiom systems we are aware of come with the *Barcan Formula*,[16] which, when coupled with other axioms familiar to economists, will have undesirable implications. To illustrate this, let's consider what would have happened had we adopted the possibilitist axioms as well. By this, we mean replacing our Axiom **E1** with the Barcan Formula:

BF for any formula α and variables x, $\forall x L_i \alpha \rightarrow L_i \forall x \alpha$ is an axiom.

and replacing our Axiom **E2** with:

E2' for any formula α and variables x and y, $\forall x \alpha \rightarrow \alpha[y/x]$ is an axiom.

The class of structures that is axiomatized by this new axiom system is exactly those object-based unawareness structures where $\mathcal{E}(w) = W$ for every $w \in W$. If we further add Axioms **LA1** and **LA2**—which, as we explained in

[16] It is named after the philosopher and logician Ruth Barcan Marcus, the founder of first-order modal logic.

Section 5, appeal to many economists—then the resulting subclass of $\mathcal{M}$ will have a very undesirable feature. In any structure within this subclass, an agent either knows for sure that there exists something he is unaware of, or knows for sure that there is nothing he is unaware of—but he can never be uncertain. If he were to assign a probability to the event that there exists something he is unaware of, then that probability would have to be either 0 or 1—it could not lie strictly between 0 and 1.

As Sillari (2006) points out, this very problem arises in Halpern & Rego (2006): "The second-order logic of Halpern and Rego also requires the Barcan to be validated, hence does not lend itself to model knowledge as high-probability operators." That is, once Halpern & Rego (2006) incorporate axioms analogous to **LA1** and **LA2** into their axiom system, sentences such as "the agent is not sure whether or not there are still things that he is unaware of" or "I am not sure whether or not you are aware of something that I am not" will become contradictory in their resulting structures—they must be false in every world of every structure.

Although Sillari (2006) also adopts the possibilitist axioms for the existential quantifier, his axiom system is an exception in that it does not contain the the Barcan formula, as he has very different axioms for implicit knowledge. His weaker axioms on implicit knowledge lead to a class of structures very different from our OBU structures. Roughly speaking, our OBU structures are generalizations of Kripke structures, which are more familiar to economists; while his structures are generalizations of the neighborhood semantics.

Another, more important, difference between Halpern & Rego (2006) and Sillari (2006) and our current paper lies in the way unawareness is modelled. Both of them use the same approach as Fagin & Halpern (1988) by introducing an unawareness function that assigns to each agent in each possible world a list of those formulas that agent is unaware of—we call this the *semi-syntactic approach*. In our object-based approach, on the other hand, we provide a foundation for awareness of formulas in terms of awareness of objects.

We believe the object-based approach offers an advantage over the semi-syntactic approach. Logicians like to preserve a clear distinction between the extra-linguistic reality (which we can think of in our structure as W, D, $\mathcal{E}$, $\mathcal{P}_i$, and $\mathcal{A}_i$) as the *semantics* (the truth-value assignment rule) which maps the language into this reality. This distinction is cut by the semi-syntactic approach, which explicitly uses the language to represent part of the reality (specifically, the awareness of the agents). Why does this matter? In the semi-syntactic

approach, any restrictions that are imposed on the awareness function (Halpern & Rego (2006) consider several) must of course be expressed linguistically, and correspond closely to equivalent axioms in the axiom system. But in the object-based approach, the (non-linguistic) awareness function and the assumptions we make about it look very different from the corresponding axioms governing the behavior of the awareness operator: this gives us two different viewpoints from which to assess the reasonableness of our underlying model of awareness.

At the risk of setting up a strawman, consider as an analogy two different ways of modelling knowledge: first, the standard approach, where we have a set of possible worlds, and a possibility correspondence for each agent describing, in each world, which worlds the agent consider possible; second, a semi-syntactic approach, where instead of the possibility correspondence we have a knowledge function which simply lists the set of formulas each agent knows in each world. Just as various assumptions about the possibility correspondence (that it is reflexive, transitive, etc.) correspond to various axiomatizations of the properties of knowledge (in some appropriate language), restrictions could be imposed on the knowledge function to derive similar equivalence results. But it is clear that the standard approach offers us two distinct perspectives on the concept of knowledge, and potentially a better understanding of it, while the semi-syntactic approach offers only one.

Finally, we should also mention the contribution of Feinberg (2004), who adopts an ingenious meta-approach to the problem: instead of attempting to express unawareness directly within the formal language, he describes unawareness implicitly by describing which subsets of the language make up each agent's subjective world view.

7. CONCLUSION

In this paper we have proposed a user-friendly model that allows us to express sentences such as "the agent is not sure whether or not there are still things that he is unaware of". Instead of trying to assign truth values to these sentences within existing unawareness models in the literature, and worrying about whether or not the truth-value-assignment rule is consistent with some set of "reasonable" axioms, we started with an explicit list of axioms, and then constructed the class of structures (together with a truth-value-assignment rule) that is axiomatized by exactly those axioms. As an application, we explained

how our structures can be used to model the actions of those American founding fathers who were opposed to the inclusion of the Bill of Rights into the constitution.

Appendix A:

In this appendix we shall prove Theorem 1. Throughout this proof, **PC** and **MP** will be used too often for us to acknowledge every time. Hence we shall often refrain from citing their names when we use them.

As usual in this literature, the proof involves two steps: the soundness part and the completeness part. The soundness part says that all provable formulas of *AWARE* are valid in $\mathcal{M}$. The completeness part says the converse is also true.

Lemma 3. *Every provable formula $\alpha \in \mathcal{L}$ of AWARE is valid in $\mathcal{M}$.*

Proof. We shall prove that each axiom is valid in every $M \in \mathcal{M}$, at every w, and under every V; and that each inference rule preserves validity. We shall, however, skip the parts of **PC** and **MP**.

For **E1**, notice that for every x-alternate V' of V such that $V'(x) \in D_w$, we have $(M, w, V') \models E(x)$, which implies $(M, w, V) \models \forall x E(x)$.

For **E2**, suppose $(M, w, V) \models \forall x\alpha$ and $(M, w, V) \models E(y)$ but $(M, w, V) \not\models \alpha[y/x]$. Let V' be the x-alternative of V such that $V'(x) = V(y)$. Then we have both $(M, w, V') \not\models \alpha$ and $V'(x) \in D_w$, contradicting that $(M, wV) \models \forall x\alpha$.

For **E3**, suppose $(M, w, V) \models \forall x(\alpha \to \beta)$ and $(M, w, V) \models \forall x\alpha$. Then, for any x-alternative V' of V such that $V'(x) \in D_w$, we have both $(M, w, V') \models \alpha \to \beta$ and $(M, w, V') \models \alpha$, which implies $(M, w, V') \models \beta$, which in turn implies $(M, w, V) \models \forall x\beta$.

For **E4**, notice that if x is not free in α, then $(M, w, V) \models \alpha$ iff $(M, w, V') \models \alpha$ for any x-alternative V' of V.

For **UG**, suppose $(M, w, V) \not\models \forall x\alpha$. Then there exists some x-alternative V' of V such that $V'(x) \in D_w$ and $(M, w, V') \not\models \alpha$, which implies that the formula α is not valid in $\mathcal{M}$.

For **K**, it follows directly from the truth-value-assignment rule in Section 2.3.

For **A1**, notice that if α contains no free variables, then $V(x) \in \mathcal{A}(w)$ for every x free in α, and hence we have $(M, w, V) \models A_i\alpha$.

For **A2**, suppose $(M, w, V) \models A_i\alpha$ and $(M, w, V) \models A_i\beta$. Then $V(x) \in \mathcal{A}(w)$ for every free x in α and every free x in β, and hence also for every free x in $\alpha \wedge \beta$, and hence we have $(M, w, V) \models A_i(\alpha \wedge \beta)$.

For **A3**, suppose $(M, w, V) \models A_i\alpha$. Then $V(x) \in \mathcal{A}(w)$ for every free x in α, and hence also for every free x in β, and hence we have $(M, w, V) \models A_i\beta$.

For **L**, suppose $(M, w, V) \models L_i(\alpha \rightarrow \beta)$ and $(M, w, V) \models L_i\alpha$ but $(M, w, V) \not\models L_i\beta$. Then there exists some $w' \in \mathcal{P}_i(w)$ such that $(M, w', V) \models \alpha \rightarrow \beta$ and $(M, w', V) \models \alpha$ but $(M, w', V) \not\models \beta$, contradicting the truth-value-assignment rule in Section 2.3.

For **LN**, suppose $(M, w, V) \not\models L_i\alpha$. Then there exists some $w' \in \mathcal{P}_i(w)$ such that $(\mathcal{M}, w', V) \not\models \alpha$, which implies the formula α is not valid in $\mathcal{M}$.

For **UGL**, suppose $(M, w, V) \not\models \alpha_1 \rightarrow L_i\big(\alpha_2 \rightarrow \cdots \rightarrow L_i(\alpha_h \rightarrow L_i\forall x\beta)\cdots\big)$. Then there is a sequence $w_1, \ldots, w_{h+1}$ such that $w_1 = w$, $(M, w_k, V) \models \alpha_k$ for $1 \leq k \leq h$, and $(M, w_{h+1}, V) \not\models \forall x\beta$. Moreover, there eixsts some x-alternative V' of V such that $V'(x) \in D_{w_{h+1}}$ and $(M, w_{h+1}, V') \not\models \beta$. Since x is not free in each α_k, we have $(M, w_k, V') \models \alpha_k$ for $1 \leq k \leq h$, which implies $(M, w, V') \not\models \alpha_1 \rightarrow L_i\big(\alpha_2 \rightarrow \cdots \rightarrow L_i(\alpha_h \rightarrow L_i\beta)\cdots\big)$, which in turn implies the formula $\alpha_1 \rightarrow L_i\big(\alpha_2 \rightarrow \cdots \rightarrow L_i(\alpha_h \rightarrow L_i\beta)\cdots\big)$ is not valid in $\mathcal{M}$. □

By the truth-value-assignment rule in Section 2.3, the formula $\alpha \wedge \neg\alpha$ is not valid in $\mathcal{M}$, and hence by Lemma 3 is not a provable formula in *AWARE*. That there are some formulas that are not provable in *AWARE* means that the system is "consistent" in logicians' terminology. More importantly, it implies that it cannot be the case that both α and $\neg\alpha$ are provable formulas of *AWARE*. This observation will be used in subsequent proofs.

As usual, the proof of the completeness part involves the construction of a structure $M \in \mathcal{M}$, called the canonical structure, and a valuation V, such that every formula $\alpha \in \mathcal{L}$ that is valid in M under V is a provable formula of *AWARE*. Completeness then follows from the fact that any formula $\alpha \in \mathcal{L}$ that is valid in $\mathcal{M}$ must also be valid in M under V.

We say that a formula $\alpha \in \mathcal{L}$ is *AWARE-consistent* if $\neg\alpha$ is not a provable formula of *AWARE*. We say that a finite list of formulas $\{\alpha_1, \ldots, \alpha_k\} \subset \mathcal{L}$ is *AWARE*-consistent if the formula $\alpha_1 \wedge \ldots \wedge \alpha_k$ is *AWARE*-consistent. We say that an infinite list of formulas is *AWARE*-consistent if every finite sublist of it is *AWARE*-consistent.

We say that a list of formulas is *maximal* if, for every formula $\alpha \in \mathcal{L}$, either α or $\neg\alpha$ is in the list. We say that a list of formulas is *maximal*

AWARE-consistent if it is both maximal and *AWARE*-consistent.

It is a standard result that if α is a provable formula of *AWARE*, then it is in every maximal *AWARE*-consistent list.

We say that a list Γ of formulas possesses the $L\forall$-property if

1. for every formula α and variable x, there is some variables y such that the formula $E(y) \wedge (\alpha[y/x] \rightarrow \forall x\alpha)$ is in Γ;

2. for any formulas $\beta_1, \ldots, \beta_h$ ($h \geq 0$) and α, and every variable x that is not free in $\beta_1, \ldots, \beta_h$, there is some variable y such that the formula $L_i\Big(\beta_1 \rightarrow \cdots \rightarrow L_i\Big(\beta_h \rightarrow L_i\big(E(y) \rightarrow \alpha[y/x]\big)\Big)\cdots\Big) \rightarrow L_i\big(\beta_1 \rightarrow \cdots \rightarrow L_i(\beta_h \rightarrow L_i\forall x\alpha)\cdots\big)$ is in Γ.

Lemma 4. *If formula α is AWARE-consistent, then there is an AWARE-consistent list Γ of formulas with the $L\forall$-property such that $\alpha \in \Gamma$.*

To prove Lemma 4, we need another lemma first.

Lemma 5. *The formula $\exists y(\theta[y/x] \rightarrow \forall x\theta)$ is a provable formula of AWARE.*

Proof. By **E2**, the formula $\big(E(x) \wedge \forall y\theta[y/x]\big) \rightarrow (\theta[y/x])[x/y]$ is a provable formula. Notice that $(\theta[y/x])[x/y]$ gives us back θ. Therefore, by **UG** and **E3**, the formula $\forall xE(x) \rightarrow \forall x(\forall y\theta[y/x] \rightarrow \theta)$ is a provable formula. By **E1** and **E3**, the formula $\forall x\forall y\theta[y/x] \rightarrow \forall x\theta$ is a provable formula. But x is not free in $\forall y\theta[y/x]$ anymore, and hence by **E4**, the formula

$$\forall y\theta[y/x] \rightarrow \forall x\theta \tag{1}$$

is a provable formula.

Given (1), it suffices to prove that the formula

$$\forall y\neg(\theta[y/x] \rightarrow \forall x\theta) \rightarrow \neg(\forall y\theta[y/x] \rightarrow \forall x\theta) \tag{2}$$

is a provable formula.

By **PC**, both formulas $\neg(\theta[y/x] \rightarrow \forall x\theta) \rightarrow \theta[y/x]$ and $\neg(\theta[y/x] \rightarrow \forall x\theta) \rightarrow \neg\forall x\theta$ are provable formulas. By **UG** and **E3**, both formulas $\forall y\neg(\theta[y/x] \rightarrow \forall x\theta) \rightarrow \forall y\theta[y/x]$ and $\forall y\neg(\theta[y/x] \rightarrow \forall x\theta) \rightarrow \forall y\neg\forall x\theta$ are also provable formulas. Since y is not free in $\neg\forall x\theta$, by **E4**, we have (2) as needed. □

Proof of Lemma 4: Assume that all variable x are enumerated, and similarly for all formulas of the form $\forall x\theta$, and similarly for all formulas of the form $L_i\big(\xi_i \to \cdots \to L_i(\xi_h \to L_i\forall x\theta)\cdots\big)$ with $h \geq 0$ and x not free in $\xi_1, \ldots, \xi_h$.

Define a sequence of lists of formulas $\Gamma_0, \Gamma_1, \ldots$ as follows: $\Gamma_0 = \{\alpha\}$. Given Γ_n, we define Γ_{n+1} in two steps.

Step 1: We first extend Γ_n to Γ_n^+. Let $\forall x\theta$ be the $n+1$st formula of this form. Let y be the first variable that does not appear in Γ_n and θ, and define

$$\Gamma_n^+ = \Gamma_n \cup \{E(y), \theta[y/x] \to \forall x\theta\}.$$

We claim that, as long as Γ_n is *AWARE*-consistent, Γ_n^+ is *AWARE*-consistent as well. Suppose not. Then the formula $\beta \to \big(E(y) \to \neg(\theta[y/x] \to \forall x\theta)\big)$, where β denote the (finite) conjunction of all formulas in Γ_n, is a provable formula. By **UG**, **E3**, and **E4** (applicable because y does not occur in β), the formula $\beta \to \big(\forall y E(y) \to \forall y\neg(\theta[y/x] \to \forall x\theta)\big)$ is a provable formula. By **E1**, the formula $\beta \to \forall y\neg(\theta[y/x] \to \forall x\theta)$ is a provable formula. By Lemma 5, the formula $\neg\beta$ is a "theroem," contradicting the presumption that Γ_n is *AWARE*-consistent.

Step 2: We next extend Γ_n^+ to Γ_{n+1}. Let $L_i\big(\xi_i \to \cdots \to L_i(\xi_h \to L_i\forall x\theta)\cdots\big)$ be the $n+1$st formula of this form. Let y be the first variable that does not appear in Γ_n^+ and $\xi_1, \ldots, \xi_h$, and define $\Gamma_{n+1} = \Gamma_n^+ \cup \{L_i\Big(\xi_1 \to \cdots \to L_i\Big(\xi_h \to L_i\big(E(y) \to \theta[y/x]\big)\Big)\cdots\Big) \to L_i\big(\xi_1 \to \cdots \to L_i(\xi_h \to L_i\forall x\theta)\cdots\big)\}$.

We claim that, as long as Γ_n^+ is *AWARE*-consistent, Γ_{n+1} is *AWARE*-consistent as well. Suppose not. Then both formulas

$$\beta \to L_i\Big(\xi_1 \to \cdots \to L_i\Big(\xi_h \to L_i\big(E(y) \to \theta[y/x]\big)\Big)\cdots\Big) \tag{3}$$

and

$$\beta \to \neg L_i\big(\xi_1 \to \cdots \to L_i(\xi_h \to L_i\forall x\theta)\cdots\big), \tag{4}$$

where β denotes the (finite) conjunction of all formulas in Γ_n^+, are provable formulas.

Since y does not appear in Γ_n^+, by **UGL** (putting $n = h + 1$), from (3) we infer that the formula

$$\beta \rightarrow L_i\Big(\xi_1 \rightarrow \cdots \rightarrow L_i\Big(\xi_h \rightarrow L_i\forall y\big(E(y) \rightarrow \theta[y/x]\big)\Big)\cdots\Big) \tag{5}$$

is a provable formula.

By **E3**, **LN**, and **L**, the formula $L_i\forall y\big(E(y) \rightarrow \theta[y/x]\big) \rightarrow \big(L_i\forall y E(y) \rightarrow L_i\forall y\theta[y/x]\big)$ is a provable formula. Since by **E1** and **LN**, the formula $L_i\forall y E(y)$ is also a provable formula, we infer that the formula $L_i\forall y\big(E(y) \rightarrow \theta[y/x]\big) \rightarrow L_i\forall y\theta[y/x]$ is a provable formula. By using **LN** and **L** repeatedly (for h times to be exact), we infer that the formula $L_i\Big(\xi_1 \rightarrow \cdots \rightarrow L_i\Big(\xi_h \rightarrow L_i\forall y\big(E(y) \rightarrow \theta[y/x]\big)\Big)\cdots\Big) \rightarrow L_i\big(\xi_1 \rightarrow \cdots \rightarrow L_i(\xi_h \rightarrow L_i\forall y\theta[y/x])\cdots\big)$ is a provable formula. From this, together with (5), we infer that $\beta \rightarrow L_i\big(\xi_1 \rightarrow \cdots \rightarrow L_i(\xi_h \rightarrow L_i\forall y\theta[y/x])\cdots\big)$ is a provable formula. From this, together with (4), we infer that $\neg\beta$ is a provable formula, contradicting the presumption that Γ_n^+ is *AWARE*-consistent.

We can now let Γ be the union of all Γ_n's. Since Γ_0 is *AWARE*-consistent, Γ is also *AWARE*-consistent. And Γ will have the $L\forall$-property by construction. □

Lemma 6. *If an AWARE-consistent list Γ of formulas possesses the $L\forall$-property, then there is a maximal AWARE-consistent list Δ of formulas with the $L\forall$-property such that $\Gamma \subseteq \Delta$.*

Proof. Assume all the formulas in $\mathcal{L}$ are enumerated. Define a sequence of lists of formulas $\Delta_0, \Delta_1, \ldots$ as follows: $\Delta_0 = \Gamma$. Given Δ_n, let α be the $n+1$st formula in $\mathcal{L}$, and let $\Delta_{n+1} = \Delta_n \cup \{\alpha\}$ if $\Delta_n \cup \{\alpha\}$ is *AWARE*-consistent, or $\Delta_{n+1} = \Delta_n \cup \{\neg\alpha\}$ if not. In either case Δ_{n+1} is *AWARE*-consistent if Δ_n is. We can now let Δ be the union of all Γ_n. □

The construction of the canonical structure is as follows. $\mathcal{W}$ is the set of all maximal *AWARE*-consistent lists of formulas with the $L\forall$-property. D is the set of all variables in $\mathcal{L}$; or equivalently, $D = X$. For every state w, which by construction is a list of formulas,

- D_w is the set of all variables x such that $E(x) \in w$;
- $\mathcal{P}_i(w)$ is the set of all states w' such that $L_i^-(w) \subseteq w'$, where $L_i^-(w)$ is the set of all formulas α such that $L_i\alpha \in w$;

- $\mathcal{A}_i(w)$ is the set of all variables x such that $A_iE(x) \in w$; and
- for every k-ary predicate P, $\pi(w)(P)$ is the set of all k-tuples $(x_1, \ldots, x_k)$ such that $P(x_1, \ldots, x_k) \in w$.

Notice that, for any list $w \in \mathcal{W}$, since w satisfies the part 1 of the $L_i\forall$-property, there must be at least one variable y such that $E(y) \in w$, and hence D_w is non-empty. Therefore the canonical structure is indeed an instance of the object-based unawareness structures.

Lemma 7. *If Γ is a maximal AWARE-consistent list of formulas with the $L\forall$-property, and α is a formula such that $L_i\alpha \notin \Gamma$, then there is a maximal AWARE-consistent list Δ of formulas with the $L\forall$-property such that $L_i^-(\Gamma) \cup \{\neg\alpha\} \subseteq \Delta$.*

Proof. Assume that all variables x are enumerated, and similarly for all formulas of the form $\forall x\theta$, and similarly for all formulas of the form $L_i\big(\xi_1 \to \cdots \to L_i(\xi_h \to L_i\forall x\theta)\cdots\big)$ with $h \geq 0$ and x not free in $x_1, \ldots, x_h$.

Define a sequence of formulas $\delta_0, \delta_1, \ldots$ as follows: δ_0 is $\neg\alpha$. Given δ_n, we define δ_{n+1} in two steps.

Step 1: We first extend δ_n to δ_n^+. Let $\forall x\theta$ be the $n+1$st formula of this form. Let y be the first variable such that $L_i^-(\Gamma) \cup \{\delta_n \wedge E(y) \wedge (\theta[y/x] \to \forall x\theta)\}$ is *AWARE*-consistent, and let δ_n^+ be $\delta_n \wedge E(y) \wedge (\theta[y/x] \to \forall x\theta)$.

We claim that, as long as $L_i^-(\Gamma) \cup \{\delta_n\}$ is *AWARE*-consistent, such a variable y must exist. Suppose not. Then for every variable y there is a finite sublist $\{L_i\beta_1, \ldots, L_i\beta_k\} \subset \Gamma$ such that $(\beta_1 \wedge \ldots \wedge \beta_k) \to \Big(E(y) \to \big(\delta_n \to \neg(\theta[y/x] \to \forall x\theta)\big)\Big)$ is a provable formula of *AWARE*. Therefore, by **LN** and **L**, the formula

$$(L_i\beta_1 \wedge \ldots \wedge L_i\beta_k) \to L_i\Big(E(y) \to \big(\delta_n \to \neg(\theta[y/x] \to \forall x\theta)\big)\Big)$$

is also a provable formula of *AWARE*. Since Γ is maximal *AWARE*-consistent and $L_i\beta_1, \ldots, L_i\beta_k \in \Gamma$, we have $L_i\Big(E(y) \to \big(\delta_n \to \neg(\theta[y/x] \to \forall x\theta)\big)\Big) \in \Gamma$ as well. And this is so for *every* variable y.

Since Γ has the $L\forall$-property, there is a variable y such that the formula

$$L_i\Big(E(y) \to \big(\delta_n \to \neg(\theta[y/x] \to \forall x\theta)\big)\Big) \to L_i\forall z\big(\delta_n \to \neg(\theta[z/x] \to \forall x\theta)\big)$$

is in Γ, where the variable z is chosen so that it does not occur in δ_n or in θ. Since $L_i\Big(E(y) \rightarrow \big(\delta_n \rightarrow \neg(\theta[y/x] \rightarrow \forall x\theta)\big)\Big)$ is in Γ for *every* variable y, the formula

$$L_i\forall z\big(\delta_n \rightarrow \neg(\theta[z/x] \rightarrow \forall x\theta)\big)$$

is in Γ. But z does not occur in δ_n or θ, and so by **E3** and **E4**, the formula

$$L_i\big(\delta_n \rightarrow \forall z\neg(\theta[z/x] \rightarrow \forall x\theta)\big)$$

is also in Γ.

However, by Lemma 5, the formula

$$\exists z(\theta[z/x] \rightarrow \forall x\theta)$$

is a provable formula of *AWARE*. So the formula

$$L_i\neg\delta_n$$

must also be a provable formula of *AWARE*, and hence is in Γ, or equivalently, $\neg\delta_n \in L_i^-(\Gamma)$. And this would make $L_i^-(\Gamma) \cup \{\delta_n\}$ *AWARE*-inconsistent, a contradiction.

Step 2: We next extend δ_n^+ to δ_{n+1}. Let $L_i\big(\xi_1 \rightarrow \cdots \rightarrow L_i(\xi_h \rightarrow L_i\forall x\theta)\cdots\big)$ be the $n+1$st formula of this form. We may assume that x is not free in δ_n^+ or in $\xi_1, \ldots, \xi_h$ since if it is we may choose a bound alphabetic variant of $\forall x\theta$ in which the variable that replaces x is not free in these formulas. Let y be the first variable such that $L_i^-(\Gamma) \cup \{\delta_n^+ \wedge \Big(L_i\Big(\xi_1 \rightarrow \cdots \rightarrow L_i\Big(\xi_h \rightarrow L_i\big(E(y) \rightarrow \theta[y/x]\big)\Big)\cdots\Big) \rightarrow L_i\big(\xi_1 \rightarrow \cdots \rightarrow L_i(\xi_h \rightarrow L_i\forall x\theta)\cdots\big)\Big)\}$ is *AWARE*-consistent, and let δ_{n+1} be $\delta_n^+ \wedge \Big(L_i\Big(\xi_1 \rightarrow \cdots \rightarrow L_i\Big(\xi_h \rightarrow L_i\big(E(y) \rightarrow \theta[y/x]\big)\Big)\cdots\Big) \rightarrow L_i\big(\xi_1 \rightarrow \cdots \rightarrow L_i(\xi_h \rightarrow L_i\forall x\theta)\cdots\big)\Big)$.

We claim that, as long as $L_i^-(\Gamma) \cup \{\delta_n^+\}$ is *AWARE*-consistent, such a variable y must exist. Suppose not. Then for every variable y there is a finite sublist $\{L_i\beta_1, \ldots, L_i\beta_k\} \subset \Gamma$ such that $(\beta_1 \wedge \ldots \wedge \beta_k) \rightarrow \Big(\delta_n^+ \rightarrow$

$\neg\Big(L_i\Big(\xi_1 \to \cdots \to L_i\big(\xi_h \to L_i(E(y) \to \theta[y/x])\big)\cdots\Big) \to L_i\big(\xi_1 \to \cdots \to L_i(\xi_h \to L_i\forall x\theta)\cdots\big)\Big)\Big)$ is a provable formula of *AWARE*. Therefore, both

$$(\beta_1 \wedge \ldots \wedge \beta_k) \to \Big(\delta_n^+ \to L_i\Big(\xi_1 \to \cdots \to L_i\Big(\xi_h \to L_i\big(E(y) \to \theta[y/x]\big)\Big)\cdots\Big)\Big) \tag{6}$$

and

$$(\beta_1 \wedge \ldots \wedge \beta_k) \to \Big(\delta_n^+ \to \neg L_i\big(\xi_1 \to \cdots \to L_i(\xi_h \to L_i\forall x\theta)\cdots\big)\Big) \tag{7}$$

are provable formulas of *AWARE*. From (6), by **LN** and **L**,

$$(L_i\beta_1 \wedge \ldots \wedge L_i\beta_k) \to L_i\Big(\delta_n^+ \to L_i\Big(\xi_1 \to \cdots \to L_i\Big(\xi_h \to L_i\big(E(y) \to \theta[y/x]\big)\Big)\cdots\Big)\Big) \tag{8}$$

is also a provable formula of *AWARE*. Since formulas $L_i\beta_1, \ldots, L_i\beta_k$ are all in Γ, so, from (8), the formula

$$L_i\Big(\delta_n^+ \to L_i\Big(\xi_1 \to \cdots \to L_i\Big(\xi_h \to L_i\big(E(y) \to \theta[y/x]\big)\Big)\cdots\Big)\Big)$$

is also in Γ. And this is true for *every* variable y.

Since Γ has the $L_i\forall$ property, by a similar argument as in Step 1, the formula

$$L_i\Big(\delta_n^+ \to L_i\big(\xi_1 \to \cdots \to L_i(\xi_h \to L_i\forall x\theta)\cdots\big)\Big)$$

is also in Γ, or equivalently, the formula

$$\delta_n^+ \to L_i\big(\xi_1 \to \cdots \to L_i(\xi_h \to L_i\forall x\theta)\cdots\big)$$

is in $L_i^-(\Gamma)$. This, together with (7), would make $L_i^-(\Gamma) \cup \{\delta_n^+\}$ *AWARE*-inconsistent, a contradiction.

Since $L_i^-(\Gamma) \cup \{\delta_0\}$ is *AWARE*-consistent, $L_i^-(\Gamma) \cup \{\delta_n^+\}$ and $L_i^-(\Gamma) \cup \{\delta_{n+1}\}$ are *AWARE*-consistent for all n.

Let Δ^- be the union of $L_i^-(\Gamma)$ and all the δ_n's. Δ^- is *AWARE*-consistent, and by construction, also possess the $L_i\forall$ property. Therefore, by Lemma 6, Δ^- can be extended into a maximal *AWARE*-consistent list Δ with the $L_i\forall$ property such that $L_i^-(\Gamma) \cup \{\neg\alpha\} \in \Delta$. □

Lemma 8. *Let M be the canonical structure, and V be the valuation such that $V(x) = x$ for every variable/object $x \in X = D$. Then, for every maximal AWARE-consistent list $w \in \mathcal{W}$ of formulas with the $L\forall$-property, and for every formula $\alpha \in \mathcal{L}$, $(M, w, V) \models \alpha$ iff $\alpha \in w$.*

Proof. The proof proceeds by induction on the length of the formulas. For any atomic formula of the form $E(x)$, $(M, w, V) \models E(x)$ iff $V(x) \in D_w$ iff $x \in D_w$ iff $E(x) \in w$.

For any other atomic formula of the form $P(x_1, \ldots, x_k)$, $(M, w, V) \models P(x_1, \ldots, x_k)$ iff $\big(V(x_1), \ldots, V(x_k)\big) \in \pi(w)(P)$ iff $(x_1, \ldots, x_k) \in \pi(w)(P)$ iff $P(x_1, \ldots, x_k) \in w$.

For any formula of the form $\neg\alpha$, $(M, w, V) \models \neg\alpha$ iff $(M, w, V) \not\models \alpha$ which, by the induction hypothesis, is true iff $\alpha \notin w$ which, by the maximality of the list w, is true iff $\neg\alpha \in w$.

For any formula of the form $\alpha \wedge \beta$, $(M, w, V) \models \alpha \wedge \beta$ iff $(M, w, V) \models \alpha$ and $(M, w, V) \models \beta$ which, by the induction hypothesis, are true iff $\alpha \in w$ and $\beta \in w$ which, by the maximal *AWARE*-consistency of the list w, are true iff $\alpha \wedge \beta \in w$.

For any formula of the form $\forall x\alpha$, suppose $\forall x\alpha \in w$. Consider any x-alternative V' of V such that $V'(x) = y \in D_w$. Since $y \in D_w$, we have $E(y) \in w$. By **E2** and the maximal *AWARE*-consistency of the list w, we have $\alpha[y/x] \in w$. By the induction hypothesis, we have $(M, w, V) \models \alpha[y/x]$, which in turn implies $(M, w, V') \models \alpha$. Since this is true for every x-alternative V' of V such that $V'(x) \in D_w$, we have $(M, w, V) \models \forall x\alpha$.

Conversely, suppose $\forall x\alpha \notin w$. Since the list w possesses the $L\forall$-property, there is some variable y such that $E(y) \wedge (\alpha[y/x] \rightarrow \forall x\alpha) \in w$. By the maximal *AWARE*-consistency of the list w, we have $E(y) \in w$ (making $y \in D_w$) and $\alpha[y/x] \notin w$. By the induction hypothesis, the latter implies that $(M, w, V) \not\models \alpha[y/x]$, which in turn implies $(M, w, V') \not\models \alpha$, where V' is the x-alternative of V such that $V'(x) = y$. But $V'(x) \in D_w$, and hence we have $(M, w, V) \not\models \forall x\alpha$.

For any formula of the form $A_i\alpha$, let $\{x_1, \ldots, x_k\}$ be the free variables in α. If $k = 0$, then we have $(M, w, V) \models A_i\alpha$ and, by **A1** and the maximal *AWARE*-consistency of the list w, $A_i\alpha \in w$ as well. So let's assume $k \geq 1$. Since $V(x) \in \mathcal{A}_i(w)$ iff $x \in \mathcal{A}_i(w)$ iff $A_iE(x) \in w$, we have $(\mathcal{M}, w, V) \models A_i\alpha$ iff $A_iE(x) \in w$ for every $x \in \{x_1, \ldots, x_k\}$. By **A2**, **A3**, and the maximal *AWARE*-consistency of the list w, we have $A_iE(x) \in w$ for every $x \in \{x_1, \ldots, x_k\}$ iff $A_i\big(E(x_1) \wedge \ldots \wedge E(x_k)\big) \in w$ iff $A_i\alpha \in w$.

For any formula of the form $L_i\alpha$, suppose $L_i\alpha \in w$. Then we have $\alpha \in L_i^-(w)$, which implies $\alpha \in w'$ for every $w' \in \mathcal{P}_i(w)$. By the induction hypothesis, we have $(M, w', V) \models \alpha$ for every $w' \in \mathcal{P}_i(w)$, which implies $(M, w, V) \models L_i\alpha$.

Conversely, suppose $L_i\alpha \notin w$. By Lemma 7, there is an $w' \in \mathcal{W}$ such that $L_i^-(w) \cup \{\neg\alpha\} \subseteq w'$. Since $L_i^-(w) \subseteq w'$, we have $w' \in \mathcal{P}(w)$. Since $\neg\alpha \in w'$, by the induction hypothesis, we have $(M, w', V) \not\models \alpha$. Combining the two, we have $(M, w, V) \not\models L_i\alpha$.

For any formula of the form $K_i\alpha$, $(M, w, V) \models K_i\alpha$ iff $(M, w, V) \models A_i\alpha$ and $(M, w, V) \models L_i\alpha$ which, by the induction hypothesis, are true iff $A_i\alpha \in w$ and $L_i\alpha \in w$ which, by **K** and the maximal *AWARE*-consistency of the list w, are true iff $K_i\alpha \in w$. □

Proof of Theorem 1: That every provable formula of *AWARE* is valid in $\mathcal{M}$ follows from Lemma 3. To prove the converse, suppose formula $\alpha \in \mathcal{L}$ is not a provable formula of *AWARE*. Then $\neg\alpha$ is *AWARE*-consistent, and hence by Lemmas 4 and 6, there exists a maximal *AWARE*-consistent list $w \in \mathcal{W}$ with the $L\forall$-property that contains it. By Lemma 8, $(M, w, V) \models \neg\alpha$. Therefore α is not valid in the canonical structure M under the valuation V. Since the canonical structure is one instance of the object-based unawareness structures, this proves that α is not valid in $\mathcal{M}$. □

References

Aumann, R. (1976). Agreeing to disagree. *Annals of Statistics*, *4*, 1236–9.

Board, O. J., & Chung, K.-S. (2008). Object-based unawareness: Theory and applications. *University of Pittsburgh, Working Paper*.

Chung, K.-S., & Fortnow, L. (2016). Loopholes. *Economic Journal*, *126*, 1774–1797.

Dekel, E., Lipman, B., & Rustichini, A. (2016). Standard state space models preclude unawareness. *Econometrica*, *66*, 159–173.

Fagin, R., & Halpern, J. (1988). Belief, awareness, and limited reasoning. *Artificial Intelligence*, *34*, 39–76.

Feinberg, Y. (2004). Subjective reasoning. *Games with Unawareness Research Paper, Stanford Graduate School of Business.*, *1875*.

Halpern, J. Y. (2001). Alternative semantics for unawareness. *Games and Economic Behavior*, *37*, 321–339.

Halpern, J. Y., & Rego, L. C. (2006). Reasoning about knowledge of unawareness. In *Proceedings of the tenth international conference on principles of knowledge representation and reasoning* (pp. 14–24).

Heifetz, A., Meier, M., & Schipper, B. C. (2006a). A canonical model of interactive unawareness. *The University of California, Davis*, *Working Paper*.

Heifetz, A., Meier, M., & Schipper, B. C. (2006b). Interactive unawareness. *Journal of Economic Theory*, *130*, 78–94.

Hughes, G. E., & Cresswell, M. J. (1996). *A New Introduction to Modal Logic*. Routledge, London.

Li, J. (2006). Informational structures with unawareness. *University of Pennsylvania*, *Working Paper*.

Modica, S., & Rustichini, A. (1994). Awareness and partitional information structures. *Theory and Decision*, *37*, 107–124.

Modica, S., & Rustichini, A. (1999). Unawareness and partitional information structures. *Games and Economic Behavior*, *27*, 265–298.

Sillari, G. (2006). Models of unawareness. In G. Bonanno, W. van der Hoek, & M. Wooldridge (Eds.), *Logic and the Foundations of Game and Decision Theory, Proceedings of the Seventh Conference* (pp. 209–40). Amsterdam University Press.

Tirole, J. (2009). Cognition and incomplete contracts. *American Economic Review*, *99*, 265–294.

Journal of Mechanism and Institution Design
ISSN: 2399-844X (Print), 2399-8458 (Online)
DOI:10.22574/jmid.2021.12.002

AXIOMS CONCERNING UNCERTAIN DISAGREEMENT POINTS IN 2-PERSON BARGAINING PROBLEMS

Youngsub Chun

Seoul National University, Korea
ychun@snu.ac.kr

ABSTRACT

We consider 2-person bargaining situations in which the feasible set is known, but the disagreement point is uncertain. We investigate the implications of various axioms concerning uncertain disagreement points and characterize the family of linear solutions, which includes the egalitarian, lexicographic egalitarian, Nash, and Kalai-Rosenthal solutions. We also show that how the important subfamilies (or members) of this family can be singled out by imposing additional axioms or strengthening the axioms used in the characterizations.

Keywords: Axiomatic approach to bargaining problems, uncertain disagreement point, linear solutions.

JEL Classification Numbers: C71, C78, D70.

1. INTRODUCTION

As formulated by Nash (1950), the bargaining problem consists of a feasible set and a disagreement point, represented in utility space. The agents can achieve any point in the feasible set if they agree on it. Otherwise, they end up at the disagreement point. We are interested in the existence of solutions for such problems which satisfy a certain set of axioms.[1]

This is a revised version of my paper circulated as "Axioms concerning uncertain disagreement points for 2-person bargaining problems" (Working Paper No.99, University of Rochester) and "The role of uncertain disagreement points in 2-person bargaining" (Discussion Paper No. 87-17, Southern Illinois University). I am grateful to William Thomson, Hans Peters, Shiran Rachmilevitch, and Ching-jen Sun for their comments.

[1] For a survey of the literature, see Thomson (1994, 1998).

In the traditional formulation, it is assumed that the disagreement point is fixed. The possibility of a varying disagreement point has also been the subject of a number of articles (Thomson, 1987; Livne, 1986, 1989; Peters, 2010; Bossert, 1994; Rachmilevitch, 2011a, 2011b; Anbarci & Sun, 2013). Moreover, bargaining situations in which the feasible set is known but the disagreement point is uncertain have been studied extensively (Chun, 1989, 1990; Chun & Thomson, 1990a, 1990b, 1990c; Livne, 1988, 1989; Peters & van Damme, 1991). This paper is also focused on bargaining situations in which the feasible set is known, but the disagreement point is uncertain.

Suppose that bargaining takes place today, without precise knowledge of the location of the disagreement point, this uncertainty being resolved tomorrow. Under what conditions can agents reach an agreement today? A minimal requirement is that each agent should be guaranteed at least the minimum of what she receives when the uncertainty is lifted tomorrow. Otherwise, the agent is definitely better off by waiting until tomorrow. We require that all agents should be guaranteed this minimum. This requirement of *disagreement point quasi-concavity* was introduced in Chun & Thomson (1990b) and variants are studied by Chun & Thomson (1990a), Livne (1988), Peters (2010), and Peters & van Damme (1991). The purpose of this paper is to explore the implication of this axiom for 2-person bargaining problems.

We provide characterizations of the family of *linear solutions* studied by Chun (1990). They are defined as follows. Let δ be a function associating with each problem a non-negative direction. Moreover, if another problem with the same feasible set has a disagreement on the line passing through the disagreement point of the first problem in the direction assigned by the function δ, then that problem has the same direction δ. The *linear solution relative to* δ is defined by choosing as a solution outcome of each problem at the maximal feasible point such that the vector of utility gains from the disagreement point is in the direction determined by applying δ to the problem. This family of solutions, which we call the *linear family,* is fairly large, including many well-known solutions such as the egalitarian, lexicographic egalitarian, Nash, and Kalai-Rosenthal solutions.

By imposing disagreement point quasi-concavity in conjunction with the standard conditions of weak Pareto optimality, individual rationality, and continuity (with respect to the disagreement point and the feasible set), we characterize the continuous members of the linear family. Also, by strengthening weak Pareto optimality to Pareto optimality and weakening continuity (to

continuity with respect to the disagreement point), we characterize the Pareto optimal members of the family. Other characterizations of the family can be obtained by using axioms related to disagreement point quasi-concavity. We also show how well-known subfamilies or members of the family can be singled out by imposing additional axioms or strengthening the axioms used in the characterizations.

The paper is organized as follows. Section 2 contains some preliminaries and introduces the basic axioms. Section 3 states our main axiom of disagreement point quasi-concavity and characterizes the linear family. Section 4 discusses axioms related to disagreement point quasi-concavity and establishes alternative characterizations of the linear family. Finally, Section 5 characterizes various subfamilies including the egalitarian, lexicographic egalitarian, Nash, and Kalai-Rosenthal solutions.

2. PRELIMINARIES

Let $N = \{1, 2\}$ be the set of agents. A 2-person bargaining problem, or simply a *problem,* is a pair (S, d), where S is a subset of $\mathbb{R}^2$ and d is a point in S such that

(1) S is convex and closed,

(2) $a_i(S) = \max\{x_i | x = (x_1, x_2) \in S\}$ exists for all $i \in N$,

(3) S is *comprehensive,* i.e., for all $x \in S$ and all $y \in \mathbb{R}^N$, if[2] $y \leq x$, then $y \in S$,

(4) there exists $x \in S$ such that $x > d$.

The set S is the *feasible set*. Each point x of S is a *feasible alternative*. The coordinates $x \in S$ are the von Neumann-Morgenstern utility levels attained by the agents through the choice of some joint action. The point d is the *disagreement point* (or "status quo"). The intended interpretation of (S, d) is as follows: the agents can achieve any point in S if they unanimously agree on

[2] Vector inequalities: given $x, y \in \mathbb{R}^N$, $x \geqq y$ means $x_i \geq y_i$ for all $i \in N$, $x \geq y$ means $x \geqq y$ and $x \neq y$, $x > y$ means $x_i > y_i$ for all $i \in N$.

it. If they do not agree on any point, they end up at d. Let Σ^2 be the class of all problems and Γ^2 be the class of all feasible sets satisfying (1), (2) and (3).

A *solution* is a function $F\colon \Sigma^2 \to \mathbb{R}^2$ such that for all $(S,d) \in \Sigma^2$, $F(S,d) \in S$. The value taken by the solution F when applied to the problem (S,d), $F(S,d)$, is the *solution outcome of* (S,d).

Now we introduce standard axioms in the literature. *Weak Pareto optimality* requires that there should be no feasible alternative at which all agents are strictly better off than at the solution outcome. *Pareto optimality* requires that the solution outcome should exhaust all gains from cooperation.

Weak Pareto optimality (WPO): For all $(S,d) \in \Sigma^2$ and all $x \in \mathbb{R}^2$, if $x > F(S,d)$, then $x \notin S$.

Pareto optimality (PO): For all $(S,d) \in \Sigma^2$ and all $x \in \mathbb{R}^2$, if $x \geq F(S,d)$, then $x \notin S$.

Let $WPO(S) = \{x \in S |$ for all $x' \in \mathbb{R}^2$, $x' > x$ implies $x' \notin S\}$ be the set of *weakly Pareto optimal points of* S. Similarly, let $PO(S) = \{x \in S |$ for all $x' \in \mathbb{R}^2$, $x' \geq x$ implies $x' \notin S\}$ be the set of *Pareto optimal points of* S.

Individual rationality requires that no agent should be worse off at the solutions outcome than at the disagreement point. *Strong individual rationality* requires that all agents should be strictly better off than at the disagreement point.

Individual rationality (IR): For all $(S,d) \in \Sigma^2$, $F(S,d) \geqq d$.

Strong individual rationality (SIR): For all $(S,d) \in \Sigma^2$, $F(S,d) > d$.

Let $IR(S,d) = \{x \in S | x \geqq d\}$ be the set of *individually rational points of* (S,d).

Next are two continuity properties. *d-continuity* (respectively, *S-continuity*) requires that a small change in the disagreement point (respectively, the feasible set) cause only a small change in the solution outcome.

d-continuity (d-CONT): For all sequences $\{(S^k, d^k)\} \subset \Sigma^2$ and all $(S,d) \in \Sigma^2$, if $S^k = S$ for all k and $d^k \to d$, then $F(S^k, d^k) \to F(S,d)$.

In the following, convergence of a sequence of sets is evaluated in the Hausdorff topology.

S-continuity (S-CONT): For all sequences $\{(S^k, d^k)\} \subset \Sigma^2$ and all $(S, d) \in \Sigma^2$, if $S^k \to S$ and $d^k = d$ for all k, then $F(S^k, d^k) \to F(S, d)$.

Our final two axioms require that the solution outcome should dominate or be dominated by a reference point. In *mid-point domination 1*, the reference point is the average of the disagreement point and the maximal feasible utility level of each agent guaranteeing the other agent the utility level at the disagreement point. In *mid-point domination 2*, the reference point is the average of the disagreement point and the maximal feasible utility level of each agent.

Mid-point domination 1 (MPD1): For all $(S, d) \in \Sigma^2$, $F(S, d) \geqq \frac{d+a(S,d)}{2}$, where for all $i \in N$, $a_i(S, d) = \max\{x_i | x \in IR(S, d)\}$.

Mid-point domination 2 (MPD2): For all $(S, d) \in \Sigma^2$, $F(S, d) \geqq \frac{d+a(S)}{2}$ or $F(S, d) \leqq \frac{d+a(S)}{2}$.

The following notation and terminology will be used frequently. Given $x_1, \ldots, x_k \in \mathbb{R}^2$, $comp\{x_1, \ldots, x_k\}$ is the *comprehensive hull* of these points x (the smallest comprehensive set containing them). Given $A \subset \mathbb{R}^2$, $Int(A)$ is the relative interior of A. Δ^{n-1} is the $(n-1)$-dimensional simplex. Given $x \in \mathbb{R}^2$ and $\delta \in \Delta^1$, $\ell(x, \delta)$ is the line passing through x in the direction δ. Finally, given $x, y \in \mathbb{R}^2$ such that $x \neq y$, $\ell(x, y)$ is the line passing through x and y.

3. DISAGREEMENT POINT QUASI-CONCAVITY. THE MAIN CHARACTERIZATION

The main objective of this paper is to explore the implication of the following axiom introduced by Chun & Thomson (1990b).

Disagreement point quasi-concavity (D.Q-CAV): For all $(S^1, d^1), (S^2, d^2) \in \Sigma^2$, all $i \in N$, and all $\alpha \in [0, 1]$. if $S^1 = S^2 = S$, then $F_i(S, \alpha d^1 + (1-\alpha)d^2) \geq \min\{F_i(S, d^1), F_i(S, d^2)\}$.

Note that $(S, \alpha d^1 + (1-\alpha)d^2)$ is a well-defined element of Σ^2.

This axiom can be motivated on the basis of timing of bargaining. To illustrate, consider two agents who will face one of two equally likely problems

(S, d^1) and (S, d^2) tomorrow, having the same feasible set but different disagreement points. The agents have two options: either they wait until tomorrow for the uncertainty to be lifted and solve then whatever problem has come up, or they consider the problem obtained by taking as a disagreement point the average of d^1 and d^2 and solve that problem today. If for some agent i, $F_i(S, \frac{d^1+d^2}{2})$ is smaller than the minimum of $F_i(S, d^1)$ and $F_i(S, d^2)$, then the agent will definitely prefer waiting. For agent i to be persuaded that the problem should be solved today, she should be guaranteed at least the minimum of $F_i(S, d^1)$ and $F_i(S, d^2)$. D.Q-CAV provides this minimum incentive to all agents.

We are interested in the following family of solutions (Chun, 1990), which generalizes the egalitarian, lexicographic egalitarian, Nash, and Kalai-Rosenthal solutions.

Definition Let $\delta : \Sigma^2 \to \Delta^1$ be a function such that for all $S \in \Gamma^2$ and all $y \in Int(S)$, $y \in \ell(d, \delta(S, d))$ implies that $\delta(S, y) = \delta(S, d)$ and is continuous with respect to y. The *linear solution relative to* δ, F^δ, is defined by setting, for all $(S, d) \in \Sigma^2$, $F^\delta(S, d)$ equal to $\ell(d, \delta(S, d)) \cap WPO(S)$.

Note that for F^δ to be well-defined, it should be that for all $S \in \Gamma^2$ and all d^1, $d^2 \in Int(S)$, if $\delta(S, d^1) \neq \delta(S, d^2)$, then $\ell(d^1, \delta(S, d^1)) \cap \ell(d^2, \delta(S, d^2)) \cap Int(S) = \emptyset$.

We now turn to the results. The proof of Lemma 1 is the same as the proof of Lemma 1 in Chun & Thomson (1990b).

Lemma 1. Let F be a solution satisfying WPO, IR, and D.Q-CAV. Also, let $(S, d) \in \Sigma^2$ be such that $F(S, d) \in PO(S)$. Then, for all $x \in [d, F(S, d)[$, $F(S, x) = F(S, d)$.

Proof. First, note that for all $x \in [d, F(S, d)]$, $(S, x) \in \Sigma^2$. Let $x \in [d, F(S, d)]$ be given. Let $\bar{\lambda} \in [0, 1]$ be such that $x = \bar{\lambda} d + (1 - \bar{\lambda}) F(S, d)$, and $\{\lambda^k\} \subset [0, 1]$ be such that $\lambda^k < \bar{\lambda}$ for all k and $\lambda^k \to \bar{\lambda}$. Also, let $x^k = \frac{x - \lambda^k d}{1 - \lambda^k}$ for all k. Then, $(S, x^k) \in \Sigma^2$ for all k. By D.Q-CAV, $F_i(S, x) \geq \min\{F_i(S, x^k), F_i(S, d)\}$ for all i and all k. As $k \to \infty$, $x^k \to F(S, d)$ and since $F(S, d) \in PO(S)$, it follows from IR that $F(S, x^k) \to F(S, d)$. Therefore, $F(S, x) \geqq F(S, d)$. Since $F(S, d) \in PO(S)$, we conclude that $F(S, x) = F(S, d)$. □

Lemma 2. Let F be a solution satisfying WPO, IR, d-CONT, and D.Q-CAV. Also, let $(S,d) \in \Sigma^2$ be such that $F(S,d) \in Int(PO(S))$. Then, for all $x \in \ell(d, F(S,d)) \cap Int(S)$, $F(S,x) = F(S,d)$.

Proof. Let F and $(S,d) \in \Sigma^2$ satisfying the hypothesis of the lemma be given. From Lemma 1, we know that for all $x \in [d, F(S,d)]$, $F(S,x) = F(S,d)$. Now, suppose by way of contradiction that there exists $y \in Int(S)$ such that $d \in [y, F(S,d)]$ and $F(S,y) \neq F(S,d)$. Since $F(S,d) \in Int(PO(S))$, from WPO and d-CONT, we can assume that $F(S,y) \in Int(PO(S))$.

(a) We consider the case when $\ell(d, F(S,d))$ is neither horizontal nor vertical. Suppose that $F_1(S,y) > F_1(S,d)$. Let $z = (y_1, d_2)$.

Claim 1. $F_1(S,z) \leq F_1(S,d)$.
Otherwise, from WPO and d-CONT, there exists $z^1 \in]z,d[$ such that $F_1(S,d) < F_1(S,z^1) \leq F_1(S,z)$ and $F(S,z^1) \in PO(S)$. From Lemma 1, for all $x \in [z^1, F(S,z^1)]$, $F(S,x) = F(S,z^1)$. Since $F_2(S,z^1) \geq z^1_2 = d^2$ by IR, there exists $\bar{x} \in [d, F(S,d)] \cap [z^1, F(S,z^1)]$ which is a contradiction.

Claim 2. $F_1(S,y) > F_1(S,d)$ is impossible.
Since $F_1(S,z) \leq F_1(S,d)$, by d-CONT, there exists $z^2 \in [z,y]$ such that $F(S,z^2) = F(S,d)$. From Lemma 1, for all $x \in [z^2, F(S,d)]$, $F(S,x) = F(S,d)$. Also, from WPO, IR and Lemma 1, for all $x \in [z^2, d]$, $F(S,x) = F(S,d)$. Now define the sequence of problems $\{(S,z^k)\}$ by setting $z^{k+1} = \frac{1}{2}(z^k + y)$ for all $k \geq 2$. Also, for all $k \geq 3$, let ℓ^k be the line passing through z^k and d, and $a^k = \ell^k \cap WPO(S)$. For all $x \in]z^2, z^3]$, if $F_1(S,d) < F_1(S,x) \leq \min\{F_1(S,y), a^3_1\}$, then there exists z' such that $z' \in \ell(x, F(S,x)) \cap \ell(z^2, d)$. Since we assumed that $F(S,x) \neq F(S,d)$, this is impossible. Therefore, for all $x \in]z^2, z^3]$, we have $F_1(S,x) \leq F_1(S,d)$ or $F_1(S,x) > \min\{F_1(S,y), a^3_1\}$. By d-CONT, we have $F_1(S,x) \leq F_1(S,d)$ for all $x \in [z^2, z^3]$. Repeating the same procedure, for all $x \in [z^2, y[$, we obtain $F_1(S,x) \leq F_1(S,d)$. Therefore, $F_1(S,y) > F_1(S,d)$ contradicts d-CONT.

By a similar argument, we obtain a contradiction to $F_1(S,y) < F_1(S,d)$.

(b) Now suppose that $\ell(d, F(S,d))$ is horizontal and there exists $y \in Int(S)$ such that $d \in [y, F(S,d)]$ and $F(S,y) \neq F(S,d)$. By IR and d-CONT, there exists $z^1 \in [y,d[$ such that $F(S,z^1) \neq F(S,d)$ and $\ell(z^1, F(S,z^1))$ is positively sloped. From (a), for all $z \in \ell(z^1, F(S,z^1)) \cap Int(S)$, $F(S,z) = F(S,z^1)$.

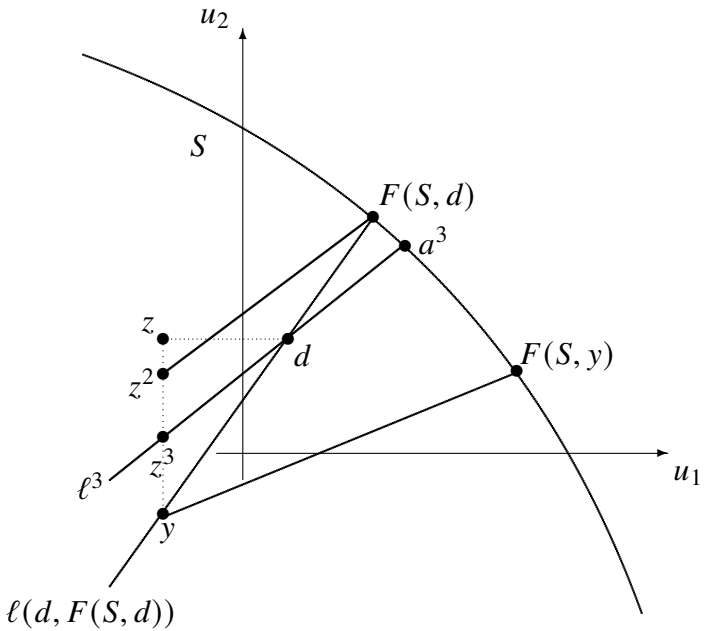


Figure 1: Proof of Claim 2 in Lemma 2.

Now let a^* be the Pareto optimal point of S on the line passing through d parallel to $\ell(z^1, F(S,z^1))$. For some $z \in [z^1, d]$, say z^2, if $\ell(z^2, F(S,z^2))$ is flatter than $\ell(z^1, F(S,z^1))$, then there exists $z' \in Int(S)$ such that $z' \in \ell(z^1, F(S,z^1)) \cap \ell(z^2, F(S,z^2))$, which is impossible. Therefore, for all $z \in [z^1, d]$, $F_1(S,z) < a_1^*$. This is incompatible with d-CONT. A similar argument can be established when $\ell(d, F(S,d))$ is vertical. □

Remark 3. Lemma 1 can easily be generalized to n-person problems. However, it remains an open question whether Lemma 2 can be generalized to such problems.

Now we present our main results.

Theorem 4. A solution satisfies PO, IR, d-CONT, and D.Q-CAV if and only if it is a linear solution F^δ with the additional property that for all $(S,d) \in \Sigma^2$, $\ell(d, \delta(S,d)) \cap WPO(S) \backslash PO(S) = \emptyset$.

Proof. It is obvious that all F^δ satisfy IR, d-CONT and D.Q-CAV, and if δ satisfies the additional property, PO. Conversely, let F be a solution satisfying the four axioms. For all $(S,d) \in \Sigma^2$, let $\delta(S,d) = \frac{F(S,d)-d}{||F(S,d)-d||}$. Since PO and IR together imply that $F(S,d) \geq d$, δ is a well-defined function from Σ^2 to Δ^1. It is enough to show that for all (S^1,d^1), $(S^2,d^2) \in \Sigma^2$, if $S^1 = S^2 = S$ and $d^2 \in \ell(d^1, F(S,d^1))$, then $\delta(S,d^2) = \delta(S,d^1)$. If $F(S,d^1) \in Int(PO(S))$, then the desired conclusion follows from Lemma 2. Suppose now that $F(S,d^1) \notin Int(PO(S))$ and that $\delta(S,d^1) \neq \delta(S,d^2)$. From Lemma 1, for all $d \in [d^1, F(S,d^1)]$, $F(S,d) = F(S,d^1)$ and for all $d \in [d^2, F(S,d^2)]$, $F(S,d) = F(S,d^2)$. By PO and d-CONT, there exists $d' \in [d^1,d^2]$ such that $F(S,d') \in Int(PO(S))$, $F(S,d') \neq F(S,d^2)$ and that either $\ell(d', F(S,d')) \cap [d^1, F(S,d^1)] \neq \emptyset$ or $\ell(d', F(S,d')) \cap [d^2, F(S,d^2)] \neq \emptyset$. Since $F(S,d') \neq F(S,d^1)$ and $F(S,d') \neq F(S,d^2)$, it is a contradiction.

Finally, we note that PO implies that for all $(S,d) \in \Sigma^2$, $\ell(d, \delta(S,d)) \cap WPO(S) \backslash PO(S) = \emptyset$. □

Remark 5. The family of solutions characterized in Theorem 4 is fairly large, including the lexicographic egalitarian, Nash, and Kalai-Rosenthal solutions. However, the egalitarian solution is excluded, since it violates PO.

Theorem 6. A solution satisfies WPO, IR, d-CONT, S-CONT, and D.Q-CAV if and only if it is a linear solution F^δ with the additional property that δ is continuous with respect to S.

Proof. It is obvious that all F^δ satisfy WPO, IR, d-CONT, and D.Q-CAV, and if δ is continuous with respect to S, S-CONT. Conversely, let F be a solution satisfying the five axioms. For all $(S,d) \in \Sigma^2$, let $\delta(S,d) = \frac{F(S,d)-d}{||F(S,d)-d||}$. Since WPO and IR together imply that $F(S,d) \geq d$, δ is a well-defined function from Σ^2 to Δ^1. It is enough to show that for all $(S,d) \in \Sigma^2$, if there exists $d' \in \ell(d, F(S,d)) \cap Int(S)$, then $\delta(S,d') = \delta(S,d)$. If $F(S,d) \in Int(PO(S))$, then the desired conclusion follows from Lemma 2. Otherwise, let $\{(S^k,d)\} \subset \Sigma^2$ be a sequence of problems such that for all k, $F(S^k,d) \in Int(PO(S^k))$ and $d \in Int(S^k)$ and that $S^k \to S$. By the previous argument, $F(S^k,d) = F^\delta(S^k,d)$ for all k, and by S-CONT, $F(S,d) = F^\delta(S,d)$.

Finally, we note that S-CONT implies the continuity of $\delta(\cdot,x)$ with respect to S in the Hausdorff topology. □

Remark 7. The family of solutions characterized in Theorem 6 includes the egalitarian, Nash, and Kalai-Rosenthal solutions. However, the lexicographic egalitarian solution is excluded, since it violates S-CONT.

4. VARIANTS OF THE MAIN RESULT

Bargaining situations in which the feasible set is known but the disagreement point is uncertain have been studied extensively in the literature. Moreover, several axioms related to D.Q-CAV have appeared. Here we discuss how the linear family can be characterized using these axioms.

The first axiom, which we call *weak disagreement point linearity,* was introduced by Livne (1988) in his study of the Nash solution.[3]

Weak disagreement point linearity (W.D.LIN): For all $(S^1, d^1), (S^2, d^2) \in \Sigma^2$ and all $\alpha \in [0,1]$, if $S^1 = S^2 = S$ and $F(S, d^1) = F(S, d^2) = x$, then $F(S, \alpha d^1 + (1-\alpha)d^2) = x$.

This axiom can be interpreted on the basis of timing of bargaining. To see this, consider two agents who will face one of two equally likely problems (S, d^1) and (S, d^2) tomorrow, having the same feasible set, but different disagreement points. Suppose that the solution outcome of the two problems coincides. Since all agents receive the same amount tomorrow irrespective of the uncertainty, it is natural to require that they should receive the same amount when the problem is solved today by taking as a disagreement point the average of d^1 and d^2. W.D.LIN provides such an incentive to agents.

Now we explore the implication of this axiom for 2-person bargaining problems. By replacing D.Q-CAV by W.D.LIN in Theorems 4 and 6, we obtain the same conclusions. In addition, by using the following weak condition, a characterization of the linear family can be established.

Boundary (BOUND): For all sequences $\{(S^k, d^k)\} \subset \Sigma^2$ and all $(S, d) \in \Sigma^2$, if $S^k = S$ for all k, $F(S, d) = x$ and $d^k \to x$, then $F(S^k, d^k) \to x$.

For a solution satisfying Pareto optimality, BOUND is just a considerable weakening of IR. For a solution satisfying only WPO, BOUND is a continuity property which requires that if the disagreement point is closer to the boundary of the feasible set, then the solution outcome is also closer to the disagreement point. It is a weak condition satisfied by all well-known solutions.

Now we have the following result.

[3] Under the name of *independence of convex combination of equivalent conflict outcomes*.

Lemma 8. Let F be a solution satisfying WPO, IR, d-CONT, BOUND, and W.D.LIN. Also, let $(S,d) \in \Sigma^2$ be given. Then, for all $x \in [d, F(S,d)[$, $F(S,x) = F(S,d)$.

Proof. First, note that for all $x \in [d, F(S,d)[$, $(S,x) \in \Sigma^2$. We assume that $WPO(S)$ contains a vertical segment. The case when $WPO(S)$ contains a horizontal (or both a vertical and horizontal) segment can be dealt with similarly. Now suppose by way of contradiction that there exists $d^1 \in [d, F(S,d)]$ such that $F(S,d^1) \neq F(S,d)$. Two cases are possible:

(i) $F_2(S,d^1) > F_2(S,d)$.

Note that if $F_2(S,d^1) > F_2(S,d)$, IR implies that $\ell(d, F(S,d))$ is not vertical. Let $d^2 \in Int(S)$ be such that $d_1^2 = d_1$ and that for all $a \in IR(S,d^2)$, $a_2 > F_2(S,d^1)$. By WPO, $F(S,d^2) \in WPO(S)$ and by IR, $F_2(S,d^2) > F_2(S,d^1)$. By d-CONT, there exists $d^3 \in [d^2,d]$ such that $F(S,d^3) = F(S,d^1)$. By W.D.LIN, for all $d' \in [d^1,d^3]$, $F(S,d') = F(S,d^1)$.

Now let $d(\lambda)$ be a parametrization of $[d^1, F(S,d)]$ such that $d(0) = d^1$ and $d(1) = F(S,d)$. By d-CONT, $F(S,d(\lambda))$ moves continuously. By BOUND, there exists $\bar{\lambda} \in [0,1]$ such that $F_2(S,d^1) > F_2(S,d(\bar{\lambda})) \geq F_2(S,d)$. Let $d(\bar{\lambda}) = d^4$. Also, by d-CONT, there exists $d^5 \in [d^3,d]$ such that $F(S,d^5) = F(S,d^4)$. By W.D.LIN, for all $d' \in [d^4,d^5]$, $F(S,d') = F(S,d^4)$. Then, $[d^1,d^3]$ and $[d^4,d^5]$ intersect. Let d^6 be the intersection point. Clearly, $d^6 \in Int(S)$. Since $F(S,d^1) \neq F(S,d^4)$, it is a contradiction.

(ii) $F_2(S,d^1) < F_2(S,d)$.

From the same argument as in (i), for all $d' \in [d^1, F(S,d^1)[$, $F_2(S,d') \leq F_2(S,d^1)$. Let d^2 be a point in $]d^1, F(S,d^1)[$.

Let $d(\lambda)$ be a parametrization of $[d^2, F(S,d)]$ such that $d(0) = d^2$ and $d(1) = F(S,d)$. By d-CONT, $F(S,d(\lambda))$ moves continuously. By BOUND, there exists $\bar{\lambda} \in [0,1[$ such that $F_2(S,d) \geq F_2(S,d(\bar{\lambda})) > F_2(S,d^1)$. Let $d(\bar{\lambda}) = d^3$. Also, by d-CONT, there exists $d^4 \in [d^2,d]$ such that $F(S,d^4) = F(S,d^3)$. By W.D.LIN, for all $d' \in [d^3,d^4]$, $F(S,d') = F(S,d^3)$. Then, $[d^1, F(S,d^1)]$ and $[d^3,d^4]$ intersect. Let d^5 be the intersection point. Clearly, $d^5 \in Int(S)$. Since $d^5 \in [d^1, F(S,d^1)]$, $F_2(S,d^5) \leq F_2(S,d^1)$, and since $d^5 \in [d^3,d^4]$, $F_2(S,d^5) = F_2(S,d^3) > F_2(S,d^1)$. This is a contradiction. □

Theorem 9. A solution satisfies WPO, IR, d-CONT, BOUND, and W.D.LIN if and only if it is a linear solution.

Proof. It is obvious that all F^δ satisfy the five axioms. Conversely, let F be a solution satisfying the five axioms. First, we know from Lemma 8 that for all $(S,d) \in \Sigma^2$, and all $x \in [d, F(S,d)]$, $F(S,x) = F(S,d)$. It remains to extend the conclusion of Lemma 8 to all $x \in \ell(d, F(S,d)) \cap Int(S)$. Since the proof is similar to that of Lemma 2, we omit it. □

The second axiom was introduced by Peters & van Damme (1991) in their study of the Nash solution.[4]

Disagreement point linearity (D.LIN): For all $(S,d) \in \Sigma^2$ and all $\alpha \in [0,1]$, $F(S, \alpha d + (1-\alpha)F(S,d)) = F(S,d)$.

Let (S,d) be given and consider a new problem obtained by taking the same feasible set and a different disagreement point, which is a convex combination of the old disagreement point and its solution outcome. Then, D. LIN, a strengthening of W.D.LIN, requires that the solution outcome be the same in two problems. If we extend our domain of bargaining problems to allow the disagreement point to lie on the boundary of the feasible set and define the solution outcome of such problems be the disagreement point, then the motivation similar to W.D.LIN can be given.

Now we explore the implication of this axiom for 2-person problems. Again, by replacing D.Q-CAV by D.LIN in Theorems 4 and 6, we obtain the same conclusion. In addition, the following theorem can be established.

Theorem 10. A solution satisfies WPO, IR, d-CONT, and D.LIN if and only if it is a linear solution.

Proof. It is obvious that all F^δ satisfy the four axioms. The converse statement can be established by exploiting the logical implications of these axioms. Indeed, it can easily be shown that (i) WPO and D.LIN together imply W.D.LIN, and that (ii) d-CONT and D.LIN together imply BOUND. Therefore, by Theorem 9, we obtain the desired conclusion. □

Remark 11. If IR is dropped from the list in Theorem 10, then the following *generalized linear solutions* are permissible. Let $B^1 = \{x \in \mathbb{R}^2 | \sum |x_i| = 1$ and $-x \notin \mathbb{R}^2_+\}$ and given $x \in \mathbb{R}^2$ and $\delta \in B^1$, let $\bar{\ell}(d,\delta)$ be the line passing through d in the direction δ. Also, given $(S,d) \in \Sigma^2$, let $\bar{\ell}(d,\delta) \cap WPO(S)$

[4] Under the name of *convexity*.

be the weakly Pareto optimal point of S on the half-line passing through d in the direction δ.

Definition Let δ be a function such that, for all $(S,d) \in \Sigma^2$, $\delta(S,d) \in B^1$ and that for all $S \in \Gamma^2$ and all $y \in Int(S)$, $y \in \bar{\ell}(d, \delta(S,d))$ implies that $\delta(S,y) = \delta(S,d)$ and that $\delta(S,\cdot)$ is continuous with respect to d. Given the function δ, the *generalized linear solution relative to* δ is defined by setting for each $(S,d) \in \Sigma^2$, $F^\delta(S,d)$ equal to $\bar{\ell}(d, \delta(S,d)) \cap WPO(S)$.

5. FURTHER CHARACTERIZATIONS

In this section, we discuss how important subfamilies of the linear family can be characterized by imposing additional axioms or strengthening the axioms used in the Theorems 4 and 6.

5.1. The Egalitarian Solution

First, we consider a subfamily of the linear family, which generalizes the well-known egalitarian solution (Kalai, 1977; Thomson & Myerson, 1980).

Definition. Given a continuous function $\delta : \Gamma^2 \to \Delta^1$, the *directional solution relative to* δ, E^δ, is defined by setting for all $(S,d) \in \Sigma^2$, $E^\delta(S,d)$ equal to $\ell(d, \delta(S)) \cap WPO(S)$. Given $\alpha \in \Delta^1$, the weighted egalitarian solution with weights α, E^α, is defined by setting for all $(S,d) \in \Sigma^2$, $E^\alpha(S,d)$ equal to $\ell(d,\alpha) \cap WPO(S)$. The *egalitarian solution* is obtained by choosing $\alpha_1 = \alpha_2$.

This family can be characterized by the following axiom, which strengthens D.Q-CAV.

Disagreement point concavity (D.CAV): For all (S^1,d^1), $(S^2,d^2) \in \Sigma^2$ and all $\alpha \in [0,1]$, if $S^1 = S^2 = S$, then $F(S, \alpha d^1 + (1-\alpha)d^2) \geqq \alpha F(S,d^1) + (1-\alpha)F(S,d^2)$.

This axiom, introduced and studied in Chun & Thomson (1990a), gives an even stronger incentive to all agents to reach an agreement today than D.Q-CAV does. To illustrate, consider two agents who will face one of two equally likely problems (S,d^1) and (S,d^2) tomorrow, having the same feasible set, but different disagreement points. The agents have two options: either they

wait until tomorrow for the uncertainty to be lifted and solve then whatever problem has come up, or they consider the problem obtained by taking as a disagreement point the average of d^1 and d^2 and solve that problem today. The expected payoff associated with the contingent agreements of the first option is $\frac{F(S,d^1)+F(S,d^2)}{2}$ and that associated with the second option is $F(S, \frac{d^1+d^2}{2})$, since $\frac{d^1+d^2}{2}$ is the corresponding "expected" disagreement point. If either $F(S, \frac{d^1+d^2}{2})$ weakly dominates $\frac{F(S,d^1)+F(S,d^2)}{2}$ or the reverse holds, all agents agree on when to do. A conflict may arise if neither of these inequalities holds. If a solution satisfies D.CAV, then all agents agree to solve the problem today.

It can easily be checked that D.CAV implies D.Q-CAV. D.CAV can be regarded as a dual to an axiom considered by Myerson (1981) concerning uncertainty in the feasible set (variants of which are studied by Perles & Maschler (1981), Peters (1986), and Chun (1988)).

The following result, which can be generalized to n-person bargaining problems, is due to Chun & Thomson (1990a). We note that d-CONT is not needed.

Theorem 12. (Chun & Thomson, 1990a) A solution satisfies WPO, IR, S-CONT, and D.CAV if and only if it is a directional solution.

The family of weighted egalitarian solutions can be characterized by strengthening IR to the following axiom.

Independence of Non-Individually Rational Alternatives (INIR): For all $(S, d) \in \Sigma^2$, $F(S, d) = F(comp\{IR(S, d)\}, d)$.

This axiom, introduced by Peters (2010), says that the non-individually rational alternatives are irrelevant to the determination of the solutions outcome. It is a natural condition since agents are guaranteed their utilities at the disagreement point. It can easily be checked that WPO, INIR and S-CONT (or PO and INIR) together imply IR.

Theorem 13. (Chun & Thomson, 1990a) A solution satisfies WPO, INIR, S-CONT, and D.CAV if and only if it is a weighted egalitarian solution.

Alternative characterizations of the weighted egalitarian solutions can be found in Chun & Thomson (1990a, 1990c). In particular, they show that the weighted egalitarian solutions can be characterized by additionally imposing the following axiom of *contraction independence* (Nash, 1950)[5] to Theorem

[5] Under the name of *independence of irrelevant alternatives*.

12.

Contraction independence (CI): For all (S^1, d^1), $(S^2, d^2) \in \Sigma^2$, if $S^2 \subseteq S^1$, $d^2 = d^1$, and $F(S^1, d^1) \in S^2$, then $F(S^2, d^2) = F(S^1, d^1)$.

CI requires that if an alternative has been judged superior to all others in some feasible set, then it should be judged superior to all others in any subset (to which it belongs) provided the disagreement point is kept constant.

Theorem 14. (Chun & Thomson, 1990a, 1990c) A solution satisfies WPO, IR, *S*-CONT, D.CAV, and CI if and only if is a weighted egalitarian solution.

The egalitarian solution is the only weighted egalitarian solution satisfying the following axiom.

Symmetry (SY): For all $(S, d) \in \Sigma^2$ and all permutations $\pi : \{1, 2\} \to \{1, 2\}$, if $S = \pi(S)$ and $d = \pi(d)$, then $F_1(S, d) = F_2(S, d)$.

SY says that if the only information available on the conflict situation is contained in the mathematical description of (S, d), and (S, d) is a symmetric problem, then there is no ground for favoring one agent at the expense of another.

Corollary 15. (Chun & Thomson, 1990a, 1990c) A solution satisfies WPO, INIR, *S*-CONT, D.CAV, and SY (or WPO, IR, *S*-CONT, D.CAV, CI, and SY) if and only if it is the egalitarian solution.

Let $\tilde{\Sigma}^2$ be the class of problems satisfying (1), (2) and (3) which allows the disagreement point to be on the boundary of the problem. On $\tilde{\Sigma}^2$, Bossert & Peters (2021) provide an alternative characterization of the weighted egalitarian solutions by weakening D.CAV to *individual disagreement point concavity,* dropping *S*-CONT, and additionally imposing *translation invariance* and *disagreement point sensitivity*. *Translation invariance* requires that adding a constant to an agent's utility function should change the solution outcome by the constant. *Disagreement point sensitivity* requires that the solution outcome should respond to certain changes in the disagreement point. Finally, *individual disagreement point concavity* requires that the conclusion of D.CAV should hold if there is only one agent whose utility at the disagreement point is uncertain.

Translation invariance (T.INV): For all $(S, d) \in \Sigma^2$ and all $b \in \mathbb{R}^2$, $F(S+b, d+b) = F(S, d) + b$.

Disagreement point sensitivity (D.SEN): For all (S, d), $(S, d') \in \Sigma^2$ such that $d \not\geq d'$, $F(S, d) \neq F(S, d')$.

Individual disagreement point concavity (I.D.CAV): For all (S^1, d^1), $(S^2, d^2) \in \Sigma^2$, all $i \in N$, and all $\alpha \in [0, 1]$, if $S^1 = S^2 = S$ and for all $j \in N\backslash\{i\}$, $d^1_j = d^2_j$, then $F_i(S, \alpha d^1 + (1-\alpha)d^2) \geq \alpha F_i(S, d^1) + (1-\alpha)F_i(S, d^2)$.

Theorem 16. (Bossert & Peters, 2021) A solution satisfies WPO, T.INV, IR, IIA, D.SEN, and I.D.CAV if and only if is a weighted egalitarian solution.

Once again, the egalitarian solution can be characterized by additionally imposing SY to the list appearing in Theorem 16.

5.2. The Lexicographic Egalitarian Solution

The directional solutions often violate PO. The following extension, called the *lexicographic egalitarian solution,*[6] is an adaptation of the egalitarian solution that satisfies PO. On the other hand, this solution does not satisfy S-CONT.

Definition. The lexicographic egalitarian solution, L, is defined by setting, for all $(S, d) \in \Sigma^2$, $L(S, d) = E(S, d)$ if $E(S, d) \in PO(S)$ and $L(S, d) = \{x \in PO(S) | x_1 = E_1(S, d) \text{ or } x_2 = E_2(S, d)\}$, otherwise.

Chun & Thomson (1990a) showed that no solution satisfies PO, IR and D.CAV together. However, the following weakening of D.CAV is compatible with PO and IR (Chun, 1990).

Restricted disagreement point concavity (R.D.CAV): For all (S^1, d^1), $(S^2, d^2) \in \Sigma^2$ and all $\alpha \in [0, 1]$, if $S^1 = S^2 = S$ and $F(S, d^1)$, $F(S, d^2) \in Int(PO(S))$, then $F(S, \alpha d^1 + (1-\alpha)d^2) \geqq \alpha F(S, d^1) + (1-\alpha)F(S, d^2)$.

The motivation for this axiom is same as for D.CAV, except that the conclusion is required to hold for the interior of the Pareto optimal set. For a solution satisfying PO, if it chooses the boundary point of the Pareto optimal set as the solution outcome, then the solution outcome becomes less sensitive to changes

[6] This solution has been studied by Imai (1983) and Thomson & Lensberg (1989).

in the disagreement point. Therefore, it is unreasonable to require that the solution behave well even on the boundary. It can easily be checked that PO, d-CONT and R.D.CAV (or PO, IR and R.D.CAV) together imply D.Q-CAV.

To characterize the lexicographic egalitarian solution, S-CONT is weakened to the following condition.

Pareto-continuity (P-CONT): For all sequences $\{(S^k, d^k)\} \subset \Sigma^2$ and all $(S, d) \in \Sigma^2$, if $S^k \to S$, $PO(S^k) \to PO(S)$, and $d^k = d$ for all k, then $F(S^k, d^k) \to F(S, d)$.

P-CONT requires that a small change in the feasible set and the Pareto optimal set cause only a small change in the solution outcomes. It can easily be checked that S-CONT implies P-CONT.

Now we are ready to present a characterization of the lexicographic egalitarian solution.

Theorem 17. (Chun, 1989) A solution satisfies PO, SY, INIR, d-CONT, P-CONT, and R.D.CAV if and only if it is the lexicographic egalitarian solution.

5.3. The Nash Solution

Now we discuss the best-known solution in the axiomatic bargaining theory, the *Nash solution*. This solution, introduced and characterized by Nash (1950), has been extensively studied in the literature. Properties which describe its behavior with respect to changes in the disagreement point have been investigated by Chun & Thomson (1990b), Peters (2010), and Peters & van Damme (1991).

Definition. Given $\alpha \in Int(\Delta^1)$, the weighted Nash solution with weights α, N^α, is defined by setting for all $(S, d) \in \Sigma^2$, $N^\alpha(S, d)$ to be the maximizer of the product $\Pi(x_i - d_i)^{\alpha_i}$ over $IR(S, d)$. The *Nash solution, N*, is the member of this family obtained by choosing $\alpha_1 = \alpha_2$.

To characterize the Nash solution, we introduce an invariance property. A *positive affine transformation* is a function $\lambda : \mathbb{R}^2 \to \mathbb{R}^2$ given by $a \in \mathbb{R}^2_{++}$ and $b \in \mathbb{R}^2$ such that for all $x \in \mathbb{R}^2$, $\lambda(x) = (a_1x_1 + b_1, a_2x_2 + b_2)$.

Scale invariance (S.INV): For all $(S, d) \in \Sigma^2$ and all positive affine transformations $\lambda : \mathbb{R}^2 \to \mathbb{R}^2$, $F(\lambda(S), \lambda(d)) = \lambda(F(S, d))$.

S.INV can be justified by the fact that agents' utility functions are von Neumann-Morgenstern types, which are unique up to positive affine transformations. Of course, S.INV implies T.INV.

Nash (1950) showed that his solution is the unique solution satisfying PO, SY, CI, and S.INV. Now we establish an alternative characterization of the Nash solution by investigating logical implications of CI and other axioms.

Lemma 18. Let F be a continuous linear solution characterized in Theorem 6. Then, the solution satisfies CI if and only if it satisfies INIR.

Proof. It is clear that for a solution satisfying IR, CI implies INIR. To prove the converse statement, let (S^1, d), $(S^2, d) \in \Sigma^2$ be two problems such that $S^2 \subseteq S^1$ and $F(S^1, d) \in S^2$. Now define the sequence of problems $\{(S^k, d^k)\}$ such that $S^2 \subseteq S^k \subseteq S^1$, $S^k \to S^2$, $d^k \in [d, F(S^1, d)]$ and $IR(S^k, d^k) = IR(S^1, d^k)$ for all k. By INIR, $F(S^k, d^k) = F(S^1, d^k)$ for all k. Since $d^k \in [d, F(S, d)]$ and F belongs to the linear family, $F(S^k, d^k) = F(S^k, d)$ and $F(S^1, d^k) = F(S^1, d)$ for all k. Altogether, we have $F(S^k, d) = F(S^1, d)$ for all k. Since F is continuous, we conclude that $F(S^2, d) = F(S^1, d)$. □

Variants of the following theorem can be found in Chun & Thomson (1990b), Peters (2010), and Peters & van Damme (1991). Note that S-CONT and S.INV together imply d-CONT.

Theorem 19. (Chun & Thomson, 1990b) A solution satisfies PO, INIR, S-CONT, D.Q-CAV, and S.INV if and only if it is a weighted Nash solution.[7]

Remark 20. By dropping S-CONT form Theorem 19, the following solutions are permissible.

Definition. Given i, the i^{th} *benevolent dictatorial solution,* D^i, is defined by setting for all $(S, d) \in \Sigma^2$, $D^i(S, d)$ equal to the point of $IR(S, d) \cap PO(S)$ preferred by agent i.

In fact, we can show that a solution satisfies PO, INIR, d-CONT, D.Q-CAV and S.INV if and only if it is a weighted Nash solution or a benevolent dictatorial solution. Since its proof is similar to that of Theorem 1 in Peters (2010), we omit it.

[7] Note that the Nash solution does not satisfy D.Q-CAV in the bargaining problem with more than 2-agents, as discussed in Chun & Thomson (1990b).

Remark 21. D.Q-CAV in Theorem 19 can be replaced by the following axiom:

Restricted disagreement point linearity (R.D.LIN): For all (S^1, d^1), $(S^2, d^2) \in \Sigma^2$ and all $\alpha \in [0, 1]$, if $S^1 = S^2 = S$, $\alpha F(S, d^1) + (1-\alpha)F(S, d^2) \in PO(S)$, and S is smooth at both $F(S, d^1)$ and $F(S, d^2)$, then $F(S, \alpha d^1 + (1-\alpha)d^2) = \alpha F(S, d^1) + (1-\alpha)F(S, d^2)$.

This result can be generalized to n-person bargaining problems. For details, we refer to Chun & Thomson (1990b).

On the other hand, Peters & van Damme (1991) characterize the weighted Nash solutions by imposing D.LIN.

Theorem 22. (Peters & van Damme, 1991) A solution satisfies SIR, INIR, d-CONT, S.INV, and D.LIN if and only if it is a weighted Nash solution.

It is well-known that the Nash solution is the only weighted Nash solution satisfying the symmetry axiom.

Corollary 23. (Chun & Thomson, 1990b; Peters & van Damme, 1991) A solution satisfies PO, INIR, S-CONT, D.Q-CAV, S,INV and SY (or SIR, INIR, d-CONT, S,INV, D.LIN, and SY) if and only if it is the Nash solution.

An alternative characterization of the Nash solution can be obtained by imposing MPD1.

Theorem 24. (Chun, 1990) A solution satisfies WPO, d-CONT, S-CONT, D.Q-CAV, and MPD1 if and only if it is the Nash solution.

Equivalently, Theorem 24 can be understood as stating that among the continuous linear solutions characterized in Theorem 6, the Nash solution is the only one satisfying MPD1. Also, Peters (2010) shows how the Nash solution can be singled out from the linear family.

5.4. The Kalai-Rosenthal Solution

Finally, we discuss the Kalai & Rosenthal (1978) solution.

Definition. The *Kalai-Rosenthal solution, KR*, is defined by setting, for all $(S, d) \in \Sigma^2$, $KR(S, d)$ be the maximal point of S on the line segment connecting d and $a(S)$, where for each i, $a_i(S) = \max\{x_i | x \in S\}$.

To characterize the Kalai-Rosenthal solution, we introduce two additional axioms. For all $(S,d) \in \Sigma^2$, let $T(S_d) = comp\{(d_1, a_2(S)), (a_1(S), d_2)\}$.

Independence of strongly individually rational outcome (ISIR): For all $(S,d) \in \Sigma^2$ and all $x \in \mathbb{R}^2$, if $S = comp\{IR(S,d)\}$, $x \leq d$ and $F(T(S_d), x) = F(T(S_d), d)$, then $F(S,x) = F(S,d)$.

Strict disagreement point monotonicity (S.D.MON): For all $(S^1, d^1), (S^2, d^2) \in \Sigma^2$ and all i, j such that $i \neq j$, if $S^1 = S^2$, $d^1_i = d^2_i$, $d^1_j < d^2_j$, and $a(S) \notin S$, then $F_j(S^2, d^2) > F_j(S^1, d^1)$.

ISIR, introduced by Peters (2010), is interpreted as a weak form of path independence. S.D.MON, introduced by Livne (1989), requires that if an agent's utility at the disagreement point increases while the other's remains fixed, then the agent should gain strictly. This is a strengthening of the condition introduced by Thomson (1987).

The next theorem follows from Peters (2010).

Theorem 25. A solution satisfies PO, IR, d-CONT, D.Q-CAV, SY, S.INV, ISIR, and S.D.MON if and only if it is the Kalai-Rosenthal solution.

Proof. It is clear that KR satisfies all eight axioms. Conversely, let F be a solution satisfying the eight axioms. From Theorem 4, it is a Pareto-optimal member of the linear solution. Now by borrowing the proof of Theorem 2 in Peters (2010), we can obtain the desired conclusion.[8] □

An alternative characterization of the Kalai-Rosenthal solution can be obtained by imposing MPD2.

Theorem 26. (Chun, 1990) A solution satisfies WPO, d-CONT, S-CONT, D.Q-CAV, and MPD2 if and only if it is the Kalai-Rosenthal solution.

As in Theorem 24, Theorem 26 can be understood as stating that among the continuous linear solutions characterized in Theorem 6, the Kalai-Rosenthal solution is the only one satisfying MPD2. Therefore, it is interesting to note that among the linear solutions, the Nash and Kalai-Rosenthal solutions can be singled out by imposing two different mid-point domination conditions.

[8] It remains to check whether all axioms are independent.

References

Anbarci, N., & Sun, C. (2013). Robustness of intermediate agreements and bargaining solutions. *Games and Economic Behavior*, *77*, 367-376.

Bossert, W. (1994). Disagreement point monotonicity, transfer responsiveness, and the egalitarian bargaining solution. *Social Choice and Welfare*, *11*, 381-392.

Bossert, W., & Peters, H. (2021). Individual disagreement point concavity and the bargaining problem. *International Journal of Economic Theory*, *Forthcoming*.

Chun, Y. (1988). Nash solution and timing of bargaining. *Economics Letters*, *28*, 27-31.

Chun, Y. (1989). Lexicographic egalitarian solution and uncertainty in the disagreement point. *Zeitschrift für Operations Research*, *33*, 259-266.

Chun, Y. (1990). Minimal cooperation in bargaining. *Economics Letters*, *34*, 311-316.

Chun, Y., & Thomson, W. (1990a). Bargaining with uncertain disagreement points. *Econometrica*, *58(4)*, 951-959.

Chun, Y., & Thomson, W. (1990b). Nash solution and uncertain disagreement points. *Games and Economic Behavior*, *2*, 213-223.

Chun, Y., & Thomson, W. (1990c). Egalitarian solutions and uncertain disagreement points. *Economics Letters*, *33*, 29-33.

Imai, H. (1983). Individual monotonicity and lexicographic maxmin solution. *Econometrica*, *51*, 389-401.

Kalai, E. (1977). Proportional solutions to bargaining situations: Interpersonal utility comparisons. *Econometrica*, *45*, 1623-1630.

Kalai, E., & Rosenthal, R. (1978). Arbitration of two-party disputes under ignorance. *International Journal of Game Theory*, *7*, 65-72.

Livne, Z. (1986). The bargaining problem: Axioms concerning changes in the conflict point. *Economics Letters*, *21*, 131-134.

Livne, Z. (1988). The bargaining problem with an uncertain conflict outcome. *Mathematical Social Sciences*, *15*, 287-302.

Livne, Z. (1989). On the status quo sets induced by the Raiffa solution to the two-person bargaining problem. *Mathematics of Operations Research*, *14*, 688-692.

Myerson, R. B. (1981). Utilitarianism, egalitarianism, and the timing effect in social choice problems. *Econometrica*, *49*, 883-897.

Nash, J. F. (1950). The bargaining problem. *Econometrica*, *18*, 155-162.

Perles, M. A., & Maschler, M. (1981). The super-additive solution for the Nash bargaining game. *International Journal of Game Theory*, *10*, 163-193.

Peters, H. (1986). Simultaneity of issues and additivity in bargaining. *Econometrica*, *54*, 153-169.

Peters, H. (2010). Characterizations of bargaining solutions by properties of their

status quo sets. In A. V. Deeman & A. Rusinowska (Eds.), *Collective Decision Making: Views from Social Choice and Welfare* (p. 231-247). Springer.

Peters, H., & van Damme, E. (1991). Characterizing the Nash and Raiffa bargaining solutions by disagreement point axioms. *Mathematics of Operations Research*, *16*, 447-461.

Rachmilevitch, S. (2011a). Disagreement point axioms and the egalitarian bargaining solution. *International Journal of Game Theory*, *40*(1), 63-85.

Rachmilevitch, S. (2011b). A characterization of the Kalai–Smorodinsky bargaining solution by disagreement point monotonicity. *International Journal of Game Theory*, *40*(4), 691-696.

Thomson, W. (1987). Monotonicity of bargaining solutions with respect to the disagreement point. *Journal of Economic Theory*, *42*, 50–58.

Thomson, W. (1994). Cooperative models of bargaining. In R. J. Aumann & S. Hart (Eds.), *Handbook of Game Theory with Economic Applications* (Vol. 2). North-Holland.

Thomson, W. (1998). *Bargaining Theory: The Axiomatic Approach*. (Forthcoming.) Academic Press.

Thomson, W., & Lensberg, T. (1989). *Axiomatic Theory of Bargaining with a Variable Number of Agents*. Cambridge, U.K.: Cambridge University Press.

Thomson, W., & Myerson, R. (1980). Monotonicity and independence axioms. *International Journal of Game Theory*, *9*, 37-49.

Journal of Mechanism and Institution Design
ISSN: 2399-844X (Print), 2399-8458 (Online)
DOI:10.22574/jmid.2021.12.003

A DEFERRED ACCEPTANCE MECHANISM FOR DECENTRALIZED, FAST, AND FAIR CHILDCARE ASSIGNMENT

Tobias Reischmann

University of Münster, Germany
tobias.reischmann@uni-muenster.de

Thilo Klein

Pforzheim University and ZEW – Leibniz-Centre for European Economic Research, Germany
thilo.klein@hs-pforzheim.de

Sven Giegerich

University of Oxford, United Kingdom
sven.giegerich@oii.ox.ac.uk

ABSTRACT

We design a program-proposing deferred acceptance mechanism with ties (DAT) and apply it to childcare assignment in two German cities. The mechanism can accommodate complementarities in providers' preferences, is fast to terminate even in larger cities, is difficult to manipulate in practice, and produces stable allocations. It can be further sped up by introducing two new features. First, allowing for an arbitrary share of facilities who participate in a centralized manner by submitting a rank-order-list over applicants. Second, by breaking ties in applicants' rank-order-lists on a first-come-first-serve basis, which sets incentives for programs to propose faster. We provide and evaluate simulation results.

All authors have contributed equally. We acknowledge helpful comments from Péter Biró, Inácio Bó, Tobias Riehm and audiences at the 14th and 15th Matching-in-Practice Workshops in Cologne and Mannheim. We are thankful to the youth welfare office of the district of Steinfurt for their cooperation in the project. We acknowledge funding from the Leibniz Association as part of project K125/2018: Improving school admissions for diversity and better learning outcomes.

Keywords: Childcare assignment, deferred acceptance algorithm, simulation.

JEL Classification Numbers: C78, D02, D47, D82, I24.

1. INTRODUCTION

In many German cities, the allocation of available childcare placements is not transparent and carries considerable costs for parents and childcare providers alike. While there is a recognized shortage of childcare placements, inefficient allocation procedures have made the shortage seem more acute than it actually is.[1] One possible blanket solution to the problem would be to introduce a central allocation system that relies on tried and tested matching algorithms, thus ensuring a well-designed matching system.

Online platforms already facilitate the registration for childcare placements in Germany.[2] Their scope, however, is often limited to the registration of applications and their forwarding to childcare providers. Very few platforms allow for a coordinated offer process that takes into account both parents' and providers' preferences. Instead, childcare providers send out offers independently, in an uncoordinated fashion. Problems thus arise when (i) parents feel forced to accept an early unattractive offer for the sake of security or (ii) when they temporarily hold and thus block placements for other families in anticipation of a better offer. The first aspect can lead to what we will refer to as an *unfair* admission. The second aspect *slows down* the admission process, resulting in uncertainty on the side of parents and employers.

One solution to these problems would be to establish a "central clearinghouse" for admissions. Such clearinghouses have proven their worth in grade school and university admissions in Germany and other countries worldwide. However, they take time to develop. For example, five years after its creation, the German clearinghouse for university admissions only had a market share of 13%.[3] Furthermore, similar institutions have yet to emerge for childcare markets, a fact that we attribute to small care-group sizes, the large share of private (rather than public) providers, and complementarities in provider preferences.

[1] Several illustrative newspaper articles raise these issues (e.g. Bös, 2017; Völker, 2018).

[2] An example is the 'Kita-Navigator' software, which is widely used in the state of North Rhine-Westphalia.

[3] According to Konegen-Grenier (2018) in 2017 only 1,080 out of 8,097 academic programs with restricted admissions used the central clearing house for universities in Germany (Gehlke et al., 2017).

Private providers are often not able or willing to contribute full ranking lists of children to a central clearinghouse. To account for these market details, we depart from the literature by proposing an admission mechanism that allows private providers to make decentralized offer decisions while providing fast and fair results. The matching literature has only recently reflected an interest in such decentralized mechanisms; see for example, Bó & Hakimov (2016), who focus on strategy-proof design for college admissions, and Grenet et al. (2019), who focus on applicant-side complementarities (i.e. friends preferring to study at the same university). By contrast, the IDAT mechanism proposed in this paper focuses on a setting in which complementarities are on the provider side. The IDAT accommodates these complementarities and speeds up the deferred acceptance mechanism in this decentralized setting. We believe that this model is a good intermediate step in moving from a decentralized market to a fully centralized mechanism. In the municipalities that adopted the mechanism, market coverage was 100% from year one.

As a starting point, we take the current allocation practice in the city of Münster in North Rhine-Westphalia (NRW), the most populous German state. This allocation practice is the most commonly used method in large cities in NRW and other German states (Klein & Herzog, 2018). Admissions are for a fixed start date at the beginning of the school year in August, when most childcare placements free up. While we take into consideration dynamic aspects that arise when parents move or change placements (Kennes et al., 2014), these factors are of secondary importance in this context.

The matching literature has established speed and fairness as conflicting desiderata when allocation mechanisms are used in a decentralized context. Fairness can be defined in terms of the stability of an allocation. An allocation is stable if and only if it is non-wasteful and no market participant has justified envy of another participant (Kamada & Kojima, 2020). Justified envy occurs if an applicant would prefer a place at facility A but receives a childcare place at facility B, while later learning that an applicant with a lower priority received an offer from A.[4] One of mechanisms that satisfy this requirement is the deferred acceptance (DA) algorithm (Roth & Sotomayor, 1992). This algorithm allows applicants to defer the acceptance of an offer until all higher-ranked childcare facilities have been considered. While widely

[4] In the childcare context, priorities are established based on admissions criteria, such as geographical distance, socio-economic status, single parenthood, siblings attending the same facility and the parents' occupational status.

used in centralized markets (Biró, 2017, for a recent overview), the DA is slow to complete matching in decentralized markets. To speed up the process, many childcare providers make exploding offers that only remain valid for two weeks. However, exploding offers force applicants to accept early, unattractive offers (for the sake of security) and are thus considered unfair.

In this paper, we introduce a *deferred acceptance mechanism with ties* (DAT). The mechanism runs in multiple iterations to include the decentralized offer decisions of private facilities. Compared to running a standard DA in a decentralized context, we accelerate the process by:

(i) automating parents' acceptance decisions based on their submitted rank-order preference of childcare facilities,

(ii) automating public facilities' offer decisions based on pre-specified rules (e.g. admissions criteria), and

(iii) incentivizing private facilities not to delay making offers.

The third element is implemented by allowing parents to state weak preferences (i.e. in-differences) and prioritizing competing offers from the same indifference class on a first-come, first-served basis.

The new mechanism allows one to cater to the interests of all involved stakeholders, including parents, childcare providers, and cities. For parents, fast assignment is desirable for planning reentry into the labor market. Furthermore, fairness is desirable for parents who value transparency. For childcare providers, a fast procedure reduces planning uncertainties, and fairness ensures that providers' admissions criteria are respected. Finally, for the city, a fast process has positive labor market effects, and fairness improves the perception of an efficient public administration, while also reducing the risk of lawsuits.

The mechanism, which is implemented as a software application, was tested in the cities of Saerbeck and Greven (NRW). We also evaluate the applicability and scalability of the mechanism in larger markets.

The remainder of the paper is organized as follows. Section 2 illustrates the context of childcare assignment in the city of Münster and presents the four challenges identified for the design of a revised assignment process. In section 3, we review the literature while focusing on our identified design

restrictions. Section 4 presents the proposed mechanism, its properties and software implementation. In section 5, we evaluate the mechanism with respect to its implementation in the two cities and we present simulation evidence on its scalability for larger cities. Section 6 concludes.

2. THE GERMAN CHILDCARE MATCHING PROBLEM

2.1. Context

In 2013 legal changes in Germany granted children under the age of three a legal right to a childcare placement. This change in legislation has generated extremely high demand and strong competition for available placements. In Germany, cities are responsible for designing their childcare market and implementing admission processes in consultation with providers. One of the characteristics of the German childcare market is provider autonomy, which gives the provider the right to choose their own admission criteria and decide which children are granted a placement at their facility. The admission process thus differs between cities and regions (Klein & Herzog, 2018). In the following, we outline the process in the city of Münster and highlight the problems associated with designing an admission process that fulfills the requirements of all stakeholders.

The city of Münster has a population of 314,000 and 190 childcare facilities. About 15% of the facilities are public (i.e. held and managed by the city). The other facilities are run by different social service providers, such as the church, the German Red Cross, and parental initiatives. Münster conducts one single admissions process every year. The process begins in November for placements starting in August. In November, most facilities organize an open day for parents to get to know their facility and staff. Parents can then register a place for their child online up to the registration deadline on February 1st. This is done by listing up to seven acceptable facilities, without ranking them.[5] The city has published universal admission criteria to guide providers. However, providers neither need to comply nor publish their own admission criteria on the internet platform, which creates uncertainty for parents when registering their choices. After the registration deadline, facilities decide which applicants to send an offer to. This is an uncoordinated process, although some facilities

[5] In the last years, this was extended to 12 acceptable facilities including in-home daycare providers.

may coordinate their offers on an informal basis. Offers are exploding, and a decision is generally expected within two weeks. If parents accept the offer, they are removed from the system and marked as no longer available for other facilities. If the offer expires or is declined, facilities send an offer to the next applicant in line. This process takes about four months.

2.2. Method

The following analysis of the childcare market in Münster gathers the insights drawn from a case study that comprises multiple information sources. These sources include direct information about the process from the city, semi-structured interviews with a representative selection of childcare providers, a survey among parents, as well as several newspaper articles.

Five interviews were conducted with individuals responsible for the application process at the facilities. A subset of individuals among all providers was chosen due to the heterogeneity between providers in terms of the number of childcare facilities under management, the provider's organizational structure, and hierarchical level at which application decisions are made. The interview group contained the head of a municipal childcare facility, the head of a Protestant childcare facility, the head of a Catholic community, a divisional director of the German Red Cross, and the head of a parental initiative. The interview questions focused on the current admission process as well as requirements for the redesign of the mechanism.

The survey was conducted among parents who had participated in the childcare assignment in past years or who planned to do so during the next assignment. The survey contained questions about their preferences. They were asked about their registration decisions during the assignment, as well as their actual preferences. Furthermore, we asked about the importance of factors influencing these preferences, such as a facility's geographical distance and quality. Finally, the parents were asked to report on different characteristics known from the interviews to influence the facilities' priority rankings, such as the age of the child or the parents' denomination. Unfortunately, despite a large marketing campaign with flyers and posters in every facility, only 295 complete survey responses were received. Considering that over 3,500 parents apply for a childcare placement every year, this sample is not large enough to be used in statistical analysis or simulations. However, it can still provide insight into the preferences held by parents. When specific questions from the

survey are mentioned in the following, they are referenced in the footnotes.

2.3. Problem Description

From the gathered information sources four major problems and challenges could be identified in total. The first one is perceived or effective unfairness. This problem arises when there is a lack of transparency concerning how admissions criteria are applied or how parent preferences are taken into account. The second challenge is associated with the lengthy time requirement of the current process, which causes insecurities on both sides of the market. A third challenge is posed by provider autonomy, which makes it impossible to centralize the entire application process, and which establishes important requirements for decentralized decision-making. Finally, the heterogeneity of the childcare placements represents the fourth challenge. More specifically, requirements regarding group composition and variety complicate the application of standard theory, and make adjustments necessary.

2.3.1. Fairness

Req 1.1 – Perceived Unfairness: One reason for perceived unfairness in the admission process stems from intransparency concerning how admissions criteria are applied. Although the admissions criteria of most facilities are accessible online, if and how the facilities implement these criteria is not certain or transparent. Thus, there is a risk that decisions will be unduly influenced by subjective factors, which is perceived as unfair by the parents. Indeed, in a 2017 court case, the city of Münster was criticized by the court for a lack of transparency (OVG NRW, 2017).

Req 1.2 – Importance of Rankings: Parents are not able to accurately state their preferences. In current practice, parents are only able to specify an unranked set of childcare providers or in-home daycare. It is only possible to differentiate between the options by specifying a priority for childcare or in-home daycare. Thus, parents cannot indicate from which childcare facility they might prefer to receive an offer. This problem was raised throughout the interviews and in local newspapers.

Req 1.3 – Exploding Offers: The only other way for parents to express their preferences in the admissions process is by deciding if they want to accept

a proposed offer made by a childcare facility. However, in this way, parents are not always able to voice their true preferences concerning facilities, since often they are confronted with incomplete information. In particular, when receiving an exploding offer with a two-week response deadline, they have no idea if a better offer might arrive at a later point in time. One of the interviewed individuals reported a case in which parents would have received a better offer after accepting a worse one. Their early acceptance of the inferior offer was due to their urgent need of a placement and uncertainty as to whether they would receive another offer if they declined the first.

Req 1.4 – Ability to Specify Rankings: In order to apply the parents' preferences efficiently during the admissions process, we would need to gather their true preferences a priori. Our survey revealed that parents often state several preferred facilities in their application and have no problem ranking them. Furthermore, parents often diverge in terms of the cardinal utility they derive from different childcare facilities.[6]

2.3.2. Speed

Req 2.1 – Time Requirement: The lengthy time requirement of the current process is yet another challenge. This time requirement is inconvenient for all stakeholders and also causes a substantial administrative burden for the facilities. The length of the process is mainly caused by the lack of coordination between the providers, combined with a reliance on exploding offers. One interview participant reported that they actively advised parents to make use of the full duration of the exploding offer since they might receive a better option within that time frame. Such behavior, of course, delays the entire application process, which often stretches over several months in practice.

[6] In our survey, the parents were asked within two questions about their preference structure. First, they were asked to order their facilities in terms of preference. Second, they were asked to assign a point value from 0 to 100 to each facility, representing how they would value a placement at each facility. The first priority was fixed at 100 points. On average, the parents ranked 3.4 facilities while assigning the lowest-ranked facility a point value of 66. This average also included many parents that only listed one facility. Others differentiated strongly in terms of cardinal utility, using the full range down to a value of 0. It also could be observed that parents often chose divergently sized "steps" in cardinal utility when ranking multiple facilities.

Req 2.2 – Uncertainty: Ultimately, the extensive duration of the process leads to uncertainty for parents and childcare providers alike. Parents need to know if they have received a placement for their child and thus if they can go back to work. The providers suffer financial insecurity since the number of vacant placements directly influences their financial resources. Consequently, they need to know the number of accepted children, in order to calculate their staff requirements accurately. This problem was mentioned by the two smaller childcare providers, which are most affected by financial insecurity.

Req 2.3 – Administrative Burden: In general, providers complained about a massive administrative burden during the application phase. A new mechanism could reduce this workload by automating decisions based on a priori ranking lists. Such automation could both reduce the duration of the process and administrative costs.

Req 2.4 – Problems with Existing Mechanisms: However, existing mechanisms from theory, such as the DA, require the preference rankings of both the parents and the providers to be known a priori by the central clearinghouse. In the interviews, concerns were raised that it is not possible or practical for some facilities to place all their applications in one ranking list. One such argument was made by the parental initiative. Due to high parental participation at the facility, they seek to interview each family in advance of the assignment process. Therefore, it is impractical for these providers to rank the full set of applicants. Furthermore, each application interview takes time, thus influencing how the facilities' preferences can be obtained and used.

2.3.3. Provider Autonomy

Req 3.1 – Provider Autonomy: Provider autonomy, that is, the right of each provider to define and use their own application criteria, poses the third challenge. According to this principle, providers may independently decide on which children they wish to accept. This autonomy, which is granted to all privately run providers, is enshrined in law, and thus represents an essential requirement. Thus, we cannot force the facilities to provide global preference lists.

Req 3.2 – Adoption Probability: Furthermore, it will not be possible to implement a new system without the approval of the providers. The opinions among the interviewed persons in this regard were quite diverse. The prospect

of an automated system was particularly welcomed by providers that strive to implement entirely objective admissions criteria. Others expressed various reservations, ranging from a perceived loss of power over the decision process to difficulties in handling certain group compositions. They also cited difficulties connected to edge cases, such as the requirement to match children with special needs to the right facilities.

2.3.4. Heterogeneity of Placements

Req 4.1 – Scope of Daycare: Finally, the heterogeneity of offered placements represents a fourth challenge. Specifically, providers offer different forms of childcare. In Münster, these are 25h, 35h and 45h per week. As a consequence, providers have several types of childcare placements. The preferences of parents might differ for the same facility depending on the type of placement. Also, the ranking lists of the providers might differ for different placements – for instance, due to the group composition, if the placements belong to different playgroups within the facility.

Req 4.2 – Group Composition: Group composition is another argument as to why ranking lists are not known a priori. All five interviewed providers mentioned this issue, although it varied in importance from major to minor. Providers seek to guarantee that every child has a playmate; in this regard, age and gender play an important role. While childcare providers run by the municipality have a legally binding requirement to accept older children first, the others (such as the parental initiative and the Protestant church) stated that they considered both the age and gender of the child during the acceptance process for a specific placement. Some providers even aim to reach a 1:1 gender ratio within a playgroup. Thus, if group composition influences the formation of ranking lists, those lists might change during the allocation process, depending on current status of placement allocations at the same facility. This conflicts with the application of a standard DA, which requires unchanging ranking lists stated a priori.

3. RELATED LITERATURE

In this section, we briefly review the literature, focusing on studies that discuss mechanisms that fulfill one or more of the requirements identified in our problem analysis.

The childcare market is a two-sided market (Roth & Sotomayor, 1992), where both parents and facilities care about the outcome of the matching process. Requirement 2.3.4 showed that facilities care not only about respecting their admissions criteria (which seek to ensure stable placement allocations), but they also have strong preferences concerning group composition.

There is a growing literature on two-sided matching markets in childcare assignment (Veski et al., 2017; Carlsson & Thomsen, 2014; Kennes et al., 2014). The workhorse for most applications is the well-studied DA. While the applicant-proposing DA is strategy-proof for applicants (but not for programs), the program-proposing DA that we build on in this paper is neither strategy-proof for applicants nor for programs (Roth & Sotomayor, 1992). In practice, however, the latter is also difficult to manipulate and widely used, e.g. in German university admissions (Braun et al., 2010). Azevedo & Budish (2018) substantiate this observation theoretically by arguing that manipulation incentives disappear in the DA as the market size grows, such that the DA is "strategy-proof in the large" (SP-L).

There are various ways to allow facilities to maintain control over group composition when using the DA mechanism (see Requirement 4.2). The simple solution is to use strict admissions quotas. These are implemented in the form of so-called slot-specific priorities (Kominers & Sönmez, 2016). To achieve, say, a gender-balanced intake, the childcare facility splits its slots (i.e. placements) into two halves. In the first half, girls have priority over boys. In the second half, priority is given to boys. This guarantees a balanced gender intake if there is sufficient demand from both sexes.[7]

In general, the DA can incorporate more complex distributional requirements (Gonczarowski et al., 2019). Refined lower and upper quotas for slots – for example, to balance assignments according to socioeconomic characteristics – have been discussed in the literature with a view to affirmative action policies (Nguyen & Vohra, 2019; Hafalir et al., 2013; Kojima, 2012). Furthermore, Sönmez & Yenmez (2020) show that so-called choice functions for facility preferences can guarantee maximum compliance in the event that applicants qualify for multiple reserved slots (e.g. a disabled girl in combination with slots reserved for gender and disability).

Another topic is the inclusion of capacity constraints in the mechanism. Kamada & Kojima (2020) discuss the issue of staff headcount requirements

[7] Aygün & Turhan (2020) show how to design a strategy-proof transfer scheme, if capacities are not fulfilled in certain slots

based on the age of the children. This criterion can be modeled in a mechanism such that one child occupies several slots depending on its age. In Germany, regulations define the required staff to child ratio.[8] However, the number of placements each childcare facility can provide per age category is determined in cooperation with the municipality prior to the allocation, since numerous factors influence this decision.[9] If the legal restrictions applied to slot assignment are eased at some point in Germany, it would be interesting to integrate into our research flexible slot management based on the work of Kamada & Kojima. Until then, however, we must consider as discrete the markets for children above and below the age of three.

The requirements described in our problem analysis, however, go beyond the implementation of quota rules. Firstly, Requirement 3.1 states that private facilities are reluctant to disclose their preferred group composition to a central clearinghouse. Therefore, the implementation of choice rules only seems feasible for public providers. Second, preference complementarities arise from staff-to-child ratio requirements based on a child's age. A simple example of a complementary preference is when matching is for a single applicant or for a couple (Roth & Peranson, 1999). In the case of childcare, singles are equivalent to children above age three, and couples are equivalent to children below three (as the latter require more intensive care, and can thus be viewed as occupying two slots). Even in this basic setting, the existence of stable matches is no longer guaranteed.[10] The analysis of matching based on more general complementarities also arrived at several negative results (Kamada & Kojima, 2018; Delacrétaz et al., 2016).

Integer programming is an alternative approach to solve matching problems with distributional constraints (e.g. Ágoston et al., 2018; Geitle et al., 2020). However, in our context the constraints of private childcare providers are not known to the matchmaker, making integer programming not viable (see Requirement 3.1).

[8] Children aged one to three require a ratio of 1 childcare worker per 10 children, while children above the age of three require only a ration of 1 to 20.

[9] Younger children require a place to sleep over noon within the facility. Thus, the available slots depend on the planning of the facilities room capacities, which has to be documented and approved by the municipality, first.

[10] The Roth Peranson algorithm (Roth & Peranson, 1999) and an SAT-solver (Drummond et al., 2015) can find a stable matching in this setting if it exists. However, the existence of a stable matching is unlikely for markets with a large number of couples (Kojima et al., 2013), and the existing algorithms do not fit the requirement of a decentralized matching procedure.

In light of the limited ability of centralized admissions schemes to accommodate general complementarities, we thus consider decentralized mechanisms. The matching literature has only recently reflected an interest in such mechanisms. Bó & Hakimov (2016) demonstrate how an iterative version of the applicant-proposing DA – in which students make applications one at a time – helps students to learn about their feasible set of schools. Similarly, the DoSV mechanism used for German university admission also combines dynamic steps with a final DA phase, but is based on the program-proposing DA. The DoSV differs from the DA in that the centralized DA is preceded by a decentralized phase of 34 days, where programs make admission offers to their preferred students in real time. Students, in turn, can decide to accept an offer and exit the process, to retain all offers, or to keep just a subset of offers. In the subsequent centralized phase, students who have not yet exited the process can reconsider their rank-order lists and participate in the program-proposing DA (see Grenet et al., 2019, for a detailed description and analysis).

The DoSV mechanism has several advantages. It relaxes the assumption of fixed student preferences by allowing students to form preferences during the admissions process and can accommodate (to some degree) complementarities in the form of friends preferring to study at the same university. Instead, the focus of the IDAT mechanism – as proposed in this paper – is a setting where complementarities are on the program-side and a centralized phase is ruled out by institutional constraints. IDAT accommodates these complementarities and speeds up the DA in this setting.

A series of papers also make use of a family of mechanisms characterized by Chen & Kesten (2017) that result in allocations between the DA and the immediate acceptance (IA) algorithm. One example is the adaptive IA (Mennle & Seuken, 2017), which allows schools in each round to condition their admissions based on admissions already confirmed in the previous rounds, while at the same time making it safer for students to report their true preferences. Several other papers also study dynamic matching procedures inspired by college admissions practices (Echenique et al., 2016; Gong & Liang, 2017; Klijn et al., 2019). The IDAT mechanism proposed in the next sections has, to the best of our knowledge, not yet been studied or applied in practice.

An additional feature of the adaptive DA mechanism we propose below is that it allows for weak preferences, i.e. indifference classes, which are used to incentivize providers not to delay their offers (see Requirements 2.3.2). Indifferences result in an efficiency-loss because providers could be made

better-off by swapping their assigned applicants. This source of inefficiency can be improved on by allowing for such swaps using so-called "stable improvement cycles" (Erdil & Ergin, 2008).[11] We will not discuss this issue below because a shortcoming of such post-hoc improvement cycles is that they induce additional strategic behaviour.

Furthermore, there is literature about matching with contracts, which shows that market agents can be matched using different kinds of assignments (Hatfield & Milgrom, 2005). At first glance, the outlined model of the childcare assignment problem qualifies as a form of matching with contracts, since the different scopes of care that are offered by the providers could be mapped as different contracts with the same facility. However, in German cities, the number of placements a facility can offer for each scope of care has to be fixed a priori in consultation with the city due to various factors (e.g. security regulations or room planning). For this reason, we chose the more comprehensible alternative for representing divergent scopes of care programs by separating the placements in each facility, which is possible since each care program has its own capacity limit and might also have a distinct preference ranking. Also, the advantages inherent in the theory about matching with contracts, such as including monetary transfers, are not relevant to our specific use case. If the IDAT were to be applied in the future to a market in which placements regarding the scope of care are not fixed but interchangeable, our model could be extended to include matching with contracts.

4. SOLUTION

4.1. Mechanism

The market for childcare placements represents a two-sided matching market. In our model, children are on one side of the market. All children $I = \{i_1, ..., i_N\}$ are single individuals. On the other side of the market, we have the providers $S = \{s_1, ..., s_N\}$. Each facility is run by a provider. These providers can be categorized into those who want to provide a complete list of priorities (for automation) and those that prefer to make offers conditional on the status of previous offers. For simplicity, we refer to the facilities run by the first kind of provider as public providers S_{Pub} and the facilities run

[11] Erdil & Ergin (2017) also extend this concept to "stable improvement chains", which allow for indifferences on both sides of the market.

by the second type as private providers S_{Pr} so that $S = S_{Pub} \cup S_{Pr}$ as well as $S_{Pub} \cap S_{Pr} = \emptyset$.

For educational purposes, providers usually separate their children into multiple play groups. Often, each group has different characteristics, such as the scope of care (in hours per week) or an age limit. These characteristics influence the preferences of the parents and can also result in a differing priority list among one facility type. Additionally, each play group has a capacity constraint in terms of a limited number of available placements. For these reasons, these groups are used as entities on the facility side within the mechanism. Further, these entities are called "programs" in the following and are labelled $C = \{c_1, ..., c_k\}$. Thus, children are assigned to placements within certain programs. We assume that each program decides on its own which children to accept, due to possibly divergent priority rankings. The fact that a program c belongs to a facility s is represented by $c \in s$. Analogously, we write $c \in S_{Pr}$ and $c \in S_{Pub}$ to indicate how a program's preferences are expressed in the market mechanism.

Each of these programs has a placement capacity of $q_c \in Q$ of places that can be filled. In the following, we use different ranked lists. The preference rankings of children are labelled P_i and the rank order preference lists of programs run by public providers are labelled R_c, respectively. The private providers play within each round manually and provide a ranked waiting list in each iteration t. The waiting lists are labelled $W_{c,t}$. In each iteration, these waiting lists can include an arbitrary number of children and are not limited by the capacity of the respective program.

Let P_i be the preference ranking of a child i concerning a subset of programs, such that $c \in P_i$ states that program c is ranked by child i. The program c is therefore strictly preferred by i over remaining unassigned. Also, $c_1 \succ_{P_i} c_2$ states that under preference ranking P_i the child i strictly prefers c_1 over c_2. Indifference between two programs is stated through $c_1 \sim_{P_i} c_2$. The same applies to R_c and $W_{c,t}$, although strict preference rankings are required here, without indifferences. During the assignment process, $\mu_t : I \to C \cup I$ denotes the current assignments at time t in which every child is either assigned to a program or itself. The latter option means that the child is unassigned. Further, let the inverse of μ be defined as $\mu_t^{-1}(c) = \{i \mid \mu_t(i) = c\}$.

Since the private programs can adjust their waiting list in each iteration depending on their information about current matches, we use $\omega_c(\mu_t) = W_{c,t+1}$ to represent the waiting lists prior to the start of matching. Formally, we define

our matching problem as $\phi(I, S_{Pub}, S_{Pr}, Q, P_I, R_C, \omega_C)$. Since the first four parameters of ϕ are fixed for a given market from the perspective of the agents, in the following, we will simplify the problem definition as $\phi(P_I, R_C, \omega_C)$.

As a first step, parents submit the preference lists P_i that rank the programs in tiers, whereby parents are indifferent between offers from the same tier. In addition to their preferred facilities, they can specify other preferences, such as the providers themselves, their opening hours, or the scope of care. In our mechanism, these formally stated preferences can be transformed into the weak preference lists P_i over the set of programs C. In this way, a child could rank a placement at facility A in a 45h care program as preferable to a 45h placement at facility B, which is in term preferable to a 35h placement at facility A. The current iteration of manual offers is denoted by t, which is initialized with 1. We further state that $|\mu_0^{-1}(c)| = 0, \forall c \in C$.

The *deferred acceptance mechanism with ties* (DAT) is a multiple-round algorithm that assigns children in the following steps:

1. For every program $c \in S_{Pr}$: They receive a list of all feasible applicants. Feasible means that
    a. $c \in P_i$: The program is contained in the preference ranking of child i and
    b. $c \succ_{P_i} \mu_{t-1}(i)$: Child i strictly prefers program c over the current match.
2. The decision-makers of program c can select a set of children with size equal to $q_c - |\mu_{t-1}^{-1}(c)|$, which is the number of currently vacant placements, to send an offer to. Additionally, they can exceed their capacity and register as many children as they want on a strictly ranked priority list. The registered offers and the priority list together form the waiting list $W_{c,t}$.
3. Run program-proposing DA with the private programs' waiting lists $W_{c,t}$, the public programs' rank order lists R_c and the children's preference lists P_i.
    a. If $t \neq 1$: Initialize the DA by assigning each child i the place they held in μ_{t-1}.
    b. Private programs: Send out their registered offers (and subsequently offers based on $W_{c,t}$).

c. Public programs: Send offers based on R_c.

d. Children: In each round of the DA, children hold the first offer for the highest tier according to P_i and reject all others. Ties are broken at random (using either multiple or single tie-breaking).

e. For rejected offers: New offers for public and private programs' are submitted repeatedly based on their R_c and $W_{c,t}$, respectively.

f. If no program makes any new offers, the resulting matching is called μ_t.

4. After each DA run, the decision-makers of private programs can review all tier-one acceptances, deferred acceptances and rejections. The mechanism continues with the next round from step 1.

5. This cycle terminates when a predefined number of rounds is exceeded, when no rejections are issued, or when no private facility registers a new offer.

4.2. Game-theoretical Considerations

Stability and strategy-proofness are typical performance desiderata considered in the theoretical literature on two-sided matching. In our context, stability is relevant for preventing justified envy, and, by extension, perceived unfairness, as specified in Requirement 1.1. Meanwhile, strategy-proofness is a necessary design element. We also have to consider the time to complete the matching process (see Requirement 2.3.2) and allow for complementarities in the form of preferences concerning group composition (see Requirement 4.2). In the following, we discuss the properties of the IDAT in view of these requirements.

The program-proposing DA is stable but not strategy-proof on the child's side (Roth & Sotomayor, 1992). Roth (1982) generally shows that a DA cannot be stable and at the same time strategy-proof on both sides. Most practical applications implement the applicant-proposing version. The applicant-proposing DA is strategy-proof in theory. In practice, however, it is often subject to modifications that break strategy-proofness. Two of the most common modifications are to limit the number of programs to which a child may apply (see section 2.3) or to assign placements several times a year (e.g. Kennes et al., 2014). Also, the applicant-proposing DA is less suitable in our iterative context, as it

allows programs to infer students' preferences based on the order in which proposals are received. This information induces strategic behaviour in applicants if programs consider this information when accepting proposals. We therefore consider the program-proposing DA and discuss its properties.

Azevedo & Budish (2018) show that DA is "strategy-proof" in the large (SP-L). As the number of agents in a market grows, a mechanism is SP-L for children if for any $i \in I$ it is unlikely that there is a preference list P_i' such that $\phi((P_i', P_{I_{-i}}), R_C, \omega_C) \succ_{P_i} \phi(P_I, R_C, \omega_C)$. In words, SP-L ensures that truthful reporting is approximately optimal for every agent. The childcare market at the municipal level can be considered large, and the proportion of agents who can successfully manipulate the mechanism should diminish in size when information (e.g. about capacities) is limited (Budish & Cantillon, 2012). Further, Azevedo and Budish base these findings for DA being SP-L on the analysis of a weakening of envy-freeness by considering mechanisms that include any kind of tie-breaking lottery. They call this concept "envy-free but for tiebreaking." Their results state that a mechanism that includes tie-breaking can still be SP-L, provided no participant envies another simply due to a lower lottery number. Thus, our adaptation to allow for indifference within the children's preference lists, which is a form of tie-breaking, should not influence DAT being SP-L. In any case, successful manipulation requires broad information about the market, for which the probability diminishes if the market grows large.

Moreover, we argue that the potential loss in strategy-proofness of the program-proposing DA is tolerable in our scenario. From the children's side, the outcome can be manipulated through misrepresentation or truncation of the preference lists. Both will result in the realization of a different stable outcome, which is no longer program-optimal. Thus, manipulability is given if there are multiple stable outcomes in the market, which is observed to be unlikely in practice (Roth & Peranson, 1999; Pathak & Sönmez, 2008; Hitsch et al., 2010). In addition, the program side is bound to fixed admissions criteria (at least in the case of public providers), and thus cannot engage in strategic behavior to manipulate the market.

The DAT mechanism is stable, as the DA does not produce blocking pairs (Kojima & Manea, 2010). Our iterative approach does not change this. Technically, a pair (c, i) is a blocking pair for μ if

(i) i and c list each other in their ranking lists,

(ii) i is unassigned or prefers c to $\mu(i)$, and

(iii) c is undersubscribed or prefers i to at least one member of $\mu(c)$.

Thus, there are no justified incentives for either applicant to leave their assigned matches. This also means that there is no justified envy between the applicants. When we allow for indifference on the applicant side, stability only holds if the applicants accept the strict preference rankings created through tiebreaking.

The presence of complementary preferences on the applicant side is widely studied in the context of couples in labor assignment problems (Roth & Peranson, 1999; Kojima et al., 2013). In the market for childcare, siblings generate complementary preferences if parents want them to attend the same childcare facility together. Although complementarities on the program side have received little attention in the literature to date, they are particularly relevant to group composition. From a pedagogical point of view, it is generally accepted that programs aim to balance their intake in terms of age and gender. Aldershof & Carducci (1996) show theoretically that in markets with complementary preferences, stable allocations are not guaranteed to exist. Annual intakes per program are often too small to add so-called slot-specific priorities, as suggested by Kominers & Sönmez (2016). Research in the field of combinatorial auctions has shown that iterative mechanisms, in which offers can be made in rounds, allow participants to use a simple preference language for stating preferences regarding bundles (Parkes & Ungar, 2000). Thus, the iterative approach of the DAT not only solves the challenges posed by provider autonomy (see Requirement 3.1), but also facilitates complementary preferences for group composition (see Requirement 4.2).

Nevertheless, the class of iterative DA mechanisms has been generally slow to converge (Bó & Hakimov, 2016). Furthermore, any private facility has a clear incentive to delay the mechanism if it allows for complementary preferences. In particular, such a program would benefit from letting all other programs to move first, in order to choose freely from the remaining applicants in the last round (i.e. all applicants who would accept an offer from the program at that point). To speed up the mechanism, we induce facilities to send out offers fast by allowing for indifference tiers (weak preferences) on the applicant side. However, when combined with immediate acceptance, i.e. breaking weak preferences early, this results in efficiency losses (Erdil & Ergin, 2008).

We conclude that our modifications do in fact impair efficiency. Nonetheless, these losses are more than offset by the larger welfare gain of a faster process that ensures fairness. Overall, the proposed mechanism meets all desired performance requirements in the context of German childcare allocation.

4.3. Information System

To implement the revised mechanism in a real-life setting, we constructed the open-source matching platform *Kitamatch*.[12] The software provides a comprehensive approach to the childcare matching problem, and is the first solution to make the decentralized deferred acceptance mechanism accessible (see Figure 3 for a screenshot).

The matching platform consists of four components: (i) a preference module for families, (ii) an administrative unit for municipal authorities, (iii) an interface for childcare facilities, and (iv) a matching mechanism unit. The software is designed to allow discrete use of a matching module and the importation of ranking lists from alternative software systems that register applications.[13]

The municipality hosts the assignment process and is therefore also responsible for the software. It registers all public providers by their different programs, capacities, and criteria catalogs. In general, a criteria catalog assigns weights to each criterion in order to rank applicants objectively. For all public programs, this ranking R_c is binding.

After registering on the website, families need to fill out a survey questionnaire about their preferred scope of care for their child i and provide data relevant to the facilities' criteria catalogs. Most importantly, parents construct a tiered-based preference list P_i via drag and drop over the programs C. From the perspective of the parents, each such program consists of the facility s and the corresponding scope of care. The acceptance or rejection of potential offers is then fully automated. The applicant does not see temporary offers during the matching rounds and is only informed about his or her final allocation μ_T at the end of the matching process.

Each private childcare facility needs to register its different programs by

[12] The software is available under https://github.com/svengiegerich/kitamatch.

[13] See for example Kitanavigator, https://www.itk-rheinland.de/, or KVJS, https://www.kitaweb-bw.de.

name, capacity, and scope of care. Afterward, there is the option for every facility to create an individual criteria catalogue for the automatic presorting of applicants. As the matching process starts, each private program sees its feasible applicants presorted, and can add children to the waiting list $W_{c,t}$. After every round, a temporary matching μ_t is computed, and the software interface gives detailed feedback about acceptances and rejections. The programs can then decide once again to add applicants to the updated waiting list $W_{c,t+1}$. During every round, the waiting list can be shorter than the capacity of the program, it can exceed the capacity, or the facility can send no offer at all. This iterative process makes it possible to take group composition into account by conditioning offers in later rounds on previously formed matches.

The matching process takes several rounds. Within these rounds, private programs register offers manually to the waiting list $W_{c,t}$, while the preferences of children P_i and public programs R_c, as they stick to their criteria catalogues, are automated. A round t is closed either after a specific time interval (e.g. each day) or manually. The round-based computed matching μ_t, as well as the rejected offers, are returned to the database. Afterwards, a new round starts. As soon as no private program sends any more offers, the matching process is considered complete, and the applicants are informed of the outcome μ_T. The corresponding DAT mechanism is also available in the statistical R package *matchingMarkets* (Klein, 2021).

5. EVALUATION

To evaluate the proposed solution, we consider two aspects in this section. First, we document the implementation of our solution to provide evidence on how it resonates with stakeholders, including in particular childcare providers. Second, we provide simulation evidence on how well the mechanism scales for larger market sizes, different public/private provider shares, and different admissions criteria. This second aspect is relevant due to the number of rounds necessary for the mechanism to terminate and reach a stable matching. If the number of rounds grows too large, our approach would be too time-consuming and thus impractical.

5.1. Implementation

We discuss the empirical results of our DAT implementation in the cities of Saerbeck and Greven. Comparing these two cities – which have divergent population sizes and preference profiles – yields particularly interesting results. To avoid repetition, when the results are similar or comparable, we portray our findings exclusively for Greven. Table 2 provides an overview of the results for both municipalities.

Greven, which has some 40,000 inhabitants, is a representative mid-sized city in Germany. The decentralized immediate acceptance (IA) process previously took four months to complete, and the matching repeatedly caused various problems (see section 2.1). Due to a new legal judgement (see Requirement 1.1), the city decided to reform its immediate acceptance (IA) process. The second city, Saerbeck, which has 7,000 inhabitants, used the same system and faced similar issues. In 2018, we were approached by the cities and asked to assist in redesigning their assignment systems. This request, in tandem with an ongoing case study in Münster, led to the development of the DAT mechanism. In the first implementation of the childcare year 2019/20, the mechanism was implemented as described in section 4.1, without the feature of indifference tiers. The following analysis, therefore, is silent on the effects of this feature. The simulations in section 5.2, however, shed some light on this issue.

The Greven childcare setting consisted of 479 children and 26 childcare facilities, including one public facility. The facilities differentiated their care offers between the age cohorts $U2$ (under two years old), 2 (two years old) and $O3$ (older than three years), which resulted in 65 programs. All childcare facilities constructed point-based criteria catalogs that assigned points to rank applicants. These catalogs were quite heterogeneous between facilities, both with a view to the criteria used and their number, which varied between three and eight. Without exception, all of the facilities used common criteria regarding siblings and the employment. In addition, some facilities include the denomination of the child, and others consider more abstract criteria such as family emergency (illness), inclusion, and cultural diversity. Saerbeck's institutions use a similar range of criteria, but apply the criteria in strictly lexicographical order.

In both municipalities, children's preferences were collected using a written questionnaire. On average, parents indicated 3.29 programs in Greven and

Market	# Children	# Programs	# Providers	Avg.Preference Length
Greven	479	65	26	3.29
Saerbeck	100	18	7	2.73

Table 1: Greven's and Saerbeck's childcare assignment setting.

2.73 in Saerbeck in their preference list.

For the actual matching process in Greven, all heads of the childcare institutions were invited to the youth welfare office in January 2020. The matching in Saerbeck took place in January 2019. In both meetings, participants were presented with a pre-sorted list of applicants based on points constructed by their submitted criteria catalogs. The sorting only served as a suggestion; facilities were allowed to consider group complementarities as they sent out decentralized offers through the matching platform round by round. Since all decision-makers sat together, it was possible to conduct the matching rounds flexibly instead of waiting for fixed time intervals. The final matching in Greven and Saerbeck took an hour and was completed in seven and six rounds, respectively. In total, 403 of 479 and 85 of 100 children, respectively, were assigned a place. Thus, none of the markets cleared their full capacity. This can be mainly explained by two factors. First, many parents indicated only a few – in 9.6% (31%) of the cases just one – acceptable childcare programs. Second, parents had no information beforehand on which programs had available placements. In fact, 19% (24%) of all parents listed programs that did not have a single vacant care place. Some preference lists even shared both of these problems simultaneously. Thus, these children's preference lists were practically empty.

5.1.1. Welfare analysis

This section concludes with a welfare analysis that constructs counterfactuals for the matches in Greven and Saerbeck to compare the performance of three mechanisms: (i) the previously used IA, (ii) the standard program-proposing DA, and (iii) the DAT, which allows facilities to deviate from the pre-sorted priority list. This comparison is possible because we observe for all childcare facilities both (i) their ranking of applicants based on a *criteria catalogue*, denoted by $r_{catalogue}$, and (ii) their *revealed preferences* from the sequence of

Mechanism	# Matches	Children's avg. rank	# Blocking pairs*		Group complements
			with $r_{catalogue}$	with $r_{revealed}$	
Greven					
IA	401	1.15	84	90	No
DA	406	1.34	0	19	No
DAT	403	1.33	54	0	Yes
Saerbeck					
IA	86	1.31	14	–	No
DA	85	1.36	0	–	No
DAT	85	1.33	–	0	Yes

Table 2: Comparison of the redesigned DAT with simulated results for immediate acceptance (IA) and deferred acceptance (DA) algorithms in the cities of Greven and Saerbeck. The DAT respects group complementarities as it allows round-based deviations from the preference list – used in 12% (9%) of all sent offers made in Greven (Saerbeck) from round two onwards.
(*) Blocking pairs are evaluated based on the ranking of facilities using both the criteria catalogue ($r_{catalogue}$) and facilities' revealed preferences ($r_{revealed}$). The latter could not be calculated for Saerbeck because of missing information.

offers made in the DAT, denoted by $r_{revealed}$. In what follows, we analyze four aspects: unfilled placements; applicant preferences; facility preferences and participation constraints; and fairness.

A first observation from Table 2 is that no mechanism allocates the full number of placements available.[14] In Greven (Saerbeck), the IA assigns 401 (86) children, the DA 406 (85) and the DAT 403 out of 469 (85 out of 101) available places. The places are free primarily due to the incomplete ranking lists of the parents.

The average rank of the assigned placements based on parent preferences is relatively equal across all mechanisms in both cities. The IA allocates more extreme matches, as more children get their first or their last rank, while the DA and DAT both result in a more balanced distribution.

For an analysis of facility preferences, we consider their ranking over applicants as inferred from the sequence of offers made, $r_{revealed}$. This measure is more appropriate than the ranking generated from the criteria catalog, because

[14] For the city of Greven, the available placements were determined as reported in Table 4.

it also captures deviations in the DAT that allow facilities to express richer preferences and, in particular, to account for complementarities. Group complementarities were of importance for programs, as they deviated from their priorities several times during the matching process. For example, in Saerbeck, a program with three placements sent an offer to the girl ranked fourth in round two, as it already held two boys from round one. As this girl rejected the offer, the facility skipped the following boys with rank five and six and instead sent an offer in subsequent rounds to the girls in seventh and eighth place. Thus, the program made a trade-off between priorities and group composition. Overall, in Greven (Saerbeck), 12 out of 26 programs (8 out of 18) deviated from the pre-sorted prioritization at least once. In total, we find a deviation rate from the pre-sorted rankings of 12% (9%) when considering all rounds, starting from round two to the final round seven. See Table 3 in the Appendix for an overview of results for Greven. We observe that in Greven, in 32% of all deviations, applicants were preferred because they added heterogeneity in the gender composition and in 15% due to heterogeneity in age.[15]

We measure the welfare effects enjoyed by facilities when using DAT rather than DA in terms of the number of programs that would block the implementation of DA matching. In particular, we know that no program would oppose the implementation of DA matching if $r_{catalogue}$ is identical to $r_{revealed}$. That is, if programs' revealed preferences in $r_{revealed}$ do not deviate from the critera catalog in $r_{catalogue}$. If programs, however, have complementary preferences, then these will be expressed in terms of deviation from the criteria catalog, and result in blocking pairs. Using $r_{revealed}$, we find for Greven a total of 19 blocking pairs in the DA matching. These 19 blocking pairs involve 13 programs and 14 children. This means that 13 out of the 65 programs would block the implementation of DA matching, and would presumably only agree to participate in the matching scheme if the DAT were used instead.[16]

[15] Childcare facilities sometimes prefer heterogeneity in ages because this allows them to distribute the settling-in periods for children evenly over different years, rather than having a high turnover in any particular year.

[16] In this welfare analysis for facilities, we do not consider the alternative welfare measure of the *average assigned rank* (as we did for the analysis of applicant preferences in Table 2, column 3). This is because of the following complication. In the DAT, programs that strictly follow ranking $r_{catalogue}$ in the mechanism, can have a sequence of offers $r_{revealed}$ that is a shortened version of $r_{catalogue}$. This is because some children in $r_{catalogue}$ cannot be selected by the program in the mechanism if the children already hold their first preference from another program. This complication, however, does not affect our blocking pair measure, since offers

We now turn to the welfare analysis for applicants. As seen in the analysis of facility preferences, the DAT allows facilities to deviate from the criteria catalogue, which may be a crucial feature for ensuring all facilities participate on a voluntary basis. For applicants, however, the DAT can be perceived as unfair, when $r_{revealed}$ is not identical to $r_{catalogue}$. In this case, applicants will have justified envy of other applicants. The magnitude of this problem can be measured by asking how many applicants would block the DAT matching. Using $r_{catalogue}$, we find for the city of Greven a total of 54 blocking pairs in the DAT. These 54 blocking pairs involve 13 programs and 49 children. This means that 49 out of the 403 matched applicants would oppose the DAT matching. While this number is still lower than the total of 77 blocking pairs we find when applying the previously-used IA mechanism for Greven (respectively 14 in Saerbeck),[17] it shows that fairness may have to be compromised, even in the DAT, in order to respect provider autonomy and achieve voluntary participation.

5.2. Scalability

We simulate childcare markets and evaluate the effect of different parameter settings on the required number of iterations.[18] All settings are based on the childcare market in the city of Münster, which had a total of 3,031 childcare placements in 2020 (Press and Information Office of the City of Münster, 2019). Although these placements are separated between children younger than three and older than three, there are 2,810 children below the age of three in need of a place. As in the baseline scenario, we therefore use a market consisting of 3000 applicants and 200 facilities. Each facility has, on average, three programs, resulting in a total of 600 programs. We work with 1.2 applicants per childcare place, which reflects the shortage of placements for children under the age of 3 in many German cities.[19] Finally, we randomize

to these children would have been rejected anyway.

[17] Thus, in the old procedure around 14% to 20% of all children were disadvantaged because of the unstable process.

[18] The code to reproduce all simulations and figures can be found on: https://github.com/tobiasreischmann/matchingmarkets-simulation

[19] In Germany, 43% of the parents with children aged below three apply for a childcare placement, but only 35% can be offered one, leading demand to outstrip supply by a ratio of 1.23:1. (For more see, https://www.kindergartenpaedagogik.de/fachartikel/kita-politik/bildungspolitik/1650)

programs to be private facilities with probability q, which we refer to as the private facility share.

The simulated preferences of applicants and programs combine horizontal, vertical, and idiosyncratic components,[20] and are modeled using two selection functions: one for the applicant-side vis-à-vis the program-side, and vice versa. The selection functions operate on randomly generated attributes assigned to applicants and programs. These attributes are then used to calculate and combine the three preference components into a lexicographic preference function.

5.2.1. Applicant preferences

For *horizontal* preferences, x and y coordinates are assigned to applicants and programs as geographical information. Applicants are modeled to be indifferent about programs within the same city district if the decision is solely based on distance. Thus, x and y are realizations from a uniformly distributed categorical variable from 1 to i. i is chosen depending on the number of programs, such that there are about ten programs within the same district. Horizontal preferences are then determined using the Euclidean distance.

Vertical preferences are modeled by a variable that assigns a quality level to each program, which is used directly in the applicants' selection function.

The *idiosyncratic* component of applicants' preferences assigns each applicant and each program one out of ten random types. Applicants only value a program if it is of the same type. This accounts for valuation based on religious beliefs or special forms of care.

In addition to the three preference components, each applicant receives three uniformly distributed variables that determine which of the three preference aspects carries highest importance for the respective applicant.

Finally, we include the effects of allowed indifference within the children's ranking lists in our simulation. In general, each child only ranks a small subset of the whole market. In the baseline scenario of 600 programs, each child ranks 30 of them. If indifferences are allowed, the ranking lists of the children are

[20] Applicant preferences are defined as follows. A horizontal component captures applicants' preferences for attributes of programs that are closer to their own attributes, such as the geographical distance between the applicant and program. The vertical component captures program attributes that all applicants value equally, such as a high staff ratio. Finally, the idiosyncratic component models subjective valuations.

separated into a set of predefined tiers of equal size. In the baseline scenario, these are four tiers sized three, seven, ten and ten. For other market sizes, the size of the ranking lists and the tiers are adjusted accordingly.

5.2.2. Program preferences

A similar selection function is used on the program side. The design of this function is guided by the admissions criteria of the city of Münster, which requires providers to consider the three preference components in a strict order.

For the horizontal component, each program only values applicants from the program's home district. Actual distance does not matter. For the vertical component, applicants are assigned a priority variable, which represents the family's social need or the age of the child.

The idiosyncratic component is also generated randomly for each combination of applicant and program. It models the subjective valuations of the program staff. This component only enters into the selection function of private providers. For public providers, the inclusion of subjective factors is ruled out by law.

The preference rankings of the programs are usually model criteria catalogues, which have a strict order of vertical and horizontal components. For the preference generation, we use two random variables, which provide the three preference components in a strict order for each program.

5.2.3. Simulation

In our analysis, we evaluate how different market characteristics impact the number of rounds required in the DAT. To this end, we simulate several markets under different parameter settings and obtain the necessary number of rounds until at least 95% of the matches in the final DAT results are reached.

Figure 1 shows how the assigned placements change during the matching process in the baseline scenario described above. In each iteration, we differentiate three assignment states. For unmatched placements, no assignment exists in the current iteration. While final assignments represent a match that is also part of the final iteration's match, temporary assignments are changed in future iterations. Although most of the placements quickly reach their final allocation, some placements are frequently changed, which causes a significant number of rounds, if we wish to run the process to its natural end. However, we argue that ending a matching process early when the percentage of non-final assignments

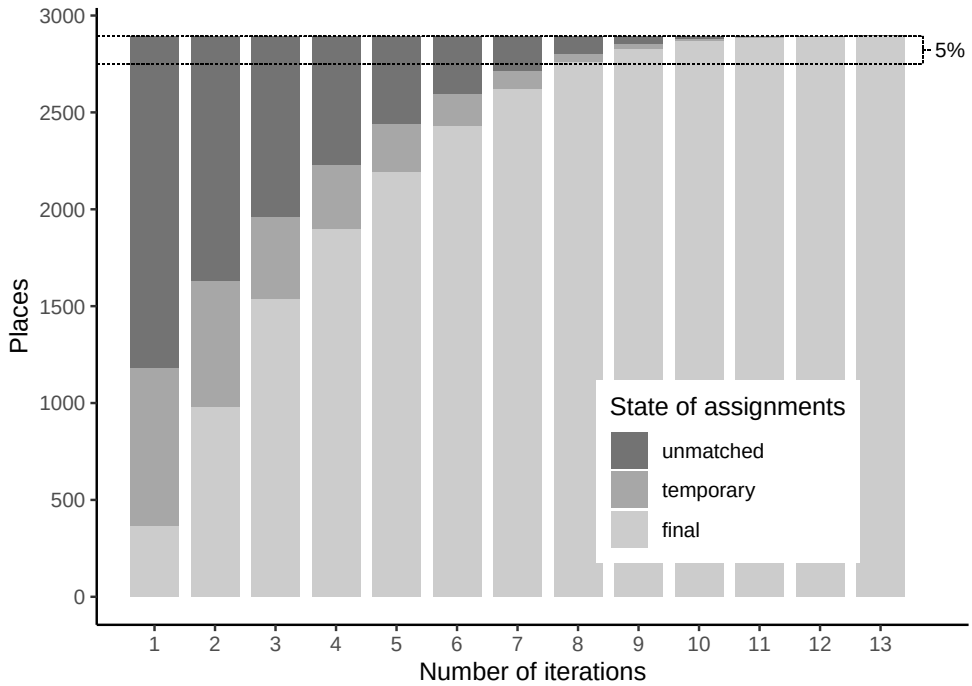


Figure 1: Matching after t iterations in comparison to a full DAT run (with 60% private facilities, occupancy rate of 1.2, 3000 applicants, and 600 programs)

is below 5% is tolerable given the advantage this brings in making the mechanism more practical. In practice, the admission criteria have to ensure that no hardship cases are among the 5% (or among the even smaller number of unmatched assignments). Since the hardship cases are usually ranked higher and thus receive offers earlier, it is unlikely that hardship cases will fail to receive offers within the first couple of rounds.

Figure 2 illustrates how different parameters affect the number of required iterations. For this evaluation, we build an average of played rounds for ten simulated markets. The evaluated characteristics are:

1. Occupancy rate (number of children over the number of available placements), ranging from 0.2 to 3. Baseline scenario: 1.2.

2. Share of private facilities, ranging from $q = 0\%$ to 100%. Baseline scenario: 0.8.

3. Applicants can rank facilities in tiers (yes/no). Baseline scenario: yes.

4. Number of applicants (keeping the number of programs and the occupancy rate fixed), ranging from 600 to 5,000. Baseline scenario: 3,000.

5. Length of children's ranking lists. Baseline scenario: 30 separated into four tiers of 3, 7, 10, and 10 programs.

6. Threshold for the percentage of the stable matches to be reached, ranging from 93% to 100%. Baseline scenario: 95%.

7. Market size, ranging from 250 applicants and 50 programs to 6,000 and 1,200. Baseline scenario: 3,000 and 600.

8. Selection function of the programs, including categories "mixed preferences"; "vertical only"; "horizontal only"; "both sides vertical only". Baseline scenario: mixed preferences.

Our motivation for inclusion of the last characteristic is the controversial discussion in the city of Münster as to whether children should receive priority at a facility if they live in the same district. In addition, this final evaluation serves as a proof of concept for our simulation. We represent this through the inclusion of different combinations of preference aspects within the selection functions. While mixed preferences are a feature of the scenario described above, the next three scenarios restrict the programs' preferences to only one aspect. The last scenario assumes that both the children and the facilities only care about vertical preferences. This scenario represents an edge case, which leads logically to a significant increase in necessary rounds.

Within the eight analyzed characteristics, each city has scope for decision-making. While the characteristics 1, 2, 4 and 7 can be assumed to be fixed within a given market, each city can freely choose their ranking strategies for children, the applied threshold for the mechanism, and the selection function of the programs. They can also influence the number of programs a child can rank via the implemented admissions system. Additionally, a city can strive to convince childcare providers to provide full ranking lists through objective admission criteria, thus reducing the private facility share. Also, in the long run, the cities can provide more placements for the children, which will change the occupancy rate. Moreover, the number of rounds that are deemed applicable depends on the way the rounds are played. If it is possible

to place all decision-makers into one room and play all rounds in one day, a higher number of rounds might be tolerable. If the childcare providers agree on playing one round each week, each additional round will delay the time when the parents are informed about the final assignments. We argue that a suitable number of rounds might be around six to nine. This area is highlighted in Figure 2 through the grey areas.

The results show that neither the occupancy rate (Figure 2.1) nor the number of applicants (Figure 2.5) has a significant influence on the number of rounds played. With Figure 2.5, the number of placements was scaled, since the occupancy rate and the number of providers stayed fixed. At the same time, the number of rounds played grows linear with a higher private facility share (Figure 2.2) and a lower threshold (Figure 2.4). However, we can conclude that a lower threshold and thus a higher convergence to the stable matching might be applicable in cities with a lower share of private facilities. Also, letting the children rate facilities in tiers (Figure 2.3) has a positive impact on the number of rounds played. The reason is that longer preference lists on the children's side of the market will cause a higher number of rounds. With tiers, the preference lists are shortened more rapidly during the matching process.

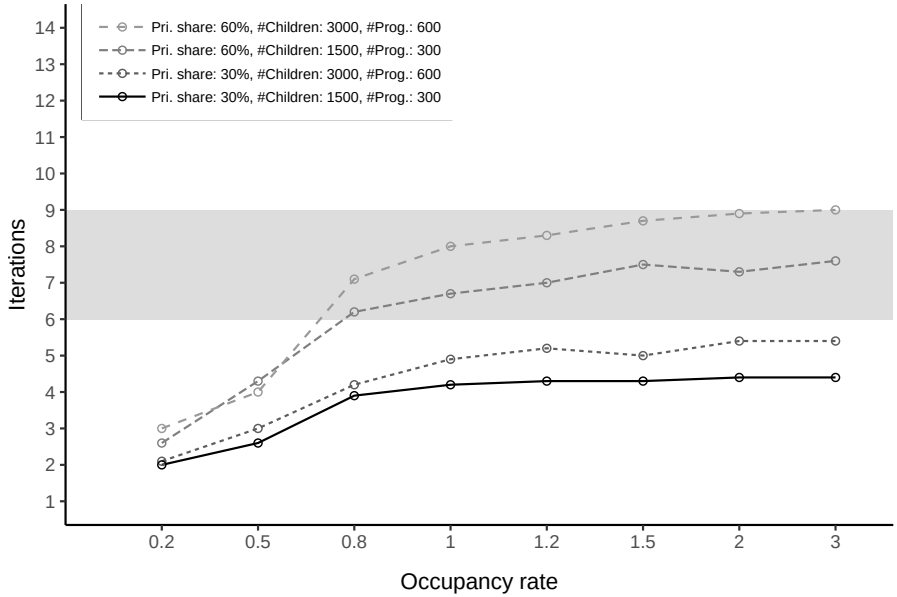
As mentioned, the length of the children's ranking lists has a significant effect on the number of rounds played (Figure 2.6). In practice, children are highly unlikely to rank the full market. According to the survey, the average of facilities ranked was 3.4, which does not even exhaust the list of seven facilities for which parents were allowed to apply.

Figure 2.7 reveals the impact of the market size on the number of rounds played. While the number of rounds increases with the size of the market, the number of rounds generally remains in a manageable range.

Finally, Figure 2.8 highlights two intriguing results. First, the actual admissions criteria that the facilities apply seem to have almost no effect on the number of rounds played. Thus, the facilities do not have to restrict their method of ranking children for the mechanism to be viable in practice. Furthermore, we also find in Figure 2.8 evidence for the hypothesis that the mechanism will break if the preferences are solely vertical. In theory, this would cause as many rounds as the number of placements a facility has. In our case, this effect is limited due to the short preference lists of the children. Through our survey, we expect that the distance to a facility plays a crucial role in the decision-making of the parents.[21] Thus, a scenario in which both

[21] In a survey, parents were asked about the characteristics that influence their preference rankings.

2.1: Occupancy Rate



2.2: Share of private facilities

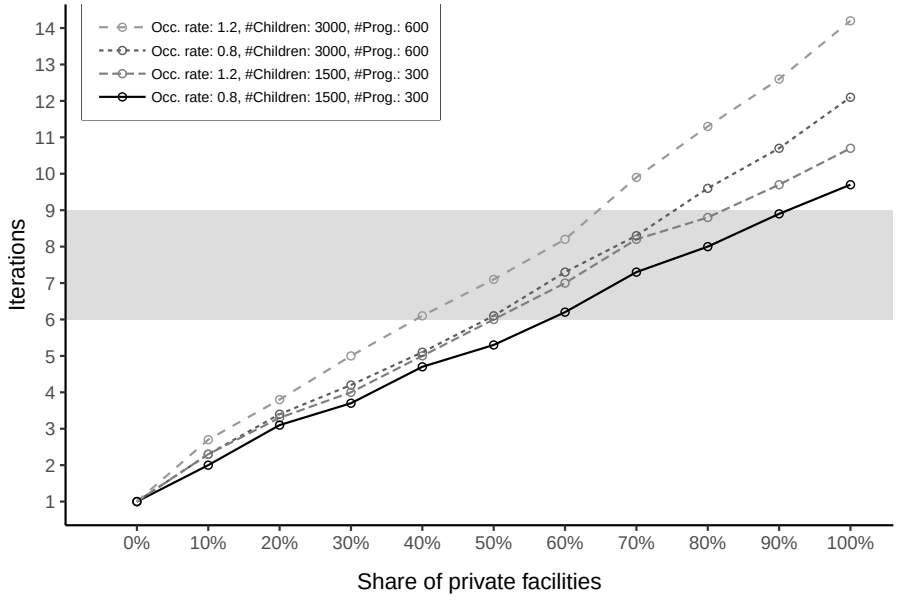
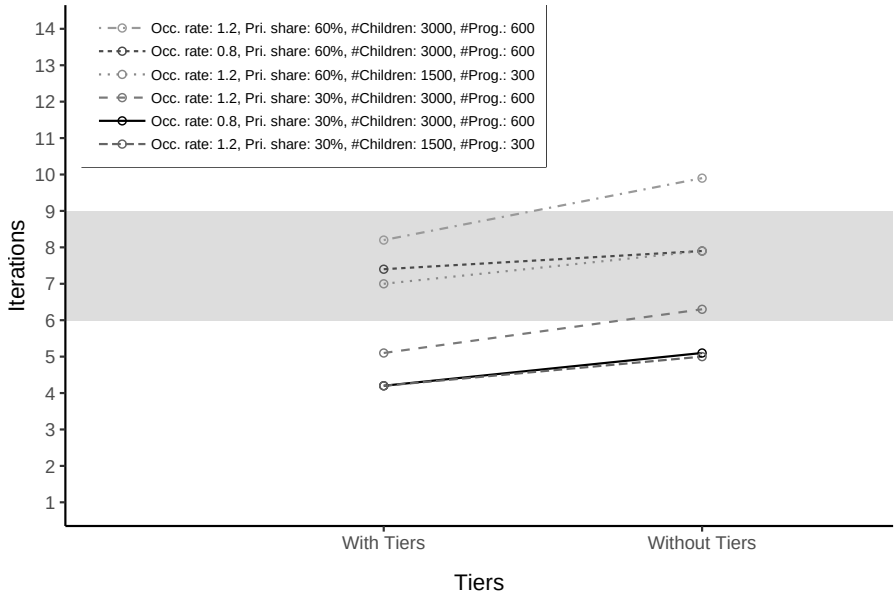


Figure 2: Effects of different parameters on the number of iterations.

2.3: Tiers, with or without



2.4: Threshold

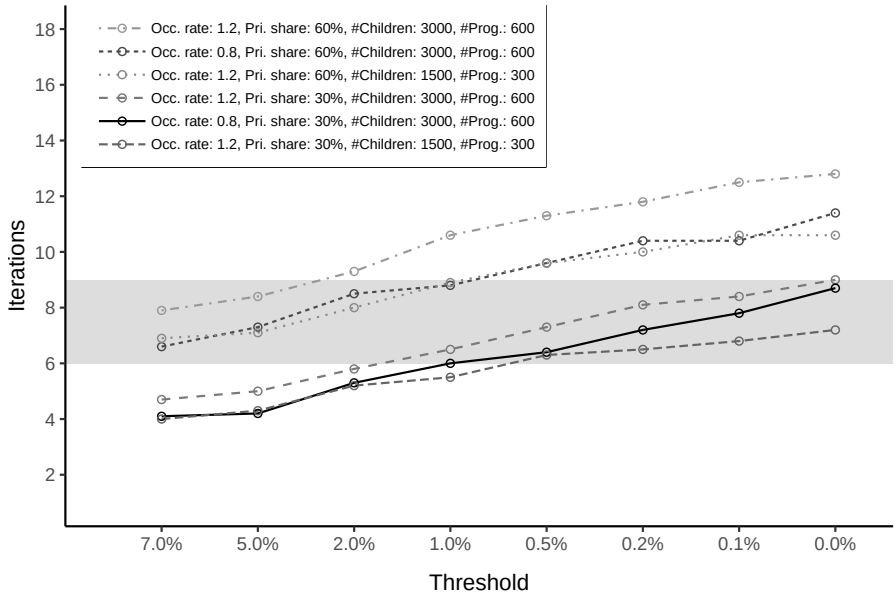
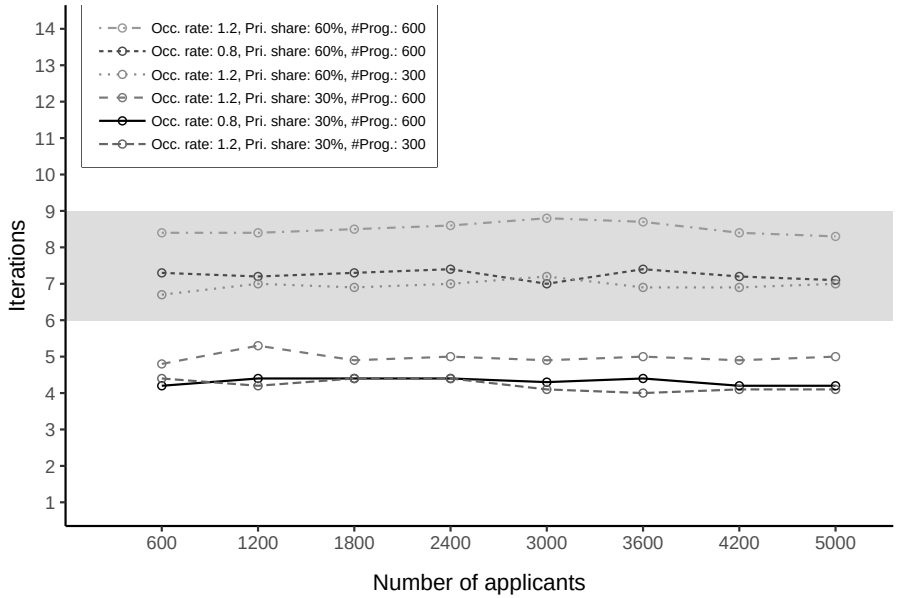


Figure 2: Effects of different parameters on the number of iterations (cont.).

2.5: Number of applicants



2.6: Length of ranking list

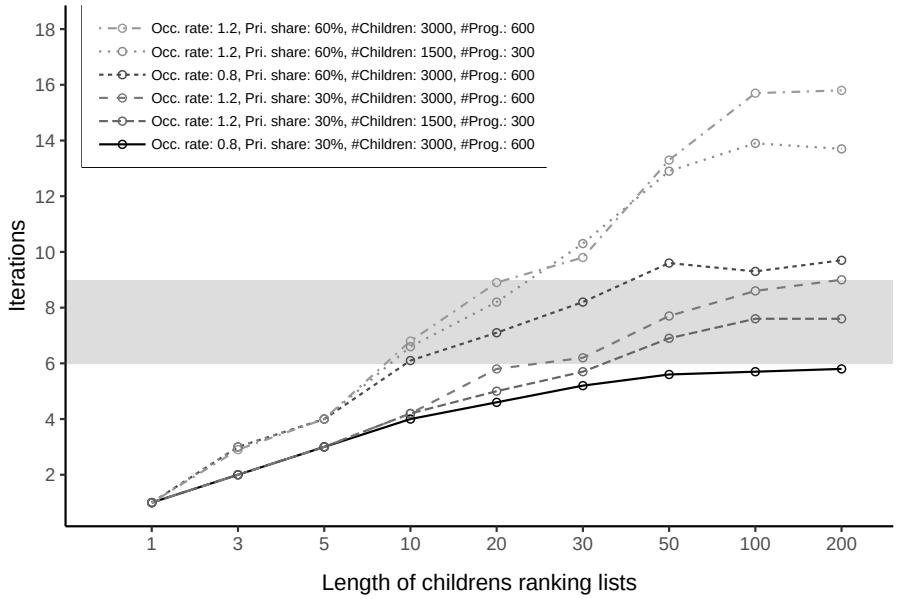
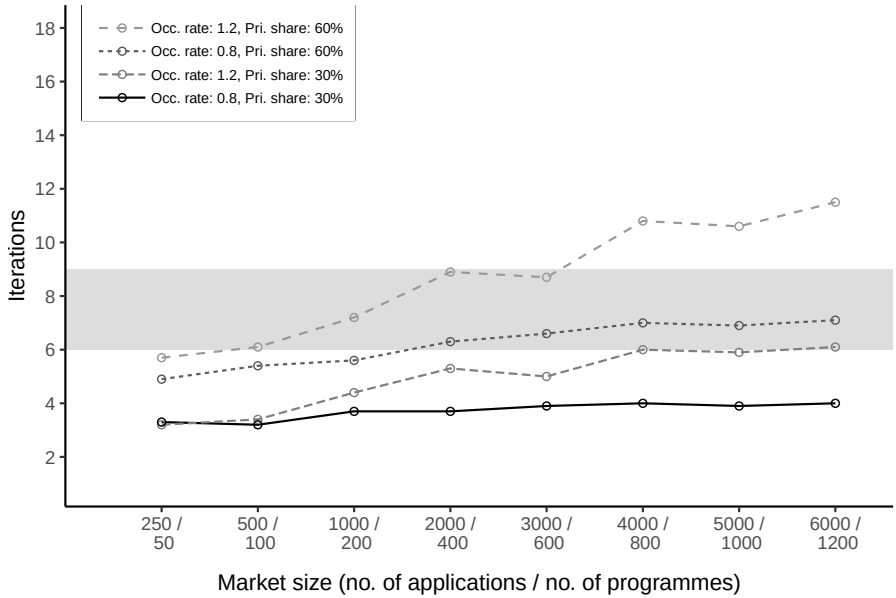


Figure 2: Effects of different parameters on the number of iterations (cont.).

2.7: Market size



2.8: Preference scenarios

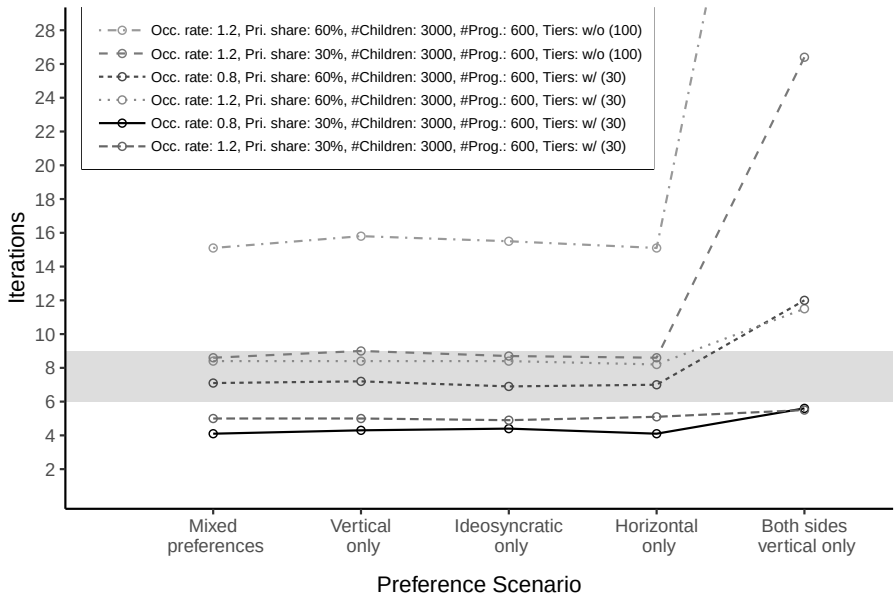


Figure 2: Effects of different parameters on the number of iterations (cont.).

sides of the market only exhibit vertical preferences seems unlikely.

6. CONCLUSION

In this paper, we analyzed the challenges attendant to a redesign of the childcare market in German cities, including in particular the requirements that the allocation mechanism needs to fulfill. In a case study, we identified fairness and speed as two significant issues in current allocation practice. While the matching literature has successfully addressed various matching problems, it is not perfectly applicable to the German childcare market, for the reasons discussed in the foregoing. To solve these issues, we introduced a deferred acceptance mechanism with ties (DAT), which is tailored to the challenges we identified and which ensures a decentralized, fast, and fair allocation process.

The applicability of the new mechanism was also tested in two German cities. Furthermore, we performed a simulation of the mechanism in different market settings, showing which market characteristics affect its performance. We find that while the new mechanism appears to be applicable to various markets, it has to be adjusted to specific market needs.

Although our solution certainly solves problems shared by many German cities, other cities might have additional requirements that need to be incorporated into the mechanism's design. One such requirement is moving from an annual to a monthly allocation schedule, which adds strategic issues if parents postpone their applications to secure a better placement. In the end, the successful implementation of a redesigned mechanism is dependent on local policy and stakeholder support for reform. Further cooperative activities with other cities for the future implementation of the mechanism are already underway.

They could select from a set of eight preselected characteristics (e.g. facility distance, quality, opening hours) and were asked to provide a ranking for them. The parents could indicate a maximum of 6 characteristics and were also able to provide other characteristics using a free text field. The distance from home was selected by 79% of all parents as a relevant characteristic. Also, 50% of all parents ranked it at first or second place in 50% of the cases. Distance to work and trip to work scored significantly lower, with only 23% and 35% of parents mentioning it as an influencing characteristic.

References

Ágoston, K. C., Biró, P., & Szántó, R. (2018). Stable project allocation under distributional constraints. *Operations Research Perspectives*, *5*, 59–68.

Aldershof, B., & Carducci, O. M. (1996). Stable matchings with couples. *Discrete Applied Mathematics*, *68*(1-2), 203–207.

Aygün, O., & Turhan, B. (2020). Dynamic reserves in matching markets. *Journal of Economic Theory*, 105069.

Azevedo, E. M., & Budish, E. (2018). Strategy-proofness in the large. *Review of Economic Studies*, *86*(1), 81–116.

Biró, P. (2017). Applications of matching models under preferences. In U. Endriss (Ed.), *Trends in Computational Social Choice.* COST: European Cooperation in Science and Technology.

Bó, I., & Hakimov, R. (2016). Iterative versus standard deferred acceptance: Experimental evidence. *Working paper*, *SP II 2016-209*.

Braun, S., Dwenger, N., & Kübler, D. (2010). Telling the truth may not pay off: An empirical study of centralized university admissions in Germany. *BE Journal of Economic Analysis & Policy*, *10*(1).

Budish, E., & Cantillon, E. (2012). The multi-unit assignment problem: Theory and evidence from course allocation at Harvard. *American Economic Review*, *102*(5), 2237–71.

Bös, N. (2017). Raus aus der Kita-Warteschlange. *Frankfurter Allgemeine Zeitung*. Retrieved 2021-10-10, from https://www.faz.net/aktuell/wirtschaft/kinderbetreuung-raus-aus-der-kita-warteschlange-15053793.html

Carlsson, S., & Thomsen, S. (2014). Nicht ausgeschöpfte Potenziale in der Kita-Platzvergabe. *Vierteljahrshefte zur Wirtschaftsforschung*, *83*(1), 183–198.

Chen, Y., & Kesten, O. (2017). Chinese college admissions and school choice reforms: A theoretical analysis. *Journal of Political Economy*, *125*(1), 99–139.

Delacrétaz, D., Kominers, S. D., & Teytelboym, A. (2016). Refugee resettlement. *Working paper*.

Drummond, J., Perrault, A., & Bacchus, F. (2015). SAT is an effective and complete method for solving stable matching problems with couples. *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Echenique, F., Wilson, A. J., & Yariv, L. (2016). Clearinghouses for two-sided matching: An experimental study. *Quantitative Economics*, *7*(2), 449–482.

Erdil, A., & Ergin, H. (2008). What's the matter with tie-breaking? Improving efficiency in school choice. *American Economic Review*, *98*(3), 669–89.

Erdil, A., & Ergin, H. (2017). Two-sided matching with indifferences. *Journal of*

Economic Theory, *171*, 268–292.

Gehlke, A., Hachmeister, C.-D., Hüning, L., & de Vries, L. (2017). Der CHE Numerus Clausus-Check 2017/18. Eine Analyse des Anteils von NC-Studiengängen in den einzelnen Bundesländern. *CHE Centre for Higher Education*.

Geitle, A., Johnsen, Ø., Ruud, H., Fagerholt, K., & Julsvoll, C. (2020). Kindergarten allocation in Norway: An integer programming approach. *Journal of the Operational Research Society*, 1–10.

Gonczarowski, Y. A., Nisan, N., Kovalio, L., & Romm, A. (2019). Matching for the Israeli "Mechinot" gap-year programs: Handling rich diversity requirements. *Proceedings of the 2019 ACM Conference on Economics and Computation*, 321–321.

Gong, B., & Liang, Y. (2017). A dynamic college admission mechanism in Inner Mongolia: Theory and experiment. *Working paper*.

Grenet, J., He, Y., & Kübler, D. (2019). Decentralizing centralized matching markets: Implications from early offers in university admissions. *WZB Discussion Paper*.

Hafalir, I. E., Yenmez, M. B., & Yildirim, M. A. (2013). Effective affirmative action in school choice. *Theoretical Economics*, *8*(2), 325–363.

Hatfield, J. W., & Milgrom, P. R. (2005). Matching with contracts. *American Economic Review*, *95*(4), 913–935.

Hitsch, G. J., Hortaçsu, A., & Ariely, D. (2010). Matching and sorting in online dating. *American Economic Review*, *100*(1), 130–63.

Kamada, Y., & Kojima, F. (2018). Stability and strategy-proofness for matching with constraints: A necessary and sufficient condition. *Theoretical Economics*, *13*(2), 761–793.

Kamada, Y., & Kojima, F. (2020). Fair matching under constraints: Theory and applications. *Working paper*.

Kennes, J., Monte, D., & Tumennasan, N. (2014). The day care assignment: A dynamic matching problem. *American Economic Journal: Microeconomics*, *6*(4), 362–406.

Klein, T. (2021). Analysis of stable matchings in R: Package matchingMarkets. Vignette to R package matchingMarkets. *The Comprehensive R Archive Network*.

Klein, T., & Herzog, S. (2018). Matching practices for childcare in Germany. Retrieved 2021-10-10, from http://www.matching-in-practice.eu/matching-practices-for-childcare-germany/

Klijn, F., Pais, J., & Vorsatz, M. (2019). Static versus dynamic deferred acceptance in school choice: Theory and experiment. *Games and Economic Behavior*, *113*, 147–163.

Kojima, F. (2012). School choice: Impossibilities for affirmative action. *Games and Economic Behavior*, *75*(2), 685–693.

Kojima, F., & Manea, M. (2010). Axioms for deferred acceptance. *Econometrica*, *78*(2), 633–653.

Kojima, F., Pathak, P. A., & Roth, A. E. (2013). Matching with couples: Stability and incentives in large markets. *Quarterly Journal of Economics*, *128*(4), 1585–1632.

Kominers, S. D., & Sönmez, T. (2016). Matching with slot-specific priorities: Theory. *Theoretical Economics*, *11*(2), 683–710.

Konegen-Grenier, C. (2018). Wer bekommt einen Studienplatz? Die Regelung des Hochschulzugangs im Umbruch. *IW-Report*.

Mennle, T., & Seuken, S. (2017). Trade-offs in school choice: Comparing deferred acceptance, the classic and the adaptive Boston mechanism. *Working paper*.

Nguyen, T., & Vohra, R. (2019). Stable matching with proportionality constraints. *Operations Research*, *67*(6), 1503–1519.

OVG NRW. (2017). Entscheidung des Oberverwaltungsgericht Nordrhein-Westfalen, 12 B 930/17.

Parkes, D. C., & Ungar, L. H. (2000). Iterative combinatorial auctions: Theory and practice. *17th National Conference on Artificial Intelligence*, 74-81.

Pathak, P. A., & Sönmez, T. (2008). Leveling the playing field: Sincere and sophisticated players in the Boston mechanism. *American Economic Review*, *98*(4), 1636–52.

Press and Information Office of the City of Münster. (2019). 3697 Kinder im Kita-Navigator vorgemerkt. *Pressemitteilung der Stadt Münster*. Retrieved 2021-10-10, from https://www.muenster.de/stadt/presseservice/pressemeldungen/web/frontend/index.php?show=1010440

Roth, A. E. (1982). The economics of matching: Stability and incentives. *Mathematics of Operations Research*, *7*(4), 617–628.

Roth, A. E., & Peranson, E. (1999). The redesign of the matching market for American physicians: Some engineering aspects of economic design. *American Economic Review*, *89*(4), 748–780.

Roth, A. E., & Sotomayor, M. (1992). Two-sided matching. *Handbook of Game Theory with Economic Applications*, *1*, 485–541.

Sönmez, T., & Yenmez, M. B. (2020). Affirmative action with overlapping reserves. *Working paper*.

Veski, A., Biró, P., Pöder, K., & Lauri, T. (2017). Efficiency and fair access in kindergarten allocation policy design. *Journal of Mechanism and Institution Design*, *2*(1), 57–104.

Völker, K. (2018). Verdruss bei der Kitaplatz-Vergabe: Eltern drohen mit weiteren Klagen. *Westfälische Nachrichten*. Retrieved 2021-10-10, from https://www.wn.de/Muenster/3203115-Verdruss-bei-der-Kitaplatz-Vergabe-Eltern-drohen-mit-weiteren-Klagen

Appendix

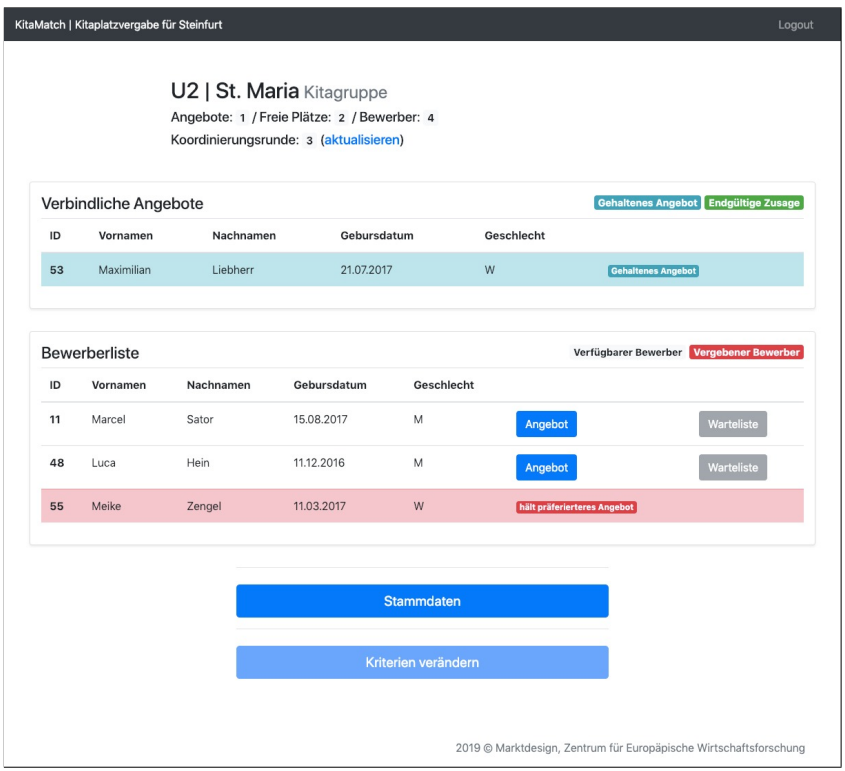


Figure 3: A screenshot of the Kitamatch application. In this view, childcare providers can register offers, they see the matches and rejections per round, and can modify their criteria catalog. The data presented here are anonymized.

Round	Age cohort	Program	Skipped child ID	Preferred child ID	Reason	# Skipped
2	U2	46	803	677	gender	3
2	U2	25	515	603	–	1
2	U2	28	522	500	–	3
2	U2	43	803	537	–	4
2	U2	46	803	746	–	8
2	U2	49	500	511	–	2
2	2	65	210	297	gender	3
2	O3	33	33	56	gender	5
2	O3	33	33	37	gender	2
2	O3	45	12	33	gender	1
2	O3	12	88	103	–	1
2	O3	12	88	826	–	2
2	O3	75	831	48	–	1
2	O3	75	831	145	–	12
3	2	41	359	454	gender	2
3	2	65	359	381	–	2
3	O3	24	815	54	age	2
4	O3	24	815	836	age	1
4	O3	24	815	132	age	2

Table 3: List of deviations of programs' sequence of offers (revealed preferences) from the pre-sorted ranking implied by the criteria catalog for the city of Greven. For each deviation, the table reports the ID of the first skipped child on the presorted ranking and the child ID that was preferred by the program. It also gives the reason and the total number of children skipped. Of the 19 deviations, 6 are to achieve gender balance and 3 to achieve heterogeneity in age.

Age cohort	Program	Excess capacity	Adjusted capacity
U2	52	1	2
2	29	1	6
2	35	2	15
2	41	1	7
2	56	1	16
2	62	1	6
O3	24	1	13
O3	42	1	1
O3	75	6	15

Table 4: Capacity adjustment for Greven. For administrative purposes, in the Greven match, nine programs deliberately stated a capacity in excess of their actual vacancies. The capacities entered by programs were thus adjusted by subtracting these excess capacities.

Journal of Mechanism and Institution Design
ISSN: 2399-844X (Print), 2399-8458 (Online)
DOI:10.22574/jmid.2021.12.004

ON THE DEGREE OF DISTORTIONS UNDER SECOND-DEGREE PRICE DISCRIMINATION

Ram Orzach

Oakland University, USA
orzach@oakland.edu

Miron Stano

Oakland University, USA
stano@oakland.edu

ABSTRACT

This paper highlights the limitations and applicability of results developed by Chao & Nahata (2015) for nonlinear pricing. Although Chao and Nahata appear to provide necessary and sufficient conditions for general utility functions, we show that one of their results leads only to a restatement of two constraints, and another result may not be valid when consumers can freely dispose of the good. Their model allows for the possibility that higher quantities will have a lower price than smaller quantities. We provide conditions under free disposal that preclude this anomaly. Our analysis suggests that further research on violations of the single-crossing condition should be encouraged.

Keywords: Second-degree price discrimination, nonlinear pricing, single-crossing condition.

JEL Classification Numbers: D42, D61, D82, L12.

1. INTRODUCTION

An extensive literature builds on Maskin & Riley's (1984) characterization of nonlinear monopoly pricing.[1] Maskin and Riley consider the general

We are grateful to the editor and two anonymous referees for their many helpful comments and suggestions. The usual disclaimer applies.

[1] Under nonlinear pricing, also known as a nonlinear tariff, a seller with monopoly power price discriminates by offering a menu of bundles with varying quantities and prices where the

case involving N types of consumers with differing demand functions where the monopoly seller cannot identify the individual's consumer type. The technique to maximize profits under nonlinear pricing, for the simplified example of two consumer types, is one where the firm offers two bundles of quantity and price that satisfy the incentive compatibility (IC) and individual rationality (IR) constraints. The IC constraints require that neither type will pretend to be the other type and buy the other's quantity; while IR requires that consumers buy the product only if their utilities from the exchange are non-negative.

Tirole (1988)(p.149) provides a succinct description of the relevant Maskin-Riley results for two consumer types: low demand and high demand. The firm will have an optimal tariff where: i) the low demand consumer has zero surplus while the high demand has a positive one; ii) the high demand consumer will be indifferent between buying his own designated bundle and the low one (therefore, he will buy his own); and iii) the high demand type buys a quantity that corresponds to marginal cost, while the low demand consumer buys the designated quantity that is less than the one corresponding to marginal cost.

Those results are achieved under the assumption of monotonicity in utilities, a property that is known as the Sorting Condition, the Spence-Mirrlees Condition, or the Single-Crossing Condition (SCC).[2] We will use SCC because the term is more revealing in that the utility functions cross only once. The SCC will be satisfied if the demand curves lie one above the other, i.e., if, at any given price, the quantity for the high demand consumer will be larger than that of the low demand consumer. In cases where the demand curves cross, as in Figure 1 in Section 2.3, the SCC may be violated.

Although applications that satisfy the SCC dominate the nonlinear pricing literature, analyses of nonlinear tariffs when the SCC is violated have appeared. Andersson (2005) develops conditions under which relaxation of the SCC will still produce the Maskin-Riley results, but Andersson (2008) further provides a counterexample to the main Maskin-Riley result. In his counterexample, the quantities for both types can correspond to those under marginal-cost pricing with zero gains for both consumer types. Chao & Nahata (2015) (hereafter CN) also consider a violation of the SCC to determine the regions in which

per-unit price typically diminishes as the bundle quantity increases. These bundles are often seen as providing quantity discounts. Nonlinear pricing has long been a fixture for many products and services.

[2] See Tirole (1988)(p.148, fn 28). Within the context of this paper, any distinctions among the three conditions are not relevant.

the quantities correspond to the efficient quantities (marginal-cost pricing) or exceed (fall short) of the efficient quantity. Our work will concentrate only on the case where the two quantities correspond to marginal cost.

CN appear to provide necessary and sufficient conditions for general utility functions. After describing their model in Section 2.1, Section 2.2 shows that their Proposition 2(ii) follows from a well-known mathematical method that leads only to a restatement of two constraints. Furthermore, CN's Proposition 3 provides conditions for efficient quantities, but we use a numerical example in Section 2.3 to illustrate that a portion of the proposition will have limited economic relevance when consumers can freely dispose of the good. In particular, CN's model can generate a pricing scheme where the larger quantity has a lower price than the smaller quantity. With Proposition 3 as a special case of Proposition 2, the same anomaly will also apply to Proposition 2.

To deal with this limitation, Section 3 develops an alternative to CN's Proposition 3(ii) where consumers can freely dispose of the good. In that section, we impose the conditions on the demand functions that would enable the firm to offer two menu quantities and corresponding prices so that profits are maximized and the consumer surplus is zero. The conditions also guarantee that the larger quantity will always have a higher (or equal) price than the smaller quantity a result that does not necessarily hold in the CN model.

2. ARGUMENTS

2.1. The CN Model

A monopolist serves two consumer types: Type i utility from quantity q is $u_i(q)$, $i \in \{1, 2\}$. The seller cannot *a priori* distinguish between the types but knows their utilities and ratio $\gamma = n_2/n_1$ where n_i is the number of Type i consumers. Under the Revelation Principle, the monopolist can maximize profits by using nonlinear tariffs with (q_i, T_i), $i \in \{1, 2\}$, where T_i is the tariff for quantity q_i. The firm has a linear cost function with marginal cost $c \geq 0$. CN state that, without any loss of generality, the tariffs and the utilities can be normalized by the marginal cost c so that $t_i \equiv T_i - c \cdot q_i$, and $v_i(q) \equiv u_i(q) - c \cdot q$.

With t_i as CN's net-of-cost tariff and $v_i(q)$ as their net-of-cost valuation, the model assumes that $v_i(q)$ has a unique optimization: $q_i^e \equiv \arg\max_q v_i(q) \in (0, +\infty)$, where q_i^e is their efficient quantity for the Type i. The following is

CN's constrained maximization problem (p. 209).

$$\begin{array}{lll} \max_{(t_i,q_i),i=1,2} & t_1 + \gamma \cdot t_2 & ([\mathrm{P}]) \\ \text{s.t.} & v_i(q_i) - t_i \geq v_i(q_j) - t_j,\ i,\ j = 1,\ 2 & (\mathrm{IC}_i) \\ & v_i(q_i) - t_i \geq 0,\ i = 1,\ 2 & (\mathrm{IR}_i) \end{array}$$

Although not defined by CN, IC and IR represent the incentive compatibility and individual rationality constraints described in our introduction. IC requires that neither consumer type will pretend to be the other type and buy the other's quantity; while IR requires that each type buy his respective bundle only if the utility resulting from the purchase is non-negative. With marginal cost normalized to zero, the objective function ($[P]$), also not defined in CN, is equivalent to total profit divided by the number of Type 1 consumers, i.e., multiplying $[P]$ by $n1$ results in a profit function that is further described in our Section 3. CN also define $\widehat{q}$ as the quantity where $v_1(\widehat{q}) = v_2(\widehat{q})$ (see Proposition 1, p. 209).

2.2. CN's Proposition 2

CN's Proposition 2, found on p. 209, states:
(i) (**The higher peak type gets efficient quantity**) $q_i^* = q_i^e$, where $i = \arg\max_j v_j(q_j^e)$.
(ii) (**A sufficient and necessary condition for overall efficiency**) $q_i^* = q_i^e$ and $t_i^* = v_i(q_i^e)$ $(i = 1, 2)$ if and only if $v_i(q_i^e) \geq v_j(q_i^e)$ $(i, j = 1, 2$ and $i \neq j)$.

As our concern is with the second part of this proposition, the following provides CN's proof of (ii).
(ii) (if Part) When $v_i(q_i^e) \geq v_j(q_i^e)$ $(i, j = 1, 2$ and $i \neq j)$, offering a menu $T = \{(q_i^e, v_i(q_i^e))\}$ $(i = 1, 2)$ satisfies all constraints and extract all surplus from consumers.
(Only if Part) When $q_i^* = q_i^e$ and $t_i^* = v_i(q_i^e)$ $(i = 1, 2)$, this part follows from IC_i $(i = 1, 2)$

2.2.1. Our Argument with Proposition 2(ii)

We will claim that (ii) is equivalent to the following: maximize the profit for each type separately (which is straightforward) and then check if the ICs hold. If they hold, the quantities form the solution.

To elaborate, optimization for the firm occurs when quantity for each type corresponds to marginal cost (as CN normalize the cost to zero, it is q_i^e), and the firm captures the entire consumer surplus, i.e., $t_i^* = v_i(q_i^e)$, $i \in \{1, 2\}$. It remains to be determined whether the ICs hold. However, the condition under Proposition 2 that $v_i(q_i^e) \geq v_j(q_i^e)$ is exactly the IC when cost normalizes to zero. To see our claim, reduce both sides of the inequality with the price t_i^* so that: $v_i(q_i^e) - t_i^* \geq v_j(q_i^e) - t_i^*$ or $0 \geq v_j(q_i^e) - t_i^*$. If Type j deviates to the quantity of Type i, then his gain is negative. This negative inequality is not more restrictive than the IC. Consider that the gain of Type j from his quantity q_j^e is zero as $t_j^* = v_j(q_j^e)$. Substitute the zero on the left-side of the previous inequality so that $v_j(q_j^e) - t_j^* \geq v_j(q_i^e) - t_i^*$. Therefore, $v_i(q_i^e) \geq v_j(q_i^e)$ is equivalent to $v_j(q_j^e) - t_j^* \geq v_j(q_i^e) - t_i^*$ or IC_j (IC_i in CN).

Under established mathematical practice, if the objective function is separated by some parameters but some of the constraints are not, one can solve for each set of parameters that are separated in the objective function, and then check if the mixed constraints hold. Consequently, CN′s short proof on their p. 209 states only that the constraints hold and it is left to the reader to determine whether their claim to "provide a simple necessary and sufficient condition for overall efficiency" is actually informative. In fact, the following section illustrates that normalizing cost to zero greatly simplifies the ICs but the simplification may lead to an anomaly.

2.3. CN's Proposition 3

For Proposition 3, CN introduce two general quadratic evaluation functions: $v_1(q) = q \cdot (1 - b \cdot q/2)$ and $v_2(q) = q \cdot (\alpha - b\beta \cdot q/2)$, that are equivalent to two linear inverse demands: $p_1(q) = 1 - b \cdot q$ and $p_2(q) = \alpha - b\beta \cdot q$. For the two demands to cross, they restrict $0 < \beta < \alpha < 1$, and derive $q_1^e = 1/b < q_2^e = \alpha/b\beta$, $v_1(q_1^e) = 1/(2b)$ and $v_2(q_2^e) = \alpha^2/(2b\beta)$. The v_i functions cross once at $\widehat{q} = \frac{2}{b}\frac{1-\alpha}{1-\beta}$.

CN's Proposition 3, found on p. 210, states:
(i) (**Who gets efficient quantity**) If $\alpha^2 \leq \beta$ then $q_1^* = q_1^e$; if $\alpha^2 \geq \beta$ then $q_2^* = q_2^e$.
(ii) (**Overall efficiency**) $q_i^* = q_i^e$ and $t_i^* = v_1(q_i^e)$ $i = 1, 2$ if and only if $q_1^e \leq \widehat{q} \leq q_2^e$.
(iii) (**Oversizing**) $q_1^* = q_1^e$ and $q_2^* > q_2^e$ if and only if $q_1^e < q_2^e < \widehat{q}$.
(iv) (**Undersizing**) $q_1^* < q_1^e$ and $q_2^* = q_2^e$ if and only if $\widehat{q} < q_1^e < q_2^e$.

As we will introduce an example that relates only to part (ii) of this proposition, the proof of parts (i) and (ii) follow directly from the corresponding parts of Proposition 2.

2.3.1. Our Argument with Proposition 3(ii)

We claim that Proposition 3(ii) can produce an example that defies economic sense. Begin with footnote 7 in CN's Conclusion: "Many convenience stores located along the US interstate highways (catering primarily to drivers on the go) charge more for a smaller cup of coffee than the larger one" (pp. 212-213). In the example we introduce below that is based on the CN model, the firm sets a price of 2 for 4 units and 1.25 for 10 units. We also explain how this paradox can be created in a simple CN model that includes IC.

We concentrate on CN's part (ii) above: (**Overall efficiency**) $q_i^* = q_i^e$ and $t_i^* = v_i(q_i^e)$, $i = 1, 2$ if and only if $q_1^e \leq \widehat{q} \leq q_2^e$.

Consider first the case where $c = 0$. Let $p_1(q) = 1 - 0.25 \cdot q$ and $p_2(q) = 0.25 - 0.025 \cdot q$, so that $b = 0.25$, $\alpha = 0.25$, $\beta = 0.1$. It is easy to see that their condition is satisfied.

$$\begin{aligned} q_1^e \;&\leq\; \widehat{q} \;\leq\; q_2^e \quad \text{as } q_1^e = 1/b = 4 \text{ and } q_2^e = \alpha/b\beta = 10 \\ 4 \;&\leq\; \widehat{q} \;\leq\; 10 \quad \text{as } \widehat{q} = \frac{2}{0.25}\left(\frac{1-0.25}{1-0.1}\right) = 6\frac{2}{3} \end{aligned}$$

It is also easy to verify (see Figure 1) that the prices are: $T_1 = t_1 = v_1(q_1^e) = 2$ for $q_1 = 4$, and $T_2 = t_2 = v_2(q_2^e) = 1.25$ for $q_2 = 10$.

The utility of Type 1 equals the utility of Type 2 at $\widehat{q} = 6\frac{2}{3}$ meaning that it is in the negative price territory for Type 1. The solution to the puzzle is that these prices will not be mathematically correct unless there is an increasing disposal cost at a rate of 0.25 per unit for Type 1 so that $\widehat{q} = 6\frac{2}{3}$ with total disposal cost of 0.89. At $\widehat{q}$, $v_1(6\frac{2}{3}) = v_2(6\frac{2}{3}) = 1.11$.

This example shows that CN's Proposition 3(ii) depends on the inverse demand function for Type 1 as having the same slope in the negative price region. CN's deviation from conventional practice explains why Type 1's IC is so easily mathematically satisfied.

The coffee example, previously described as part of the long footnote 7 in their Conclusion on pp. 212-213 suggests that CN were aware of this abnormality. Although we have not observed this phenomenon for the same product, e.g., regular brewed coffee as opposed to regular vs. espresso, one

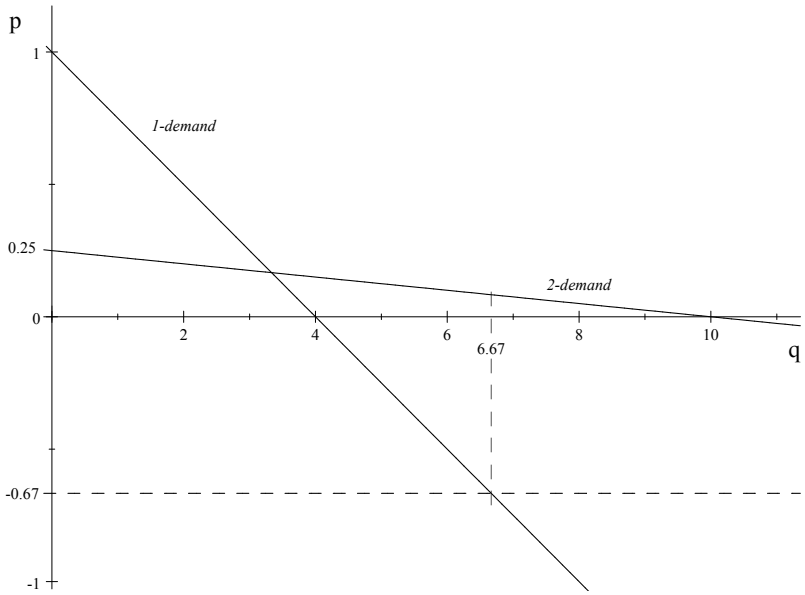


Figure 1: Valuations for the two consumers are equal at 6.67 units.

can reasonably assume that the disposal cost of coffee is zero.[3] Thus, when the negative territory is excluded, the theoretical solution will be harder but more relevant to most economic applications.

We argue that the CN model is even more limited. By normalizing the cost to be zero, their model cannot distinguish between the negative and positive regions. To illustrate, let the cost of production in our example be $c = 0.1$. It is easy to show that the prices of 4 and 10 units are 2.4 and 2.25 respectively. The anomaly remains.[4] However when $c = 1$, the anomaly disappears (6 for 4 units and 11.25 for 10 units).

[3] CN claim that there are "many" such pricing schemes, but the prices of the smaller and larger quantities in the two additional examples described in their footnote 7 are the same. We agree that such cases are not unusual but CN's two additional examples do not fit their model or explain the anomaly.

[4] As CN derive Proposition 3 specifically for two crossing linear demand curves, this anomaly will also apply to Proposition 2 that deals with general demand functions.

3. RESULTS UNDER FREE DISPOSAL

Consider CN's model only for their case where the marginal cost is zero, i.e., $c = 0$, as this will allow us to maintain the previous notation. The model has two crossing linear demand functions: $p_1 = 1 - bq$ for the Type 1 consumers and $p_2 = \alpha - b\beta q$ for the Type 2 consumers, where $0 < \beta < \alpha < 1$ (see Figure 2). There are n_t Type t consumers, $t \in \{1, 2\}$.

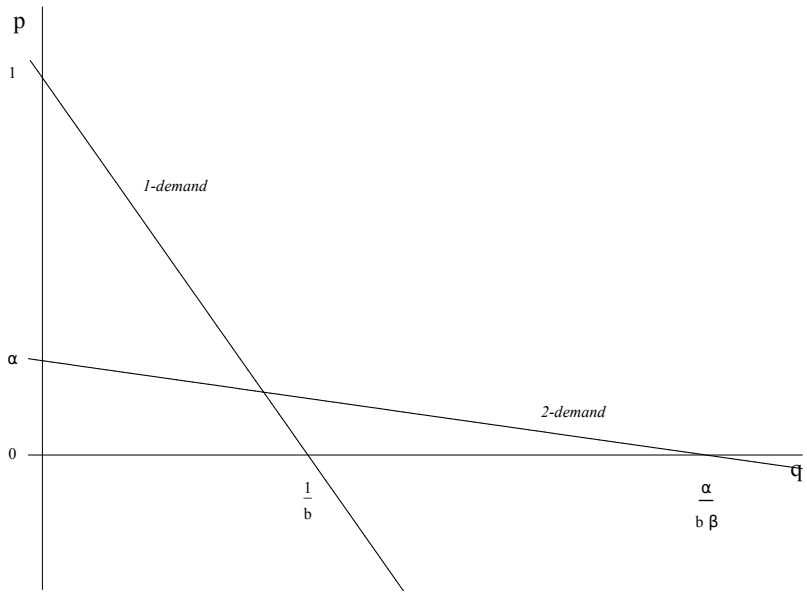


Figure 2: The model has two crossing linear demand functions: $p_1 = 1 - bq$ for the Type 1 consumers and $p_2 = \alpha - b\beta q$ for the Type 2 consumers, where $0 < \beta < \alpha < 1$.

CN provide valuation functions, $v_1(q) = q(1 - bq/2)$ and $v_2(q) = q(\alpha - b\beta q/2)$, that will correspond to utility when $c = 0$. If we are restricted to positive prices (i.e., no disposal costs), then $v_1^+(q)$ and $v_2^+(q)$ can be defined as:

$$v_1^+(q) = \begin{cases} v_1(q) = q(1 - \frac{bq}{2}) & \text{if } q < \frac{1}{b} \\ v_1(\frac{1}{b}) = \frac{1}{2b} & \text{if } q \geq \frac{1}{b} \end{cases}$$

$$v_2^+(q) = \begin{cases} v_2(q) = q(\alpha - \frac{b\beta q}{2}) & \text{if } q < \frac{\alpha}{b\beta} \\ v_2(\frac{\alpha}{b\beta}) = \frac{\alpha^2}{2b\beta} & \text{if } q \geq \frac{\alpha}{b\beta} \end{cases}$$

Under the Revelation-Principle (Myerson, 1979), for any Bayesian equilibrium defined by a game, there exists a Bayesian equilibrium of an incentive compatible mechanism that will yield the same payoff. We can thus concentrate only on the case where the seller offers menu prices $T(q_t)$ for the quantities q_t, $t \in \{1,2\}$.

Let $v_t^+(q_j)$ be the utility of Type t buyers from consuming q_j; $t \in \{1,2\}$, $j \in \{1,2\}$. The firm's objective function is to maximize its profit subject to individual rationality constraints and the incentive compatibility constraints.

$$\begin{array}{lll} \max_{T(q_t)} & \pi = n_1 \cdot T(q_1) + n_2 \cdot T(q_2) & \\ \text{s.t.} & v_1^+(q_1) - T(q_1) \geq 0 & \text{IR}_1 \\ & v_2^+(q_2) - T(q_2) \geq 0 & \text{IR}_2 \\ & v_1^+(q_1) - T(q_1) \geq v_1^+(q_2) - T(q_2) & \text{IC}_1 \\ & v_2^+(q_2) - T(q_2) \geq v_2^+(q_1) - T(q_1) & \text{IC}_2 \end{array}$$

where π represents the profit as the cost function is linear and marginal cost $c = 0$. As before, IR_i are the individual rationality constraints, and IC_i are the incentive compatibility requirements, $i \in \{1,2\}$.

Following the arguments described in Section 2.2.1, to maximize π, the firm should maximize the profit for each type and then check if the ICs hold. The firm should thus set the quantity for each type that corresponds to marginal cost which here is 0 (CN label these quantities as q_i^e). This strategy enables the firm to capture the maximum consumer surplus from both types at quantities, $q_1^e = \frac{1}{b}$ and $q_2^e = \frac{\alpha}{b\beta}$ with prices $T(\frac{1}{b}) = v_2^+(\frac{1}{b}) = 1/(2b)$ and $T(\frac{\alpha}{b\beta}) = v_2^+(\frac{\alpha}{b\beta}) = \alpha^2/(2b\beta)$. As $T(q_i^*) = v_i^+(q_i^*)$, the first two constraints (IR_1 and IR_2) are satisfied. It remains to be determined for which values of b, α and β the ICs hold. The following develops an alternative to CN's Proposition 3(ii).

Proposition

If $2\alpha - 1 \leq \beta \leq \alpha^2$, then $q_1^* = q_1^e = 1/b$ and $q_2^* = q_2^e = \alpha/(b\beta)$ are the two menu quantities together with prices $T(\frac{1}{b}) = 1/(2b)$ and $T(\frac{\alpha}{b\beta}) = \alpha^2/(2b\beta)$, the firm's profits are maximized and the consumer surplus is zero.

Proof Let us start with IC_1, namely $v_1^+(q_1^*) - T(q_1^*) \geqq v_1^+(q_2^*) - T(q_2^*)$. As $T(q_i^*) = v_i^+(q_i^*)$, the left-hand side is 0. It is left to show that $0 \geqq v_1^+(q_2^*) - T(q_2^*)$, or $0 \geqq v_1^+(\frac{\alpha}{b\beta}) - T(\frac{\alpha}{b\beta})$. As $v_1^+(\frac{\alpha}{b\beta}) = \frac{1}{2b}$ and $T(\frac{\alpha}{b\beta}) = \frac{\alpha^2}{2b\beta}$, then $0 \geqq \frac{1}{2b} - \frac{\alpha^2}{2b\beta} \Longrightarrow \frac{\alpha^2}{\beta} \geqq 1$. Therefore, IC_1 requires that $\alpha^2 \geqq \beta$.

For IC_2, $v_2^+(q_2^*) - T(q_2^*) \geqq v_2^+(q_1^*) - T(q_1^*)$. Again, the left-hand side is 0 so that $0 \geqq v_2^+(q_1^*) - T(q_1^*)$, or $0 \geqq v_2^+(\frac{1}{b}) - T(\frac{1}{b})$. As $v_2^+(\frac{1}{b}) = (q \cdot (\alpha - b\beta \cdot q/2))$ at $q = \frac{1}{b} \Longrightarrow v_2^+(\frac{1}{b}) = \frac{\alpha}{b} - \frac{\beta}{2b}$ and $T(\frac{1}{b}) = \frac{1}{2b}$, this leads to $0 \geqq (\frac{\alpha}{b} - \frac{\beta}{2b}) - \frac{1}{2b}$, or $0 \geqq 2\alpha - \beta - 1$. Therefore, IC_2 requires that $\beta \geqq 2\alpha - 1$.

Note that the proposition is independent of b as CN were able to normalize the quantity, namely, $b = 1$. Therefore, $q_1^* = q_1^e = 1$ and $q_2^* = q_2^e = \frac{\alpha}{\beta}$ with respective prices $T(1) = 0.5$ and $T(\frac{\alpha}{\beta}) = \alpha^2/(2\beta)$.

To more clearly illustrate the role of the ICs, consider extreme values of α, for example, 0.1 and 0.9. If $\alpha = 0.1$, then $0 \leq \beta \leq 0.01$. Choosing $\beta = 0.005$ leads to $q_2^* = 20$ and $T(20) = 1$, i.e., the quantity for Type 2 is 20 times that for Type 1. With $\alpha = 0.1$, $2\alpha - 1$ is negative so that IC_2 becomes irrelevant and IC_1 is the only binding constraint. Consistent with intuition, the price of the larger quantity must be higher than the price of the smaller one (otherwise Type 1 will buy the larger quantity and dispose any amount above 1). As $\alpha = 0.1$, and the price for Type 2 must be higher, $T(q_2^*) = v_i(q_2^*) = T(\frac{\alpha}{\beta}) = \frac{\alpha^2}{2\beta} \geq T(q_1^*) = \frac{1}{2}$, it follows that β must be small, i.e., $\frac{\alpha^2}{2\beta} \geq \frac{1}{2}$.

If $\alpha = 0.9$, then $0.8 \leq \beta \leq 0.81$. Choosing $\beta = 0.805$, it follows that $q_2^* = 1.118$ and $T(1.118) \simeq 0.503$. Unlike the previous example, both ICs determine the range of β. The condition IC_2 ($2\alpha - 1 \leq \beta$) determines that the utility of Type 1 is higher than the utility of Type 2 from the smaller quantity $q_1^* = q_1^e = 1$. Therefore, β must be large as the utility of Type 1 is $v_1^+(1) = T(1) = \frac{1}{2}$, while the utility of Type 2 is $v_2^+(1) = q \cdot (\alpha - b\beta \cdot q/2) = 1 \cdot (0.9 - 1 \cdot 0.805 \cdot 1/2) = 0.4975$. Similar to the previous example, IC_1 will require that the price of $q_2^* = 1.118$ must be more than $T(1) = 0.5$. Therefore, the price $v_2^+(1.118) = T(1.118) \simeq 0.503$.

Figure 3 shows the two IC constraints and the non-negative region between them, labelled as A, where β satisfies: $Max\{0, 2\alpha - 1\} \leq \beta \leq \alpha^2$.

As seen in Figure 3, with $\alpha = 0.5$, the condition $(2 \cdot \alpha - 1) = 0 \leq \beta \leq 0.25 = \alpha^2$ will allow β to have larger ranges between the constraints IC_1 and IC_2. In contrast, as we have explained using $\alpha = 0.1$ (which is far from the intercept of Type 1's demand function in Figure 2), the slope of Type 2's demand function is much more horizontal than that of Type 1, and the profit maximizing quantities will vary greatly between the types. On the other hand, when $\alpha = 0.9$, which is near the intercept for Type 1's demand, the slopes of the two demands are similar, and the profit maximizing quantities will be close.

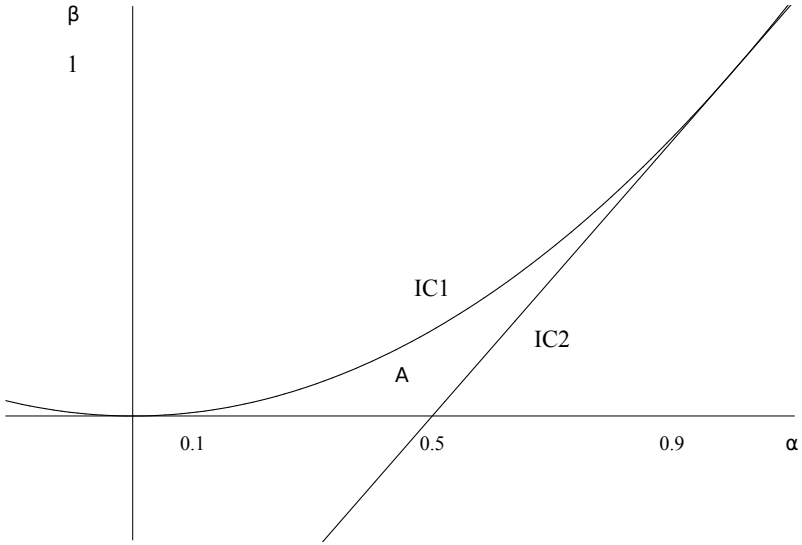


Figure 3: The region A where the incentive compatibility constraints are satisfied.

4. CONCLUSION

Price discrimination, especially in the fast food sector, is of considerable academic and policy interest. Scholars across a range of disciplines have focused on the pricing, marketing and health effects of supersized portions of "junk food".[5] Economic advances in our understanding of this "value-size pricing" through models of price discrimination should, in principle, contribute to the development of more effective policy and regulatory actions to limit consumption of such products.

CN's theoretical approach appears to provide new insights by easily solving the nonlinear tariff challenge for general utility functions. However, given their normalization of cost to zero, their condition for Proposition 2(ii) is just

[5] "Junk food" is often associated with calorie-dense food that has high levels of sugar, glycemic starch, and saturated fat. Vermeer, Steelhuis, and Vermeer et al. (2014) summarize some of the policy interventions to reduce portion sizes. As for strategies where firms charge the same price for any sized beverage (see our footnote 3), Haws et al. (2020) recently examined the effects of this strategy on consumption of soft drinks. In addition to increasing beverage size choices compared to standard pricing and, unlike standard pricing, this strategy nullifies the effectiveness of calorie postings in reducing larger sizes.

a variant of the IC. CN's subsequent restriction to linear demand functions that violate the SCC, while failing to exclude the negative price region, may seriously limit the economic relevance of Propositions 2 and 3. Analyses involving violations of the SCC should not be considered as settled and, as such, further research for these cases is well warranted.

References

Andersson, T. (2005). Profit maximizing nonlinear pricing. *Economics Letters*, *88*(1), 135-139.

Andersson, T. (2008). Efficiency properties of non-linear pricing schedules without the single-crossing condition. *Economics Letters*, *99*(2), 364-366.

Chao, Y., & Nahata, B. (2015). The degree of distortions under second-degree price discrimination. *Economics Letters*, *88*(1), 135–9.

Haws, K., Liu, P., Dallas, S., Cawley, J., & Roberto, C. (2020). Any size for a dollar: The effect of any-size-same-price versus standard pricing on beverage size choices. *Journal of Consumer Psychology*, *30*(2), 392-401.

Maskin, E., & Riley, J. (1984). Monopoly with incomplete information. *RAND Journal of Economics*, *15*(2), 171-196.

Myerson, R. (1979). Incentive-compatibility and the bargaining problem. *Econometrica*, *47*(1), 61-73.

Tirole, J. (1988). *The Theory of Industrial Organization.* Cambridge: MIT Press.

Vermeer, W., Steenhuis, I., & Poelman, M. (2014). Small, medium, large or supersize? The development and evaluation of interventions targeted at portion size. *International Journal of Obesity*, *38*(1), 513-518.

Journal of Mechanism and Institution Design
ISSN: 2399-844X (Print), 2399-8458 (Online)
DOI:10.22574/jmid.2021.12.005

DREAM TEAMS AND THE APOLLO EFFECT

Alex Gershkov

The Hebrew University of Jerusalem, Israel and University of Surrey, UK
alexg@huji.ac.il

Paul Schweinzer

Alpen-Adria-Universität Klagenfurt, Austria
paul.schweinzer@aau.at

ABSTRACT

We model leadership selection, competition, and decision making in teams with heterogeneous membership composition. We show that if the choice of leadership in a team is imprecise or noisy—which may arguably be the case if appointment decisions are made by non-expert administrators—then it is not necessarily the case that the best individuals will be selected as team members. On the contrary, and in line with what has been called the "Apollo effect," a "dream team" consisting of unambiguously higher-performing individuals may perform worse in terms of team output than a group composed of lower performers. We characterize the properties of the leadership selection and production processes that lead to the Apollo effect. Finally, we clarify when the opposite effect occurs in which supertalent performs better than comparatively less qualified groups.

Keywords: Team composition, leadership, mistakes.

JEL Classification Numbers: C70, D70, J80.

We thank Mike Borns, Philipp Hungerländer, Alberto Versperoni, and seminar participants at the University of Vienna for helpful comments and discussions. This project was supported by funds of the Oesterreichische Nationalbank (Austrian Central Bank, Anniversary Fund, project number: 17663).

1. INTRODUCTION

The "Apollo Syndrome" is a phenomenon first described and popularized in the management literature by Belbin (1981). It describes situations in which teams of highly capable individuals, collectively, perform badly. The phenomenon is named after the mission teams in NASA's Apollo space program and refers to situations in which one team is composed of unambiguously more capable individuals than the teams with which it is compared. Contrary to intuition, in the experiments Belbin conducted in the sixties at what is now Henley Business School, the Apollo teams often finished near the bottom among the competing teams.[1] One of the reasons Belbin gives for the Apollo teams' failure is that Apollo team members "spent a large part of their time engaged in abortive debate, trying to persuade the other members of the team to adopt their own particular, well-stated point of view. No one seemed to convert another or be converted. However, each seemed to have a flair for spotting the weak points of the other's argument. [. . .] Altogether, the Apollo company of supposed supertalent proved an astonishing disappointment" (Belbin, 1981, p. 15).[2]

For our main result, we model a team production problem in which an executive or administrator (either a principal or the team itself) appoints a single leader and subsequently all team members produce joint output by exerting individual efforts. We assume that the administrator is more likely to select a "wrong" or suboptimal leader if the skills of the candidates are similar. The model represents the administrator's selection capabilities through a symmetric black-box function (for which we supply micro-justifications) that with some probability selects individuals for leadership positions on the basis of their innate leadership skills, which are unknown to the executive. The

[1] "Of 25 companies that we constructed according to our Apollo design, only three became the winning team. The favourite finishing position out of eight was sixth (six times), followed by fourth (four times)" (Belbin, 1981, p. 20). The performance data of the remaining Apollo teams is not available. If we allocate the remaining 12 teams with equal probability to each remaining rank, the resulting hypothetical expected Apollo rank is 4.6.

[2] The general observation itself is not novel. For instance, it finds expression in the description of the sinking of the *Mary Rose*: "it chanced unto this gentleman, as the common proverb is, — *the more cooks the worse potage*, he had in his ship a hundred marines, the worst of them being able to be a master in the best ship within the realm; and these so maligned and disdained one the other, that refusing to do that which they should do, were careless to do that which was most needful and necessary, and so contending in envy, perished in forwardness" (Hooker, J., *The Life of Sir Peter Carew*, 1575).

higher the skill differences, the easier it is to find the better team leader. We show that in this environment the Apollo effect—which we define as a team of highly skilled individuals being outperformed by a team consisting entirely of lower-qualified members—is generally inescapable and arises from any noisy selection process.

The process of the selection of candidates for leadership roles is as follows. The (human resources) executive or administrator charged with assigning tasks to workers and managers is not an expert in the production processes for which the appointments under consideration are made. She collects information on the performance of the individuals according to some standardized management selection protocol. Although she may perform her job admirably, she occasionally makes the wrong leadership assignment.

The narrative offered in this Introduction explains the Apollo effect based on competition for leadership. This need not be taken literally, however. Any potential for conflicting opinions, differential styles of conducting business, management philosophies, etc, can be similarly thought of as the basis for the frictions that are modeled through our black-box assignment function.[3] In section 4 we define and describe the properties of a task-matching model in which the single-leadership feature is replaced by a function that matches workers to differentially productive tasks. In this extension of the model, the assignment function models the potential for mistakenly assigning the wrong worker to a given task. Although the Apollo effect is less prevalent in this environment than in the leadership game, we show that there are always skill profiles of workers for which the Apollo effect can arise for suitably noisy task-selection technologies. While we assume in the main body of our analysis that workers know each other's skills, we show that the Apollo effect persists under incomplete skill information among workers. Finally, we show that the Apollo effect exists regardless of the introduction of a profit-maximizing principal into the pure team environment.

The rest of the paper is organized as follows. After a short overview of the related literature we define our model in Section 2. Section 3 presents and illustrates our main result, the ubiquity of the Apollo effect. Section 4 discusses several extensions, alternative interpretations, and the robustness of the main model. In the concluding Section 5 we discuss a further set of potential applications and extensions. Proofs of all the results and details of

[3] A simple "lost production complementarities" explanation of the Apollo effect which is similar to Belbin's own story is illustrated in example 4.

some of the derivations can be found in the Appendix.

Literature

Belbin (1981) introduces a "team role" theory designed to enhance team composition based on a series of business school training games.[4] The Apollo syndrome is described as an effect of team composition and as such it is distinct from the "Ringelmann-type" free-riding (or social loafing) due to moral hazard in teams (Gershkov et al., 2016).

Cyert & March (1963), Marschak & Radner (1972), and Holmström (1977) generated a rich literature on the economics of organizations. We are unaware, however, of any attempt in the theoretical literature to introduce systematic errors into (team) decision-making processes and analyze their effect on team performance and team composition. There are accounts of cognitive biases and heuristics in the management literature (e.g., Schwenk, 1984; Gary, 1998), psychology (e.g., Kahneman, 2003; Gigerenzer & Gaissmaier, 2011), sports (e.g., Lombardi et al., 2014), and administrative science (e.g., Tetlock, 2000), but we know of no directly related explorations in economics.

The existing economic literature on team composition problems consists of only a handful of papers. Chade & Eeckhout (2020) analyze problems of team composition when teams compete subsequent to the matching stage.[5] Their matching setup results in a model in which the externalities that affect sorting patterns differ substantially from those of the standard case.[6] Eliaz & Wu (2018) use an all-pay auction to model the competition between two teams. In their setup, the competing teams may differ in size and have

[4] For recent management surveys on team composition and pointers to empirical work, see, for example, Aritzeta et al. (2007) or Mathieu et al. (2013). There is a topical link to the literatures on collective intelligence in organizations (Woolley et al., 2015) and on swarm intelligence/stupidity (Kremer et al., 2014).

[5] In their motivation, Chade & Eeckhout (2020) ask whether or not a single "superstar" team would have been able to confirm the existence of the Higgs boson quicker than the competing ATLAS and CMS teams at CERN's Large Hadron Collider.

[6] In the settings we analyze, the optimal allocation is usually given by assortative matching, that is, the more talented team member should be assigned the leadership position or the higher productivity task. However, as the administrator (or organization) assigns leadership positions based on imprecise skill information, this results in a noisy allocation (for bounds on efficiency in the case of coarse matching see McAfee, 2002). The main departure from this literature is that, in our analysis, the matching procedure is taken into account in the specified compensation scheme.

incomplete information on the prize the opposing team receives as a group-specific public good. They explore the interplay of the effort aggregation (or team production) function's curvature with individual incentives and analyze endogenous team formation from the angles of aggregation and differing team size. Palomino & Sákovics (2004) discuss a model of revenue sharing when sports teams competitively bid to attract talent. They find that the organization of the league(s) is key to the optimal design of remuneration schemes and the resulting availability of talent. In a paper on board composition, Hermalin & Weisbach (1988) discuss how firm performance and CEO turnover determine the choice of directors. None of these papers develops the core issue of our paper, namely, leadership selection under assignment errors.

The endogenous emergence of team leadership is modeled explicitly in several recent papers. In Kobayashi & Suehiro (2005), each of two players obtains imperfect, private signals on team productivity. The individual incentives to lead by example (as in Hermalin, 1998) give rise to a coordination problem. Andreoni (2006) analyzes a public goods provision game in which a team can learn the project type by individually expending a small amount of goods and the investing "leader" faces free-riding incentives. Huck & Rey-Biel (2006) analyze teams of asymmetrically productive agents biased towards conformism. They find that the less productive of two equally biased agents should lead. By contrast, our paper does not model a particular leadership game but employs a black-box assignment function yielding selection probabilities based on idiosyncratic skills that should, in principle, be compatible with a large set of selection procedures.

Finally, there are many issues in the organizational design literature that are touched on in this paper, such as the concept of leadership (Hermalin, 1998; Lazear, 2012), battles for control (Rajan & Zingales, 2000), sequentiality of production (Winter, 2006), transparency of effort (Bag & Pepito, 2012), and repetition (Che & Yoo, 2001). For other aspects of organizational theory see the excellent recent overviews by Bolton et al. (2010); Hermalin (2012); Waldman (2012); Garicano & Van Zandt (2012).

2. THE MODEL

There is a team consisting of two members $\{1, 2\}$. Each team member is supposed to exert unobservable effort that contributes to joint output. In addition, each team member $i \in \{1, 2\}$ is attributed with managerial or leadership skill

$\theta_i \in \mathbb{R}_+$. The team's output depends on the assigned leader and on the efforts of all team members.[7] Denote by $y(\theta_i, e_1, e_2)$ the team output when agent $i \in \{1, 2\}$ is assigned to lead the team, agent 1 exerts effort e_1 and agent 2 exerts effort e_2. The cost of exerting effort e_i is the same for both agents, $c(e_i)$, with $c'(0) > 0$, $c' > 0$, and $c'' > 0$. The effect of the agents' effort exertion on output is symmetric, that is, for any θ_i, e_1 and e_2 the team generates

$$y(\theta_i, e_1, e_2) = y(\theta_i, e_2, e_1). \tag{1}$$

We assume that y is differentiable with

$$y_1 = \frac{\partial y}{\partial \theta_i} > 0,\ y_{j+1} = \frac{\partial y}{\partial e_j} > 0,\ y_{j+1,j+1} = \frac{\partial^2 y}{\partial e_j^2} < 0,\ y_{j+1,\mathbf{1}} = \frac{\partial^2 y}{\partial e_j \partial \theta_i} > 0 \tag{2}$$

for any $j \in \{1, 2\}$. The time structure of the modeled events is as follows. At the first stage of the interaction, one of the agents is appointed the team leader. At the second stage, the agents exert uncontractible efforts after observing the chosen leader and his leadership skill.[8] The resulting output is divided equally between the team members.[9] Monotonicity of output y with respect to the leader's skill attribute implies that it is optimal to choose the agent with the highest leadership skills as a team leader.

The main premise of the paper, however, is that selecting a team leader (or decision making in general) is a complex process that sometimes involves mistakes.[10] More precisely, we denote by $f(\theta_i, \theta_j)$ the probability that agent i is appointed to the leadership position when i's leadership skills are represented by parameter θ_i, while the other team member's skill is θ_j. With probability

[7] The leadership position creates a (sufficiently high) private and non-monetary benefit to the appointed leader, which renders the trivial (and potentially first-best) solution of "selling the project to the manager" infeasible. For empirical justifications of such benefits including "self-dealing," see, for instance, Tirole (2006, p. 17).

[8] Similarly to the sequential game outlined above, the Apollo effect can be shown to exist in a simultaneous production version of the model in which all players choose their respective strategies at the same time.

[9] The paper's results hold regardless of the chosen output division rule. In particular, it is unimportant for the occurrence of the Apollo effect whether incentives are provided to exert (constrained) efficient efforts or not (Gershkov et al., 2016).

[10] As detailed in our assumptions (3) and (4) below, in order for the Apollo effect to arise, the function $f(\theta_i, \theta_j)$ must decrease sufficiently in θ_j: a "simple" constant probability of making mistakes is not sufficient.

$1 - f(\theta_i, \theta_j)$ player j is assigned the leadership position. We assume that the assignment function is symmetric:

$$f(\theta_i, \theta_j) = 1 - f(\theta_j, \theta_i), \tag{3}$$

responsive:

$$\frac{\partial f(\theta_i, \theta_j)}{\partial \theta_i} > 0, \tag{4}$$

and satisfies appropriate probability limit behavior, in particular $f(0, \hat{\theta}) = 0$ for[11] $\hat{\theta} > 0$.

In the Introduction, we informally motivate how this function f may arise from some management selection processes. We now give two more formal micro-justifications for the main properties of the black-box function we use throughout the paper. In the first formalization, we think of the appointing executive as having access to a test that is potentially capable of ranking the candidates: if one candidate is below and the other candidate is above the test location, then the test returns the ranking. If both candidates are below or above the test location, then one candidate is picked at random. Being less than perfectly well informed, however, the executive can choose the location of the test only probabilistically. Assume that the test realizes at threshold $\hat{\theta}$ with positive density $t(\hat{\theta})$. Then the probability of player 1 with skill θ_1 being chosen under this test is

$$\frac{1}{2}\left[\int_0^{\min(\theta_1,\theta_2)} t(\hat{\theta})d\hat{\theta} + \int_{\max(\theta_1,\theta_2)}^{1} t(\hat{\theta})d\hat{\theta}\right] + \mathbf{1}_{\{\theta_1 \geq \theta_2\}} \int_{\min(\theta_1,\theta_2)}^{\max(\theta_1,\theta_2)} t(\hat{\theta})d\hat{\theta}. \tag{5}$$

The derivatives for any realization of $\theta_1 > \theta_2$ are as required by our assumptions.

Our second micro-foundation is based on the idea that the administrator can make noisy observations of the two agents' types $\theta_i + \varepsilon_i$ and knows only that ε_i is distributed independently and identically according to any continuous distribution H (for a complete model development, see Lazear & Rosen, 1981). The administrator then bases a decision on her noisy observation of leadership abilities. In this environment, the probability that agent 1 will be appointed is

$$\Pr(\theta_1 + \varepsilon_1 > \theta_2 + \varepsilon_2) \; = \Pr(\varepsilon_2 - \varepsilon_1 < \theta_1 - \theta_2) \equiv P, \tag{6}$$

where the difference between the two independently distributed random variables is itself a continuously distributed random variable. The derivatives of this assignment probability P satisfy the required properties of $f(\theta_1, \theta_2)$.

[11] The implied discontinuity at $f(0, 0)$ does not play a role in our analysis.

3. THE MAIN RESULTS

This section presents the principal finding of this paper, the ubiquity of the Apollo effect. Before we start the formal analysis we would like to point out that the first-best efficient selection in which the better-qualified player is always appointed the team leader by an uninformed administrator is generally unattainable in the specified game based on selection capabilities f. We start the discussion by means of a simple illustrative example of the main idea.

Example 1: In the following comparative static arguments we distinguish between two teams $j \in \{\mathrm{A}, \mathrm{B}\}$ and typically assume that team members' abilities are ranked $\theta_1^A \geq \theta_1^B$ and $\theta_2^A \geq \theta_2^B$, where team A consists of unambiguously higher-ability players than team B. For leadership selection, an administrator employs a black-box function based on ability ratios according to which the probability of player $i \in \{1, 2\}$ being selected as leader is[12]

$$f(\theta_i, \theta_j) = \frac{\theta_i^r}{\theta_1^r + \theta_2^r}, \; r > 0. \tag{7}$$

If player $i \in \{1, 2\}$ is selected as team j's leader ($j \in \{A, B\}$), then $\hat{\theta} = \theta_i^j$ and the team generates simple linear output

$$y(\hat{\theta}, e_1, e_2) = \hat{\theta}(e_1 + e_2). \tag{8}$$

As either player 1's or player 2's ability is employed exclusively for leadership, we refer to this case as "exclusive" management or production.[13] Following the time structure outlined above, workers know whether or not they are assigned leadership roles before exerting effort, i.e., any mistakes are made during a first leadership-assignment stage while unobservable efforts are exerted by perfectly informed agents at a second stage. More specifically, player i's stage-2 objective, given that the player with type $\hat{\theta}$ is chosen as leader and output is shared equally, is

$$\max_{e_i} \; \frac{y(\hat{\theta}, e_i, e_j)}{2} - c(e_i). \tag{9}$$

[12] In different environments similar functions have been called "logistic" or "sigmoid" functions. The contest literature refers to a variant of (7) as "ratio," "power," or "Tullock" contest success function (Jia et al., 2013). Note that—as there are no strategies involved at this stage—our use of this function for leadership selection is purely descriptive and does not constitute a game.

[13] We later extend our model to task-matching in order to also capture shared production aspects in teams with complementary skills where individuals are matched to tasks.

Assuming quadratic effort costs $c(e) = e^2$, symmetric equilibrium efforts are simply

$$e_1(\hat{\theta}) = e_2(\hat{\theta}) = \hat{\theta}/2. \tag{10}$$

At the leadership selection stage, the administrator selects either player 1 with probability $f(\theta_1, \theta_2)$ or player 2 with probability $1 - f(\theta_1, \theta_2)$ as the team leader. Hence, the first-stage expected equilibrium team output is

$$\begin{aligned} Y(\theta_1, \theta_2) &= f(\theta_1, \theta_2) y(\theta_1, e(\theta_1), e(\theta_1)) + (1 - f(\theta_1, \theta_2)) y(\theta_2, e(\theta_2), e(\theta_2)) \\ &= \theta_2^2 + f(\theta_1, \theta_2)(\theta_1^2 - \theta_2^2) \\ &= \frac{\theta_1^{r+2} + \theta_2^{r+2}}{\theta_1^r + \theta_2^r}. \end{aligned}$$

We now implicitly define an "isoquant" function $\theta_2(\bar{y}, \theta_1)$, which determines the type θ_2 that achieves the constant output level $\bar{y}$ for some type θ_1. An example is shown in Figure 1: low precision $r = .25$ is shown on the left, moderate precision $r = 2$ in the middle, and high precision $r = 15$ on the right.[14] We restrict attention (without loss of generality) to $\theta_1 \geq \theta_2$, and so only the subset under the diagonal is relevant in the figure. Team compositions "under the isoquant," i.e., to the left of the isoquant $\theta_2(\bar{y}, \theta_1)$, produce lower output than $\bar{y}$. Skill pairs "above the isoquant," i.e., to the right of isoquant $\theta_2(\bar{y}, \theta_1)$, produce higher output than $\bar{y}$. The Apollo effect arises here because, for any point $(\hat{\theta}_1, \hat{\theta}_2)$ on a positively sloped part of an isoquant, we can find a point $(\theta_1 > \hat{\theta}_1, \theta_2 > \hat{\theta}_2)$ *under this isoquant* (close to where it is vertical), such that $y(\hat{\theta}_1, \hat{\theta}_2) > y(\theta_1, \theta_2)$. Note that one would not expect a positive slope of the isoquants in Figure 1 without the defined possibility of making mistakes in leadership assignment. In this example, the Apollo effect crops up for all selection precisions, provided that the spread $\theta_1 - \theta_2$ is sufficiently high. ◁

One may, however, ask how pervasive the occurrence of the Apollo effect is in the above example. In order to answer this question, we start the formal argument by defining the Apollo effect in a general production environment with two teams.

[14] We refer to the exponent r in (7) as the "selection precision" of player 1 because it parameterizes the derivative of the assignment function with respect to θ_1. The comparison case of no mistakes is obtained for $r \to \infty$, i.e., $f(\theta_1, \theta_2) = 1$ iff $\theta_1 \geq \theta_2$. In this case, the level sets in the right panel of Figure 1 become a perfectly rectangular map. By contrast, if $r = 0$, we have $f(\theta_1, \theta_2) = 1/2$ for any θ_1 and θ_2.

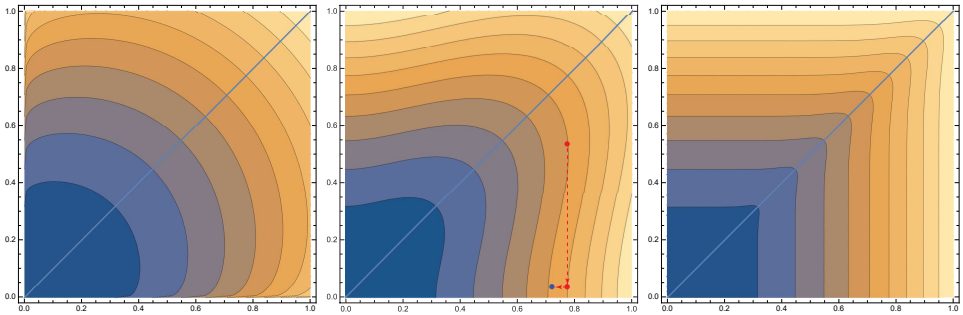


Figure 1: "Isoquant" expected team output level sets with θ_1 on the horizontal and θ_2 on the vertical axis for $r = .25$ on the left, $r = 2$ in the middle, and $r = 15$ on the right.

Definition 1. *The environment expresses the Apollo effect, if there exist two teams* $\{A, B\}$ *with leadership skills* $\left(\theta_1^A, \theta_2^A\right) >> \left(\theta_1^B, \theta_2^B\right)$ *with* $y^A < y^B$*, where* y^A *is the equilibrium output of team A and* y^B *is that of team B.*

Without loss of generality, we assume that $\theta_1 \geq \theta_2$. Observing the appointed leader of type $\hat{\theta}$ and assuming equal sharing of output, team member i maximizes effort stage utility

$$\max_{e_i} \frac{y(\hat{\theta}, e_1, e_2)}{2} - c(e_i). \tag{11}$$

Taking the derivative with respect to e_i, we define symmetric equilibrium effort $e^* = e_1 = e_2$ as

$$y_{i+1}(\hat{\theta}, e_i, e_j) - 2c'(e_i) = 0, \tag{12}$$

where subscripts on functions denote derivatives. The assumed curvature of the output and cost functions guarantee that $e^*(\hat{\theta})$ is non-decreasing. We substitute these equilibrium efforts into output that determines equilibrium team output as

$$\begin{aligned} Y(\theta_1, \theta_2) = {} & f(\theta_1, \theta_2) y(\theta_1, e^*(\theta_1), e^*(\theta_1)) \\ & + (1 - f(\theta_1, \theta_2))\, y(\theta_2, e^*(\theta_2), e^*(\theta_2)). \end{aligned} \tag{13}$$

It turns out that an analytically convenient way to demonstrate that the Apollo effect exists is to show that there exists a skill combination (θ_1, θ_2) such that

$y(\theta_1, \theta_2)$ has a positive gradient, i.e., there exist $(\eta_1, \eta_2) >> 0$ such that

$$\frac{\partial Y(\theta_1, \theta_2)}{\partial \theta_1}\eta_1 + \frac{\partial Y(\theta_1, \theta_2)}{\partial \theta_2}\eta_2 < 0. \tag{14}$$

Our main claim is that there exists a team endowed with skills θ^A for which the equilibrium team output shrinks if both types are increased infinitesimally.

Lemma 2. *The Apollo effect arises if and only if*

$$-\frac{f_2(\theta_1, \theta_2)}{1 - f(\theta_1, \theta_2)} > \frac{e'^*(\theta_2)2y_e(\theta_2, e^*(\theta_2), e^*(\theta_2)) + y_1(\theta_2, e^*(\theta_2), e^*(\theta_2))}{y(\theta_1, e^*(\theta_1), e^*(\theta_1)) - y(\theta_2, e^*(\theta_2), e^*(\theta_2))}. \tag{15}$$

The lemma allows us to specify when the Apollo effect is plausible. It shows that increasing the difference between the team members increases the chance of observing the Apollo effect (if $y(\theta_1, e^*(\theta_1), e^*(\theta_1)) - y(\theta_2, e^*(\theta_2), e^*(\theta_2))$ is high, $f(\theta_1, \theta_2)$ is high and hence it is easier to satisfy the condition of the last Lemma). Moreover, the probability of misallocation must be responsive to the skills, that is, $f_2(\theta_1, \theta_2)$ must be substantially low (and negative).

An immediate implication of the lemma and its proof is that while it is always beneficial for the best team member to improve her leadership skills, this is certainly not the case for the lower-qualified team member. We proceed to state a general property of exclusive production.

Lemma 3. *For exclusive production $y(\theta, e(\theta), e(\theta))$ and any $\hat{\theta} > 0$, we have*

$$\begin{aligned} &f(\hat{\theta}, \hat{\theta})y(\hat{\theta}, e(\hat{\theta}), e(\hat{\theta})) + (1 - f(\hat{\theta}, \hat{\theta}))y(\hat{\theta}, e(\hat{\theta}), e(\hat{\theta})) \\ &= f(\hat{\theta}, 0)y(\hat{\theta}, e(\hat{\theta}), e(\hat{\theta})) + (1 - f(\hat{\theta}, 0))y(0, e(0), e_2(0)). \end{aligned} \tag{16}$$

Therefore, for any $\hat{\theta}$, the points $(\hat{\theta}, \hat{\theta})$ and $(\hat{\theta}, 0)$ belong to the same isoquant. Note that for symmetric functions f, the isoquants' slope at $\theta_1 = \theta_2$ must be -1 at the diagonal of our level sets. Together, these observations imply the following general result.

Proposition 1. *The Apollo effect arises under an exclusive leadership assignment for every feasible continuous function f.*

This result shows that the only case in which the Apollo effect cannot arise is if the defined possibility of making mistakes in leadership selection is

entirely absent. For concave production technology, and any conceivable not infinitely accurate continuous leadership selection technology f, there will be skill profiles that give rise to the Apollo effect, i.e., where unambiguously better-qualified teams must be expected to produce lower output than a set of "underdogs." Before we proceed to explore the implications of our main result through a series of applications and direct extensions in the form of remarks we should point out that this result is an implication of the introduced possibility for errors in the leadership selection process. This will not be the case in the "task-matching" generalization of the following section.

Remark 4 (Dispersion & conflict)**.** *Why do Apollo teams have more problems in appointing the right leader than less-qualified teams? "A possible answer lies in the very pressures that our educational system and culture exert on clever people. Those who at school are 'top of the class,' or who have it within their reach, are continually being judged in terms of their scholastic preeminence. To come second is to fail. Beating the next person is the name of the game. Difficult problems excite the greatest rivalry and so destroy the bonds of mutual co-operation and complementary functioning upon which the success of a team ultimately depends. In other words, overconcentration on coming top of the class provides an unconscious training in anti-teamwork." Belbin (1981, p. 18)*

In our model the probability of being appointed as leader is a function of the skill-dispersion of the team: more dispersed teams have fewer problems in selecting the better leader. The Apollo-effect arises if and only if the gain in expected output due to a more dispersed non-Apollo team (and therefore higher probability of selecting the right type as leader) can more than compensate for the potential loss of erroneously selecting the wrong leader in the better Apollo team. The above quotation suggests that there may be good reasons to expect this higher probability of agreement in lower-quality teams.

We reiterate and emphasize that we do not claim that less talented teams make better decisions in choosing a leader. However, it is the case that, in a more dispersed team, both an outsider's task of identifying the better qualified individual is easier and team members themselves will be less eager to fight for leadership if the difference in abilities is stark.

Remark 5 (Labor market)**.** *The environment of Proposition 1 can be used to study the effect of imprecise leadership selection on the optimal assignment*

of agents to several teams. We keep the same informational assumptions as in the rest of this section but are here only interested in characterizing the optimal team composition, rather than a game capable of bringing it about. In particular, we ask which agent types from the ordered set $\theta_1 > \theta_2 > \cdots > \theta_n$, $n \geq 3$ *optimally self-select and for what team structure.*

We assume that the firm wishes to create $k < n/2$ *teams of two agents each. Subsequent to the creation of the teams, a leader will be chosen in each team following the procedure described above. How should the hiring and team creation process take the later noisy leadership selection into account? To answer this question, we have to identify the optimal hiring and matching strategies assuming that the types are observable at this stage. This illustrates which types should be targeted and the information that needs to be collected on candidates.*

Absent the possibility of making mistakes in subsequent leadership selection, an optimal matching is to form k *teams with team* j *led by agent* θ_{2j-1}*, i.e., one of the* k *agents with the highest leadership ability with any second agent chosen from the lower half of the types. If the lower-skill partners' types have an arbitrarily small output contribution, then the lowest type(s)* $(\theta_{2k+1}, \ldots, \theta_n)$ *will never be employed.*[15] *Therefore, it is important to identify and exclude the lowest ability types. Yet, if leadership assignment is imprecise, an implication of the Apollo effect is that a set of workers strictly better qualified than these "worst" types will be optimally excluded.*

Consider, for example, the ordered set of $n = 5$ *agent types* $\theta_i = (n-i)/(n-1)$ *with identical, linear production* $y(\theta, e_1, e_2) = \theta(e_1 + e_2)$*, quadratic costs* $c(e) = e^2/2$*, and ratio assignment* $f(\theta) = \theta_i^r/(\theta_i^r + \theta_j^r)$*,* $r > 0$*. Assume that the organization needs two teams and hence seeks to exclude one agent. For* $r \geq 1$ *it is optimal to exclude the agent with median ability* θ_3*. The intuition of "dropping the middle" types for sufficiently precise assignment* f *can be generalized and has implications for the labor market: firms demand the* right *types, and not necessarily the highest available types. In the example, given a sufficient precision of* f*, the middle types are left unemployed whereas the lowest type* θ_n *is employed in all optimal matchings!*

Remark 6 (Project selection)**.** *Consider a manager's choice between two*

[15] This positive influence can be made precise and formalized by an infinitesimally small positive multiplier t_l, as discussed in the task-matching environment of Section 4.1. In general, such a task-matching construction gives qualitatively similar results to exclusive production only for intermediate precisions of the assignment function f.

projects of unknown quality θ_1, θ_2, guided by the imperfect selection technology $f(\theta_1, \theta_2)$. In this application, project output $y(\theta_i, K, L)$ is increasing in θ_i, satisfies the equivalents of Assumptions (1) and (2), and the symmetric factors K and L are chosen by strategic project employees who privately observe quality θ_i. Proposition 1 shows that there are situations in which improving both individual projects to $\theta_1' > \theta_1$ and $\theta_2' > \theta_2$ actually decreases the firm's expected revenue relative to the original, unambiguously worse project environment.

Remark 7 (Larger teams)**.** *Whereas our other results are stated for assignment functions defining selection probabilities for just two players, we now analyze the consequences of increasing the team size.*[16]

For example, consider an n-player version of our model governed by the usual linear production $y(\theta, e_1, \ldots, e_n) = \theta(e_1 + \cdots + e_n)$ and quadratic efforts cost $c(e) = e^2/2$. We adopt a ratio-assignment function that gives the probability of (the highest-type) player 1 being selected as

$$f(\theta_1, \ldots, \theta_n) = \frac{\theta_1^r}{\theta_1^r + \cdots + \theta_n^r}, \; r > 0. \tag{17}$$

Provided that all team members share output equally, this results in type-contingent equilibrium efforts of $e = \theta/n$, whereas a benevolent planner would dictate the efficient $e^* = \theta$. The Apollo effect also arises in this example as in the two-agent case.

4. FURTHER RESULTS

4.1. Task Matching

The main result of this paper rests on an interpretation of conflict (for leadership) to explain the Apollo effect since every team member's management skills enter the production process exclusively. Only one of the team members is appointed the leader and the rest's leadership skills are completely discarded.

[16] Amazon's Jeff Bezos is reported to employ a "two pizza rule": if a team cannot be fed by two pizzas, then that team is too large. The idea is that having more people work together is less efficient, i.e., team output decreases beyond the optimal size. This is the case in Shellenbarger (2016) who argues that participants tend to feel less accountable in crowded meetings and doubt that any contribution they make will be rewarded, and hence reduce effort.

Deviating from this interpretation, we now assume that the production technology requires that all workers be matched to their "correct" tasks and therefore all individuals' skills enter production.[17] That is, we consider an environment in which the organization or its executives must assign team members to different tasks and, after the assignment, the agents apply their skills and exert effort on the allocated tasks. However, this assignment may involve mistakes or misallocations of agents to tasks. We employ the following output function

$$y(\theta_i, \theta_j, e_i, e_j) = y^h(\theta_i, e_i) + y^l(\theta_j, e_j) \tag{18}$$

where both $y^h(\theta, e)$ and $y^l(\theta, e)$ are weakly concave and increasing in both arguments. That is, each worker is matched either with task h or with task l. Each worker uses "leadership" skills and exerts effort in executing the allocated task. Function f chooses the assignment of workers to tasks. Otherwise, the model is the same as in the previous section. Without loss of generality, we assume that $\theta_1 \geq \theta_2$. We assume that for any $e \geq 0$

$$y_1^h(\theta, e) \; > y_1^l(\theta, e) > 0. \tag{19}$$

Therefore, the efficient assignment is that higher-ability agent 1 is assigned task h, while agent 2 is assigned task l. Given an allocation, the agents will exert effort, as dictated by first-order conditions $(e_i^h(\theta_i), e_j^l(\theta_j))$:

$$y_2^h(\theta_i, e_i) \; = 2c'(e_i), \; y_2^l(\theta_j, e_j) = 2c'(e_j). \tag{20}$$

Assuming, in addition to (19), that

$$y_2^h(\theta, e) \; > y_2^l(\theta, e) > 0, \; y_{12}^h(\theta, e) > y_{12}^l(\theta, e) > 0 \text{ and } 0 > y_{22}^h(\theta, e) > y_{22}^l(\theta, e) \tag{21}$$

implies that equilibrium effort in both tasks is increasing in type and that both $e^h(\theta) > e^l(\theta) > 0$ and $e^{h\prime}(\theta) > e^{l\prime}(\theta) > 0$. At the selection stage, the expected team output under task matching is

$$Y(\theta_i, \theta_j) \; = f(\theta_i, \theta_j) z(\theta_i, \theta_j) + \big(1 - f(\theta_i, \theta_j)\big)\, z(\theta_j, \theta_i), \tag{22}$$

[17] Referring back to our motivational example of the NASA Apollo missions, the Apollo team members were selected to fulfill distinct roles. The Apollo 11 team, for instance, consisted of mission commander Neil Armstrong, command module pilot Michael Collins, and lunar module pilot Edwin Aldrin. Hence, team performance depended on each member of the crew being selected for and performing a very specific task.

where we assume that $z(\theta_i, \theta_j)$ is the equilibrium output if agent i is assigned to task h and agent j is assigned to task l, i.e.,

$$z(\theta_i, \theta_j) = y(\theta_i, \theta_j, e_i^h(\theta_i), e_j^l(\theta_j)). \tag{23}$$

Our assumptions above imply that $z(\theta_1, \theta_2) > z(\theta_2, \theta_1)$. We introduce our result by means of a simple example.

Example 2: We assume that team output is created by the simple production function

$$y\left(\theta_i, \theta_j, e_i, e_j\right) = t_h\theta_i e_i + t_l\theta_j e_j, \text{ with } t_h \geq t_l. \tag{24}$$

Similarly to the previous example, we assume that costs are quadratic, $c(e) = e^2$, and that the allocation technology is

$$f(\theta_i, \theta_j) = \frac{\theta_i^r}{\theta_1^r + \theta_2^r}, \; r > 0, \tag{25}$$

which specifies the probability that agent i is assigned task h. Then, the task-specific equilibrium efforts are $e^x(\theta) = t_x\theta$, $x \in \{h, l\}$, and the expected equilibrium team output is

$$\begin{aligned} Y(\theta_i, \theta_j) &= f(\theta_i, \theta_j) z(\theta_i, \theta_j, e_i^h(\theta_i), e_j^l(\theta_j)) \\ &\quad + (1 - f(\theta_i, \theta_j)) z(\theta_j, \theta_i, e_j^h(\theta_j), e_i^l(\theta_i)) \\ &= \frac{f(\theta_i, \theta_j)(\theta_i^2 - \theta_j^2)(t_h^2 - t_l^2) + \theta_j^2 t_h^2 + \theta_i^2 t_l^2}{2} \\ &= \frac{t_h^2\left(\theta_i^{r+2} + \theta_j^{r+2}\right) + t_l^2\left(\theta_j^2\theta_i^r + \theta_i^2\theta_j^r\right)}{2\left(\theta_i^r + \theta_j^r\right)}. \end{aligned}$$

Figure 2 shows the isoquants under the different selection precisions r. As in the exclusive leadership case (see Figure 1), low precision $r = .25$ is shown on the left, moderate precision $r = 2$ in the middle, and high precision $r = 15$ on the right. The task values are $t_h = 2/3$, $t_l = 1/3$.

The figure illustrates that under task matching and for a given pair (t_h, t_l), the Apollo effect only crops up in cases where the subsequent selection precision r is below the minimal threshold, which, in the present example, is implicitly given by

$$\frac{2\theta_2^2(\theta_1^r + \theta_2^r)}{(\theta_1^2 - \theta_2^2)(\theta_1\theta_2)^r} = r\frac{t_h^2 - t_l^2}{\theta_1^r t_h^2 + \theta_2^r t_l^2} \tag{26}$$

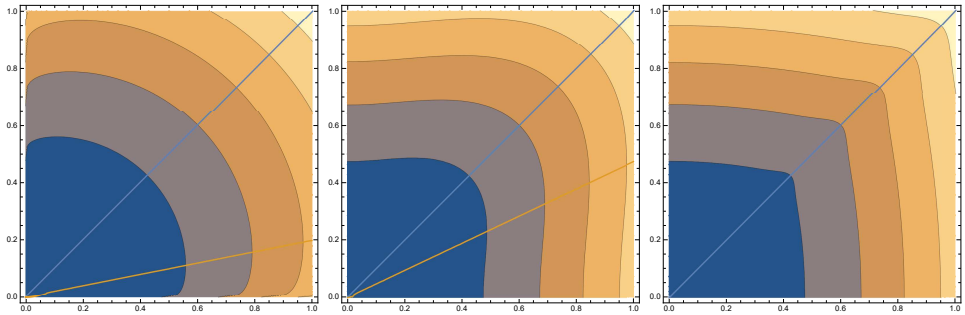


Figure 2: Task-matching level sets showing expected output for $t_h = 2/3$ and $t_l = 1/3$. The level sets are drawn for $r = .25$ on the left, $r = 2$ in the middle, and $r = 15$ on the right. The solid golden line represents condition (26).

or, plugging in the example values, $r < 3.3545$. This threshold condition expresses that the less it matters who is assigned to which task, i.e., the closer t^h and t^l are, the more likely it is that assignment mistakes must be made in order for the Apollo effect to arise. ◁

Intuitively, we can decompose the second player's marginal output contribution into two components: productive and disruptive. For the moment, consider the (efficient) case of an infinitely precise allocation function f, where the disruptive effect does not arise. Starting at any interior point $\hat{\theta} = \theta_1 = \theta_2$ on the diagonal in Figures 2, 3, and 4, a decrease in θ_2 results in lower output and hence must be compensated by an increase in θ_1 in order to stay on the same isoquant $I(\cdot)$. Hence, the isoquants in the middle panel of the top row of Figure 3 now become "triangular" in the sense that the point on the diagonal where $\theta_1 = \theta_2 = \hat{\theta}$ is connected by a negatively sloped curve with the point on the horizontal axis where $(\tilde{\theta}_1 > \hat{\theta}_1, \theta_2 = 0)$. This latter point is to the right of the point $(\hat{\theta}_1, \theta_2 = 0)$ directly under the diagonal from which we started. The horizontal shift of the isoquant depicts the marginal productive influence of player 2, which we call the "productive effect" (which includes, generally speaking, the "synergies" created by teamwork).

The isoquant maps discussed above are illustrated in Figure 3, and their detailed decomposition into productive and disruptive marginal effects is shown in Figure 4. The latter figure displays the productive marginal effect (the negative vertical slope of the blue isoquant $I(a', b')$) and the total marginal effect (the vertical slope of the red isoquant $I(a'', b'')$) for the task-matching case of

$t_h > t_l$ and intermediate selection precision f.

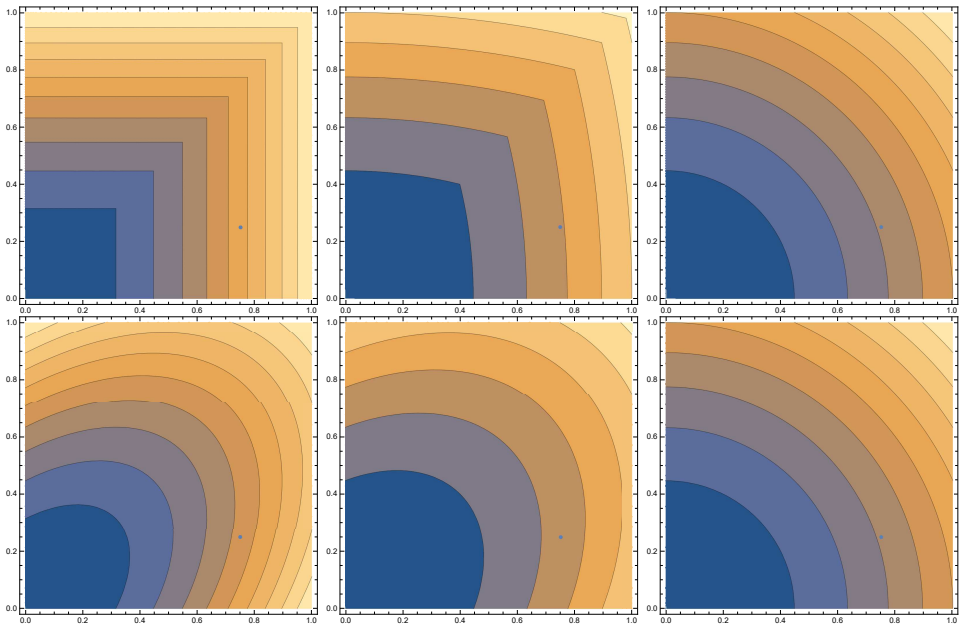


Figure 3: Isoquants for infinitely precise f in the first row (illustrating the pure productive effect) and $r = 1$ in the second row (illustrating both productive and disruptive effects). Plotted are the cases of $t_h = 1$ and $t_l = 0$ (left), $t_l = 1/2$ (middle), and $t_l = 1$ (right).

Any assignment function f that satisfies our assumptions introduces allocative inefficiency, thereby shifting all points of the efficient-assignment blue isoquant—except for the two points on the diagonal and horizontal axis just pinned down—further to the right, resulting in the red isoquant of Figure 4. This is what we call the "disruptive effect." The disruptive effect tends to shift points (θ_1, θ_2) close to the diagonal (where the chance of mistakes is highest) further to the right than those with lower θ_2. But the Apollo effect arises only in the extreme case in which the disruptive effect causes an isoquant to become positively sloped. More precisely, it arises if the (negative) marginal disruptive effect—described by $f_2(\theta_1, \theta_2)$—outweighs the (positive) marginal productive effect of a marginal increase of θ_2.

Compare this to the exclusive leadership case considered in the previous section. There, as illustrated in the two left-hand panels of Figure 3 and the

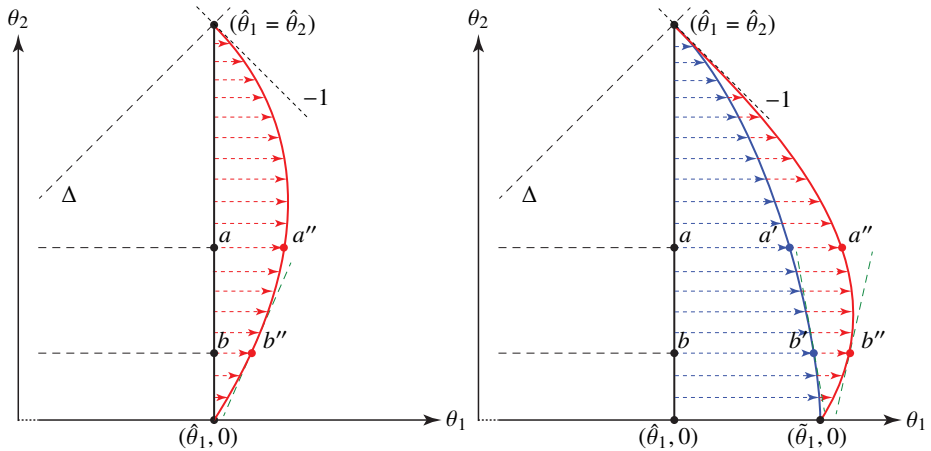


Figure 4: Three isoquants are shown on the right: infinitely precise f: $I(a, b)$ under exclusive leadership (black), infinitely precise f: $I(a', b')$ for task matching (blue), and finite f: $I(a'', b'')$ under task matching with $t_h > t_l$ (red). The marginal positively sloped (total) Apollo effect is represented by the dashed tangent through b''. The necessity of the Apollo effect under exclusive leadership for finite f is illustrated on the left.

black isoquant of Figure 4, the efficient isoquant map is perfectly rectangular and any imprecision of f leads to disruption. In particular, all points of the isoquant (except for those on the diagonal and horizontal axis) shift to the right. Hence, the Apollo effect is always present in the simpler exclusive leadership environment of Proposition 1. The existence of the Apollo effect in the task-matching environment of this section, however, depends on the marginal output of player 2 (1)—her productive contribution—being smaller than the disruptive effect introduced through the possibility of wrongly assigning her to the more (less) productive task h (l). Our next result summarizes this intuition and generalizes the previous example by identifying a condition on the assignment function f that guarantees that the Apollo effect arises also in the task-matching environment.

Proposition 2. *For equilibrium task-matching production $z(\theta_i, \theta_j)$, a sufficient condition for the Apollo effect to arise for some type profile $\theta_1 > \theta_2$ is that the*

selection technology $f(\theta_1, \theta_2)$ *satisfies*

$$f_2(\theta_1, \theta_2) < \frac{z_2(\theta_1, \theta_2) + z_1(\theta_2, \theta_1)}{2z(\theta_2, \theta_1) - 2z(\theta_1, \theta_2)}. \tag{27}$$

Notice that the condition of this Proposition holds if $f_2(\theta_1, \theta_2)$ is sufficiently low (and negative). To get a better understanding of the last condition, observe that for f infinitely precise, we have $f_2(\theta_1, \theta_2) = 0$ for any $\theta_1 > \theta_2$.

Example 3: We remain in the same environment as in the previous example with

$$y\left(\theta_i, \theta_j, e_i, e_j\right) = t_h\theta_i e_i + t_l\theta_j e_j, \text{ with } t_h \geq t_l.$$

For quadratic costs and task-specific linear production (24), the equilibrium production is $y^x(\theta, e(\theta)) = t_x\theta^2$, $x \in \{h, l\}$. The condition for an isoquant to have a positive slope (inequality (45) in the proof of Proposition 2) is

$$\frac{t_h^2}{t_h^2 - t_l^2} < f(\theta_1, \theta_2) - f_2(\theta_1, \theta_2)\frac{\theta_1^2 - \theta_2^2}{2\theta_2}. \tag{28}$$

For the general ratio assignment function (7), this condition (28) equals

$$\frac{t_h^2}{t_h^2 - t_l^2} < \frac{\theta_1^r}{\theta_1^r + \theta_2^r} - (r\theta_2^{r-2})\frac{\theta_1^r\left(\theta_2^2 - \theta_1^2\right)}{2\left(\theta_1^r + \theta_2^r\right)^2}, \tag{29}$$

where the term $r\theta_2^{r-2}$ goes to infinity as $\theta_2 \to 0$ for $0 < r < 2$, irrespective of $\theta_1 > \theta_2$. Hence the claimed inequality holds for some spread of types. This is confirmed by the sufficient condition (27) which equals in this example

$$-\frac{r\theta_1^r\theta_2^{r-1}}{\left(\theta_1^r + \theta_2^r\right)^2} < -\frac{\theta_2\left(t_h^2 + t_l^2\right)}{\left(\theta_1^2 - \theta_2^2\right)\left(t_h^2 - t_l^2\right)}. \tag{30}$$

At the arbitrary point $\theta_1 = 3/4, \theta_2 = 1/4$ (indicated in the below figure) and task multipliers $t_h = 1, t_l = 1/4$, this condition amounts to

$$-\frac{r}{\cosh(r\log(3)/2)^2} < -\frac{17}{30} \Leftrightarrow r \in [0.64, 2.5]. \tag{31}$$

Figure 5 shows examples of the corresponding output contour sets for different task multipliers and selection precisions. ◁

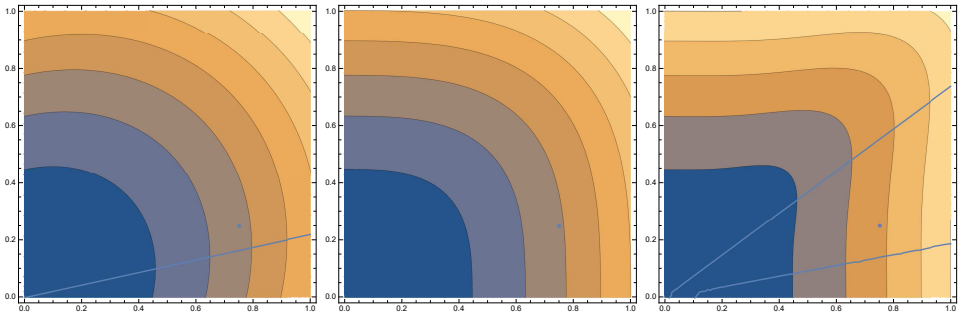


Figure 5: Left and center: type combinations for $t_h = 1, t_l = 3/4$ producing the same output $Y(\theta_1, \theta_2)$. The left panel shows the case of $r = 1$, and the center panel showns the same isoquants for $r = 2$. The blue lines show type-pairs for which the isoquants are vertical. The right panel illustrates a case of multiple critical locations ($t_h = 1$, $t_l = 1/4$, and $r = 4$).

Remark 8 (Black box assignment)**.** *There are many alternative interpretations of the black box assignment function f that differ from the "mistakes" employed for both our leadership and task matching stories of this and the previous sections. The following example illustrates that the Apollo phenomenon may be alternatively explained by "lost complementarities" in production akin to elements of Belbin's motivation quoted in the Introduction. That is, the production function exploits heterogeneity in the skills of the team members. These benefits diminish as team members become more similar.*

Example 4: ("Lost complementarities"). We simplify the task matching environment to $t_l = t_h = 1$ and define output as

$$\check{y}(\theta_1, \theta_2, e_1, e_2) = \theta_1 e_1 + \theta_2 e_2 - g(\theta_1, \theta_2)\phi, \tag{32}$$

in which the "lost complementarities" are constant ϕ, costs $c(e_i)$ are quadratic, and the "damage" function is

$$g(\theta_1, \theta_2) = \frac{1 - (\theta_1 - \theta_2)^{2r}}{2}. \tag{33}$$

Notice that the function $\check{y}(\theta_1, \theta_2, e_1, e_2)$ satisfies all requirements to exhibit the Apollo effect. Given unchanged equilibrium efforts, $e^*(\theta_i) = \theta_i/2$, the example of reference team $\theta^A = (.9, .1)$ is illustrated in Figure 6 that shows strictly inferior output for Apollo teams $(\theta_2^A = 0.9, \tilde{\theta}_2)$ with $\tilde{\theta}_2 \in (0.1, 0.5) > \theta_2^A$. ◁

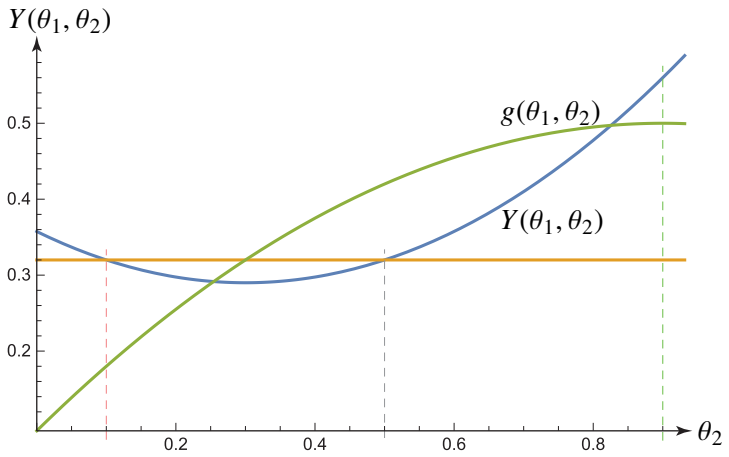


Figure 6: Lost complementarities example for $r = 1$ and $\phi = 1/2$. The blue curve is Apollo team output which is below the gold reference output of team $\theta^A = (.9, .1)$ for $\theta_2 \in (.1, .5)$. The green curve shows the "assignment" function $g(\theta_1, \theta_2)$ defined in (33).

4.2. Incomplete Information Among Agents

In this section we illustrate the robustness of our previous results by relaxing the assumption that agents know each other's types. To do so, we assume that the types distribute independently and identically according to distribution function G with density g on support $[a, b]$. When exerting effort, each agent knows only his own type and whether or not (s)he was assigned as a leader. Therefore, a symmetric equilibrium is characterized by two functions: $e^L(\theta)$, the effort function of the agent who was selected to be the team leader, and $e^F(\theta)$, the effort function of the agent who was not selected to be the team leader.

Proposition 3. *A pair of necessary conditions for agent equilibrium effort under incomplete information about agents' skills is*

$$\int_a^b y_2\left(\theta, e^L(\theta), e^F\left(\theta'\right)\right) f\left(\theta, \theta'\right) g\left(\theta'\right) d\theta' = 2c'(e^L(\theta)) \int_a^b f\left(\theta, \theta'\right) g\left(\theta'\right) d\theta' \tag{34}$$

and

$$\begin{aligned}&\int_a^b y_3\left(\theta', e^L\left(\theta'\right), e^F(\theta)\right) f\left(\theta', \theta\right) g\left(\theta'\right) d\theta' \\ &= 2c'(e^F(\theta)) \int_a^b f\left(\theta', \theta\right) g\left(\theta'\right) d\theta'.\end{aligned} \tag{35}$$

We illustrate this result for the same environment as in the previous examples. Assume that $c(e) = e^2/2$ and $y(\theta, e_1, e_2) = \theta\,(e_1 + e_2)$; then first-order conditions (34) and (35) become

$$\begin{aligned}e^L(\theta) &= \theta/2, \\ e^F(\theta) &= \frac{\int_a^b \theta' f\left(\theta', \theta\right) g\left(\theta'\right) d\theta'}{2\int_a^b f\left(\theta', \theta\right) g\left(\theta'\right) d\theta'} = \mathbb{E}_{\theta'|\text{follower has type } \theta}\left[\theta'\right] \\ &= \frac{r+1}{2(r+2)} \frac{{}_2F_1\left(1, \frac{r+2}{r}; 2+\frac{2}{r}; -\theta^{-r}\right)}{{}_2F_1\left(1, 1+\frac{1}{r}; 2+\frac{1}{r}; -\theta^{-r}\right)},\end{aligned} \tag{36}$$

where ${}_2F_1(x)$ is the ordinary hypergeometric function (representing the hypergeometric series).[18] Figure 7 gives an example of the uniform distribution. These equilibrium efforts yield expected team output

$$\begin{aligned}Y(\theta_1, \theta_2) = f(\theta_1, \theta_2) y(\theta_1, e^L(\theta_1), e^F(\theta_2)) \\ + (1 - f(\theta_1, \theta_2))\, y(\theta_2, e^F(\theta_1), e^L(\theta_2)).\end{aligned} \tag{38}$$

Figure 7 shows isoquants for precisions $r \in \{0.25, 2, 8\}$. As the positively sloped parts of the isoquants illustrate, the Apollo effect is present in this example with incomplete information as well.

4.3. Principal-Agent Model

Contrary to the team production environment used for all other results of this paper, in this section we explore the robustness of our findings to the presence

[18] The ordinary hypergeometric function is defined as

$${}_2F_1(a, b; c; z) = \sum_{n=0}^{\infty} \frac{(a)_n (b)_n}{(c)_n} \frac{z^n}{n!}. \tag{37}$$

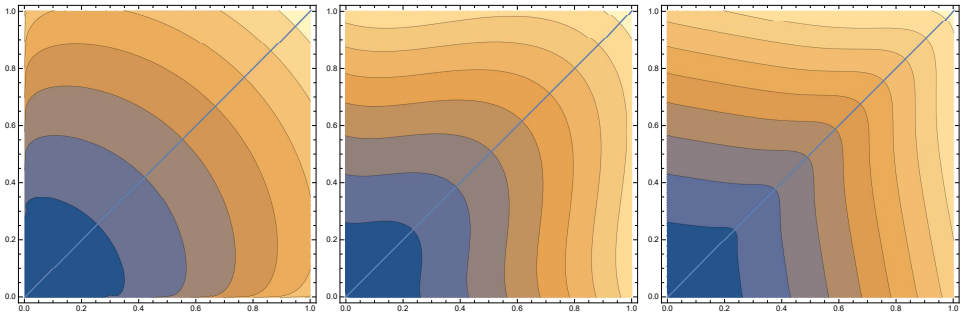


Figure 7: Expected team output level sets for uniformly distributed partner types for $r = .25$ on the left, $r = 2$ in the middle, and $r = 8$ on the right.

of a profit-maximizing principal who may act as a budget breaker and can therefore discipline the team members engaged in production. While we assume that the agents' efforts remain unobservable, we assume that final output is observable and contractible. Moreover, we assume that the principal—although she does not observe the attributes of the chosen leader—knows the skill composition in the team. Therefore, the contract that the principal specifies may depend on the produced output and the composition of the leadership skills in the team (but not on the skills of the assigned leader).

We analyze the same production setup as before in an environment in which a board (the principal) appoints a manager to a team of heterogeneous agents. We model the situation in which this principal may make mistakes in assigning the "correct" leader to the team by assuming that the principal observes ranking information only on agent types' θ, summarized by function f in (7).

Example 5: In the exclusive production environment, assume that the principal pays a fixed wage[19] w and that agents' efforts are observable by the principal (but types are not), and that wages can be conditioned on these efforts. Finally, we assume the same linear production function (8) as in the previous examples.

[19] A similar example for the principal-agent model under task matching exhibits qualitatively comparable effects and is available from the authors.

Then the principal and agents solve the problem

$$\begin{array}{ll} \underset{w(e)}{max} & y = f\left[\theta_1\left(2e_i^1\right) - 2w(e_i^1)\right] + (1-f)\left[\theta_2\left(2e_i^2\right) - w(e_i^2)\right] \\ \text{s.t.} & u_i^j = w(e_i^j) - c(e_i^j) \geq 0. \end{array} \tag{39}$$

Under standard quadratic costs, this is solved by

$$e_i^j(\theta_j) \ = \theta_j, \ w(e_i^j) = \frac{(\theta^j)^2}{2}. \tag{40}$$

Since efforts can be observed by the principal, (s)he can ex-post invert the observed efforts to learn the agents' types. However, this information is not available to her at the ex-ante stage when she makes the leadership assignment. Taking into account the assumed ratio-assignment mistakes (7), the expected equilibrium team output is

$$2\frac{\theta_1^{r+2} + \theta_1^{r+2}}{\theta_1^r + \theta_2^r}. \tag{41}$$

Our usual example confirms the possibility of the Apollo effect in this environment. Figure 8 shows that the principal's equilibrium profit exhibits the Apollo effect in all cases (the team output would show the same effect).

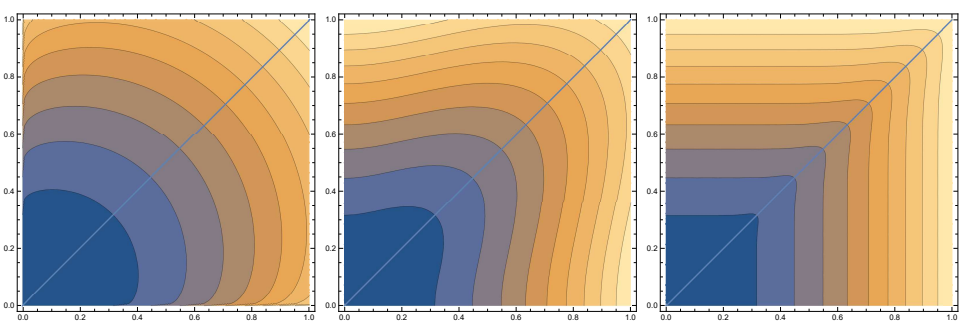


Figure 8: Expected profits in the principal-agent environment with observable efforts. The panels show selection precisions $r \in \{1/4, 2, 15\}$ from left to right.

We proceed to the case of unobservable efforts. We stay in the linear production environment with $y = \theta(e_1 + e_2)$, quadratic costs $c(e) = e^2/2$, and only two possible assignments: $\theta_1 > \theta_2$. The principal sets the wage w based

on the observed output y. Without loss of generality we can assume that the principal pays equal amounts to both agents. The principal wants to induce effort of $e(\theta_1)$ when the assignment is θ_1, and $e(\theta_2)$ when the assignment is θ_2. Therefore, along the equilibrium path, the principal expects to see either $y(\theta_1) = 2\theta_1 e(\theta_1)$ or $y(\theta_2) = 2\theta_2 e(\theta_2)$. Without loss of generality we can assume that there are two wage levels: $w(y(\theta_1))$ and $w(y(\theta_2))$; for any other output, the principal pays a wage of zero.

Hence, the joint problem of the principal and the two agents is[20]

$$\begin{aligned}
\max_{e(\theta),w(y(\theta))} \quad & f\left[y(\theta_1) - 2w(y(\theta_1))\right] + (1-f)\left[y(\theta_2) - 2w(y(\theta_2))\right] \\
\text{s.t.} \quad & (\text{IR}_1): \; w(y(\theta_1)) - c(e(\theta_1)) \geq 0, \\
& (\text{IR}_2): \; w(y(\theta_2)) - c(e(\theta_2)) \geq 0, \\
& (\text{IC}_1): \; w(y(\theta_1)) - c(e(\theta_1)) \geq w(y(\theta_2)) - c\left(\frac{y(\theta_2)}{\theta_1} - \frac{y(\theta_1)}{2\theta_1}\right), \\
& (\text{IC}_2): \; w(y(\theta_2)) - c(e(\theta_2)) \geq w(y(\theta_1)) - c\left(\frac{y(\theta_1)}{\theta_2} - \frac{y(\theta_2)}{2\theta_2}\right).
\end{aligned} \tag{43}$$

The wages (60) and the efforts (61) that solve this problem are derived in the appendix. We insert them into the principal's problem and plot the level sets of the principal's expected profit in Figure 9 for different precision levels of the assignment function f. As isoprofit curves have positive slopes for some type profiles in all cases, we confirm the Apollo effect also in the principal-agent environment.◁

[20] Switching into the standard principal-agent model of i.i.d. types in which $\lambda = \Pr(\theta_1)$ and $1-\lambda = \Pr(\theta_2)$ changes the principal's objective to

$$\begin{aligned}
\max_{e(\theta),w(y(\theta))} \; & 2\lambda(1-\lambda)\left(f[y(\theta_1) - 2w(y(\theta_1))] + (1-f)[y(\theta_2) - 2w(y(\theta_2))]\right) \\
& + \lambda^2(y(\theta_1) - 2w(y(\theta_1))) + (1-\lambda)^2(y(\theta_2) - 2w(y(\theta_2)))
\end{aligned} \tag{42}$$

in which the principal's assignment capability f only matters if non-constant agent type-profiles become realized (which happens with probability $2\lambda(1-\lambda)$). In the other two cases which happen with probability λ^2 and $(1-\lambda)^2$, respectively, the two types competing for leadership are identical. Note that this change leaves the constraints in (43) and therefore agent behavior qualitatively unaffected. Similar but weaker Apollo effects can be observed in such an extended model.

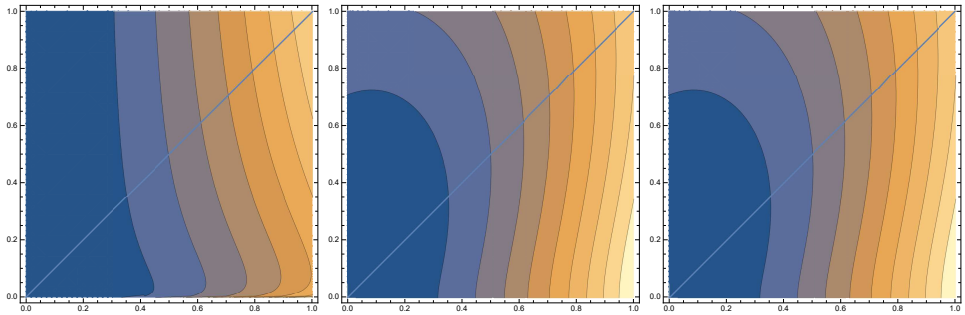


Figure 9: Expected PAM-profits for unobservable efforts that exhibit the Apollo effect. The level sets are drawn for $r \in \{0.25, 2, 15\}$; only region $\theta_1 > \theta_2$ below the diagonal is relevant.

4.4. Simulated Incidence

How much of a problem is the Apollo effect? Because they strive to hire the brightest graduates, by definition, successful law firms, medical or accounting partnerships, etc, are all Apollo teams consisting of competitive individuals whose professional training may not always have emphasized lateral relationship skills. On average, the "real" Apollo teams documented by Belbin (referenced in footnote 1) ranked 4.6^{th} out of 6. Our paper confirms that a theoretical basis for the Apollo effect exists but—anecdotal evidence notwithstanding—we cannot say much about how often it occurs in reality.

Hence, the purpose of this section is to present simulation results which may serve as a partial answer to this question. Our most stringent test of Apollo-plausibility is shown in Figure 7. There, we normalize $\theta_1^A = 1$ and randomly draw θ_2^A. Then, for each such pair, we randomly draw θ_1^B from $[0, \theta_1^A]$ and θ_2^B from $[0, \theta_2^A]$. This is done 100,000 times for each precision-r-step of $1/8$ for $r \in [0, 5]$. In the case of exclusive leadership (Example 1) shown on the left, the Apollo effect occurs at most in 3% of these random draws and is highest at a precision of $r \approx 9/8$. In the task-matching environment (Example 2) shown on the right, the Apollo effect occurs at most in 1.2% of the draws and is highest at a precision of $r \approx 6/8$. In both scenarios, therefore, the Apollo effect will usually be considered a surprise.

If more is known about the team composition—in particular about the high types—then more precise questions are possible. We now consider

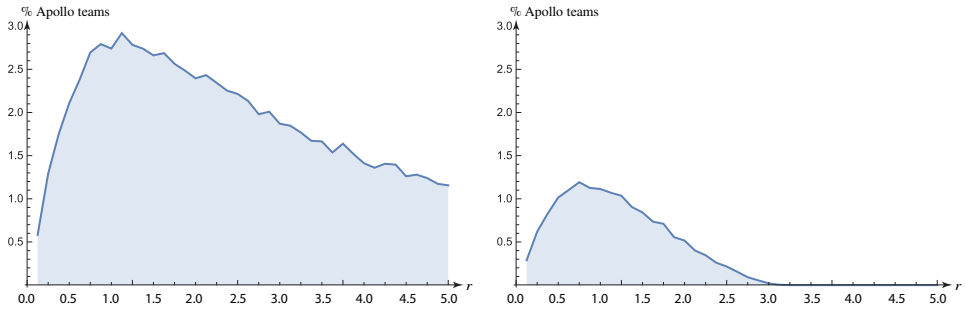


Figure 10: Simulated incidence of the Apollo effect for the exclusive leadership case (Example 1) on the left and task-matching (Example 2) on the right as a function of the selection precision r.

Apollo teams $\theta^A = (1, \theta_2^B)$, set $\theta_1^B = 0.95$ and uniformly drew one hundred thousand partners $\theta_2^B \in [0, \theta_2^A]$. The percentage at which the Apollo effect (here loosely redefined as an underdog team θ^B producing higher output than the reference team θ^A) occurs is shown for a range of selection precisions $r \in \{1/4, 1/2, 1, 2, 3, 4\}$. The results are presented in Figure 11 for exclusive leadership (Example 1) on the left and task-matching (Example 2) on the right. In both cases, team A is composed of $\theta_1^A = 1$, $\theta_2^A \in \left\{\frac{1}{10}, \frac{2}{10}, \frac{3}{10}, \frac{4}{10}, \frac{5}{10}, \frac{6}{10}, \frac{7}{10}, \frac{8}{10}, \frac{9}{10}, 1\right\}$, team B is $\theta_1^A = 0.95$, together with 100,000 uniformly drawn $\theta_2^B \sim U_{[0,\theta_2^A]}$.

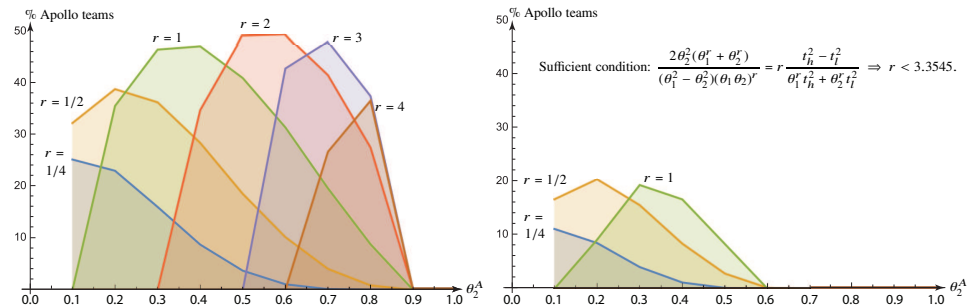


Figure 11: Simulated incidence of Apollo effect: Exclusive leadership (Example 1) on the left and task-matching (Example 2) on the right.

Whether these simulations bear any interesting correlation with what hap-

pens in "real life" depends, of course, on the assumptions we made in Examples 1 and 2, and the chosen simulation parameters. If the observations reported by Belbin (1981) are taken as a yardstick, however, these seem conservative.

5. CONCLUDING REMARKS

Successful law firms, medical or accounting partnerships, etc. strive to hire the brightest graduates for their organizations. By definition, these firms are Apollo teams, consisting of competitive individuals whose professional training may not always have emphasized lateral relationship skills. This paper provides a model for systematically thinking about the implications of this observation.

At its core, the present paper analyzes the influence of the available skill profile on team production in the presence of promotion or other task assignment decisions. To do so, we model team members' skills as exogenous and let an official who has only statistical information on the workers' skills match the team members to tasks or positions. The baseline analysis shows that mistakes of this kind inevitably lead to what is called the Apollo effect: the property that teams composed of weaker individuals may outperform teams of unambiguously higher-qualified individuals in terms of team output. Our model's extensions allow for more complex task assignment or production modes, private information concerning the skills of the workers, and the presence of a profit-maximizing principal. We show that in all cases, to some extent, the Apollo effect cannot be avoided.

Many other economically interesting situations can be modeled with the methodology developed in this paper. For instance, a standard electoral competition model could be enriched through politicians choosing platforms (their "types" in our model) and voters who are unable to perfectly discriminate between these platforms may make mistakes in choosing their candidates. This would presumably counteract the tendency of candidates to move toward the median as such a convergence would maximize the probability of mistakes by the electorate. Another application of a similar idea is the possibility of making mistakes when identifying the "best" bid in general auction environments when (potentially multi-dimensional) bids are close.

This paper presents an analytically rigorous way of generating the Apollo effect in a variety of production environments. The resulting way of thinking about organizations has, in our view, important implications. Effects similar

to those we report for leadership selection are at work for imperfect project selection with unobserved quality and training investments in human capital. Looking beyond the production environment, it can be seen that selecting a speaker from competing party officials, choosing the most promising of several architectural designs, or picking a substitute goalkeeper from sets of alternatives in a soccer team may all give rise to similarly negative effects in terms of expected overall performance.[21]

Appendix

Proof of Lemma 2. We show that while $\partial Y(\theta_1, \theta_2)/\partial\theta_1 > 0$ always holds, it is the case that $\partial Y(\theta_1, \theta_2)/\partial\theta_2 < 0$ if and only if the condition of the lemma holds. In this latter case, there exist $(\eta_1, \eta_2) >> 0$ such that (14) holds. Taking the derivative of (13) with respect to θ_2 gives the change in output for an increase in type θ_2 as

$$\begin{aligned} & f_2(\theta_1, \theta_2)(y(\theta_1, e^*(\theta_1), e^*(\theta_1)) - y(\theta_2, e^*(\theta_2), e^*(\theta_2))) \\ & + (1 - f(\theta_1, \theta_2)) \left(e'^*(\theta_2) \left(y_3(\theta_2, e^*(\theta_2), e^*(\theta_2)) \right.\right. \\ & \left.\left. + y_2(\theta_2, e^*(\theta_2), e^*(\theta_2))\right) + y_1(\theta_2, e^*(\theta_2), e^*(\theta_2))\right), \end{aligned} \tag{44}$$

where $y_e(\theta_2, e^*(\theta_2), e^*(\theta_2)) = y_2(\theta_2, e^*(\theta_2), e^*(\theta_2)) = y_3(\theta_2, e^*(\theta_2), e^*(\theta_2))$. This change is negative if (15) holds. As claimed, the derivative of $Y(\theta_1, \theta_2)$ with respect to θ_1 is

$$\begin{aligned} & f_1(\theta_1, \theta_2) \left[y(\theta_1, e^*(\theta_1), e^*(\theta_1)) - y(\theta_2, e^*(\theta_2), e^*(\theta_2))\right] \\ & + f(\theta_1, \theta_2) \left[e'^*(\theta_1) 2 y_e(\theta_1, e^*(\theta_1), e^*(\theta_1)) + y_1(\theta_1, e^*(\theta_1), e^*(\theta_1))\right] > 0. \end{aligned}$$ □

Proof of Lemma 3. By assumption of symmetry and $f(0, \hat{\theta}) = 0$. □

Proof of Proposition 1. From Lemmata 1 and 2 and the intermediate value theorem, every feasible continuous function f has a range in which the slope of the isoquant is positive. □

[21] See Woolley et al. (2015) for a particularly interesting example of the performance of the Russian (Apollo) ice hockey team at the 2014 Sochi olympics. For an account of other recent dream team failures, see Martinez (2013).

Proof of Proposition 2. The condition for the isoquant to have positive slope, i.e., for the derivative of output $Y(\theta_1, \theta_2)$ from (22) with respect to θ_2 to be negative, is

$$\frac{z_1(\theta_2, \theta_1)}{z_1(\theta_2, \theta_1) - z_2(\theta_1, \theta_2)} < f(\theta_1, \theta_2) - f_2(\theta_1, \theta_2)\frac{z(\theta_1, \theta_2) - z(\theta_2, \theta_1)}{z_1(\theta_2, \theta_1) - z_2(\theta_1, \theta_2)}. \quad (45)$$

Assumptions (19) and (21) imply single-crossing of z_1 and z_2 since

$$\begin{aligned} z_1(\theta_2, \theta_1) - z_2(\theta_1, \theta_2) = e_1^h(\theta_2) y_2^h(\theta_2, e^h(\theta_2)) - e_1^l(\theta_2) y_2^l(\theta_2, e^l(\theta_2)) \\ + y_1^h(\theta_2, e^h(\theta_2)) - y_1^l(\theta_2, e^l(\theta_2)) > 0, \end{aligned} \quad (46)$$

where the second line of the last expression is positive due to the assumption that $y_1^h(\theta, e) > y_1^l(\theta, e) > 0$ and $e^h(\theta_2) > e^l(\theta_2)$. The first line is positive since $e^{h\prime}(\theta_2) > e^{l\prime}(\theta_2)$ and $y_2^h(\theta_2, e^h(\theta_2)) > y_2^l(\theta_2, e^l(\theta_2))$, which, in turn, follows from

$$y_2^h(\theta_2, e^h(\theta_2)) = 2c'\left(e^h(\theta_2)\right) \text{ and } y_2^h(\theta_2, e^l(\theta_2)) = 2c'\left(e^l(\theta_2)\right), \quad (47)$$

$e^h(\theta_2) > e^l(\theta_2)$, and $c'' > 0$. Thus, the left-hand side of (45) exceeds 1 while the term multiplied with $f_2(\theta_1, \theta_2)$ on the right-hand side of (45) is positive. Hence, as $f(\theta_1, \theta_2) \in [1/2, 1]$, a sufficient condition for the Apollo effect to arise for some type profile $\theta_1 > \theta_2$ is (27). □

Proof of Proposition 3. Equilibrium effort functions must satisfy

$$\begin{aligned} e^L(\theta) &\in \arg\max_e \mathbb{E}_{\theta'|\text{ leader has type } \theta}\left[\frac{y\left(\theta, e, e^F\left(\theta'\right)\right)}{2}\right] - c\left(e\right), \\ e^F(\theta) &\in \arg\max_e \mathbb{E}_{\theta'|\text{ follower has type } \theta}\left[\frac{y\left(\theta', e, e^L\left(\theta'\right)\right)}{2}\right] - c\left(e\right). \end{aligned} \quad (48)$$

We calculate the conditional expectations as

$$\begin{aligned} \Pr\left(\Theta \leq \theta' | \text{leader has type } \theta\right) &= \frac{\Pr\left(\Theta \leq \theta' \,\&\, \text{leader has type } \theta\right)}{\Pr\left(\text{leader has type } \theta\right)} \\ &= \frac{\int_a^{\theta'} f\left(\theta, \theta''\right) g\left(\theta''\right) d\theta''}{\int_a^b f\left(\theta, \theta''\right) g\left(\theta''\right) d\theta''}. \end{aligned} \quad (49)$$

Therefore, the density of $(\theta'|\text{leader has type } \theta)$ is

$$\frac{f(\theta,\theta')\,g(\theta')}{\int_a^b f(\theta,\theta'')\,g(\theta'')\,d\theta''}. \tag{50}$$

Therefore,

$$\mathbb{E}_{\theta'|\text{leader has type } \theta}\frac{y\left(\theta,e,e^F(\theta')\right)}{2}=\frac{\int_a^b y\left(\theta,e,e^F(\theta')\right)f(\theta,\theta')\,g(\theta')\,d\theta'}{2\int_a^b f(\theta,\theta'')\,g(\theta'')\,d\theta''}. \tag{51}$$

The first-order condition is given by

$$\frac{\int_a^b y_2\left(\theta,e,e^F(\theta')\right)f(\theta,\theta')\,g(\theta')\,d\theta'}{2\int_a^b f(\theta,\theta'')\,g(\theta'')\,d\theta''}-c'(e)=0. \tag{52}$$

Therefore, $e^L(\theta)$ must satisfy (34). Calculating the conditional expectations for the second case gives

$$\begin{aligned}\Pr(\Theta\le\theta'|\text{follower has type }\theta) &= \frac{\Pr(\Theta\le\theta'\ \&\ \text{follower has type }\theta)}{\Pr(\text{follower has type }\theta)}\\ &= \frac{\int_a^{\theta'} f(\theta'',\theta)\,g(\theta'')\,d\theta''}{\int_a^b f(\theta'',\theta)\,g(\theta'')\,d\theta''}\\ &= \frac{\int_a^{\theta'}\left[1-f(\theta,\theta'')\right]g(\theta'')\,d\theta''}{\int_a^b\left[1-f(\theta,\theta'')\right]g(\theta'')\,d\theta''}.\end{aligned} \tag{53}$$

Therefore, the density of $(\theta'|\text{follower has type } \theta)$ is

$$\frac{f(\theta',\theta)\,g(\theta')\,d\theta'}{\int_a^b f(\theta'',\theta)\,g(\theta'')\,d\theta''}. \tag{54}$$

Therefore,

$$\mathbb{E}_{\theta'|\text{follower has type } \theta}\frac{y\left(\theta',e,e^L(\theta')\right)}{2}=\frac{\int_a^b y\left(\theta',e^L(\theta'),e\right)f(\theta',\theta)\,g(\theta')\,d\theta'}{2\int_a^b f(\theta'',\theta)\,g(\theta'')\,d\theta''}. \tag{55}$$

The first-order condition is given by

$$\frac{\int_a^b y_3\left(\theta', e^L(\theta'), e\right) f(\theta', \theta)\, g(\theta')\, d\theta'}{2\int_a^b f(\theta', \theta)\, g(\theta')\, d\theta'} - c'(e) = 0. \tag{56}$$

Therefore, we know that $e^F(\theta)$ must satisfy (35). □

Derivation of equilibrium efforts and wages in Example 5.
Assume that both (IR_2) and (IC_1) are binding. Combining (IR_2) and (IC_1) gives

$$e(\theta_2) = \sqrt{2}\sqrt{w(y(\theta_2))},\ e(\theta_1) = \frac{\theta_1^2(w(y(\theta_1)) - w(y(\theta_2))) + 4\theta_2^2 w(y(\theta_2))}{2\sqrt{2}\theta_1\theta_2\sqrt{w(y(\theta_2))}}. \tag{57}$$

Inserting these into (IR_1) gives

$$w(y(\theta_1)) = w(y(\theta_2))\frac{(\theta_1 + 2\theta_2)^2}{\theta_1^2}. \tag{58}$$

Inserting this back into the principal's problem in (43) gives her the following unconstrained objective:

$$2\sqrt{w(y(\theta_2))}\left(\sqrt{2}\theta_2 + \frac{(\theta_1 + \theta_2)\theta_1^{r-2}\left(\sqrt{2}\theta_1^2 - 4\theta_2\sqrt{w(y(\theta_2))}\right)}{\theta_1^r + \theta_2^r} - \sqrt{w(y(\theta_2))}\right). \tag{59}$$

Taking the derivative with respect to $w_2 = w(y(\theta_2))$ and solving results in the pair of wages

$$w(y(\theta_1)) = \frac{\theta_1^2(\theta_1 + 2\theta_2)^2\left((\theta_1 + 2\theta_2)\theta_1^r + \theta_2^{r+1}\right)^2}{2\left((\theta_1 + 2\theta_2)^2\theta_1^r + \theta_1^2\theta_2^r\right)^2}, \tag{60a}$$

$$w(y(\theta_2)) = \frac{\theta_1^4\left((\theta_1 + 2\theta_2)\theta_1^r + \theta_2^{r+1}\right)^2}{2\left((\theta_1 + 2\theta_2)^2\theta_1^r + \theta_1^2\theta_2^r\right)^2} \tag{60b}$$

implying efforts of

$$e(\theta_1) = \frac{\theta_1(\theta_1 + 2\theta_2)\left((\theta_1 + 2\theta_2)\theta_1^r + \theta_2^{r+1}\right)}{(\theta_1 + 2\theta_2)^2\theta_1^r + \theta_1^2\theta_2^r}, \tag{61a}$$

$$e(\theta_2) = \frac{\theta_1^2\left((\theta_1 + 2\theta_2)\theta_1^r + \theta_2^{r+1}\right)}{(\theta_1 + 2\theta_2)^2\theta_1^r + \theta_1^2\theta_2^r}. \tag{61b}$$

References

Andreoni, J. (2006). Leadership giving in charitable fund-raising. *Journal of Public Economic Theory*, *8*(1), 1–22.

Aritzeta, A., Swailes, S., & Senior, B. (2007). Belbin's team role model: Development, validity and applications for team building. *Journal of Management Studies*, *44*(1), 96–118.

Bag, P., & Pepito, N. (2012). Peer transparency in teams: Does it help or hinder incentives? *International Economic Review*, *53*(4), 1257–86.

Belbin, R. M. (1981). *Management Teams: Why They Succeed or Fail* (3rd ed.). Oxford, UK: Butterworth-Heinemann.

Bolton, P., Brunnermeier, M., & Veldkamp, L. (2010). Economists' perspectives on leadership. In N. Nohria & R. Khurana (Eds.), *Handbook of Leadership Theory and Practice* (pp. 239–264). Boston, MA: Harvard Business Press.

Chade, H., & Eeckhout, J. (2020). Competing teams. *Review of Economic Studies*, *87*, 1134–73.

Che, Y.-K., & Yoo, S.-W. (2001). Optimal incentives for teams. *American Economic Review*, *91*(3), 525–41.

Cyert, R. M., & March, J. G. (1963). *A Behavioral Theory of the Firm* (2nd ed.). Cambridge, MA: Blackwell Publishers.

Eliaz, K., & Wu, Q. (2018). A simple model of competition between teams. *Journal of Economic Theory*, *176*, 732–92.

Garicano, L., & Van Zandt, T. (2012). Hierarchies and the division of labor. In R. Gibbons & J. Roberts (Eds.), *Handbook of Organizational Economics* (pp. 604–54). Princeton University Press.

Gary, L. (1998, April). Cognitive Bias: Systematic Errors in Decision Making. *Harvard Management Update*.

Gershkov, A., Li, J., & Schweinzer, P. (2016). How to share it out: The value of information in teams. *Journal of Economic Theory*, *162*, 261–304.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*, 451–382.

Hermalin, B. E. (1998). Toward an economic theory of leadership: Leading by example. *American Economic Review*, *88*(5), 1188–206.

Hermalin, B. E. (2012). Leadership and corporate culture. In R. Gibbons & J. Roberts (Eds.), *Handbook of Organizational Economics* (pp. 432–78). Princeton University Press.

Hermalin, B. E., & Weisbach, M. S. (1988). The determinants of board composition. *RAND Journal of Economics*, *19*(4), 589–606.

Holmström, B. (1977). On Incentives and Control in Organizations. *Stanford University*, *PhD Thesis*.

Huck, S., & Rey-Biel, P. (2006). Endogenous leadership in teams. *Journal of Institutional and Theoretical Economics*, *162*(2), 253–61.

Jia, H., Skaperdas, S., & Vaidya, S. (2013). Contest functions: Theoretical foundations and issues in estimation. *International Journal of Industrial Organization*, *31*, 211–22.

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review*, *93*(5), 1449–75.

Kobayashi, H., & Suehiro, H. (2005). Emergence of leadership in teams. *Japanese Economic Review*, *56*(3), 295–316.

Kremer, I., Mansour, Y., & Perry, M. (2014). Implementing the "wisdom of the crowd". *Journal of Political Economy*, *122*(5), 988–1012.

Lazear, E. P. (2012). Leadership: A personnel economics approach. *Labor Economics*, *19*(1), 92–101.

Lazear, E. P., & Rosen, S. (1981). Rank order tournaments as optimal labor contracts. *Journal of Political Economy*, *89*, 841–64.

Lombardi, R., Trequattrini, R., & Battistan, M. (2014). Systematic errors in decision making processes: The case of the Italian Serie A football championship. *International Journal of Applied Decision Sciences*, *7*, 239–54.

Marschak, J., & Radner, R. (1972). *Economic Theory of Teams*. New Haven, CT: Yale University Press.

Martinez, J. (2013, 10-Apr). A recent history of failed dream teams. *Complex*. Retrieved from http://www.complex.com/sports/2013/04/a-history-of-failed-dream-teams/

Mathieu, J., Tannenbaum, S., Donsbach, J., & Alliger, G. (2013). A review and integration of team composition models: Moving toward a dynamic and temporal framework. *Journal of Management*, *40*(1), 130–60.

McAfee, R. P. (2002). Coarse matching. *Econometrica*, *70*(5), 2025–34.

Palomino, F., & Sákovics, J. (2004). Inter-league competition for talent vs. competitive balance. *International Journal of Industrial Organization*, *22*(6), 783–97.

Rajan, R. G., & Zingales, L. (2000). The tyranny of inequality. *Journal of Public*

Economics, *76*(3), 521–58.

Schwenk, C. R. (1984). Cognitive simplification processes in strategic decision making. *Strategic Management Journal*, *5*(2), 111–28.

Shellenbarger, S. (2016, 20-Dec). A manifesto to end boring meetings. *The Wall Street Journal*. Retrieved from http://www.wsj.com/articles/a-manifesto-to-end-boring-meetings-1482249683

Tetlock, P. E. (2000). Cognitive biases and organizational correctives: Do both disease and cure depend on the politics of the beholder? *Administrative Science Quarterly*, *45*, 293–326.

Tirole, J. (2006). *The Theory of Corporate Finance*. Princeton, New Jersey: Princeton University Press.

Waldman, M. (2012). Theory and evidence in internal labor markets. In R. Gibbons & J. Roberts (Eds.), *Handbook of Organizational Economics* (pp. 520–74). Princeton University Press.

Winter, E. (2006). Optimal incentives for sequential production processes. *RAND Journal of Economics*, *37*(2), 376–90.

Woolley, A. W., Aggarwal, I., & Malone, T. W. (2015). Collective intelligence in teams and organizations. In T. W. Malone & M. S. Bernstein (Eds.), *Handbook of Collective Intelligence* (pp. 143–168). Cambridge, MA: Massachusetts Institute of Technology Press.

How to prepare a paper for submission

Before submitting a paper to **this Journal**, the authors are advised to follow closely the following instructions to prepare the paper.

1. Papers submitted to this Journal must be unpublished and original work that is neither under review elsewhere nor will be submitted elsewhere for publication without withdrawing from this Journal.

2. This Journal requires that all results (experimental, empirical and computational) be replicable. All underlying data necessary to replicate results must be made available to the Journal.

3. Papers must be submitted in electronic form (preferably as pdf files) and the size of the text font must be at least 12 point. Each figure and table must be included on the relevant page of the paper and should not be collected at the end of the paper. The list of references should appear after any appendices, as the last part of the paper.

4. There is no restriction on the number of pages of any submitted manuscript. We seek to process any first submission within 3 months. However, very long manuscripts may take considerably more time to review.

5. Each submitted paper should have an abstract of no more than 150 words containing no mathematical formulas, complete with no more than 3 suggested keywords and JEL codes. We also encourage authors to make the introduction of their submitted articles understandable to the widest audience possible.

6. The first page of each submitted paper should contain every author's e-mail address, phone number and affiliation.

7. The editors or the publishing Society will not hold any responsibility for views expressed by authors in this Journal.

How to submit a paper

Papers should be submitted electronically in PDF to the Journal of Mechanism and Institution Design through the website `http://www.mechanism-design.org/`.

Aims & scope of the journal

The Journal of Mechanism and Institution Design aims to publish original articles that deal with the issues of designing, improving, analysing and testing economic, financial, political, or social mechanisms and institutions. It welcomes theoretical, empirical, experimental, historical and practical studies. It seeks creative, interesting, rigorous, and logical research and strives for clarity of thought and expression. We particularly encourage less experienced researchers such as recent PhD graduates to submit their work to the Journal and are sympathetic towards those papers that are novel and innovative but which have been unsuccessful elsewhere. We hope that the published articles will be interesting and valuable to a broad audience from the areas of economics, finance, politics, law, computer science, management, history, mathematics, government, and related disciplines. The journal is an open-access, independent, peer-reviewed, non-profit, English-language journal with the purpose of disseminating and sharing the latest knowledge and understanding of the subject widely.