Untested Membership: Reputation, Ambiguity and International Relations

Thomas Richard Cook
University of Colorado at Boulder, TRCOOK@GMAIL.COM

Follow this and additional works at: https://scholar.colorado.edu/psci_gradetds
Part of the International Relations Commons

Recommended Citation
Cook, Thomas Richard, "Untested Membership: Reputation, Ambiguity and International Relations" (2016). Political Science Graduate Theses & Dissertations. 46.
https://scholar.colorado.edu/psci_gradetds/46

This Dissertation is brought to you for free and open access by Political Science at CU Scholar. It has been accepted for inclusion in Political Science Graduate Theses & Dissertations by an authorized administrator of CU Scholar. For more information, please contact cuscholaradmin@colorado.edu.
Untested Membership: Reputation, Ambiguity and International Relations

by

Thomas R. Cook

BA, University of Iowa, 2006

MPA, DePaul University, 2010

MA, University of Colorado, 2012

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Political Science

2016
This thesis entitled:
Untested Membership: Reputation, Ambiguity and International Relations
written by Thomas R. Cook
has been approved for the Department of Political Science

Professor David H. Bearce

Professor Jaroslav Tir

Assistant Professor Moonhawk Kim

Assistant Professor Amy Liu

Assistant Professor Joshua Strayhorn

Date _______________________

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.
International actors face a dilemma. While agreements with other states can yield substantial benefits, actors are often left uncertain as to whether their potential partners will follow-through on their commitments. Actors attempt to reduce this uncertainty by assessing their partners reputation. Scholars characterize a state's reputation with an international actor as the outcome of a specific sequence of events wherein the state (1) joins an agreement (2) demonstrates compliance or non-compliance with its obligations under the agreement (3) is observed demonstrating compliance/non-compliance by the international actor and (4) comes to have a reputation for reliability with that actor.

But what happens when step 2 never occurs? For example, if a state is in a collective security agreement that effectively deters aggression: without the occurrence of conflict, a state has few opportunities to demonstrate its commitment to the agreement. In this case, outcomes are ambiguous. I call this membership in untested agreements.

Contrary to extant models, membership in untested agreements is actually beneficial to a state's reputation. The benefit of membership in an untested agreement comes from the signal provided by the state's entry into the agreement. When a state enters an agreement, the other members of that agreement implicitly endorse the reliability of the state. I call this entry-endorsement. The benefit of entry endorsement depends on a variety of factors. Characteristics of the endorser influence the reputational benefit felt by the state, specifically, relationship between the endorser and subsequent actors. The role of the reputation observer's recent history also plays a role.

I test the implications of this argument as they apply to two distinct issue areas: international investment and security agreements. Through this, I am able to compare how entry endorsements effects differ depending on the relationship between the endorser and subsequent state actors. When the subsequent actor is another state, then the influence of entry endorsement is substantial, but moderated by the similarity between the actor and the endorser. When the subsequent actor is a sub-state actor, then the influence of entry endorsement is conditioned by the endorsing states vulnerability to making a bad endorsement (i.e. entering into agreements that later fail).
Contents

Chapter

1 Introduction 1

1.1 The argument in brief .................................................. 2

1.2 Plan for this document ................................................... 5

2 Research Question 6

2.1 Why is this important: the value of reputation ......................... 7

2.1.1 Uncertainty Over Types: A Cooperative Impediment ............... 7

2.1.2 Why Reputation? ..................................................... 11

2.2 What do we mean by reputation ......................................... 13

2.3 Our understanding of reputation is incomplete ......................... 14

2.3.1 How does reputation work? .......................................... 14

2.3.2 Ambiguous Outcomes and the question of untested membership .... 19

2.4 Research Question ..................................................... 23
3 A Theory of Reputation

3.1 Explaining the Effects of Untested Agreements: an Intuitive Approach

3.1.1 Reputation and Type Transition

3.1.2 Entry-Endorsement

3.1.3 Entry Endorsement in Different Contexts

3.1.4 Betrayal

3.2 The Model

3.2.1 The Reputation Parameter

3.2.2 Untested Agreements

3.2.3 Untested Agreements Benefit Reputation

3.2.4 Entry-Endorsement Interpretation when Observers are Similar to Endorsers

3.2.5 Entry Endorsement among Linked Actors

3.2.6 Expectations about transition

3.2.7 Betrayal hastens decay

3.3 Summary

4 Reputation and Investment

4.1 Where to look for evidence: Reputation in International Political Economy

4.1.1 Investment and Reputation
4.1.2 Observing reputation in investment ........................................... 58

4.1.3 Testable Implications from Theoretical Propositions .................... 59

4.2 Examining Reputation in Foreign Direct Investment: Design and Methods .......... 66

4.2.1 Unit of Analysis ................................................................. 66

4.2.2 Measurement: Foreign Direct Investment ..................................... 67

4.2.3 Measurement: Untested BIT membership (entry endorsement) ............... 68

4.2.4 Measurement: Endorser Vulnerability ....................................... 70

4.2.5 Measurement: Betrayal ....................................................... 71

4.2.6 Modeling Strategy ............................................................... 72

4.3 Findings .................................................................................. 74

4.3.1 Hypothesis Evaluation ........................................................... 77

4.3.2 Evaluating hypotheses about endorsement quality ............................. 80

4.3.3 Evaluating hypotheses about observer betrayal ............................... 83

4.4 Summary .................................................................................. 87

5 Reputation and Alliance Formation .................................................. 88

5.1 Where to look for evidence .......................................................... 89

5.1.1 Observing reputation in conflict ................................................. 91

5.1.2 Alliance formation and reputation ............................................. 95
5.1.3 Theoretical Expectations and Hypotheses

5.2 Examining Alliance Formation: Design and Methods

5.2.1 Unit of Analysis

5.2.2 Population of Cases

5.2.3 Measurement: Alliance Formation

5.2.4 Measurement: Untested Alliance Membership (Entry Endorsement)

5.2.5 Measurement: Observer-endorser similarity

5.2.6 Measurement: Betrayal

5.2.7 Modeling Strategy

5.2.8 Expectations and Conditions for Failure

5.3 Findings

5.3.1 Hypothesis Evaluation

5.4 Summary

6 Conclusion

6.1 Implications of this study

6.1.1 Implications for our understanding of Reputation

6.1.2 Implications for ‘cheap talk’

6.1.3 Implications for research on the frequency of agreement formation
6.2 Avenues for future research on reputation ................................................. 133

Appendix

A Theory Chapter Appendix ................................................................. 144

A.1 Extended discussion of probability of state transition over time ................. 144

A.1.1 Reputational costs of rejection ......................................................... 145
## Tables

### Table

3.1 Legend of Symbols  

4.1 Model Estimates of FDI investment decisions  

5.1 Alliance Formation Estimates. Full ENDORSER S Interaction (Model 1)  

5.2 Alliance Formation Estimates. Full ENDORSER S Interaction (Model 2)
## Figures

**Figure**

<table>
<thead>
<tr>
<th>Figure</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Cooperation with 2 types of players (B)</td>
<td>9</td>
</tr>
<tr>
<td>2.2</td>
<td>Agreements to cooperate with Ambiguous Types</td>
<td>10</td>
</tr>
<tr>
<td>2.3</td>
<td>Agreements to cooperate with Ambiguous Types</td>
<td>10</td>
</tr>
<tr>
<td>2.4</td>
<td>Assessment of reputation with and without a decay term</td>
<td>19</td>
</tr>
<tr>
<td>2.5</td>
<td>Decay model assessment of reputation assessment when outcomes are perfectly observable and when outcomes are occasionally unobserved</td>
<td>24</td>
</tr>
<tr>
<td>3.1</td>
<td>Model of decision to form an agreement</td>
<td>26</td>
</tr>
<tr>
<td>3.2</td>
<td>Simple Game</td>
<td>36</td>
</tr>
<tr>
<td>4.1</td>
<td>Conceptual examples of low and high levels of entry recency</td>
<td>69</td>
</tr>
<tr>
<td>4.2</td>
<td>Expected number of projects at levels of agreement entry recency when entry is tested and untested</td>
<td>79</td>
</tr>
</tbody>
</table>
4.3 Relative influence of entry endorsement over entry recency at levels of endorser voter vulnerability (polity) .......................................................... 81

4.4 Relative influence of entry endorsement over entry recency at levels of endorser asset vulnerability .......................................................... 82

4.5 Relative influence of entry endorsement over entry recency at extent of observer betrayal .......................................................... 84

5.1 Conceptual examples of low and high levels of entry recency .......................................................... 111

5.2 Alliance Formation on Entry Recency at levels of Ambiguity (Model 1) .......................................................... 117

5.3 Alliance Formation on Entry Recency at levels of Ambiguity (Model 2) .......................................................... 118

5.4 Alliance Formation on Entry Recency at levels of Betrayal (Model 1) .......................................................... 121

5.5 Alliance Formation on Entry Recency at levels of Betrayal (Model 2) .......................................................... 122

5.6 Alliance Formation on Endorser S at levels of Ambiguity (Model 1) .......................................................... 125

5.7 Alliance Formation on Endorser S at levels of Entry Recency (Model 2) .......................................................... 126
Chapter 1

Introduction

In November 2015, the foreign trade and investment minister of Cuba announced a goal of attracting $8.2 Billion dollars in new foreign direct investment\(^1\). A few months later, in 2016, US President Barack Obama visited Cuba, accompanied by representatives of several US firms that have a potential interest in investing in new ventures on the island. At the same time, the US is slowly loosening its restrictions on economic exchange with Cuba, thereby paving the way for such investment to occur. For many US firms, then, it would appear that they will soon face the question, “Should we invest in Cuba?”

Among the central concerns facing any potential investor will be political risk. Namely, potential investors will be concerned about the extent to which they can count on the Cuban government to uphold their claim to property. If the risk is sufficiently low, Cuba will likely refrain from expropriative activity and/or discriminatory treatment of investors. Under such circumstances, investment in Cuba could be quite lucrative (especially in natural resource sectors as well as in its emerging tourism market). If, on the other hand, the risk is high, then the investors will expect expropriative behavior or policies that otherwise impinge upon investor claims to property. Under these circumstances, the likelihood of expropriation will outweigh any potential return on investment.

To assess this risk, investors are likely to take insight from an assessment of Cuba’s reputation. In other words, investors will consider the extent to which Cuba’s prior behavior indicates its commitment to

---

\(^1\) see Reuters. November 3, 2015: Markets
protect investor rights. But this brings a strange observation: the dearth of investment since the 1960’s has left Cuba with few opportunities to demonstrate any changes in this commitment through direct interaction with foreign investors. And yet, sentiment towards Cuba’s commitment to protecting foreign investment seems to have changed. In addition to policy changes allowing foreign investment more widely, Cuba’s membership in investment treaties might be taken as indicative of its commitment to protecting foreign investors\(^2\). In other words, perhaps membership in investment treaties have improved Cuba’s reputation.

It is here, however, that we happen upon a puzzle. The low levels of investment in Cuba have left it with few opportunities to violate its investment treaties – even if it wanted to. Without these opportunities, de facto compliance with investment treaties compels few costs, and Cuba’s commitment to its investment treaties has gone untested. Conventional wisdom suggests that, under such circumstances, Cuba’s membership in these investment treaties amounts to little more than cheap talk. And yet Cuba’s membership in these agreements is cited in scholarly and media reports as indicative of Cuba’s commitment to investor protection\(^3\). Why is this the case? Could mere membership in an untested agreement benefit a state’s reputation?

The Cuban case exemplifies the central puzzle addressed in this document. Namely, “How does membership in an untested agreement influence a state’s reputation?” In answering this question, I will contribute to the broader research on the factors that determine how reputations form and change over time.

1.1 The argument in brief

How does membership in untested agreements influence a state’s reputation? Extant models of reputation do not easily address this question as they typically conceive of a state’s reputation\(^4\) with an international actor as as the outcome of a specific sequence of events wherein the state (1) joins an international

\(^2\) To date, Cuba has signed more than 70 investment treaties. Among its agreement partners are the UK, Italy, Spain, Columbia and Brazil.

\(^3\) See Knowledge at Wharton, April 7, 2015; Globe and Mail, March 22, 2016: Report on Business. See also Hallward-Driemeier (2003, p. 9), and Pérez-López and Travieso-Díaz (2000).

\(^4\) Specifically, its reputation for compliance in international agreements/ attempts at international cooperation.
agreement (2) demonstrates compliance or non-compliance with its obligations under the agreement (3) is observed demonstrating compliance/non-compliance by the potential partner and (4) comes to have a reputation for reliability with that potential partner. These models do not account for how reputation might change if step 2 never occurs. In other words, these models do not explain how mere membership in an untested agreement might influence a state’s reputation.

And yet, membership in untested agreements can be frequent. In addition to the Cuban investment example, untested membership would occur when states enter a collective security agreement that effectively deters aggression\(^5\): without the occurrence of conflict, a state has few opportunities to demonstrate its commitment to the agreement.

Extant models generally ignore membership in untested agreements when calculating state reputation. At the same time, contemporary reputation models\(^6\) employ decay in some capacity. The idea behind decay is that older information is less informative than newer information. Thus, as information relevant to a state’s reputation ages, it becomes less influential on the state’s reputation. The use of decay and ignorance of untested membership in contemporary models means that contemporary models of reputation expect membership in an untested agreements, to produce a de facto corrosive effect on state reputation.

Contrary to extant models, I argue that membership in untested agreements is actually beneficial to a state’s reputation. Broadly, the benefit of membership in an untested agreement comes from the signal provided by the state’s entry into the agreement. When a state enters an agreement, the other members of that agreement implicitly endorse the reliability and/or trustworthiness of the state. I call this entry-endorsement.

The benefit conferred by entry endorsement is conditioned on a variety of factors. Similar to recent scholarly work in this area\(^7\) I argue that characteristics of the endorser influence the quality and strength of entry-endorsement on reputation. I argue that the specific characteristics that determine the quality of

\(^6\) See Crescenzi et al., 2012, as a recent example.
\(^7\) See Gray, 2013; Gray and Hicks, 2014.
entry-endorsement are determined by the relationship between the endorser and reputation observers (i.e. potential investors or potential allies).

Two specific types of relationships between endorsers and observers are considered. The first type of endorser-observer relationship is one in which endorsers and observers are different types of actors, and where (some) observers can punish endorsers for making bad endorsements. This type of relationship typically occurs in the context of international political economy, where sub-state firms are observers and states are endorsers. Under these circumstances, the quality of entry-endorsement is determined by the endorser’s sensitivity to punishment. The quality of entry-endorsement increases as the endorser becomes more sensitive to punishment for making bad endorsements.

The other type of endorser-observer relationship is one in which endorsers and observers are the same sort but where observers are unlikely to punish endorsers for making bad endorsements. This tends to occur where the observers of reputation are both states, and is common in the context of international conflict and security. In this type of relationship, the quality of entry endorsement is determined by the similarity in strategic interests between the endorser and observer. Greater similarity leads to greater benefits from entry- endorsement.

The other major factor that influences entry endorsement is the prior experiences of the observer. This incorporates some notion that reputation is “in the eye of the beholder” – and is inspired by psychological theories of reputation. The fundamental concept that I model and test in this document is that observers are less influenced by entry endorsement if their prior experiences in international agreements have ended badly (i.e. if the observer has been extensively betrayed in the past). Observers that have been extensively betrayed will be increasingly skeptical of entry-endorsement unless the endorsement occurred very recently.

I test the implications of this argument as they apply to two distinct issue areas: international political economy and international conflict/security. Generally, I find that entry-endorsement applies across issue areas and as expected by my theory. The evidence for the role of entry endorsement is nuanced but prevalent in matters of international investment. Evidence in support of the applicability of my theory to security
agreements is even more pronounced.

1.2 Plan for this document

The remainder of this document will be structured as follows. In chapter 2, I discuss the research question in greater detail. I substantiate its importance and proceed to highlight how prior scholarship has ignored the potential reputational impact of membership in untested agreements. Following this chapter, I proceed to develop a theory of reputation in chapter 3. I develop this theory through a formal model. This chapter presents the intuition of the model in plain language and proceeds to a more detailed presentation of the formal model. Through this chapter I develop a set of propositions that follow from the theory and have testable, empirical implications. To demonstrate the general applicability of my theory in international relations, I test its implications in two different issue areas: conflict/security and international political economy. In chapter 4, I examine the applicability of my theory to matters of international political economy, focusing specifically on reputation, investment treaties and greenfield foreign direct investment. In chapter 5, I examine the applicability of my theory to conflict/security, focusing specifically on reputation and security agreements. In chapter 6, I provide concluding thoughts. Generally, I find support for my theory across issue areas, suggesting wide applicability to matters of international relations.
Chapter 2

Research Question

This dissertation addresses the question “How does membership in an untested agreement influence a state’s reputation?”, and it contributes to the broader question of how reputations form and change over time. This chapter discusses the research question in detail. Its goal is to (1) explain why an answer to the question is important and (2) to show that the question has not been properly answered. Simply put, with regard to the former, we should care about the potential influence of untested membership on reputation because we care about the broader concept of reputation and its role in international relations. With regard to the latter, the question asked here has not only been unanswered by extant scholarship, but its ignorance may produce wildly inaccurate expectations about changes in reputation over time.

The remainder of this chapter will proceed in two main sections. In section 2.1 we will discuss the importance of reputation in international relations and thus why we should care about the factors that cause reputation to form and change over time. In section 2.3 we will discuss the extant understanding of how reputation forms and changes over time and the unanswered question that is untested membership in the scholarship on reputation. In this section we will also discuss why, if we leave this question unanswered, it may cloud our understanding of how we expect reputations to form and change over time.
2.1 Why is this important: the value of reputation

In very simple terms, we should care about the impact of untested membership on reputation because, in the study of international relations, we care about reputation. The reason we care about reputation is because we frequently expect it to play a role in enabling cooperation amidst the anarchy that characterizes the international system. Put differently, since reputation can be important for cooperation, we should care about the determinants of reputation and thus we should seek to know if membership in an untested agreement is an important factor for a state’s reputation.

The remainder of this section will elaborate on why reputation is important in the context of international relations. It will begin, in section 2.1.1, by explaining the fundamental cooperative impediment for which reputation can be useful: the problem of uncertainty over types. This discussion is informative for our purposes here, but it is also informative because it characterizes the strategic context in which a reputation forms. We will return to the problem of uncertainty over types throughout this document. In section 2.1.2, we will turn to a discussion of why the mechanism of reputation is unique in enabling cooperation and is likely to be a factor in a variety of international interactions. This should establish an initial appetite for an answer to the question “how does reputation form or change over time”, and thus a desire to see if/how membership in untested agreements influence a state’s reputation.

2.1.1 Uncertainty Over Types: A Cooperative Impediment

Reputation seems like a very intuitive concept, but its value (i.e. usefulness) arises predominately from a specific cooperative impediment: uncertainty over types. Understanding this impediment will help us better understand the value of reputation.

To begin, let me clarify what we mean by ‘types’. Typically, when we talk about an actor as being one type or another, we are essentially grouping actors according to their expected behavior such as ‘aggressive’, ‘passive’, ‘compliant’ or ‘deceitful’. Functionally, we can reduce types to identifiers for different sets of
preferences. To see why this is the case, consider the basic supposition that actors behave in accordance with their preferences (to the extent that their knowledge/beliefs allow). Thus, if types describe the behavior of an actor, they also indicate the preferences of that actor. Furthermore, if we know an actor’s type, we can predict its behavior and this is not really different than saying that if we know an actor’s preferences we can predict its behavior. If, on the other hand, we are uncertain about an actor’s type, we are uncertain about its preferences and we will have difficulty predicting the actor’s behavior (or our prediction may be inaccurate). It is the risk produced by this uncertainty that presents an impediment to cooperation.

To better understand the problem posed by uncertainty over types, let us consider a more direct application of the problem to international relations. In a state of anarchy, there is no world government to enforce contracts. This means that agreements to cooperate are most likely to be successful when they are self-enforcing\(^1\). Put differently, international agreements succeed when the preferences of states mutually favor compliance over defection.

Under certain circumstances, a state’s type can determine whether or not an agreement will be self-enforcing or not. To understand why this is the case, consider two different types of states. Reliable-type, or compliant-type\(^2\) states prefer to comply with their agreements. Unreliable-type or non-compliant-type states prefer to renege on their agreements. In the context of an international agreement, then, a given actor may encounter wildly divergent outcomes of an agreement, depending on which type of partner it faces.

Figure 2.1 illustrates this notion in the form of two 2 \(\times\) 2 games in which players can choose to comply with an agreement or renege. The player (B) chooses an action among the rows (Comply, Renege). Player B’s partner (H), chooses from the columns (Comply, Renege). In both of the games, B’s preferences are the same and they indicate that B is a compliant-type player – that it would prefer to comply as long as its partner also complied. In the game on the left, B’s partner (H), is also a compliant-type. In the game on the right, however, H is a non-compliant-type – it prefers to renege under all circumstances. Now, if player B knew that its partner was an non-compliant-type, then it would know that it was playing the game on the

---

1 See Martin, 1992; Oye, 1986; Stein, 1982.
2 I will use these two types interchangeably.
right (asymmetric preferences game). Accordingly, it would know that its partner would renege and that B should play renege as a best response. If, on the other hand, player B knew its partner was a compliant-type, it would know it was playing the game on the left (harmony). Both B and H would recognize the agreement as self-enforcing and we might expect both B and H to choose to comply with the agreement.

Note then, that we see divergent outcomes based on type. But as long as states are aware of their partner’s type, cooperation (i.e. mutual compliance) is not entirely impeded; as long as a state knows its partners type, and as long as the partner’s type is compliant, we can expect cooperation.

---

**Figure 2.1** Cooperation with 2 types of players (B)

<table>
<thead>
<tr>
<th></th>
<th>Comply</th>
<th>Renege</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>B</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comply</td>
<td>3* 2</td>
<td>0 0</td>
</tr>
<tr>
<td>Renege</td>
<td>2 1</td>
<td></td>
</tr>
</tbody>
</table>

**a)** compliant player B  
(Harmony)

<table>
<thead>
<tr>
<th></th>
<th>Comply</th>
<th>Renege</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>B</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comply</td>
<td>2 0</td>
<td>3* 1*</td>
</tr>
<tr>
<td>Renege</td>
<td>0 1</td>
<td></td>
</tr>
</tbody>
</table>

**b)** Non-compliant player B  
(Asymmetric Preferences)

* Best response

Frequently, however, states face a crucial problem: they do not know, definitively, whether they face a compliant-type state or a non-compliant-type state. This lack of information incentivizes states to renege on agreements that should otherwise be self-enforcing. Figure 2.2 illustrates this dilemma. As illustrated, B enters an agreement with a partner (H), but the partner’s type is unknown to B (indicated by the dashed line). State B must choose whether to comply or renege in the agreement. Assume that players B and H will play the game illustrated in Figure 2.2 exactly once. As given, player B cannot determine (or even estimate) whether it is playing against a reliable-type or an unreliable-type player and thus player B cannot determine whether it is playing harmony (in which we might expect player H to comply) or whether player B is playing an asymmetric preference game (in which H would be expected to renege). With no further information, the least risky (though suboptimal) course of action, therefore, is for player B to play renege.
To extend the implications of the preceding paragraph, it would seem that a lack of information about types would subvert even the formation of agreements. Consider Figure 2.3, which allows player B to reject or accept an agreement, and subsequently to comply with, or renege upon, an accepted agreement. Compared to the potential benefits of cooperation, and so long as there is sufficient risk of non-compliance, player B can be enticed to reject agreements for fairly meager payoffs. Moreover, even if player B does accept an agreement, it may do so with the intent of reneging.

Considering the foregoing discussion, it would appear that lack of information about types poses a significant threat to cooperation, and the parallels between this discussion and the broader discussion about

---

3 This model (and its implications) is roughly analogous to a two phase bargaining model. See Fearon (1998).

4 Player B will reject if the payoff to rejection is between 1 and 2 and is greater than one plus the probability of compliance by player H (i.e. \( u_b(\text{reject}) > 1 + pr(H=\text{compliant})pr(\text{comply}|H=\text{compliant}) \)). If the probability of compliance is sufficiently high (2/3 or higher), then B will reject only if the payoff to rejection is higher than three times the probability of compliance (i.e. \( u_b(\text{reject}) > 3 \times pr(H=\text{compliant}) \times pr(\text{comply}|H=\text{compliant}) \).
information problems in the cooperation literature should be obvious. It is also worth noting that, just as with the broader literature on cooperation, states (in the figures above, players H) have an honesty problem with respect to revealing their type. That is, while the impediments to cooperation would be somewhat resolved if compliant type players H could communicate their type to player B, they face difficulty in doing so. This is because, non-compliant type players H have incentives to misrepresent themselves as compliant so as to profit from forming and then reneging upon agreements with players B.

2.1.2 Why Reputation?

The foregoing section discusses reasons that uncertainty over types should impede many attempts at cooperation. The literature provides a variety of mechanisms through which states might overcome impediments to cooperation. These include features of repeat play, institutional mechanisms that enable and/or coordinate distributed punishment, incorporation of side-payments, and processes of norm diffusion. Of these, however, reputation stands out in addressing the problem of uncertainty over actor types discussed above. Specifically, reputation stands out for two key reasons: (1) it enables cooperation in (infrequent) interactions among multiple actors, and (2) it provides ex ante insight into the feasibility of cooperation.

Reputation proves especially useful where many actors interact and/or the likelihood of (short-term) repeated interaction between actors is low. These are circumstances that explicitly undercut the role of repeat play and make for a rather short ‘shadow of the future’. Accordingly, these are circumstances in which we might expect different types of players to behave differently, according to their divergent short-term preferences. An intuitive illustration of these circumstances comes as we consider early internet marketplaces

---

6 This literature is extensive. See Axelrod, 1984 for the classic work on this issue, further see Morrow, 2001; Morrow, 2007 for a discussion of the role of reciprocity, for substantial refinements and further discussion, see Fudenberg and Maskin, 1986; Fudenberg and Tirole, 1991 among others. Reputation is related to repeat play, but, for my purposes here, I consider it to be its own distinct concept.
7 See Drezner, 2000; Martin, 1992.
8 See Bearce and Bondanella, 2007; Checkel, 2005; Wendt, 1999.
9 Such circumstances are hardly rare in IR, as there are frequently issues for which repeat interaction is of low value or for which repeated interaction is rare. In economic issue-areas, where sub-state actors rely interact with partner-state governments, the sheer number of sub-state actors for a state to interact with can reduce the value (to the partner state) of repeat interaction with the same sub-state actor in the future. Conversely, in security issue areas, the relative rarity of circumstances that necessitate or motivate interstate cooperation with a given state make repeated interactions with that same state in the near-term somewhat unlikely.
such as eBay\textsuperscript{10}. Similar to the international system, early internet marketplaces were characterized by de facto anarchy (i.e. having no strong, formal authority to enforce contracts) with many potential partners (i.e. sellers and buyers)\textsuperscript{11}. In these early marketplaces, potential buyers could not easily know whether they faced a trustworthy or fraudulent seller. If they faced a fraudulent seller, buyers would send money and receive nothing in return. If, on the other hand, buyers faced a trustworthy seller, buyers would send money and receive goods in trade. Potential buyers faced considerable risks when making purchases in online marketplaces and we might otherwise expect that, given these risks, the overall volume of exchange would be low.

Under these circumstances, the effect of reputation is to enable actors to assess the type of a potential partner by considering that potential partner’s interactions with other players in the past. Thus, even if two actors have not directly interacted in the past, cooperation remains viable (assuming that the potential partner has a sufficiently good reputation). To return to our eBay example, by attaching a record of a seller’s past exchanges (and the satisfaction of previous buyers with said exchange), eBay and similar online marketplaces have enabled buyers to evaluate the reputation of a seller and better determine whether a seller is trustworthy or fraudulent. Recent scholarship indicates that the use of such a reputation system improves the efficiency of exchange\textsuperscript{12}, and is essential to the viability of online marketplaces as a viable forum for commerce.

Other scholarship in IR confirms the intuition discussed here, though it does not often discuss it explicitly. Tomz\textsuperscript{13}, alludes to the relationship between multiple players and the insufficiency of retributive features of repeat play as singularly explaining patterns of of sovereign debt lending. Crescenzi et al.\textsuperscript{14} does not explicitly discuss reputation mechanisms in contrast to repeat play mechanisms, but the authors do explicitly focus on the role of reputation in allowing an actor to derive an assessment of type from a partner’s prior alliances with other actors.

\textsuperscript{10} Resnick and Zeckhauser (2001); Bolton, Katok, and Ockenfels (2002); and Miller, Resnick, and Zeckhauser (2002) for general discussion of reputation, eBay, and game-theoretic cooperation problems. See Milgrom, North, et al. (1990) for a broader discussion of reputation in large markets

\textsuperscript{11} See Miller, Resnick, and Zeckhauser, 2002; Resnick, 2001.

\textsuperscript{12} See Bolton, Katok, and Ockenfels, 2002.

\textsuperscript{13} 2007, pp. 5-9.

\textsuperscript{14} 2012.
Reputation also uniquely enables cooperation by enabling actors to identify actor types \textit{ex ante}, enabling cooperation even without high likelihood of future interaction or strong punishment for non-compliance. This is, of course, a different way of saying that reputation allows actors to deliberately select into agreements that will be self enforcing or otherwise yield an alignment of interests among actors. This is important as it provides some insight as to how international actors can achieve cooperation when the consequences of defection would be dire for one (or more) actors. In circumstances such as a collective security agreement, for example, the cost of defection by one member of an alliance can leave the remaining members vulnerable in the event of attack. By examining a partner’s prior behavior (i.e. its reputation), a state can select allies that are least likely to defect on an alliance. Mechanisms of coordinated punishment for defection, or rejection of future alliances will not ensure the compliance of allies in the same way. The damage from severe defeat by a betrayed state may render its threats of retribution as incredible or, at the very least, not renegotiation proof\textsuperscript{15}.

\section*{2.2 What do we mean by reputation}

Before proceeding, and to resolve any confusion, I wish to further clarify about my usage of the term reputation. Colloquially, we tend to use the term reputation with reference only to an individual state. For example, we might say “North Korea has a reputation for making idle threats” or, “the US has a reputation for repaying debtors”. Accordingly, we sometimes think of reputation as a property of an actor. Functionally, however, reputations are the beliefs of one actor about another actor’s type. In this sense, then, reputations constitute a relationship between two actors.

The first actor is the actor that we typically associate with the reputation, or the actor that “holds” the reputation. I will refer to these actors as \textit{reputation holders} or \textit{holders}. In the examples from the preceding paragraph, the reputation holders are North Korea and the United States. The contribution of a holder to a given reputation is constituted (generally) of its prior behavior.

\textsuperscript{15} See Downs and Jones (2002, pp. 105-106) for a helpful discussion of this. Complimentary discussion with respect to the deliberate selection of partners that are likely to comply is provided by Chayes and Chayes (1993), and Downs, Rocke, and Barsoom (1996)
The second actor is the actor that forms beliefs about the first actor. I will refer to these actors as reputation observers, or simply observers. Reference to reputation observers are omitted from most colloquial uses of the term reputation. The contribution of an observer to a given reputation is its interpretation of the prior behavior of a holder. The observers updated beliefs, based on its interpretation of prior behavior are what we generally conceive as the content of a reputation.

This document focuses primarily on the contributions of the reputation observer. That is, it is primarily focused on how an observer views the behavior of a reputation holder, and how those views are changed. Where I refer to reputation change, then, I am referring to changes in the beliefs of an observer about a holder.

2.3 Our understanding of reputation is incomplete

Our understanding of how reputation forms and changes over time is incomplete. To show this, I will first describe the contemporary understanding on reputation in IR. I will then highlight a question that is not satisfyingly addressed by the extant literature: how does membership in an untested agreement influence a state’s reputation. I will proceed to discuss why leaving this question unanswered may significantly limit our understanding of the determinants of reputation.

2.3.1 How does reputation work?

Given that reputation is useful for enabling cooperation when players are uncertain about their partner’s type, the question remains, “how, precisely, does reputation work?”

Let us consider the scholarship on reputation. Generally, the literature can be divided into two lines of scholarship. In one line, scholars discuss the holder’s prior behavior and its use as the basis of retribution in present and future iterations of a repeated interaction. This line of scholarship is focused on the notion of

\[^{16}\text{The two obvious reasons being that (1) the observer is implied to be the person speaking, and/or (2) the observer is implied to be many/all other actors.}\]
reputation costs as a driver of sustained cooperation in the context of (nearly) complete information. In the other line of research, scholars discuss reputation as a signal to observers. This line of scholars focuses on the use of reputation by observers to determine the credibility of a holder’s commitment to a course of action (i.e., the use of reputation to determine holder’s type) in the context of incomplete information.

Both lines of scholarship, however, are focused predominately on the consequences of reputation.

Indeed, only a small subset of the scholarship on reputation explores the precise way that observers interpret a holder’s prior behavior when assessing its type. In other words, there has been limited research on how reputations form and change. Broadly, we can divide the theories on how reputation forms and changes into two groups, based on the role of new information.

2.3.1.1 Psychological Theories of Reputation

One group of research on reputation formation derives a theory of reputational change that services the broader argument that states should not act to preserve their reputation (or, alternatively, that reputation is not important in international relations). These are arguments offered, most recently, by Mercer. The novelty of this approach to reputation comes primarily from its incorporation of psychological biases into the observer’s interpretation of a holder’s prior behavior.

The psychological approach to reputation has been revised at various times and presented in various forms. The psychological model of reputation is generally described as follows. Observers have a history of prior experiences and a distinct perspective on the world. Accordingly, observers, are likely to filter new information through a unique perceptual lens.

Holder states exhibit prior behavior that is potentially revelatory of type. That is, the holder takes action that would be costly to one type of player but otherwise preferable to other types of players. As

18 See Kreps and Wilson (1982), Huth (1997), and Mercer (1996), among others
19 1996; 2013
20 This line of reasoning is novel in its own right, but has roots dating at least as far back as Jervis (1968; 1976)
21 See Jervis, 1968, 1976, for complementary discussion.
applied to our discussion to this point – which as focused on reputation for reliability (i.e. compliance), the prior behavior of holders would be their record of compliance or non-compliance in prior cooperative agreements.

Additionally, this approach to reputation proposes that, under certain circumstances, a holder’s record of prior behavior may reflect interactions where the holder was forced to act against its type. For example, while a holder may be a compliant type, it may have previously reneged on an agreement because exogenous factors made compliance difficult or impossible. Note, then, that observers can interpret a holder’s prior behavior in two ways: as indicative of its type or, alternatively, as a function of exogenous circumstances.

Changes in reputation are dependent on whether observers view holders’ past behavior as revelatory of type (dispositionally driven) or driven by circumstance (situationally driven). If observers view a holder’s past behavior as situationally driven, then the observer’s assessment of the holder’s type does not change. If, on the other hand, observers view the holder’s past behavior as dispositionally driven, then observers will update their assessment of the holder’s accordingly.

Perceptual biases and prior experiences determine which interpretation observers apply when considering a holder’s prior behavior. Theories exhibit variation on this point. Some variants of this approach to reputation suggest, effectively, that the observer’s prior beliefs determine how it interprets holder prior behavior. Instances of holder behavior that conform to observer’s prior beliefs are taken as dispositionally driven while those actions that contradict prior beliefs are assessed as situationally driven. Earlier variants of the same theory suggested that the desirability of the behavior determined its interpretation by an observer. Per this theory, desirable behavior, would be assessed as dispositionally driven while undesirable behavior would be assessed as situationally driven.

---

22 Proponents of this type of reputation model (namely, Mercer) tend to focus on reputation for resolve. Accordingly, they tend to consider holders types as honest/dishonest, and view follow-through (or failure to follow-through) as preferable/costly to different types of holders. Note, however, that just as the prisoner’s dilemma can be applied to many different substantive scenarios, this model of reputation can be applied equivalently to either reputation for reliability or reputation for resolve.

23 Or, at a minimum, that observers believe that the holder was forced to act against type

24 e.g. a holder that reneges is an unreliable type

25 i.e. if observers view the holder as acting as dictated by its type, unperturbed by extraneous or unexpected circumstances. Mercer refers to this as dispositionally driven

26 Derived from various influences including feelings of animosity or affinity experienced by the observer, see Mercer, 2013
2.3.1.2 Rationalist theories of reputation

I will discuss the other class of models of reputational change as ‘rationalist’ theories. This choice of terminology is intended, primarily, to distinguish this class of theories from the psychological approach to reputation discussed above, and is common in the literature on reputation.\(^{27}\)

Generally, the rationalist approach to reputation is more common and, accordingly, there is more variation in the details of the rationalist models of reputational change. Generally speaking, however, rationalist models of reputational change feature a common theoretical foundation. Holders have types that dictate their (short-term) preferences. Observers can examine a holder’s prior behavior to determine if a holder has complied or not complied in prior attempts at cooperation.\(^{28}\) With a sufficiently long record of prior behavior, an observer can estimate a holder’s type and predict the likelihood of compliance/non-compliance in future agreements. Occurrences of non-compliance degrade a holder’s reputation for reliability among observers. Occurrences of compliance should improve or maintain a holder’s reputation for reliability. In its simplest formulation, we can specify a model of reputation as a simple mean of observed compliance:

\[
\text{reputation} = \frac{\text{total instances of compliance}}{\text{total number of interactions}}
\]

Scholars have expanded models of reputational change beyond these basic foundations to gain additional theoretical leverage and to craft models that better reflect reality. In doing so, rationalist models of reputational change have become more diverse and complex.\(^{29}\) There is one key innovation in reputation models that requires extensive discussion here: the use of a decay term to place primacy on recent events.

A basic model of reputation, such as the one illustrated in Equation 2.1 imply that the assessment of reputation grows increasingly inflexible and unresponsive over time. That is, if reputation is to be calculated as a running average of prior behavior, then the marginal effect of each successive instance of prior behavior (i.e. observed prior instances of compliance or non-compliance) will decrease towards zero. On a long enough

---

\(^{27}\) See Huth, 1997, 1999; Mercer, 1996, 2013, for examples of this distinction.

\(^{28}\) Again, this can be applied to other contexts by substituting appropriate terminology. See footnote 22.

\(^{29}\) An appreciation of this diversity is expertly discussed at length by Huth (1999), and Huth (1997). It is outside of the scope of this document to review each of the rationalist models of reputational change in detail.
time-line, for example, a holder with a reputation for reliability will essentially maintain that reputation even
it has reneged on each of its most recent agreements. There are three reasons, then, why this simple model is
unsatisfactory as a model of reputation. First, such a conception of reputation may allow unreliable players
to reliably bluff by building up a reputation as reliable and then using said reputation to enter into, and
renege upon, agreements with unsuspecting observers. Second, such a conception ignores the very likely
possibility that holder types evolve and change over time. Lastly, such a formulation of reputation does not
fit with the basic intuition that an actor’s most recent behavior should more accurately signal an actor’s
type (in the present) than the behavior of that actor in the distant past.\(^{30}\)

To remedy this, scholars have derived models of reputation that confer primacy to the most recently
observed events. The most common way that this has been done is through use of a decay term. As
applied to the calculation of reputation, the use of a decay term discounts the influence of prior behavior
as its temporal distance increases. The specific formulation of this term may vary, but the decay term
is generally structured such that, as the most recently observed behavior of a holder fades into the past,
the holder’s reputation for reliability (assessed as a probability of compliance in a new agreement) declines
towards zero.

A broad framework for models of reputation of this type is specified as follows:\(^{32}\):

\[
\text{reputation} = \frac{\sum_{t=1}^{T} \text{behavior}_t * \text{decay}^{T-t}}{\sum_{i=1}^{T} \text{decay}^i}
\]

(2.2)

Where the decay term is bounded on the interval \((0,1]\), and where \(T\) is the current time, and \(t\)
indexes the time at which an observable instance of prior behavior occurred. The application of a decay
term using this basic framework allows the assessment of reputation to be more responsive to changes in
player type. This is illustrated in Figure 2.4, which shows the assessment of a simulated holder state’s prior

---

\(^{30}\) See Mercer (2013) for a discussion of these problems. Mercer’s argument requires that we conceive of rationalist models of
reputation as only fitting the simple formulation expressed in Equation 2.1

\(^{31}\) There are other ways of doing this, as we shall see later. Other scholars, outside of the study of political science, have
addressed this problem by employing hidden Markov models of reputation. See Mailath and Samuelson (2014), Wang, Ou, and

\(^{32}\) Note that this model is more simplified than the model specified by Crescenzi et al. (2012) and is likely more simplified
than other decay-based assessments of reputation. The model as presented here is general enough to apply in various contexts
and yet illustrates the importance of a decay-based model in that it privileges the more recent observed behavior.
behavior. The dashed, grey line indicates the assessment of a holder’s reputation at a given time (with one observable instance of prior behavior in each time period) using a decay-term based reputation as specified in Equation 2.2 and the solid, black line indicates the assessment of a holder’s reputation using a simple model as prescribed in Equation 2.1. At time 100, the holder’s type changes from reliable to unreliable (its true probability of compliance changes from .8 to .3). Note that the assessment by the decay-term model adjusts to the change in holder type much more rapidly (indicated by the sharp decline in the dashed, grey line after time 100) than the simple model (indicated by the slight decline in the solid, black line after time 100). The implication of this figure is that the simple reputation model is far less capable of accurate reputation assessment in situations when a player’s type can change over time.

2.3.2 Ambiguous Outcomes and the question of untested membership

To this point, we have discussed the role of reputation and prior behavior with the presumption that the prior behavior of a holder clearly indicated whether the holder was compliant or non-compliant. In other
words, that compliance or non-compliance in prior interactions was clearly observable. Arguably, however this is not always the case.

Frequently, agreements are structured such that members are only called to act if some exogenous event occurs. When this exogenous event does not occur, members are not obliged to take action and may they may not have any ability to demonstrate their willingness to comply with the agreement. Collective security agreements provide a good example of this scenario. A member of a collective security agreement, demonstrates their compliance with the agreement by providing aid to allies that have come under attack, and it reneges by failing to come to the aid of members that have come under attack. Often, however, security agreements will effectively ward off potential attackers. When this occurs, the holder’s membership in the agreement is untested and the outcome of a holder’s membership in the agreement is neither compliance nor defection, but is instead ambiguous.

In an anarchic system where survival is by no means guaranteed, we may feel compelled to argue that observers view ambiguous outcomes skeptically. It is foolhardy to speculate on what a holder ‘would have’ done if circumstances demanded a demonstration of compliance, and lending holder states the benefit of the doubt (i.e. assuming that they ‘would have’ complied if given the opportunity) generates greater risk that observers may unwittingly enter into agreements with unreliable holder states. Accordingly, the most conservative (or skeptical) approach to ambiguous outcomes would be one in which observers assume holders are unreliable unless they clearly demonstrate compliance. This approach would view ambiguous outcomes as equivalent to an instance of reneging. A less conservative (though still skeptical) approach to ambiguous outcomes would be one in which observers simply omitted ambiguous outcomes and instead assessed the reliability of holder states on the basis of clearly observable instances of compliance or non-compliance.

Extant models of reputation assessment (especially those described in section 2.3.1.2) do not typically address the issue of ambiguous outcomes directly, but they do (implicitly) espouse the skeptical view. Authors either explicitly omit ambiguous outcomes from consideration, or they structure their analyses so that

---

34 Readers should note that the terms ambiguous outcome and untested membership are roughly equivalent and I will use them interchangeably throughout.
ambiguous outcomes are excluded from their data. In Tomz’s recent work on sovereign debt, for example, a lender (observer) only assess the behavior of a debtor (holder) once repayment of a debt has come due. Loans that have been disbursed but have not entered repayment are omitted from an observer’s calculations about debtors type. The model of reputation assessment proposed by Crescenzi et al., on the other hand explicitly consider, and then ignore ambiguous cases as uninformative. In both cases, the decision to ignore ambiguous outcomes is tantamount to considering ambiguous outcomes as an indicator of non-compliance. To see why this is the case, recall that rationalist theories of reputation assessment generally impose a decay term or some other mechanism by which they can confer primacy to the influence of recent instances of prior behavior. By setting aside ambiguous outcomes, observers are forced to look further back in time to find instances of compliance. The application of a decay term degrades the impact of these instances of compliance on holder reputation. In other words, this approach to ambiguous outcomes – because it is combined with a decay term – demands that membership in untested alliances erodes reputation.

The reputationally corrosive effect of omitting ambiguous outcomes in reputation assessment is illustrated in Figure 2.5. This diagram presents the assessed probability of compliance of a simulated reputation holder (who is compliant with a probability of .6) using two different models. One model, represented by the dashed, grey line, is the assessment of the holder’s probability of compliance if each of the holder’s prior outcomes had been unambiguous (i.e. the holder was clearly observable as compliant or non-compliant). The other model, represented by the solid, black line, is an assessment of the holder’s reputation if the holder had frequently encountered ambiguous outcomes and if those outcomes were omitted from the assessment of the holder’s reputation. This model (solid, black) reflects the approach of extant models to ambiguous outcomes. The occurrence of ambiguous outcomes are indicated by vertical black bars between the dashed, grey and solid, black lines. Each black bar represents an outcome that the solid, black line is treating as ambiguous and therefore is ignoring in the assessment of the holder’s reputation. The printed numerals 1 and 0 along each line indicate whether the holder (would have) complied (1) or (would have) reneged (0). Intuitively, we see that the ambiguous outcomes (solid, black) model assessment of repeated ambiguity

\[35\quad 2007\]
\[36\quad \text{These are agreements that are functionally, yet temporarily ambiguous}\]
\[37\quad 2012, \text{footnotes 9 and 16}.\]
follows a smooth inverse square curvature that is consistent with the nature of the decay term presented in Equation 2.2 and approaches zero with each ambiguous outcome (consecutive black bar). By comparison of the ambiguous model (solid, black) to the non-ambiguous model (dashed, grey) we see that the wholesale ignorance of ambiguous outcomes causes the ambiguous model to considerably underestimate the holder’s true probability of compliance. This is especially pronounced where the two models diverge in their predictions, which occurs any time an ambiguous outcome occurs (black bar) and the holder would have complied (indicated by a printed numeral 1). When this occurs, the non-ambiguous model (dashed, grey) adjusts its estimate of the holder’s reliability upwards and the ambiguous model (solid, black) adjusts its estimate of the holder’s reliability downwards.

Readers may, by this point, object that, although the skeptical view of ambiguous outcomes described in the preceding paragraphs tends to underestimate a holder’s actual type, it may still be an accurate representation of how international actors approach ambiguous outcomes. I do not, at the outset disagree with this sentiment. In fact, it seems quite sensible, given that a skeptical view of ambiguous outcomes would protect observers from entering into agreements with unreliable holders. But extant models do not provide evidence to support (or even test) this claim. Instead, extant models make strong assumptions about how states approach ambiguous outcomes and do not allow assessment of ambiguous outcomes to vary. In order to examine how observers approach ambiguous outcomes, we will need a different model that allows the assessment of reputation to vary in response to ambiguous outcomes.

The question, of course, remains as to whether we have good reason to suspect that observers would ever view ambiguous outcomes with anything other than skepticism. Indeed, we do. Much of the following chapter is dedicated to developing an explanation of why observers would view ambiguous outcomes as indicators of reliability. I argue, essentially, that when a holder enters into an agreement, it has been assessed by its partners as likely to be compliant, and that this assessment functions in as an endorsement of the holder’s likelihood of compliance in the future. Thus, even when an agreement ends ambiguously, the holder’s membership in an untested agreement provides a signal about its type.

---

38 This is generally consistent with recent findings by Gray and Hicks (2014), regarding the extent to which state reputations improve from membership in agreements with reputable states.
2.4 Research Question

Reputation matters for cooperation, especially under conditions of uncertainty. Extant models of reputation assessment are focused on the outcomes of prior interactions that clearly demonstrate the compliance or non-compliance of holders. Further, these models ignore or exclude prior outcomes that are ambiguous. As such, they ignore the influence of untested membership on reputation. As I have illustrated here, this imposes a strong, negative effect on the reputation of holders that have not had the opportunity to demonstrate compliance or non-compliance. As I will show, it is reasonable to expect that untested membership can actually benefit a state’s reputation. We have arrived, then, at the central question to which the remainder of this document will answer: how do observers incorporate untested membership into their assessment of the reputation of other international actors?
Figure 2.5: Decay model assessment of reputation assessment when outcomes are perfectly observable and when outcomes are occasionally unobserved.
Chapter 3

A Theory of Reputation

This chapter provides a general answer to the question: how do observers incorporate membership in untested agreements into their assessment of the reputation of other international actors? That is, it provides a theory that (1) explains how reputation should be assessed when holders experience ambiguous outcomes and (2) generates specific predictions depending on a particular context or issue area. The theoretical insights discussed here are the product of a formal model.

For the benefit of readers, I begin the chapter by presenting the general intuition of the formal model, in plain language. In section 3.2, I present the mathematics of the model itself.

3.1 Explaining the Effects of Untested Agreements: an Intuitive Approach

My concern in this document is to develop a better understanding of how observers assess the reputation of holder states. To do this, however, it is crucial that we consider the role of reputation in the broader context of an observer’s decision to accept an agreement with the holder. This is important because it emphasizes the potentially crucial role of reputation and for reasons that will become more apparent in section 3.1.2.

To begin, then, let us consider the formation of a bilateral agreement between a holder and an observer. Though it is a simplification, consider that an observer has essentially two options when deciding to form...
an agreement with a holder state: (1) accept the agreement or (2) reject it. If the observer rejects the agreement, then the observer forgoes the benefits of partnership while it looks for a new partner – in other words, the observer ‘goes alone’. If, on the other hand, the observer accepts the agreement, then at some later date, the holder will either fulfill its agreement obligations or renege upon them. The extent to which the holder state comes through on its agreement obligations is a function of its type (which is either reliable or unreliable). If the holder is reliable, then it fulfills agreement obligations, if it is unreliable, then it will renege upon them.

Figure 3.1 provides a diagram of this precise scenario. Moving from left to right, we see an initial draw of the holder’s type, and the observer’s subsequent lack of information over which type it faces (indicated by the dashed line). The tree continues to show the preferred behavior of the holder if it is accepted into an agreement (indicated by a thick black line).

**Figure 3.1 Model of decision to form an agreement**

For the observer, forming an agreement is only worthwhile if the probability that the holder is reliable surpasses a critical threshold formed by: (1) the potential savings from avoiding defection and going alone relative to (2) the cost of an agreement with an unreliable partner:
The observer can examine the holder state’s prior behavior and form a prediction of the holder’s type. The formation of this prediction is the assessment of the holder’s reputation, by the observer. Ultimately, then, if the assessment of the holder’s reputation for reliability is greater than the threshold discussed above, then the observer would expect to benefit from forming an agreement and would therefore choose to do so.

### 3.1.1 Reputation and Type Transition

Understanding how we expect reputation to influence an observer’s decision to form an agreement, we can turn our focus to how, specifically we expect an observer to interpret a holder’s past when assessing its reputation.

To begin, we focus on how time interacts with reputation. As I discussed in section 2.3.1.2, rationalist models of reputation would expect the influence of prior behavior to erode over time, or decay. But why, precisely, do we expect this to occur? If we think about information more generally, it is obvious that information does not lose its value simply because it grows older. No, instead, old information loses its value when it is replaced by new information or when the referent of the information is subject to change. Applied to reputation: old behavior loses its informative value because the holder state may have changed its type in the interim.

To capture the decay caused by transition from one type to another, then, I model the observer’s assessment of a holder state’s type as the output of a Markov process. This simply means that the observer

---

1 Note additionally, that since agreement formation is directly observable and reputation is not, the hypotheses presented and empirical analysis conducted will focus on agreement formation as reflective of a reputation for reliability.
believes that the holder state’s type today is determined by (1) whatever its type was yesterday (as revealed by its observable behavior) and (2) some probability that the holder state has changed types (i.e. from reliable to unreliable or vice versa) since then\(^2\). With increases in the period of time between the present and when an observer last knew a holder state’s type, the observer’s certainty over the holder’s type decreases (i.e. decays). This decline in certainty represents a decline in the holder’s reputation for reliability.

By modeling reputation in this way, the speed with which a holder’s reputation declines is critically dependent on the observer’s perspective. Specifically, it is dependent on the observers beliefs about the frequency and probability with which states transition from reliable to unreliable. I will return to this point momentarily.

3.1.2 Entry-Endorsement

Having established a context in which reputation is useful, and mechanisms by which holder states change type (and consequently, by which decay is applied to reputation), we turn to the observer’s interpretation of ambiguous outcomes.

I submit, generally, that observers interpret ambiguous outcomes positively – as reflective of a holder’s reliability. Consider that ambiguous agreement outcomes (i.e. membership in untested agreements) are still instances in which a holder has participated in an agreement and, as such, should provide some information about its type. Specifically, the holder’s entry into an agreement should reflect positively on the assessment of its type since it indicates that the holder’s partners in the past agreements have determined that its type, at time of entry, was sufficiently likely to be reliable. In other words, the entry of the holder into an agreement serves as an endorsement of its reliability by the other members of the agreement. I refer to this concept as entry-endorsement.

The notion of entry-endorsement as a mechanism that converts ambiguous outcomes into a positive

\(^2\) This conceptualization produces the same substantive effects as the more common imposition of a static decay term, but it is a bit more flexible for purposes of the formal model and provides a somewhat more intuitive representation of reputational decay.
influence on reputation should be somewhat intuitive and familiar to readers. At its center is a simple idea: the acceptance or admittance of an actor into an interaction suggests that it has been assessed as worthy of acceptance or admittance into that interaction. Heuristically, we employ this logic all the time. Consider, for example, a student that has been admitted to college at Oxford. Upon hearing of the student’s admittance, we may be inclined to believe that she is an intelligent student. Our basis for this belief is that Oxford is a selective school that typically only admits highly intelligent applicants. Therefore, this student must be intelligent. Note, that it is compelling to hold this belief without any additional knowledge about the student’s performance at Oxford, or even any knowledge about the student’s prospects for graduation. In academics, a similar logic applies with regard to grant applications. A scholar is more likely to receive a grant for a project if she has already received funding from a well-regarded foundation such as the NSF. This is likely the case even if she has not yet published findings from the NSF funded project, because the NSF is a well respected arbiter of projects worthy of funding.

Now, assuming that a reputation for reliability influences agreement formation (as we established at the start of section 3.1)\(^3\), we arrive at proposition 1. This proposition is plainly stated as follows:

**Proposition 1** (Simplified from page 42): Under certain conditions, a holder that has entered an untested agreement will receive a reputational benefit and thus be more likely to enter into an agreement with a given observer than similar holders that have not entered such an agreement.

### 3.1.3 Entry Endorsement in Different Contexts

While we expect entry-endorsement to benefit holder states, we would not expect entry-endorsements to be viewed equally. Some endorsements will confer a stronger benefit than others.

Further, we might not expect entry endorsement to be interpreted the same way in every context. In some instances, the strength of entry endorsement may be interpreted, intuitively, based on similarity in strategic interests between the endorser and the observer. In other contexts, however, the interpretation of

\(^{3}\) see section 3.2, and various studies that establish this fact, including Crescenzi et al. (2012), Gibler (2008) and Miller (2003)
entry endorsement may work differently.

There are two common scenarios where reputation is useful and which lend themselves to different mechanisms for interpreting entry endorsement. The first is the scenario in which both the observer and endorser are the same sort (i.e. the same unit – states, individuals, etc.). The second is a scenario in which the endorser and observer are different sorts (e.g. one is a state and the other is a sub-state actor, such as a firm) but in which the endorser faces punishment for making bad endorsements. I will discuss each of these scenarios and their mechanisms for interpreting entry-endorsement in turn.

3.1.3.1 Interpretation of Entry Endorsement among Similar Sort Actors

One scenario which is common to international relations and in which reputation is useful is when all actors are the same sort. This type of scenario is especially common in issues of security and conflict, though it certainly extends to other matters where we commonly view states as unitary actors.

In scenarios where actors are the same sort, while we expect entry-endorsement to benefit holder states, we would not expect entry-endorsements to be viewed equally. Instead, we would expect the benefit from entry-endorsement to be conditioned on the observer’s view of the endorsers themselves. Specifically, we should expect the strength of entry-endorsement to be conditioned on the similarity between the observer and the endorsing states.

The reason for this comes from the uncertainty that observers would face in evaluating entry-endorsement. Recall that entry into an agreement means that a holder’s likelihood of being reliable is at least high enough to satisfy some baseline threshold (see section 3.1). It follows that an observer would infer that entry-endorsement means that the holder’s likelihood of reliability was at least as high as the endorser’s baseline threshold at the time of the formation of the agreement. The value of entry-endorsement, then, should hinge critically on the observer’s ability to estimate of the endorser’s baseline threshold. It is intuitive to expect that the observer can make more certain estimates of the endorser’s baseline threshold
when the observer and endorser have similar strategic interests\textsuperscript{4}. Thus, the strength of entry endorsement should grow with the similarity between the observer and endorser (see section 3.2.4 for further discussion).

Again, combining the discussion here with the expectation that reputation influences agreement formation, we arrive at proposition 2. Stated simply as follows:

**Proposition 2** (Simplified from page 43): When actors are the same sort, then under certain conditions, the reputational benefits of entry endorsement will be stronger when observers and endorsers are similar and thus entry endorsement will make agreements between the holder and observer more likely when the observer is similar to the endorser.

### 3.1.3.2 Interpretation of Entry Endorsement among Different Sort, Linked Actors

A separate scenario that is very common in international relations and in which reputation is useful is one in which endorsers and observers are not the same sort, but where observers can punish endorsers for making bad endorsements. This type of scenario is common to economic interactions.

Though seemingly quite different from scenarios discussed thus far, it is nearly the same with respect to the role of reputation and its assessment. Consider, for example, a firm’s decision to build a factory in a foreign country. Generally much of the logic discussed above applies here. That is, the firm would evaluate the holder state’s reputation for reliability (which in this context, would largely amount to protecting capital and property), and it would be more likely to invest if the country’s reputation surpassed a critical threshold. The observer (firm) would evaluate the holder’s reputation based on its prior interactions, particularly its prior economic agreements with states\textsuperscript{5}, and the actions of holder states in those agreements would decay over time. Lastly, the entry of a holder state into an untested economic agreement with another state would act as a form of entry endorsement for the quality of the holder state as an economic partner. Nothing about this example, thus far, diverges from the discussion of the role of reputation in section 3.1, section 3.1.1 or even section 3.1.2.

\textsuperscript{4} see Gray and Hicks (2014), Gray (2013), and Crescenzi et al. (2012) for similar, complimentary discussion

\textsuperscript{5} This is further discussed below, in section 3.2.5
In contrast to section 3.1.3.1, however, we would expect the firm (or other observers in a similar scenario) to take a different approach to evaluating the strength of reputational benefit from entry-endorsement. The reason we expect the observer to interpret entry endorsement differently is largely due to the fact that the observer, a firm, is a different sort than the endorser, a state. As such, it will be difficult to assess the extent to which the firm and the endorser are similar since they are somewhat incommensurate. Accordingly, some other mechanism must be used to determine the strength of reputational benefit from entry endorsement. Under these, or similarly structured circumstances, the strength of entry endorsement is determined by the endorser’s susceptibility to domestic punishment for making bad endorsements. I will outline the rationale behind this in the following paragraphs.

To begin, consider the endorser to be the home state of the observer. That is, consider the holder to have entered into an untested economic agreement with the observer’s home state. Since the observer is a domestic constituent of the endorser, the observer can, and is inclined to, punish the endorser for making bad endorsements. For clarity, a bad endorsement would be an action that makes later observers think the holder has a better reputation than it actually has (put differently – an action that later leads observers to incorrectly predict the holder’s type). In the context of the example described above, a bad endorsement would be the formation of an agreement with a holder state that later betrays the observer firm and expropriates the firm’s capital. That observers would punish endorsers for this outcome should be somewhat intuitive. When a government adopts policies that produce disfavorable outcomes for a domestic group, we expect the domestic group to punish that government through electoral pressure or other means, such as additional lobbying pressure.6

Since, as we have discussed it so far, the endorser can be punished by the observer, the endorser’s benefit from an economic agreement with the holder (an endorsement) is inherently linked to the later interactions between the observer and the holder. Specifically, the threat of punishment from a betrayed observer makes the endorser more selective about the (holder) states that it enters agreements with. As the observer grows in power,7 the endorser will become increasingly selective, choosing to form agreements with

---

6 See Bearce, 2003.
7 alternatively, as the observer’s interest group – i.e. the sector or industry
only those states that are most likely to be and remain reliable\(^8\).

Since the endorser will become more selective about its partners as the observer grows in political strength (i.e. the ability to punish), an agreement between the endorser and the holder should become a greater indication of the holder’s reputation. Put differently, observers can interpret entry endorsement between their home state and a holder on the basis of the observer’s own ability to punish the home state for making a bad endorsement. When an observer is politically stronger, then entry endorsements by the home state confer a stronger reputational benefit on the holder state. When the observer is weaker, then entry endorsements confer a weaker reputational benefit\(^9\).

Now, to this point, we have been discussing circumstances where the endorser is the home state for the observer, but this approach to interpreting entry endorsement essentially carries over to circumstances where the endorser is not the home state of the observer. When the endorser is not the home state of the observer, the observer can still interpret entry endorsement based on the susceptibility of the endorser to punishment for making bad endorsements. The observer could determine this susceptibility by considering the size and political strength of the corresponding actors/interests in the endorser’s state. For example, if the observer was a tech firm, it might interpret the formation of an agreement between a holder and a (foreign) endorser by considering the political strength of the tech industry in the endorser’s state. If the endorser’s state had a powerful tech industry, then the observer would view the entry endorsement as conferring a strong reputational benefit on the holder. If the endorser had a small tech industry, the observer would interpret the entry endorsement as producing a small/insignificant signal of the holder’s reputation.

Assuming, again that agreements reflect a state’s reputation, we arrive at proposition 3. Simply stated:

**Proposition 3** (Simplified from page 46): Under certain conditions (including when observers and endorsers are different sorts), the reputational benefits of entry endorsement will be stronger when endorsers

\(^8\) Fearon (1994) provides analogous discussion with regard to conflict.

\(^9\) This follows, somewhat from the notion that, if the observer is not able to effectively punish the home state, then the home state may enter into agreements predominately for reasons that are not related to the observer and without regard for whether the holder ultimately proves reliable in its interactions with the observer.
are susceptible to domestic punishment for making bad endorsements and thus entry endorsement will make agreements between observers and holders more likely when endorsers are susceptible to domestic punishment for making bad endorsements.

3.1.4 Betrayal

Recall that the discussion in section 3.1.1 suggests that the speed with which a holder’s reputation decays is determined by the perspective of the observer\textsuperscript{10}.

We might expect the observer to formulate its beliefs about the probability of type transition (and thus the decay term of the reputation between the holder and a given observer) based upon its own experiences in prior agreements. This seems intuitive and feasible\textsuperscript{11} as a way for a state to estimate the probability of type transition.

Essentially, I submit that observers expect their initial assessment of a holder to be accurate and that they change their assessment when provided information to the contrary. In the context of an agreement, then, an observer would expect its partner to be reliable at the time of the agreement and to remain reliable for its duration\textsuperscript{12}. When the ally fails to comply with an agreement or demonstrates its unreliability, the observer attributes this to an unexpected transition in type. The unexpected transition is new information that the observer can then use to update its beliefs about the probability of type transition.

To put this more simply, observer beliefs about type transition are a function of betrayal in prior agreements. When an observer is betrayed in an agreement, it becomes increasingly skeptical that a holder that was reliable yesterday will still be reliable today. Effectively, then, the speed with which reputations decay is a function of the betrayal of the observer state. This, in turn, should have implications for the influence of entry-endorsement. Namely, if expectations about type transition (and thus, decay) are a

\textsuperscript{10} This feature of the model is, seemingly, quite similar with the notion of a perceptual biases that predominate the psychological literature on reputation. This is somewhat deliberate, though my approach here (see section 3.2.6) is decidedly more formal/rational in nature.

\textsuperscript{11} see section 3.2.6 for further discussion.

\textsuperscript{12} If the observer doubted that its partner would remain reliable for the duration of the agreement, it would elect not to form an agreement or negotiate a shorter term for the agreement.
function of observer betrayal, then the experience of betrayal by an observer should influence how recent an entry-endorsement must be to significantly influence reputation.

This brings us to proposition 4. Stated plainly

**Proposition 4** (Simplified from page 50): Entry endorsement decays faster when an observer has been betrayed in the past and (if the observer has extensively been betrayed) entry endorsement produces an agreement between the holder and observer only if the entry endorsement occurred very recently.

### 3.2 The Model

To begin, consider an interaction between two actors. One actor that holds a reputation (denoted player \( h \)) and an international actor (denoted as player \( b \)) that can observe the prior behavior of \( h \). The strategic action in the game lies in a singular decision by player \( b \) as to whether or not to enter into an agreement with player \( h \). This interaction is provided as a diagram in Figure 3.2.

Actors \( (b) \) face uncertainty over the type (denoted \( y \)) of holder \( (h) \) they face. If actors \( (b) \) form an alliance with an unreliable holder \( (h) \), then they will be worse off than if they had decided to forgo cooperation altogether. On the other hand, if \( h \) is reliable, then the actor \( (b) \) will be better off engaging in cooperation. The purpose of reputation is to help the actor, \( b \), distinguish one type of state from the other.

Player \( h \) as takes on one of two potential types. At any given time, player \( h \)'s type can be either reliable or unreliable. Formally, denote \( h \)'s type by \( y \), where \( y \in \{ \text{reliable, unreliable} \} \). In a simple model of international interaction (see Figure 3.2), nature determines \( h \)'s state \( (y) \). If \( y = \text{reliable} \), it will comply with any new agreement. If \( y = \text{unreliable} \), it will defect on any new agreement. Note that \( y \) proscribes \( h \)'s behavior in the enforcement phase of an agreement. Player \( h \) will choose to enter a new agreement regardless of its type.

Player \( b \) must choose whether to accept an agreement with \( h \) or to reject an agreement with \( h \).
Importantly, $b$ is uncertain over $y$, though she does possess some knowledge of the general distribution of $s$. Player $b$ pays a high cost and receives little benefit from being betrayed in an agreement, but does receive a net benefit from being a successful agreement (i.e. an agreement in which no member defects). Rejecting an agreement yields no particular benefit to $b$, but does not incur a cost. In other words, $b$’s payoff preferences are specified as follows:

$$u_b(accept|y = 1) > u_b(reject) > u_b(accept|y = 0)$$

(3.1)

The structure of this game is provided graphically in Figure 3.2. Recall that $h$’s type ($y$) determines its behavior under an agreement. The state-determined behavior of the holder $h$ is lined in bold in Figure 3.2.

---

**Figure 3.2 Simple Game**

This game has a simple sub-game perfect Nash equilibrium (SPNE). A SPNE strategy is a system of best responses for the game and every sub-game\(^\text{13}\). Denote $\sigma^i$ as a mapping of best responses for player $i \in \{b, h\}$. By construction, $h$’s best responses are to comply if its type is reliable($y = 1$) and defect if its

\(^{13}\) see Osborne and Rubinstein, 1994, p. 94
type is unreliable ($y = 0$). Thus:

$$\sigma^b(y) = \begin{cases} 
\text{comply} & y = 1 \\
\text{defect} & y = 0 
\end{cases}$$

Player $b$’s best response is dictated by a comparison of the utility she receives from rejecting an agreement to her expected utility from forming an agreement with $h$. This produces conditions for which $b$’s best reply to accept an agreement with $h$. Denote $\pi(\cdot) := pr(y = 1|\cdot)$. It then follows:

$$EU_b(\text{accept} | \pi(\cdot)) = \pi(\cdot)(u_b(\text{accept} | y = 1)) + (1 - \pi(\cdot))(u_b(\text{accept} | y = 0))$$

and $EU_b(\text{accept} | y) > u_b(\text{reject})$ iff:

$$\pi(\cdot) > \frac{u_b(\text{reject}) - u_b(\text{accept} | y = 0)}{u_b(\text{accept} | y = 1) - u_b(\text{accept} | y = 0)}$$

(3.2)

Denote $\pi^* := \frac{u_b(\text{reject}) - u_b(\text{accept} | y = 0)}{u_b(\text{accept} | y = 1) - u_b(\text{accept} | y = 0)}$. Then the condition for which $b$’s acceptance of an agreement with $h$ are stated more simply as $\pi(\cdot) > \pi^*$

Considering the conditions for acceptance, we can restate $b$’s SPNE strategy as a mapping, $\sigma^b$ of $\pi(\cdot)$ and $\pi^*$ to \{accept, reject\} as best replies:

$$\sigma^b(\pi^*, \pi(\cdot)) = \begin{cases} 
\text{accept} & \pi(\cdot) > \pi^* \\
\text{reject} & \pi(\cdot) < \pi^* 
\end{cases}$$

The SPNE is thus characterized by the strategy profile that combines these two best-reply functions: $\{\sigma^h(\pi(\cdot), \pi^*); \sigma^b(y)\}$. In equilibrium, the outcome of the game hinges, critically, on the values of $\pi(\cdot)$ and $\pi^*$. When $\pi(\cdot) > \pi^*$, $b$ will always enter into an agreement with $h$, and $b$ will be betrayed by player $h$ with probability $(1 - \pi)$. When $\pi(\cdot) \leq \pi^*$, player $b$ will reject any agreement with player $h$. 
To assist with reading the remainder of this section, consider Table 3.1, which provides a reference for much of the notation used in this section.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a$</td>
<td>Observed action</td>
</tr>
<tr>
<td>$y$</td>
<td>Actor type ($y=1$ indicates reliable type)</td>
</tr>
<tr>
<td>$\sigma(\cdot)$</td>
<td>Best response function</td>
</tr>
<tr>
<td>$b$</td>
<td>Player $b$ (superscript index)</td>
</tr>
<tr>
<td>$h$</td>
<td>Player $h$ (superscript index)</td>
</tr>
<tr>
<td>$ht$</td>
<td>Endorser of $b$</td>
</tr>
<tr>
<td>$t$</td>
<td>Point in time $t$ periods ago (subscript index)</td>
</tr>
<tr>
<td>$s$</td>
<td>Point in time, more recent than $t$ (subscript index)</td>
</tr>
<tr>
<td>$0$</td>
<td>The present (subscript index)</td>
</tr>
<tr>
<td>$\pi$</td>
<td>Reputation, probability that $y = 1$</td>
</tr>
<tr>
<td>$\pi^*$</td>
<td>Minimum reputation threshold for agreement</td>
</tr>
<tr>
<td>$\pi^{**}$</td>
<td>Endorser threshold for agreement</td>
</tr>
<tr>
<td>$m(\cdot)$</td>
<td>Reputation decay function</td>
</tr>
</tbody>
</table>

### 3.2.1 The Reputation Parameter

Denote $a_h^t$ as the outcome of a prior agreement to which $h$ was a member and that occurred $t$ periods ago. Also, let $a_h^t$ reveal whether $h$ complied or defected on its obligations under that agreement. That is, $a_h^t \in \{\text{comply, defect}\}$. Since $y$ determines $h$’s decision to comply or defect on an agreement, $b$ can infer from an observed $a_h^t$ the value of $y$. This translates into $b$’s beliefs about the reliability of $h$, and hence $b$’s assessment of the reputation of $h$. We denote this assessment of reputation as $\pi$:

$$\pi : \{\text{comply, defect}\} \rightarrow (0, 1)$$

$$\pi : a_h^t \rightarrow \text{pr}(y = 1)$$

### 3.2.1.1 Dynamic types

Let a state’s type ($y$) be dynamic, and allow $h$ to transition between reliable ($y = 1$) and unreliable ($y = 0$). Generally, $y$ is ‘sticky’, and transitions occur at random intervals but are generally infrequent.
Denote the present as 0 and the term $t$ as meaning a period that occurred $t$ periods ago. Accordingly, $h$’s type at a given time ($t$) is written $y_t$.

Since $y$ is subject to transition at random, we can express $h$’s type as the outcome of a Markov process and described as follows.

Let $M$ be a symmetrical transition kernel whereby $\alpha$ indicates the probability of remaining in the same state and $1 - \alpha$ indicates the probability of transitioning to a different state. In matrix form:

$$
M = \begin{bmatrix}
\alpha & (1 - \alpha) \\
(1 - \alpha) & \alpha
\end{bmatrix}
$$

If $y_0$ is a product of a Markov process, the probability that $h$ is reliable ($pr(y_0 = 1|\cdot)$, which is also $\pi_0(\cdot)$) can be derived as the product of an initial distribution of types and the transition matrix (exponentiated by the number of periods since observing the initial distribution of types). By setting the initial distribution of types according to an observed $a^h_t$, we determine $pr(y_0 = 1|\pi_t(a^h_t))$ and equivalently $\pi_0(a^h_t)$ through the following function (denoted $m$):

$$
m : \mathbb{R}^3 \to (0, 1)
$$

$$
m : (\pi_t, \alpha, t) \mapsto (\pi_0)
$$

$$
m(\pi_t, \alpha, t) = \left[ \begin{array}{c} \pi_t \\ 1 - \pi_t \end{array} \right] \left[ \begin{array}{cc} \alpha & (1 - \alpha) \\ (1 - \alpha) & \alpha \end{array} \right]^t \left[ \begin{array}{c} 1 \\ 0 \end{array} \right]
$$

which simplifies as:

$$
m(\pi_t, \alpha, t) = \left[ \begin{array}{c} \pi_t \\ 1 - \pi_t \end{array} \right] M^t \left[ \begin{array}{c} 1 \\ 0 \end{array} \right]
$$

and $m(\pi_t, \alpha, t) = \pi_0$
3.2.2 Untested Agreements

Incorporation of entry into untested agreements into the model requires only minimal modification. Revise \( a^h \) such that \( a^h \in \{ \text{Defect, Comply, Entry} \} \). For purposes of simplicity, let \( \pi_t(a^h_t = \text{comply}) = 1, \pi_t(a^h_t = \text{defect}) = 0 \). We then can incorporate ambiguous outcomes by specifying a value of \( \pi_s(a^h_s = \text{entry}) \), which presumably will be between 0 and 1.

To specify \( \pi_s(a^h_s = \text{entry}) \) endogenously in a way that captures entry-endorsement, we need to explicitly consider the role of \( h \) partners in interactions that ended ambiguously. The role of \( h \)'s former partners is that of an observer and is thus otherwise comparable to the role of \( b \). Denote former partners of \( h \) as \( b' \).

Player \( b' \) interacts with player \( h \) at some time prior to 0. It will be useful to compare interactions where outcomes were ambiguous with those where outcomes were unambiguous. Accordingly, while \( s \) is the time at which \( h \) most recently entered an untested alliance, denote \( t \) as the time at which \( h \) experiences a test of its commitment to an agreement. The quantity \( s \) can denote any period of time, however, we will primarily consider the role of ambiguous outcomes where \( s < t \) in other words, where an entry into an untested alliance occurred more recently than a test of \( h \)'s reliability.

At time \( s \), \( b' \) must choose whether to accept or reject an agreement with \( h \). The scenario faced by \( b' \) is the same as that described for \( b \) in section 3.2. The behavior of \( b' \) at \( s \) are denoted \( a^h_s \in \{ \text{accept, reject} \} \), with a best response function \( \sigma^{b'} \). And where \( \sigma^{b'} = \text{accept when } \pi'_s \geq \pi^{**} \), and reject otherwise\(^{14} \). Thus, in equilibrium, the outcome of the interaction between \( b' \) and \( h \) is critically dependent on the former partner’s \((b')\) assessment of \( \pi'_s \), and the value of \( \pi^{**} \).

Player \( b \) observes \( b' \) behavior with relation to \( h \) and uses this behavior to update its own assessment \( \pi_0 \).

It does this by deriving an expected value of \( \pi'_s \) from \( a^h_s \) and considering the probability of state-transition

\(^{14} \) Note that \( a^h_s = \text{accept only when } a^h_s = \text{Entry}. \) In this sense, then, the two conditions are synonymous. It is however, useful to retain both notations since the observer \( b' \)'s evaluation of entry relies, somewhat on properties of \( b' \) and its decision-making process. As such, it is useful to use notation that is more evocative of the entry endorser, and not the holder – hence the use of \( a^h_s \).
since the time $s$:

\[
pr(y_s = 1|a'_s) = \begin{cases} 
E(\pi'_s(\cdot)|\pi''_s) & a'_s = \text{accept} \\
m(\pi_t(a''_t), \alpha, n) & a'_s \neq \text{accept}
\end{cases}
\] (3.4)

Denoting:

\[
\overline{\pi'_s} := E(\pi'_s(\cdot)|\pi''_s, a'_s = \text{accept})
\]

and also denoting:

\[
(a^h_s = \text{entry}) \equiv (a'_s = \text{accept})
\]

Note:

\[
\pi''_s < \overline{\pi'_s} < 1
\]

\[
\lim_{\pi''_s \to 1} \overline{\pi'_s} = 1
\]

Note also that:

\[
\pi_s(a^h_s = \text{entry}) = \pi_s(a'_s = \text{accept}) = \overline{\pi'_s}
\]

From which it follows:

\[
\pi_0(a^h_s = \text{entry}) = \pi_0(a'_s = \text{accept}) = m(\overline{\pi'_s}, \alpha, s)
\]

Where $m(\pi_t(a''_t), \alpha, (t-s))$ is the assessment of $h$’s reputation at $s$, and based on the last, unambiguous, observation of $a^h$, occurring at a time $t$. And where $t$ occurs at either the same time or earlier than $s$. Further discussion of the derivation of equation (3.4) is found in appendix A.1.

### 3.2.3 Untested Agreements Benefit Reputation

As discussed above, inputs to the reputation parameter $\pi$ can include both unambiguous alliance outcomes (comply and defect) and mere membership in untested alliances. When the holder is a member of an untested alliance, entry-endorsement serves as a signal of the holder’s reliability. Further, so long as
Proposition 1: There exist conditions such that actors \( h \) that have recently entered into an untested agreement will be more likely to enter new agreements than similar actors \( h \) that have not entered into an untested agreement.

This follows directly, if we consider that for a sufficiently large value of \( \pi_s(a^h_s = \text{Entry}) > .5 \), there surely exist (positive, non-zero) values of \( t \) and \( s \) such that:

\[
m(\pi_s(a^h_s = \text{entry}), \alpha, s) > m(\pi_t(a^h_t = \text{compliance}), \alpha, t)
\]

and such that:

\[
\pi_0(a^h_s = \text{entry}) > \pi_0(a^h_t = \text{compliance})
\]

and which it must follow that \( \exists \pi_0^* : \)

\[
\sigma^b(\pi_0(a^h_s = \text{Entry})) = \text{accept}
\]

while

\[
\sigma^b(\pi_0(a^h_s = \text{compliance})) = \text{reject}
\]

3.2.4 Entry-Endorsement Interpretation when Observers are Similar to Endorsers

It should be clear from the discussion above that \( \pi^* \) hinges critically on \( \pi^*_s \). But on what basis can \( b \) estimate the value of \( \pi^*_s \)? While there may be alternatives, it would seem intuitive that \( b \) might base its evaluation of \( \pi^*_s \) on its own assessment of the constraints and incentives facing\(^{15} \) \( b' \). In other words, \( b \) might ask itself: “What would I do if I were \( b' \)?”. It follows\(^{16} \), straightforwardly, that we would expect \( b \) to be more confident about its estimate of \( \pi^*_s \) as \( b \) increases in similarity to \( b' \). Further, as \( b \) and \( b' \) become more similar, we would expect estimates of \( \pi^*_s \) to more closely resemble both the true value of \( \pi^*_s \) and \( \pi^* \).

\(^{15}\) for further discussion along these lines, see Gray and Hicks (2014) Crescenzi (2007), Crescenzi, Kathman, and Long (2007) and Crescenzi et al. (2012)
Generally, this corresponds to an increase in the reputational benefit conferred by entry endorsement.

**Proposition 2:** Let similarity between $b$ and $b'$ influence estimates of $\pi_s'$ (denoted $\hat{\pi}^{s'}$) such that, as similarity increases, $\hat{\pi}^{s'} \to \pi^*$ and $E((\pi^{s'} - \hat{\pi}^{s'})^2|\pi^{s'}) \to 0$. It follows that, increases in similarity increase $pr(\pi_s' > \pi^*)$ thereby making reputational benefit from entry endorsement more likely (i.e. producing, generally, $\frac{\pi_s'}{\pi} > \pi^*$) at sufficiently low values of $s$.

To begin, consider that, from the observer’s perspective, the endorser’s assessment of the reputation of a holder (i.e. $\pi_s'$) and the actual value of $\pi'$ are unknown. The observer can, however, estimate $\pi'$, denoted by $\hat{\pi}^{s'}$. Let $f_1(\pi'|\hat{\pi}^{s'}, \rho)$ denote the probability density of a given value of $\pi'$ relative to location and scale parameters $\hat{\pi}^{s'}$ and $\rho$, respectively. Now, consider $f_2(\pi_s'|\pi^{s'}, 1)$ to give the probability density of $\pi_s'$ on a truncated distribution with support restricted to the interval $(\pi^{s'}, 1]$.

Accordingly, we can restate the distribution of $\pi_s'$ as a compound distribution with a hyperparameter $\hat{\pi}^{s'}$. Further, we can specify the probability that $\pi_s'$ falls below $\pi^*$:

$$pr(\pi_s' > \pi^*) = 1 - F_2(\pi^*|\hat{\pi}^{s'}) = \int_{\pi^*}^{\hat{\pi}^{s'}} f_2(\pi_s'|\pi^{s'}, 1) f_1(\pi'|\hat{\pi}^{s'}, \rho)$$

From this it directly follows that $pr(\pi' > \pi^*) \to 1$ with increases in similarity between $b$ and $b'$ since it is also the case that $\hat{\pi}^{s'} \to \pi^*$ and $\rho \to 0$ with increases in similarity.

Further, since $\hat{\pi}^{s'} \to \pi^*$ as $b$ and $b'$ grow in similarity, it follows that $E[\pi^{s'}|\hat{\pi}^{s'}, \rho] \to \pi^*$ as $b$ and $b'$ grow in similarity. Consider that $E[\pi_s' | E[\pi^{s'}|\hat{\pi}^{s'}, \rho]] > E[\pi^{s'}|\hat{\pi}^{s'}, \rho]$. It follows that as $b$ and $b'$ grow in similarity, we expect $E[\pi_s' | E[\pi^{s'}|\hat{\pi}^{s'}, \rho]] > \pi^*$ with certainty. Since $\rho \to 0$ as $b$ and $b'$ increase in similarity, $E[\pi_s' | E[\pi^{s'}|\hat{\pi}^{s'}, \rho]] \to E[\pi_s'|\pi^{s'}]$, asymptotically. Therefore, as $b$ and $b'$ increase in similarity, $\hat{\pi}^{s'} > \pi^*$ obtains with certainty.

---

\[16\] To be explicit: This follows from the realization that, as $b$ becomes more similar to $b'$, its utility function will increase in similarity as well. Accordingly, then, when $b$ is similar to $b'$, it will intuitively assess the utilities of the strategic interaction between $b'$ and $h$ as similar to $b'$s own assessment, and $b$'s estimate of $\pi^{s'}$ will be accordingly similar to the actual value of $b'$.
Ultimately, this means that increases in similarity ensure estimates of $\pi_\alpha'(\cdot)$ that are greater than $\pi_\alpha^*$. Accordingly, this ensures that if an endorsement occurred sufficiently recently and the endorser and observer are sufficiently similar, then entry endorsement will confer some reputational benefit on the holder. This benefit does, however, have an upper bound set by $E[\pi'(\cdot)|\pi^*]$, which may, at times, be considerably less than 1, depending on the value of $\pi^*$.

3.2.5 Entry Endorsement among Linked Actors

In contrast to the scenario described above, consider an alternative, in which $b$ is a different sort of actor than player $b'$. Specifically, consider a scenario in which $b$ is a sub-state actor and $b'$ is a state. Generally, we might expect to see $b$ specified as a sub-state actor when we consider the game described in section 3.2 as applied in an economic setting such as a decision by a company to build a factory in a foreign state, or engage in international trade.

Immediately, readers may wonder why $b$ would not assess the reputation of $h$ from prior interactions between $h$ and other substate actors. Simply put, in a scenario where $b$ competes with other actors of the same sort (i.e. other substate actors), $b$ has reason to suspect interactions between $h$ and other substate actors are either hidden from view or deliberately misleading. This follows, straightforwardly, from consideration of the incentives facing both the $h$ and the other substate actors. In this scenario, player $h$ has strong incentives to suppress or obfuscate disreputable outcomes from interactions with other sub-state actors. In doing so, $h$ leaves only evidence of compliance, and assessments of $\pi_0$ derived from $h$’s interaction with substate actors is biased in its favor.

At the same time, substate actors – both those that have been betrayed by $h$ and those that have not – may wish to hide or misrepresent their interactions with the holder state, $h$. This follows from considering other substate actors as potential competitors or rivals. Fundamentally, other sub-state actors face an honesty problem in communicating their interactions with $h$ in the past. Where substate actors have previously had good interactions with $h$, they enjoy a competitive/distributional advantage over $b$. Accordingly, these
substate actors may conceal their interactions with \( h \). Where, on the other hand, substate actors have previously been taken advantage of by \( h \), substate actors would not want to warn \( b \) against interactions with \( h \). This follows from the fact that warning \( b \) would enable it to avoid betrayal by \( h \), and would leave previously betrayed substate actors at a disadvantage.

As an alternative, then, \( b \) can look to \( h \)'s prior interactions with other states (including its home state)\(^{17}\) to derive an assessment of \( h \)'s reputation. Specifically, I submit that substate observers will interpret the strength of entry-endorsement based on the endorser’s susceptibility to punishment from domestic actors.

To see how this works, begin, assuming that \( b' \) is the home state of substate actor \( b \). Further, assume that the agreement negotiated by \( b' \) and \( h \) at \( s \) was relevant\(^{18}\) to the agreement under review by \( b \) and \( h \) at time \( T \). Now, let \( b' \)'s utility take domestic audience costs (i.e. punishment by domestic actors) as an input such that \( b' \) suffers costs for the betrayal of any domestic actor by state \( h \), occurring at any time after \( s \) (i.e. let \( E[a_h^b | a_h^b = \text{accept}] \rightarrow u_{b'}(a_b'^s) \)). As applies to international relations, this assertion (that state leaders face domestic pressure) enjoys considerable scholarly support\(^{19}\). The consequence of such an assertion, however, is that \( b' \) is now inclined to limit \( h \)'s opportunity to betray actors in the future if it seems sufficiently likely that (1) \( h \) will betray actors in the future, and (2) the audience costs suffered by \( b' \) are sufficiently high. In principle, then, the consideration of audience costs establishes divergent preferences for \( b' \) over the path of play at all points beyond \( b' \) actions\(^{20}\) at \( s \). In other words, \( b' \) will prefer to accept or reject agreements with \( h \) strategically, depending on whether \( b' \) expects \( b \) to enter agreements with \( h \) as a result of the endorsement and whether it believes that \( h \) will remain reliable in the future.

Assume, for the moment, that \( b' \) thinks that substate actors (such as \( b \)) will interpret an agreement with \( h \) as an endorsement of \( h \)'s probability of being in a compliant state at \( s \). It follows that \( b' \), in

\(^{17}\) Here, I assume that states have incentives to publicize their agreements with one another. Casual observation suggests that this is at least true for economic agreements. For further discussion, see Mansfield, Milner, and Rosendorff (2002), Elkins, Guzman, and Simmons (2006) and Elkins and Simmons (2004).

\(^{18}\) In other words, assume that the nature of the interaction between \( b \) and \( h \) and \( b' \) and \( h \) are not wholly dissimilar. A weak version this assumption might be that both interactions involve the same issue-area (economics, security, etc.)


\(^{20}\) Previously, it was implicit that \( b' \)'s preferences over interactions between \( h \) and \( b \) were equivocal or weak. An explicit statement to this effect is found in section 3.2.4, and it is notable that we still derive conditions for the role of entry endorsement on reputation in that scenario.
rejecting an agreement with $h$, limits $h$’s ability to demonstrate compliance and withholds any endorsement of $pr(y_s = 1)$. The act of rejection, erodes $h$’s reputation ($\pi_0$) with $b$ and (barring an opportunity for $h$ to demonstrate compliance in the interim) thereby decreases the likelihood that $b$ will accept an agreement at time 0. If, on the other hand, $b'$ accepts an agreement with $h$, it provides an endorsement of $h$’s reliability and thereby increases the likelihood that $b$ will accept an agreement at 0. Since $b'$’s utility is a function of the outcome of interactions between $b$ and $h$, $b'$ has strong incentives to be strategic about its agreement partners$^{21} h$. As audience costs increase, $b'$ will reject agreements with all but the most certainly reliable players $h$.

Now, as $b'$ grows increasingly selective about its interactions with player $h$, $b$ will view entry endorsement as an increasingly strong indication of the $h$’s good reputation. From Equation 3.4 we know that an agreement at time $s$ is an indicator that of the minimum reliability of $h$ (i.e. an indicator that $\pi'_s(\cdot) > \pi''_s(\cdot)$). Moreover, if we consider equation (3.2), then it is obvious that $\pi''_s$ increases with audience costs. Therefore, as audience costs grow, so does the information provided by an agreement at $s$, and $b$ will view entry endorsement as a stronger sign of $h$’s good reputation.

Combining the preceding two paragraphs, we see that, in the face of increasing domestic punishment for bad endorsements, $b'$ has incentives to be more selective about its partners. Furthermore, as $b'$ grows more selective, $b$ interprets endorsements by $b'$ as stronger statements about $h$’s reputation. From this, it follows:

**Proposition 3:** Let the observer ($b$) punish the endorser ($b'$) for a bad endorsement, with the magnitude of the punishment denoted $p$, and let $(-p) \propto u_b(\text{accept}|y_s = 0)$ such that $\lim_{-p \to -\infty} u_b(\text{accept}|y_s = 0) = -\infty$. It follows that, as $p$ becomes large, entry endorsement, $\pi'_s \to 1$. For sufficiently low $s$, therefore, $\pi_0(a_s^h = \text{Entry})$ increases in $p$.

Proposition 3 follows straightforwardly from the preceding discussion, but it can also be demonstrated analytically by considering the utility function of $b'$ (the endorser).

---

21 $b'$ will want to avoid making bad endorsements if they are likely to result in the betrayal of $b$ and if punishment by $b$ will be severe.
To begin, apply Equation 3.2 to the decision of $b'$ to enter an agreement with $h$ at $s$, producing $\pi^{*'}_s$:

$$\pi^{*'}_s = \frac{u_{b'}(\text{reject}) - u_{b'}(\text{accept}|y_s = 0)}{u_{b'}(\text{accept}|y_s = 1) - u_{b'}(\text{accept}|y_s = 0)}$$

$b$ punishment for bad endorsement (-p)

It follows directly from this that, since $(-p) \propto u_{b'}(\text{accept}|y_s = 0)$, $\lim_{p \to \infty} \pi^{*'}_s = 1$. Therefore, as the ability of $b$ to punish the endorser ($b'$) increases, the endorser will become increasingly selective in its partners ($h$), indicated by an increasingly high value $\pi^{*'}_s$. Recall from Equation 3.4, $\lim_{\pi^{*'} \to 1} \pi^{*}_s = 1$. It follows, therefore that $\lim_{p \to \infty} \pi^{*}_s = 1$.

From this discussion it also follows that increases in $p$ yield increases in $m(\pi^{*'}, \alpha, s)$ for all values $s$ lower than the time needed for $M$ to reach its stationary state\(^{22}\) (see Equation 3.3). Thus, for sufficiently low $s$, $\pi_0(\pi^{*'}_s)$ increases in $p$.

An additional insight follows from proposition 3, regarding the duration of entry endorsement when observers can punish endorsers for bad endorsements:

**Corollary 1:** With $(\pi^{*'}_s > \pi^*)$, increases in punishment $(p)$ make agreements between holders and observers more likely for longer periods of time following endorsement.

Since $\pi^{*'}_s$ increases in $p$ (the ability of the observer to punish), the effect of endorsements $(\pi^{*'}_s)$ on the observer’s assessment of $h$’s reputation $\pi_s$ also increase in $p$. To see this, recall that $\pi_0(a^{b'}_s = \text{accept}) = m(\pi^{*'}_s, \alpha, (t+n))$. As defined in Equation 3.3, $m$ increases in $\pi^{*'}_s$ as long as the quantity $(T-t+n)$ is sufficiently low\(^{23}\), which is another way of saying that the endorsement must have occurred sufficiently recently.

Now, consider two potential levels of punishment: $\bar{p}$ and $p$, where $\bar{p} > p$. It follows from the preceding discussion that $(\pi^{*'}_s|\bar{p}) > (\pi^{*'}_s|p)$. Accordingly:

\(^{22}\) In this regard, the upper boundary for $s$ is a function of $\alpha$.

\(^{23}\) Specifically, $s$ must be lower than the number of periods required for the transition matrix, $M$, to reach a stationary state. If $s$ is too large, then no endorsement (regardless of strength/selectivity, or any action for that matter) occurring at $s$ will yield a reputation assessment greater than .5 (the stationary state of the transition matrix $M$ and the point of maximum uncertainty)
\{(t + n)|m((\pi_s^t[p], \cdot, (t + n)) > \pi^*) \supset \{(t + n)|m((\pi_s^t[p], \cdot, (t + n)) > \pi^*)} \\

3.2.6 Expectations about transition

It is reasonable and substantively compelling, to consider factors that may yield different outputs from \(m\). I introduce here an extension of the model that produces different beliefs about the likelihood of state-transition, \(\alpha\), and primarily in response to \(b\)'s own past experiences.

To gather an accurate estimate of \(\alpha\), \(b\) would need to observe\(^{24}\) \(\{y_t|t \in (0, n)\}\) for some sufficiently-sized \(n\). This, by construction is not feasible. Similarly, in practice, this is not possible as \(y\) is an inherently latent property of \(h\).

As an alternative, we might argue that \(b\) determines \(\alpha\) by examining \(h\)'s pattern of agreement entry and subsequent compliance. This too brings problems as it is not clear what \(b\) should infer from an observation of defection by \(h\). Consider a situation in which \(h\) entered into an agreement at time \(t\) and defected at time \(t + 1\). Should \(b\) consider the defection as indicative of a transition from reliable (\(y_t = 1\) to unreliable in the next period (\(y_{t+1} = 0\)), or should \(b\) consider this an instance of miscalculation by \(h\)'s (then) partner? Either interpretation is reasonable.

Instead, I propose that \(b\) determines \(\alpha\) largely by reference to its own past experiences. Player \(b\) is likely to suppose that its own assessment of the reliability of a partner is accurate\(^{25}\), and thus when \(b\) is betrayed (read: when \(b\)'s partners defect an agreement), \(b\) is likely to conclude that its partner has transitioned from a reliable state at the time of entry into the agreement to an unreliable state at the time of defection. Accordingly, \(b\) will update its beliefs about \(\alpha\) to a lower value (recall from section 3.2.1.1 that lower values of \(\alpha\) correspond to higher probabilities of state-transition). Conversely, if \(b\) enters into an agreement that is subsequently successful (i.e. all parties comply with the agreement), then \(b\) should update

\(^{24}\) perhaps more simply stated: \(b\) would need to observe the state of \(h\) over some sufficient period of time.
its beliefs about \( \alpha \) to a higher value. Formally stated, then, let \( f \) designate the updating function as follows:

\[
f : \mathbb{R}, \mathbb{N} \rightarrow (0, 1)
\]

\[
f : \alpha_{t-1}, a_t \mapsto \alpha_t
\]

where \( a_t \in \{ \text{betrayal, success} \} \)

and

\[
f(\alpha_{t-1}, \text{betrayal}) < f(\alpha_{t-1}, \text{success})
\]

The precise mechanism for updating can be a simple application of Bayes’ rule, though it may be preferable to use a more complex function that places primacy on more recent agreements involving \( b \).

3.2.7 Betrayal hastens decay

The model allows that \( b \)’s prior experiences influence its beliefs about the likelihood of state-transition (\( \alpha \)). This extension has implications for the demands \( b \) places on the recency of \( h \)’s last observed behavior\(^{26}\). I demonstrate this by developing lemma 1, which produces the more substantively important finding proposition 4.

**Lemma 1:** When \( b \) has been betrayed, evaluations of \( \pi_0(a^b_t) \) decrease more rapidly in \( t \) than when \( b \) has not been betrayed.

This follows from consideration of the following. First, as specified in section 3.2.1.1, when \( \pi_t(\cdot) > .5 \), the output of the function \( m \) will decline monotonically in \( t \), converging to the stationary state of the transition matrix \( M \) at a sufficiently large \( t \). The length of \( t \) required to reach this point of stationary declines with reductions in \( \alpha \). Second, as specified in Equation 3.5, \( \alpha \) is the output of \( f \), and \( f(\cdot, \text{betrayal}) < \)

\(^{25}\) At the very least, \( b \) can more accurately separate cases where betrayal indicates miscalculation from state-transition since \( b \) has private knowledge of its threshold for acceptance as well as its assessment of its partner’s reliability at the time of the agreement. Player \( b \) is unlikely to know these things about other partners to \( h \) in other agreements.

\(^{26}\) assuming, of course that \( h \)’s last observed behavior was \( a^h_t = \text{compliance} \)
\( f(\cdot, \text{success}). \)

It follows that \( \forall x \in (0.5, \pi_t(\cdot)] : \)

\[
(t|m(\pi_t(\cdot), f(\cdot, \text{success}), t) = x) > (t|m(\pi_t(\cdot), f(\cdot, \text{betrayal}), t) = x)
\] (3.6)

From Equation 3.6 and in consideration of the fact that, by construction,

\[
m(\pi_0(\cdot), \text{success}, 0) = m(\pi_0(\cdot), \text{betrayal}, 0)
\]

it follows that, if \( \pi_t(\cdot) > 0.5 \), then \( \frac{\partial m}{\partial t} < 0 \) and:

\[
\frac{\partial m(\pi_t(\cdot), f(\cdot, \text{success}), t)}{\partial t} > \frac{\partial m(\pi_t(\cdot), f(\cdot, \text{betrayal}), t)}{\partial t}
\]

In other words, the rate of decay in reputation \( \left( \frac{\partial m(\pi_t)}{\partial t} \right) \) is weaker (i.e. nearer to 0) when \( b \) has recently been in successful agreements and stronger (i.e. negative and further from 0) when \( b \) has recently been betrayed in its agreements. Additionally, from this we can show:

**Proposition 4:** When \( b \)'s recent experiences in agreements have ended in \( b \)'s betrayal by her partners, \( m(\pi_t(\cdot, a^h = \text{compliance}), \alpha, t) > \pi^* \) remains true for smaller ranges of \( t \) than when \( b \)'s recent experiences in alliances have been successful.

It follows directly from lemma 1 that for \( \pi_t(\cdot) > 0.5: \)

\[
(t|m(\pi_t(\cdot), f(\cdot, \text{success}), t) = \pi^*) > (t|m(\pi_t(\cdot), f(\cdot, \text{betrayal}), t) = \pi^*)
\]

thus:

\[
\{t|m(\pi_t(\cdot), f(\cdot, \text{success}), t) > \pi^*) \supset \{t|m(\pi_t(\cdot), f(\cdot, \text{betrayal}), t) > \pi^*) \}
\]

3.3 Summary

This chapter has presented a theory that explains how international actors might observe and assess reputation when forming new international agreements. Through the discussion presented here, I have developed four primary theoretical findings (propositions). Proposition 1 suggests that membership in untested
agreements should benefit assessments of holder reputation. I argue that the process of entry-endorsement is the driving force behind the benefit from membership in untested agreements and suggest that observer interpretation occurs via two different, context specific mechanisms. As indicated by proposition 2 when observers and endorsers are the same sort, we expect the strength of entry endorsement’s effect on holder reputation to be determined by the similarity between the observer and holder. When observers and holders are more similar, we expect entry endorsement to confer a stronger reputational benefit on the holder. Alternatively, as indicated in proposition 3, when observers and endorsers are different sorts (e.g. the observer is a firm and the endorser is a state), and when the endorser faces punishment for making bad endorsements, the strength of entry endorsement will be determined by the endorser’s vulnerability to punishment for bad endorsements. When endorsers face more severe punishment for bad endorsements, the reputational benefit of entry-endorsement should increase. Lastly, through proposition 4, we expect the rate of decay in reputation to depend on the observer’s prior experiences in agreements. When an observer has been betrayed in prior agreements, reputation will decay more quickly.

In the chapters that follow, I build these theoretical propositions into observable implications and test them empirically. It should be noted that proposition 2 and proposition 3 are inherently at odds with one another. Therefore, no singular empirical test can apply to both propositions. Instead, I test these propositions separately, in two different empirical chapters. In chapter 5, I test proposition 1 along with proposition 2 and proposition 4. In chapter 4, I test proposition 1 along with proposition 3 and proposition 4.
Chapter 4

Reputation and Investment

In the previous chapter, I outlined a general theory of reputation that specifies conditions under which reputation should improve and decay. In this chapter, I test this theory empirically, as it applies to international political economy. Specifically, in this chapter, we will test the implications of proposition 1, proposition 3 and proposition 4 for international investment and investment treaties. As a reminder, here are the plain language versions of each of these propositions:

**Proposition 1** (Simplified from page 42): Under certain conditions, a holder that has entered an untested agreement will receive a reputational benefit and thus be more likely to enter into an agreement with a given observer than similar holders that have not entered such an agreement.

**Proposition 3** (Simplified from page 46): Under certain conditions (including when observers and endorsers are different sorts), the reputational benefits of entry endorsement will be stronger when endorsers are susceptible to domestic punishment for making bad endorsements and thus entry endorsement will make agreements between observers and holders more likely when endorsers are susceptible to domestic punishment for making bad endorsements.

**Proposition 4** (Simplified from page 50): Entry endorsement decays faster when an observer has been betrayed in the past and (if the observer has extensively been betrayed) entry endorsement produces an agreement between the holder and observer only if the entry endorsement occurred very recently.
The remainder of this chapter proceeds in three main sections. In section 4.1, I outline the rationale for examining international political economy as a source of evidence to test my theory. I also provide motivation for focusing more narrowly on a subset of political/economic interactions: greenfield foreign direct investment and investment treaties. Through this section, I discuss the observable implications for investment and investment treaties that follow from the general propositions presented in chapter 3, arriving finally at a set of specific hypotheses. In section 4.2, I design empirical tests of these hypotheses. I provide descriptions of the relevant measures and modeling strategies in this section. Finally, in section 4.3, I execute this design and discuss my findings. Generally, I find support for my hypotheses and the theory as a whole. Unexpectedly, however, support for proposition 4 is somewhat nuanced.

4.1 Where to look for evidence: Reputation in International Political Economy

Generally, international interactions can be divided into two substantive issue areas: matters of international conflict/security and matters of political economy. While both issue areas make for suitable sources of evidence, there are a few reasons that I have chosen to draw upon matters of international political economy (IPE) as a source of evidence in this chapter.

First, by testing my theory as it applies in both issue areas, we dispel with concerns that the scope of my theory is restricted by substantive focus. Put differently, by considering the applicability of my theory to both issue areas, I can better establish its general applicability across international relations. This, of course, necessitates that I draw upon evidence from conflict/security (see chapter 5) as well as from political economy (this chapter).

Second, matters of IPE tend to provide a more suitable context for testing proposition 3 of my theory. Recall from chapter 3 that we expect observers to evaluate the quality (or strength) of entry endorsement by considering features of the endorser. Further, recall that the observer will focus on different features of the endorser depending on the relationship between the endorser and observer. This aspect of the model
produces two divergent propositions, only one of which should be applicable at any given time. Proposition 2 only applies when observers and endorsers are the same sort of actor (e.g. both are states) and where there is limited or no potential for retribution for bad endorsements. This proposition tends to apply more broadly to matters of conflict and security (see chapter 5). Proposition 3, however, pertains to scenarios in which the endorser and observer are different sorts of actors, and when endorsers are sensitive to punishment for making bad endorsements.

It is intuitive to see that this sort of scenario applies, generally, to matters international political economy. Generally, sub-state, economic actors (such as firms and individuals) take on considerable risk when engaging in international economic activity. These economic actors consider the behavior of their own and other state governments when assessing the risk of international exchange, and the reputation of foreign governments\textsuperscript{1}. Because failed agreements are likely to precipitate negative economic consequences, and because governments face reprisals for economic shocks\textsuperscript{2}, it is reasonable to expect that governments might face punishment (through both electoral defeat and/or special interest pressure) resulting from the formation of economic agreements that later fail (bad endorsements). I will discuss this endorser-observer dynamic (as it applies to political economy) in greater detail below.

\subsection*{4.1.1 Investment and Reputation}

Generally speaking, international political economy is concerned with the politics surrounding the flow and exchange of three types of things: migration (labor), trade (goods) and investment (capital). At a pragmatic level, it is impractical to test my theory as it applies to the whole of political economy – there is simply too much heterogeneity in the politics involved in different types of exchange. Instead, this analysis will focus, on investment. Specifically, it will examine how investment decisions (taken as a reflection of a state’s reputation for protecting investor rights) change in the context of bilateral investment treaties (taken as the forum in which reputation is shaped and changed). Generally speaking, investment and investment


\textsuperscript{2} See Mansfield and Pevehouse, 2006.
treaties are well suited for testing my theory. The specific rationale for focusing on these phenomena is motivated by both theoretical and practical considerations which bear detailed discussion.

4.1.1.1 Value of reputation and foreign direct investment

Consider foreign direct investment (FDI) as an interaction between at least two actors: an (investor) firm, and a host state. Firms provide capital investment inflows into the state and receive financial returns at a later date. States, receive, in exchange job-creating, and economy boosting investment. States can choose whether to uphold investor rights (chiefly, their property rights and rights to withdraw their profits/capital from the state). It is also possible for governments to expropriate investor property or otherwise behave in ways that are financially punitive (such as using regulations to selectively impose fines and taxes upon foreign investor holdings). Obviously, it is important for investors to identify the probability that the host government will engage in this behavior. Moreover, since investor losses from expropriation will be high and legal remedies (to the extent that they exist and can be somehow enforced) may take years to secure, there are strong incentives for investors to determine the probability of expropriation before the investment takes place. As discussed in chapter 2, the influence of reputation on cooperation is unique in that it enables observers to avoid negative consequences such as expropriation (which amounts to defection or betrayal) before it occurs.

Not all investments are identical, however, and some types of investment decisions are more likely to place a greater emphasis on reputation than others. In this chapter, I will focus on the role of reputation as manifest in foreign direct investment (FDI) decisions. This is to be contrasted with portfolio investment decisions, which are less suitable for analysis here. Foreign direct investment, is direct ownership and management control over a venture. Generally, it is a long-term form of investment that demonstrates low

---

3 This is meant here to include goods or services that can be exchanged for profit
4 See Wellhausen, 2014, p. 16-17.
5 See Simmons (2014), Bürthe and Milner (2009), Hallward-Driemeier (2003) among others for robust discussion. Direct government expropriation is relatively rare, though some studies have taken up the subject, see Tomz and Wright (2010), Hajzler (2012) and Körbin et al. (1980). Increasingly, expropriation occurs through indirect, regulatory means in which withdrawal of capital/profit is limited or losses are imposed through discriminatory regulation; see Sloane and Reisman (2004), Dolzer and Stevens (1995) and Graham, Johnston, and Kingsley (2015)
liquidity. This means that direct investment decisions must consider the long-term risk of expropriation and that investors may not be able to easily escape imminent expropriation once the initial investment decision has been made. This should make FDI decisions heavily dependent on the assessment of expropriation risk (which is essentially an assessment of a state’s type).

We will further restrict our scope to a particular type of FDI called greenfield investment. Greenfield investment is characterized by the establishment new ventures (and subsequent management/operation of those ventures). It is to be contrasted with mergers and acquisitions (sometimes called brownfield investment), in which an investor takes over some management/ownership stake in a pre-existing venture. Examples of greenfield FDI would include the construction and operation of a new manufacturing plant, or investment in the discovery and extraction of minerals or other natural resources.

Compared to brownfield investment, greenfield investment decisions are more easily influenced by reputation assessment, particularly reputation assessment as informed by the host state’s international behavior. To see this, consider that, brownfield investment decisions concern merger with or acquity of established, operating ventures in the host state. The existence of these ventures can provide potential investors with considerable resources for assessing/avoiding the risk of expropriation. For example, investors may be able to rely upon an acquired venture’s existing personnel to navigate state bureaucracy. As an alternative example, investors may consider the mere existence of the venture to be an indication that the risk of expropriation is acceptably low. Alternatively, when engaging in greenfield investment, the investor cannot rely on the information or assistance provided by a pre-existing venture. This, in turn, should magnify the reliance of the investor on its assessment of host state reputation in determining expropriation risk and, in turn, deciding whether or not to engage in greenfield investment in a given country.

---

6 Portfolio investments are largely investments via a stock market or other, similar security market mechanism, and exhibit a relatively high level of liquidity. Portfolio investment decisions certainly rely upon reputation, however, they are also much more subject to influence by the speculative behavior or other market actors. This, combined with the practical difficulty of tracking individual portfolio investment decisions over time, makes it difficult to draw inferences about reputation from portfolio investment decisions.

7 This example might be considered an example of reputation as informed by entry endorsement, and compatible with the theory presented in chapter 3. It is notable, however, that this example espouses a relationship between the endorser (the existing venture) and the observer (the investor) that is more compatible with proposition 2 than proposition 3 – which is to say it is an example of similar-sort actors that are unlinked. As such, it is not suitable for analysis in this chapter and otherwise would overlap with the analysis presented in chapter 5.
4.1.1.2 Reputation Dynamics and BITs

In the previous section, I explained why greenfield investment decisions should, expectedly, rely on reputation assessment of the host state, by the investor. The question remains, however, as to what information is used in that reputation assessment and why we should consider a state’s behavior in international agreements as significantly informing the host state’s reputation with a given investor.

Let us begin by considering the most appealing, alternative source of information for reputation assessment: the experience of other firms in a state. It seems intuitive to expect that an investor might investigate the experiences of other, similar investors in a host state. If other investors have had good experiences in the host state, then that is an indication of the host state’s reputation for upholding investor protections. If, on the other hand, prior investors have experienced expropriation losses from investing in the host state, then it is an indication of the host state’s reputation for not protecting investors. As previously discussed\(^8\) however, there are strong, strategic reasons why a firm may hide or misrepresent their experiences in a host state. To this end, it may be the case that an observer (investor) cannot obtain or rely upon information about prior investment experiences. Beyond this, and as a practical matter, it may be difficult to track down and investigate the investment experiences of other, similar firms, as these investments are not necessarily recorded or publicized in any centrally located information repository.

A host state’s behavior in international agreements may, by contrast, be a more suitable, or at least more visible source of information for reputation assessment. This is particularly the case if we focus on investment agreements, such as bilateral investment treaties (BITs). Bilateral investment treaties are agreements between states that are intended to coordinate policy on investor rights/protections. Generally, BITs take on similar design characteristics\(^9\). Important for our concerns here is the specification of arbitration terms and, typically, the specification of the United Nation’s International Centre for Settlement of Invest-

---

8 See section 3.2.5 (specifically, page 44)
9 Neumayer and Spess (2005, p. 12). It should be noted that recent scholarship tends to point out differences in the design of BITs and their implications, see Allee and Peinhardt (2014). These differences, however are less important for our concerns here since we are primarily interested in the signing of bits and filing of ICSID complaints as temporal markers for reputational events (i.e. entry endorsement and test of commitment).
ment Disputes (ICSID)\textsuperscript{10} as the arbitrator for disputes.

Extant literature supports the idea that investors are sensitive to state involvement in investment agreements\textsuperscript{11}. Moreover, membership in these agreements is centrally recorded\textsuperscript{12}. To the extent that BITs specify a standing arbitration body for the resolution of disputes, alleged violations of these agreements are centrally recorded in the form of arbitration proceedings. Considering this, BITs and the arbitration proceedings they establish are ideally suited to function as reputation clearinghouses\textsuperscript{13}. Accordingly, it seems reasonable to expect that investors would closely consider BIT membership and involvement in arbitration when assessing a state’s reputation for protecting investors.

4.1.2 Observing reputation in investment

Before moving on, it is important to emphasize that we cannot observe or measure reputation directly and to further discuss how we might observe changes in reputation. Fundamentally, reputation is a latent concept that describes a relationship between, at a minimum, a holder and an observer. Indeed, much like any relationship, it is not a phenomenon that can be directly observed. If relationships could be directly observed, the entire enterprise of science would be much easier – and far less open to debate\textsuperscript{14}. Instead, we must rely on inference to determine the status of a holder’s reputation with an observer. The most direct way to do this is to consider phenomena that we expect to be causal outputs (on behalf of the observer) of a positive or improving reputation. We expect that the observation of these outputs is an indicator of a holder’s good or improving reputation with the observer.

As should be apparent from the preceding sections, the causal outputs we will focus on as an indicator of reputation will be foreign direct investment decisions. For the reasons discussed above, we expect that

\textsuperscript{10} There are other arbitration bodies that are attached to BITs, including ad-hoc tribunals. As noted by Büthe and Milner (2009), however, ICSID has emerged as the “standard designated arbitrator in BITs” (p 187).

\textsuperscript{11} See Gray, 2013; Neumayer, 2002; Tobin and Rose-Ackerman, 2011, and others.

\textsuperscript{12} the UN Conference on Trade and Development maintains a list of signed investment agreements. Other repositories, though these may not be as comprehensive. These include the UN treaty collection, and repositories of trade agreements (some of which may also include investment-related clauses)

\textsuperscript{13} See Milgrom, North, et al., 1990.

\textsuperscript{14} The underpinnings of this reasoning runs deep in the philosophical study of epistemology and metaphysics. While it is a fascinating line of inquiry, it is inappropriate to the scope of this document. See Hume (2000) for a concise and straightforward discussion.
the announcement of a new greenfield investment project indicates that an investor (read: an observer) has assessed a host state’s reputation for upholding investor protections to be sufficiently good\(^{15}\) so as to justify a new, potentially costly, investment.

Now, it should be noted that there are other potential ways to observe reputation in the context of political economy. Chiefly, credit ratings agencies (such as Moody’s and Standard and Poor’s), and risk assessment services (such as Political Risk Services) both seem like more direct ways of observing reputation. Note however, that we are interested in the judgment of a particular set of actors – investors that use reputation in strategic decision-making. As such, information provided by rating agencies or risk assessment services are not a causal output of reputation between the holder and observer (for our purposes here, the investor). Instead, these are reputational inputs – sources of information that an investor might use to inform their assessment of reputation. To test the theory laid out in chapter 3, we need an indicator of the observer’s (investor’s) assessment of the reputation of the holder (host state). Ratings and risk assessments would not provide adequate information in this regard.

4.1.3 Testable Implications from Theoretical Propositions

Considering the preceding discussion, we now proceed to a consideration of the testable implications of proposition 1, proposition 3 and proposition 4 as they apply to foreign direct investment and international investment treaties.

4.1.3.1 Hypotheses about Entry Endorsement and Untested Membership

Let us begin with proposition 1. To derive hypotheses that test this proposition, we need to relate three main concepts to their observable investment-related counterparts: (1) reputation, (2) entry endorsement and (3) untested membership – specifically the temporal markers that delineate a period of untested membership.

\(^{15}\) specifically, we expect the reputation assessment to be consistent with the threshold established in Equation 3.2
A host state’s reputation is indicated by the decision of investors to start new, greenfield investment projects in the host state. As discussed above, since we expect greenfield foreign direct investment decisions to be heavily influenced by the host state’s reputation, we can infer that new greenfield investment projects represent a positive reputation. Specifically, new projects represent a positive reputation between the investors in a home state (observers) and the host state (holder state).

Entry into a bilateral investment treaty is an action that should confer entry endorsement benefits to the state’s reputation. As discussed in chapter 3, entry endorsement is the notion that a holder state’s mere entry into an agreement is an implicit endorsement of its reputation for reliably complying with the agreement. Many agreements may confer entry endorsement, but we expect that the strongest influence of entry endorsement would follow from entry into agreements that are particularly relevant to the decision facing an observer (in this case, an investment decision). In other words, we want to focus on the type of agreement that speaks most clearly to a state’s reputation for upholding investor protections. Bilateral investment treaties are agreements that commit member states to uphold specific investor protections and engage in arbitration proceedings in the event of investor-state disputes. Accordingly, we would expect entry into a BIT to confer an entry endorsement benefit to a host state’s reputation for upholding investor protections.\footnote{To this point we have described entry endorsement as influencing a state’s reputation for reliability, which may seem different from reputation for upholding investor protections. Consider, however, that a state’s “reputation for reliability” is a short-hand way of describing its “reputation for reliably complying with its agreements”. BITs commit states to upholding specified investor rights. Transitivity, then, the entry endorsement that comes from entry into a BIT should be interpreted as influencing its “reputation for (reliably) upholding investor protections”.}

The period of time between a state’s entry into a BIT and the filing of an arbitration claim under that BIT is a period of untested membership. As we described earlier, untested membership occurs during a period of time when a state is a member of an agreement, but its compliance with the agreement cannot be conclusively observed from its actions. Such is the case for the period of time after signing a BIT but before arbitration claims have been brought under the BIT. To see this, consider that the central provisions of an investment treaty compel the state to refrain from specific actions, such as expropriation of foreign-owned capital. We would observe non-compliance if we were to observe the state engaging in these acts. Failure
to observe the state engaging in these acts, however, does not sufficiently imply compliance (i.e. that the state intends to uphold investor protections). It could be the case that the state intends to violate the agreement, but has lacked the opportunity to do so. It may also be the case that the state has engaged in prohibited behavior, but that these actions went unobserved (or under-observed)\textsuperscript{17}. Filing an arbitration claim, however, demands the state take specific action. The state must participate in the arbitration process and abide by the arbitration decision to demonstrate compliance. Failure to participate in (or comply with) arbitration indicates non-compliance with the agreement. The filing of an arbitration claim thus initiates a test which discriminates between compliance and non-compliance\textsuperscript{18}. Taken this way then, the period of untested membership is marked by two easily observable, well-documented events: the initial signing of the BIT (which initiates membership in the agreement and confers entry endorsement), and the first filing of an arbitration claim under the BIT (which initiates a test and thus ends the period of untested membership).

At this point, having identified observable counterparts for entry endorsement, reputation and untested membership, we can discuss the implications of proposition 1. We expect, given proposition 1, that entry endorsement has some influence on state reputation, even if membership an agreement is untested. This suggests that, in the context of investment, we should expect entry into an untested BIT to improve reputation (manifest in new investment) more than membership in a BIT that was subsequently tested in arbitration. This leads to our initial hypothesis:

**H 1:** States that have recently entered into an untested BIT are more likely to receive new investment than states whose membership in a BIT has been tested by arbitration.

Note that we do not expect unqualified support for this hypothesis. We would expect that the reputation of a holder state that is a member of an untested BIT to eventually decay towards uncertainty. This suggests that we are likely to find some critical recency threshold such that a state’s entry into an untested agreement prior to the threshold will receive no more investment than a state that has been tested

\textsuperscript{17} This might occur if the expropriation occurs through regulatory or indirect means. See Sloane and Reisman (2004) and Büthe and Milner (2009). It might also occur if expropriation occurs, but the state and investors reach a settlement prior to the filing of a formal arbitration claim.

\textsuperscript{18} Though this test is specific to the arbitration portions of the agreement, it is reasonable to interpret compliance as indicative of a willingness to comply with the agreement overall.
through arbitration. Accordingly, we expect support of this hypothesis to be qualified. At a minimum, we would expect that very recent entry into a BIT yields more expected investment than a state that was tested in arbitration a long time ago and has not entered any new BITs since.

In addition to the previous hypothesis, consider also that Proposition 1 suggests that reputation will change in a specific way over time. Namely, that the benefit from entry endorsement erodes over time. Accordingly, from proposition 1 we arrive at an additional hypothesis:

**H 2:** States that have more recently entered an untested BIT will receive more investment than states that have less recently entered an untested BIT (but remain untested)

This hypothesis is important since it also addresses a possible objection to the design of this analysis. It may seem that, given the nature of investment treaties, and our focus on the period of time between membership and arbitration, that membership in an untested agreement indicates membership in an unviolated agreement. Our hypothesis, a critic may argue, is a test of membership in an unviolated agreement is reputationally beneficial and not exclusively a test of our theory. This would be inaccurate. Consider that a theory that suggested that membership in an unviolated agreement indicated compliance would view the period between membership entry and initial arbitration to be one of high, observable, compliance. As such, we would expect that the longer the period of time between membership and arbitration, the stronger the expectations of compliance. As such, the expectation would be that new greenfield investments grow increasingly likely in the host country as the period between membership and initial arbitration grows longer.

This expectation runs directly contrary to the hypothesis outlined above. Because entry endorsement decays over time, we expect new investment to be less likely as the period of untested membership grows longer. In light of this, the hypothesis provided above is suitable to my theory and discriminates among competing theories.
4.1.3.2 Hypotheses about Endorsement Quality

Proposition 3 is about the strength of entry endorsement. The discussion in section 4.1.1 describes why we might expect that investors (reputation observers) would consider the interactions between a host state (holder) with other states (endorsers) as a source of information when assessing reputation. And the chapter to this point has already discussed why observers would look specifically at a host state’s involvement in investment treaties as a source of endorsement. To derive testable hypotheses, however, we need to consider the notion of endorser vulnerability and how it might apply to the context of IPE and international investment.

The logic that supports proposition 3 is that states are influenced by the economic well-being of their constituents. They may face reprisals for negative economic shocks or negative economic performance. As discussed in section 3.2.5, as the state’s sensitivity to punishment grows, it is inclined to be more selective about its choice of partners. It does this if not for fear of direct punishment for bad endorsements, then as a means of implicitly discouraging agreements that may fail bring negative consequences for observers (here: investors). A state’s sensitivity to punishment scales to the magnitude of the punishment. There are two relevant channels through which a state might suffer punishment/influence: electoral and special interest\(^{19}\).

States face electoral punishment for bad economic performance or economic shocks. If a state makes a bad endorsement, entering into a BIT with a country that later expropriates investment capital from the investors (residents of the endorser state), then the overall state economy might suffer. In more democratic states, voters would have greater opportunity to remove incumbents from office. This should make leaders in more democratic states more selective about their endorsements than leaders in less democratic states. Thus, proposition 3 may imply:

**H 3:** States that enter into untested BITs with democratic states receive more investment than states that have entered into untested BITs with non-democratic states.

\(^{19}\) See Grossman and Helpman, 1994.
Now, it may be the case that voters are not as well-informed or well organized around the technical
details of international investment\textsuperscript{20}. As such, even in democracies, states may not be terribly vulnerable to
electoral punishment for bad endorsements. We may not find much in the way of support for this hypothesis.

Even if states are not vulnerable to electoral punishment for bad endorsements, they may still be
vulnerable to special interest punishment. Generally, we would expect that the strength of an economic
sector would be proportionate to its importance to the economy. So too, would we expect the ability of an
economic sector to punish/influence the state to be a function of its importance to the economy. It follows
that a state should be more sensitive to punishment for making bad endorsements as the role of foreign
investment grows larger. Accordingly:

**H 4:** A host state that enters into an untested BIT will receive more foreign investment if its partner
(endorser) state has large foreign asset holdings (as a proportion of the economy).

A host state that enters into an untested BIT will receive less foreign investment if its partner (en-
dorser) state has small foreign asset holdings (as a proportion of the economy).

Note that hypothesis 4 is stated in two clauses for readability. The clauses are synonymous in their
description of the expected relationship. As such, a singular test will either support both clauses or fail to
support either clause.

### 4.1.3.3 Hypotheses about Betrayal

Proposition 4 is about the influence of an observer’s own experiences on its assessment of reputation.
Specifically, we expect that observers that have been betrayed will be more skeptical about the stability
of a holders type and expect transitions between type to occur more frequently. This means that entry
endorsement’s benefit to a state’s reputation with an observer should decay more quickly if the observer has
been betrayed in the past.

\textsuperscript{20} See Kono (2006) for supportive discussion
Now, in an ideal world, we might consider examining betrayal by focusing on the extent to which individual investors have experienced failed foreign investments. As a practical matter, however, this is unmanageable. Collecting data on investors might be possible, but firms may have incentives not to reveal prior betrayals. This would produce a substantial, non-ignorable selection effect in our data. Furthermore, while focusing on individual investors may reveal an observable indicator of betrayal, but it would limit our ability to observe reputation over time and in a way that is relevant to reputation.

Instead, we will consider the implications of betrayal as they might aggregate to the state level and manifest through arbitration filings. In other words, we will consider the experiences of the investors of a state, as a whole. We would not (and should not) necessarily expect every betrayal of an individual investor to produce an aggregate influence on investors. The filing of an arbitration claim, however, is a highly-visible, well documented event that an investor believes that it has been betrayed by a host state. It is an indication of betrayal that is visible to all investors and from which we can reasonably expect there to be an effect on the aggregate behavior of investors in a state. The impact of filing a claim on investors should decline over time since the composition of the investors (as a group) will naturally change (i.e. new firms will emerge, others will go out of business). As such, we can expect that the influence of an old claim on investor behavior is less than the influence of a new claim. Effectively, then, investors are less betrayed if the last filed claim occurred long ago, and more betrayed if the last filed claim occurred recently. This leads us to the following hypothesis:

**H 5:** Among states that are home to a firm that has filed an arbitration claim recently, the relative improvement in investment that attends a host state’s entry into an untested BIT will deteriorate rapidly. Among states that are home to a firm that has filed an arbitration claim less recently, Host states that have
entered an untested BIT will be less likely to receive investment from a state that is home to a firm that has recently filed an arbitration claim.

4.2 Examining Reputation in Foreign Direct Investment: Design and Methods

Having developed a set of testable hypotheses, we are left to conduct a suitable empirical analysis. The design of this analysis is nuanced and requires detailed discussion here. I will proceed by explaining the general structure of the data to be examined, the measures to be used and the modeling/estimation strategy that I will use.

4.2.1 Unit of Analysis

The unit of analysis will be directed dyad-time (state-state-month). One member of the dyad is the host (holder) state. The other member is meant to represent investors (observers). Since we are focused on investors as a societal/economic interest group, and we are not concerned with the behavior of competing interest groups, we can set the other member of the dyad at the state level as well (the investor state)\(^{24}\), without loss of differentiability across observations\(^{25}\).

Time will be considered at a monthly level. Most other analyses of investment tend to focus on investment by year. The order of events is crucial to my theory. To adequately test hypothesis 1, for example, we must be able to identify investments that occurred after entry into a BIT, but prior to arbitration filing. Sometimes, all three of these things occur in the same year; using a monthly unit of analysis gives us greater leverage over time order in these cases.

\(^{24}\) Generally, the term investor state should be taken as shorthand for the investors, taken as a group, that identify the state as their 'home' state.

\(^{25}\) In other words, the set of observations produced by setting second dyad member to the state is congruent to the observations that would be produced by setting the second dyad member to the investor group.
4.2.2 Measurement: Foreign Direct Investment

The dependent variable in this analysis should reflect a host state’s reputation as it changes over time. As discussed above, we will focus on FDI as an indicator of a state’s reputation. Specifically, we will focus on greenfield FDI decisions since these should be heavily dependent on a host state’s reputation. We will measure investment decisions as a monthly count of new greenfield investment projects started in a given host state by investors in the investor state. Higher counts indicate more investment and a better reputation of the host state with investors in the investor state. The data for this measurement was collected from FDiMarkets, a subsidiarity of Financial Times\textsuperscript{26}, that specializes in tracking and collecting data on greenfield investment.

Readers may wonder why we are using FDiMarkets data instead of more commonly found public data on FDI flows\textsuperscript{27}. Simply put, this data is not suitable for our needs. The public data on FDI is typically derived from balance of payments statistics and presented in an aggregated form. Specifically, it is calculated on the basis of the directional principle, aggregated by host country and presented in terms of net flows (inflows - outflows). The fact that the data is not widely available in terms of bilateral flows makes it unsuitable for testing our hypotheses. It is made further unsuitable by the incorporation of outflows in the reported data, which distort our ability to use FDI as a proxy for the desirableness of the host state as a home for investment. Lastly, the aggregated presentation of the FDI data includes investment for mergers and acquisitions (brownfield investment) as well as additional funds directed towards ongoing ventures (e.g. the purchase of additional ownership stake in a foreign firm). The inclusion of these other forms of investment obscure levels of greenfield investment which is likely to indicate reputation and is the focus of our analysis here.

Readers may also wonder why I have chosen to focus on a count of investment projects instead of more common metrics such as investment flows/stocks. The reason for this is that we want to focus on decisions as an indication of reputation. A count of projects is the more appropriate metric for this purpose, since

\textsuperscript{26} I was able to acquire this data through a generous grant from the University of Colorado Political Science department

\textsuperscript{27} Such data is released by the World Bank among other sources
each project can be taken as a discrete decision to invest (and hence a positive assessment of the host state’s reputation in the investor economy). If we were to focus on the monetary value of investment (i.e. stocks/flows), we would have a sense of the magnitude of investment, but not the number of investment decisions. Consider, for example that a high monetary value of investment, in a given month, might indicate one large investment by a large corporation or a number of investments from smaller firms. Measurements of the aggregated monetary value of investment will obscure this information while measurements of projects will not.

The data collected for this measure covers the time period between 2004 to 2015 and includes observations from 200 countries. Due to the limitations of other variables, a number of these observations are dropped from model estimates, leaving roughly 29,000 observations across 111 countries.

4.2.3 Measurement: Untested BIT membership (entry endorsement)

Central to the test of the hypotheses presented in this chapter is a measurement of entry endorsement. We will measure the act of entry endorsement as a state’s entry into a BIT. As indicated in proposition 1, and as specified in hypothesis 1 and hypothesis 2, we expect the influence of entry endorsement to be sensitive to its recency. The influence of entry endorsement should be moderated along two related dimensions of recency: (1) the recency of entry endorsement relative to the present and (2) the recency of entry endorsement relative to the host state’s last test (i.e. arbitration claim – the last unambiguous demonstration of compliance). To phrase this differently, we expect the influence of entry endorsement to be moderated based on the location of entry endorsement relative to two points in time (the last demonstration of compliance and the present). I measure this through a variable, \textit{ENTRY RECENCY}, specified as follows:

\[
\text{ENTRY RECENCY} = \frac{\text{last entry into BIT} - \text{last arbitration}}{\text{present} - \text{last arbitration}}
\]

Examples of this measure are illustrated in Figure 4.1. This figure groups four timelines into two
groups, according to entry recency. As should be clear from this figure, the level of recency is determined by the distance between the last arbitration and the BIT entry relative to the total temporal distance between the last arbitration and the present.

**Figure 4.1** Conceptual examples of low and high levels of entry recency

<table>
<thead>
<tr>
<th>Low Entry Recency</th>
<th>High Entry Recency</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Weak Influence of Entry Endorsement):</td>
<td>(Strong Influence of Entry Endorsement):</td>
</tr>
</tbody>
</table>

The data used in the construction of ENTRY RECENCY comes from a few sources. The data on BIT signing is collected from the UN’s directory of investment agreements. I focus on the signing date of a BIT (as opposed to the ratifying date) because the signing date is more widely visible, and is thereby likely to carry the effect of entry endorsement\(^28\). The data used to determine the last arbitration date comes from

\(^{28}\) Construction of ENTRY RECENCY using the ratification date of BITs produces negligibly different results.
ICSID arbitration filings in which the host state is named as a defendant. In the event that the state does not have a previous dispute, the last test is taken as Nov 25, 1959, which is the date of signing of the first BIT\(^{29}\).

Readers should note, given the construction of this measure, that \textit{ENTRY RECENCY} can take on positive and negative values. Negative values of \textit{ENTRY RECENCY} indicate that a state’s most recent membership in a BIT has been tested through arbitration. Positive values of \textit{ENTRY RECENCY} indicate a holder’s most recent membership in a BIT is untested – that entry endorsement is the most recent indication of a holder’s type.

Since \textit{ENTRY RECENCY} takes on two distinct meanings, depending on if it is positive or negative, I also employ a dichotomous measure called \textit{AMBIGUOUS}. This measure is scored ‘1’ if \textit{ENTRY RECENCY} is greater than zero, and 0 if \textit{ENTRY RECENCY} is less than, or equal to zero. Effectively, this tells us whether a state’s most recent membership in a BIT was tested (0) or untested (1). By interacting these two measures (\textit{ENTRY RECENCY} and \textit{AMBIGUOUS}), I can distinguish the effect of entry endorsement on FDI when entry endorsement is the most recently observable indicator of the host state’s type (and thus the most recent information that should inform reputation assessment).

\subsection*{4.2.4 Measurement: Endorser Vulnerability}

We have two separate hypotheses about endorser vulnerability as a moderating factor of the strength of entry endorsement. These hypotheses correspond to two distinct channels of influence: electoral and special interest. To capture the electoral channel, we use the polity of the host state’s most recent BIT partner (i.e. the polity of the host state’s most recent endorser).

To capture the susceptibility of the endorser state to special interest influence, we examine the size of international investment as a proportion of the state’s economy. This is done with international investment position data, focusing on state direct investment assets. This indicator, taken as a percentage of GDP,\(^{29}\) The use of this date is somewhat arbitrary. It is, however, historically fitting as the start of the era of BITs. Selecting other dates does produce different values for \textit{ENTRY RECENCY}, but it should be noted that the overall analysis presented here is not adversely impacted by them.
provides an indication of the importance of cross-border direct investment to the economy. I will refer to this measure as investment sensitivity, and it will be denoted as ENDORSER DIRECT ASSETS. Data for constructing this measure comes from the IMF’s International Financial Statistics\(^\text{30}\).

4.2.5 Measurement: Betrayal

To evaluate hypothesis 5, we must measure the extent of betrayal that investors have experienced. Because of the nature of the data we will be considering, we will need to measure this as it occurs to investors as an interest group, at the state level. As discussed above, it is not wholly unusual to expect that the betrayal of a single investor will have spillover effects on the broader group of investors in an economy\(^\text{31}\). As such, we can expect that the filing of an arbitration claim by any investor in an economy will generate an impact on reputation assessment by other investors in the economy and thus produce an aggregate effect on investment (at the state level) that can be leveraged to test hypothesis 5.

Note, however, that as discussed in section 4.1.3.3, we expect the observable impact of betrayal to reveal itself as a function of time. The reason for this is that the composition of the economy is continually shifting, with firms entering and leaving the market for cross-border investment. Newer firms that were not investors at the time of the last arbitration claim should be less influenced by its occurrence. Thus, as the composition of investors becomes increasingly populated by new firms, the impact of an arbitration claim should decline. Accordingly, we will measure the influence of betrayal in terms of the number of days since a firm headquartered in the investor state filed an arbitration claim, and as recorded in a variable named BETRAYAL. Note, consistent with section 3.2.7, that the arbitration claim can be made against any state, including the host state.

\(^{30}\) The precise series are IAP BP6 USD – direct assets, per the BoP and IIP manual (i.e. BP6)

\(^{31}\) This expectation is made frequently, by other literature on FDI and investment. See Wellhausen (2014), Allee and Peinhardt (2011), Minhas and Remmer (2015), Allee and Peinhardt (2014) and others.
4.2.6 Modeling Strategy

It is inappropriate to model the dependent variable as a linear model since the dependent variable has a hard lower-bound of 0. Instead, we will model FDI decisions as a count model, specified as follows:

\[
\text{FDI PROJECTS}_{ijt} \sim \text{Poisson}(e^{X_B + ZC})
\]

Where \( X \) represents a specification of key independent variables and \( B \) represents the vector of corresponding coefficients. And where \( Z \) represents a vector of control variables and \( C \) represents the corresponding vector of coefficients. And where \( i \) indexes the host (holder) state, \( j \) indexes the investor (observer) state and \( t \) indexes the date (month,year).

In this analysis, I estimate two separate specifications of \( X \). The first specification provides a primary set of estimates against which we evaluate hypothesis 1, hypothesis 2, hypothesis 4 and hypothesis 5. This first specification is given as follows:
To test hypothesis 3, I estimate a model with a different specification for $X$, which includes $\text{ENDORSER POLITY}$:
This second specification incorporates an additional set of interactions and as such, is more difficult to interpret. It is also computationally more difficult to work with since it adds another layer of variables to incorporate in effect simulations (I will discuss this further below). Accordingly, I will restrict my attention to Equation FDI Model 2 largely to the evaluation of hypothesis 3. It should be noted, however, that this second specification produces estimates that are largely consistent with estimates for Equation FDI Model 1 in terms of statistical significance and direction of effect.

In addition to the variables included in \( X \), model estimates will also include a vector of control variables \( Z \). These variables include the GDP from both host and investor countries, distance between the endorser and investor countries and between the investor and host countries, temporal splines, and host country fixed effects.

### 4.3 Findings

Model estimates are provided in Table 4.1. Because of the number of terms in the model, only relevant independent variable estimates are shown. If we were estimating a linear model, the estimates would be difficult to interpret due to the depth and complexity of interactions in the model. Since we are estimating a non-linear model, however, tabularized model estimates are uninterpretable in any meaningful way. As noted by Ai and Norton\(^{32}\), when involved in an in interaction, the marginal effect of a variable in a non-linear model is not calculated the same way that it is in a linear model\(^{33}\). Importantly, the size, direction and significance of the variable’s marginal effect are contingent on the values of the component terms of the interaction along with the values of every other variable in the model. This means that the model estimates presented in a standard table of estimates cannot, by themselves, be used to evaluate the influence of interaction terms. Furthermore, since each of the hypotheses described above is tied to an interaction term, a table of model estimates cannot, by itself, be used to evaluate the hypotheses specified in

\(^{32}\) 2003.

\(^{33}\) This is commonly computed as a partial derivative. In a model \( y = \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 \), the marginal effect of \( x_1 \) is: \( \partial(y)/\partial(x_1) = \beta_1 + \beta_{12} x_2 \). In a non-linear model, however, the marginal effect is calculated differently, essentially as a result of the chain-rule. Consider a non-linear model (with a mean function \( g() \)): \( y = g(\beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2) \). The marginal effect is \( \partial(y)/\partial(x_1) = g'(\beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2)(\beta_1 + \beta_{12} x_2) \).
this chapter. Readers should note, therefore, that Table 4.1 provides little actual insight here and is provided entirely in the interest of tradition.
Table 4.1: Model Estimates of FDI investment decisions

<table>
<thead>
<tr>
<th>dependent variable:</th>
<th>Greenfield Investment Projects</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(FDI Model 1)</td>
<td>(FDI Model 2)</td>
</tr>
<tr>
<td>AMBIGUOUS</td>
<td>0.310*</td>
<td>0.396*</td>
</tr>
<tr>
<td></td>
<td>(0.090)</td>
<td>(0.143)</td>
</tr>
<tr>
<td>ENTRY REGENCY</td>
<td>-0.0001</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>(0.0002)</td>
<td>(0.002)</td>
</tr>
<tr>
<td>AMBIGUOUS:ENTRY REGENCY</td>
<td>-0.331*</td>
<td>-0.441*</td>
</tr>
<tr>
<td></td>
<td>(0.094)</td>
<td>(0.146)</td>
</tr>
<tr>
<td>ENDORSER DIRECT ASSETS</td>
<td>-0.027*</td>
<td>-0.026*</td>
</tr>
<tr>
<td></td>
<td>(0.007)</td>
<td>(0.007)</td>
</tr>
<tr>
<td>AMBIGUOUS \times ENDORSER DIRECT ASSETS</td>
<td>0.015</td>
<td>0.018</td>
</tr>
<tr>
<td></td>
<td>(0.019)</td>
<td>(0.019)</td>
</tr>
<tr>
<td>ENTRY REGENCY \times ENDORSER DIRECT ASSETS</td>
<td>-0.0001</td>
<td>0.00001</td>
</tr>
<tr>
<td></td>
<td>(0.0001)</td>
<td>(0.0001)</td>
</tr>
<tr>
<td>AMBIGUOUS \times ENTRY REENCY \times ENDORSER DIRECT ASSETS</td>
<td>0.013</td>
<td>0.008</td>
</tr>
<tr>
<td></td>
<td>(0.019)</td>
<td>(0.019)</td>
</tr>
<tr>
<td>ENDORSER POLITY</td>
<td>-0.010*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.005)</td>
<td></td>
</tr>
<tr>
<td>ENTRY REENCY \times ENDORSER POLITY</td>
<td>-0.0001</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0002)</td>
<td></td>
</tr>
<tr>
<td>AMBIGUOUS \times ENDORSER POLITY</td>
<td>-0.010</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.013)</td>
<td></td>
</tr>
<tr>
<td>AMBIGUOUS \times ENTRY REENCY \times ENDORSER POLITY</td>
<td>0.013</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.013)</td>
<td></td>
</tr>
<tr>
<td>BETRAYAL</td>
<td>-0.00000</td>
<td>-0.00000</td>
</tr>
<tr>
<td></td>
<td>(0.00000)</td>
<td>(0.00000)</td>
</tr>
<tr>
<td>BETRAYAL \times AMBIGUOUS</td>
<td>-0.0002*</td>
<td>-0.0002*</td>
</tr>
<tr>
<td></td>
<td>(0.00001)</td>
<td>(0.00001)</td>
</tr>
<tr>
<td>BETRAYAL \times ENTRY REENCY</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td></td>
<td>(0.000)</td>
<td>(0.000)</td>
</tr>
<tr>
<td>BETRAYAL \times AMBIGUOUS \times ENTRY REENCY</td>
<td>0.00004*</td>
<td>0.00004*</td>
</tr>
<tr>
<td></td>
<td>(0.00001)</td>
<td>(0.00001)</td>
</tr>
<tr>
<td>observations</td>
<td>29,103</td>
<td>29,103</td>
</tr>
</tbody>
</table>

Note: Additional coefficients are suppressed here to conserve space. Model estimates included controls for the following: economic wealth (GDP), time dependence (splines), distance, regime type (Polity), direct investment assets, and host-state fixed effects.
4.3.1 Hypothesis Evaluation

To address the problems interpreting coefficient estimates just described, I will evaluate hypothesis 1 - 5 graphically. This is ideal since evaluation of the hypotheses outlined above requires interpretation of complicated interactions in the model. Moreover, it is difficult to present marginal effects (or even predicted values) in a non-graphical way that is both intuitive to the reader and adequate for evaluating the hypotheses listed above.

The figures here are generated using simulated quantities of interest. Ordinarily, simulations are used to generate predictions over a range of values of one or two covariates while the remainder of covariates are held constant (usually at their mean). This approach cannot be taken here for reasons just discussed. That is, since the effect of a variable on alliance formation is contingent on the values of the other variables in the model, we cannot simply hold control variables constant across simulations. Instead, I employ a slightly different approach as discussed in Hanmer and Ozan Kalkan34. Essentially, I generate a set of simulated coefficients and use them to predict the probability of alliance formation for every observation in the data, holding the variables of interest constant. I then record the average of these predicted values, increment the variable(s) of interest and repeat the process. Completing this for every value in a specified range of the independent variable(s) of interest constitutes one set of simulations.

To be precise, the simulated quantity of interest is the average predicted count of new greenfield projects. I will refer to this quantity throughout this section as the QI, or, as it’s interpretation is functionally similar, the predicted number of new projects. The figures generated here are each the result of 1000 sets of simulated QI. Lines indicate the average value of the QI across simulations and for a variable value corresponding to a point on the X axis. The ribbons indicate the interval between the .975 percentile estimate and the .025 percentile estimate at each value of the X axis. This is to be taken as an approximation for a 95% confidence interval.

To evaluate hypothesis 1 and hypothesis 2, consider Figure 4.2. In this figure, the X axis represents 34 2013.
levels of ENTRY RECENCY while the Y axis represents the expected number of projects to be announced by the investor state, directed toward a given host state. The figure contains two lines. One line indicates the expected number of projects announced at a given level of ENTRY RECENCY if AMBIGUOUS is held constant at 1. The other line corresponds to the expected number of projects announced at a given level of ENTRY RECENCY given the counterfactual condition\textsuperscript{35} that AMBIGUOUS is set to 0.

As anticipated, the expected number of projects is greater when AMBIGUOUS is held constant at 1, and when ENTRY RECENCY is sufficiently high (between about .7 and 1) than when AMBIGUOUS is 0. If ENTRY RECENCY is sufficiently low (below about .7), then we expect membership in a tested BIT to provide as many or more new project announcements. This suggests that states who recently entered an untested BIT benefit simply from entering the agreement, but that benefit is somewhat short lived. Recall that ENTRY RECENCY is relative to the time of last entry and the last time the holder was in a tested agreement. Accordingly, levels of ENTRY RECENCY do not correspond to fixed or absolute periods of time. Generally, however, an ENTRY RECENCY of .7 corresponds to an agreement entry occurring between 4 and 6 years prior to the year of a given observation.

Note that the figure indicates that the benefit of entry into a BIT deteriorates over time if left untested or not superseded by entry into a new BIT. This trend is expected by our theory and runs counter to an alternative explanation for why entry into a BIT is beneficial: that membership in an untested BIT reflects a lack of violations and hence an accumulating reputation for protecting investor rights. If this alternative explanation were correct, then we would expect a state's reputation to build over time – the more time between entry into a BIT and an arbitration claim, the better the reputation. Accordingly, we would expect to see the line corresponding to AMBIGUOUS = 1 to take a negative slope, indicating that long-run membership in BITs without an arbitration claim generated a reputational benefit.

Figure 4.2 provides qualified support for hypothesis 1. As discussed in section 4.1.3, estimates only fail to support hypothesis 1 if ambiguity produces a negative effect across all possible covariate profiles. This

\textsuperscript{35} We are able to generate these counterfactual estimates because the way that we generated the simulated QI. An appropriate interpretation of this line would be the expected number of projects if, at the time corresponding to the level of ENTRY RECENCY, the host state joined a BIT and was immediately subject to an arbitration claim.
figure, however illustrates support for hypothesis 1 at every point for which the lines labeled $\text{AMBIGUOUS} = 1$ are higher than the lines labeled $\text{AMBIGUOUS} = 0$. Thus, it would appear that, in at least some instances (namely when $\text{ENTRY RECENCY}$ is greater than about .7), mere entry into an (untested) alliance makes future alliance entry more likely (suggesting a benefit to reputation).

**Figure 4.2** Expected number of projects at levels of agreement entry recency when entry is tested and untested.

Additionally, Figure 5.2 and Figure 5.3 demonstrate greater support for hypothesis 2. The benefits of membership in an untested BIT on new investment projects are strongest immediately after agreement entry and fade quickly thereafter. This is indicated by the steep decline in predicted agreement formation as one moves from $\text{ENTRY RECENCY} = 1$ to $\text{ENTRY RECENCY} = .7$, along the $\text{AMBIGUOUS} = 1$ line. At levels below $\text{ENTRY RECENCY} = .7$, a holder state does as well or better by membership in a BIT under which arbitration claims have been filed.
4.3.2 Evaluating hypotheses about endorsement quality

This chapter considers the implications of proposition 3, which suggest that an endorsement is stronger when the endorsing party is sensitive to punishment for making bad endorsements. We might expect the implications to flow through two different channels of government influence: electoral and special interest. As discussed in section 4.1.3, this produces two different hypotheses for testing. We will evaluate each here.

To begin, let us consider hypothesis 3, which expects endorser sensitivity to electoral punishment to moderate the quality of entry endorsement. This hypothesis is tested through examination of the influence of \textit{ENDORSER POLITY} on greenfield project announcements. Figure 4.3 illustrates the simulated QI (Y axis) in terms relative to the number of projects expected if the holder states most recent membership in a BIT had been tested through arbitration. Functionally, this corresponds to the level of reputational \textit{benefit} expected from entry endorsement. The dashed horizontal line indicates a level of greenfield project announcements that is identical to membership in a tested BIT. The X axis indicates levels of \textit{ENTRY RECENCY}. The two lines plotted in the figure indicate the relative increase in projects expected from entry endorsement at given levels of entry recency. One line corresponds to a state with low electoral vulnerability (low \textit{ENDORSER POLITY}) while the other line corresponds to a state with high electoral vulnerability (high \textit{ENDORSER POLITY}).

Figure 4.3 provides grounds to reject hypothesis 3. At no level of entry recency do the high electoral vulnerable and low electoral vulnerable endorsers produce statistically different projections for new investment. If anything, Figure 4.3 suggests that electoral vulnerability, at certain levels of \textit{ENTRY RECENCY} produces a depressive effect on the influence of entry endorsement. This is suggested at points of \textit{ENTRY RECENCY} below about .8, at which the line corresponding to high \textit{ENDORSER POLITY} sinks below the line corresponding to low \textit{ENDORSER POLITY}, though not by a statistically significant amount.
Moving on, then, consider hypothesis 4 which expects endorser sensitivity to special interest punishment (namely from economic actors concerned with cross-border investment, taken as an economic interest) to moderate the strength of entry endorsement. This hypothesis is tested through examination of the influence of ENDORSER DIRECT ASSETS on entry endorsement. Figure 4.4 illustrates this in a similar fashion to Figure 4.3. The Y axis indicates the relative benefit of entry endorsement by showing how many additional greenfield projects would be expected for entry into an untested BIT. The X axis indicates the recency of entry endorsement. There are three lines plotted in this figure. Each corresponds to a level of ENDORSER DIRECT ASSETS. Two of these lines (corresponding to medium and high levels of ENDORSER DIRECT ASSETS) indicate a positive relative benefit to entry endorsement. This indication is made where the line and corresponding 95% confidence interval rises above the horizontal reference line indicating neutral (or null) reputational benefit for entry endorsement. The line corresponding to low ENDORSER DIRECT ASSETS does not rise above zero, indicating that entry endorsement is not beneficial to a state’s reputation if the endorser is not very vulnerable to investor punishment for making bad endorsements.
Figure 4.4 suggests general support for hypothesis 4. The divergence in QI estimates across all three levels of ENDORSER DIRECT ASSETS indicate a moderating effect for ENDORSER DIRECT ASSETS on entry endorsement. The inversion of reputational benefit between medium and low ENDORSER DIRECT ASSETS suggests that not only is endorser quality moderated by the endorser’s vulnerability to investor punishment, but that entry endorsement is actually counter-productive (producing fewer new projects) if endorsers are relatively immune from investor punishment.

**Figure 4.4** Relative influence of entry endorsement over entry recency at levels of endorser asset vulnerability.
4.3.3 Evaluating hypotheses about observer betrayal

Finally we seek to test, in this chapter, the implications of proposition 4. We do this by evaluating hypothesis 5. Our expectations are that the extensiveness of betrayal moderates the influence of entry recency on reputation (indicated by the number of announced greenfield investment projects). Which is to say that, if an investor (observer) has been betrayed, entry endorsement will only benefit the host state if it has occurred very recently. Conversely, if an observer has been less extensively betrayed, then we expect entry endorsement to be relatively beneficial to the host state’s reputation, even if the entry endorsement occurred a long time ago.

To evaluate hypothesis 5, we will examine Figure 4.5. This figure shows, on the Y axis, the increase in new greenfield investment relative to the level expected if a host state’s membership in a BIT has been tested. Accordingly, as we interpret project announcements as an indication of reputation assessment, values above the horizontal reference line indicate a reputational benefit from entry endorsement while values below the dashed line indicate that entry endorsement confers a reputational curse (degradation of reputation). The X axis indicates entry recency.
Figure 4.5 Relative influence of entry endorsement over entry recency at extent of observer betrayal

Figure 4.5 plots two lines corresponding to two different levels of the extent of betrayal. Recall that we capture BETRAYAL by measuring the time since a firm in the investor state filed a BIT arbitration claim. One of the lines in the figure indicates the change in reputation from entry endorsement if the extensiveness of betrayal is low, with the last arbitration filing occurring 10 years ago. This line is nearly horizontal, running higher than 0 on the Y axis at every level of ENTRY RECENCY, but with confidence intervals eventually crossing below 0 when ENTRY RECENCY is about .25. The other line corresponds to a state that has been more extensively betrayed, with a new arbitration claim occurring within the last year. The slope of this line is notably steeper. It runs higher than 0 on the Y axis at higher levels of entry recency. As entry recency approaches or declines below .5, however, the confidence intervals for the line overlap 0 and below ENTRY RECENCY of .2, we would expect entry-endorsement to confer a reputational curse, indicated by significantly lower levels of expected greenfield projects than a member of an agreement tested through arbitration. At levels of ENTRY RECENCY below .1, we would expect entry endorsement to produce reputation degradation that is lower for the more extensively betrayed state than the less extensively betrayed state by a statistically
significant margin.

Considering Figure 4.5, we find cautious support for hypothesis 5. In a strict sense, Figure 4.5 demonstrates a moderating effect of betrayal on the reputational benefit of entry endorsement and this moderating effect suggests that the benefit of entry endorsement erodes more rapidly when investors (observers) are more recently betrayed. We are given cause to approach this support with caution, however, since Figure 4.5 shows that, when an observer has been recently betrayed, the benefits of entry endorsement can be surprisingly strong (and positive) if it occurred sufficiently recently. If ENTRY RECENCY is high enough, the reputation benefit is greater than the expected benefit when the investor is less recently betrayed.

The unexpected strength of recent entry endorsement for the extensively betrayed observer is not explicitly contrary to hypothesis 5, but it does merit some discussion. I expect that the reason for this finding is attributable to the betrayal measure. Namely, there are some states that after filing a claim, saw a decline or stagnation in the frequency of cross-border investment. Accordingly, with lower or stagnant investment, there are less opportunities for new incidents of expropriation or investor rights violation that would mandate new arbitration claims. Attending this trend would be an overall lower level of new project announcements. These two things, taken together, would exert a depressive force on the estimates for the less betrayed state in Figure 4.5, allowing estimates for the more recently betrayed state to seem significantly higher at the upper levels of ENTRY RECENCY. In reading this explanation, however, do note a few things. First, the impact of these less active investor states is likely to be low since I include controls for observer direct investment assets in my model. Second, collinearity between the recency of the last arbitration of the observer and subsequent investment activity cannot account for the moderating effect of BETRAYAL on the effect of ENTRY RECENCY (when AMBIGUOUS is 1), which is the emphasis of hypothesis 5. Lastly, the measurement of BETRAYAL is driven by practical considerations and it is not feasible to refine the measure to remove the lingering influence of the level of investment activity. Efforts to do so would only produce a linear shift in estimates of the less betrayed observer upward, thereby reducing the discrepancy and making the support for hypothesis 5 more obvious. The fact that we continue to find support for hypothesis 5 in spite of this collinearity speaks to the robustness of the moderating effect of betrayal on decay. Efforts to
address this issue, therefore, would only introduce complications to an already complex statistical model without altering our overall evaluation of hypothesis 5.
4.4 Summary

Broadly speaking, we find support for the theory presented in chapter 3. Particularly, we find support for proposition 1, proposition 3, and proposition 4. The support we find for proposition 3 only applies to the influence of endorser vulnerability to special interest pressure. We did not find that endorser electoral vulnerability (examined in terms of regime type) produced a meaningful effect on endorsement quality. The support we find for proposition 4 carries an unexpected finding, namely that more extensively betrayed observers enjoy stronger benefits from very recent entry endorsement. This does not necessarily run counter to expectations and it may be attributable to the breadth of the measure used to capture the extensiveness of betrayal. Future studies of entry endorsement and observer betrayal in IPE might seek to clarify this finding by refining or changing the instrument used to measure the extensiveness of observer betrayal.

This chapter establishes the applicability of my theory to matters of investment and investment treaties. In doing so, it suggests that my theory applies to (at least some) matters of international political economy. Combined with chapter 5, this suggests that the generalizability of my theory is not restricted by issue area. This chapter does not, in itself, provide wholesale support for my theory. This is the case particularly because it does not test the implications of proposition 2. To assess the applicability of this proposition to matters of reputation, and to further establish the general applicability of my theory, we are better served by turning our attention to the conflict/security issue area.
Chapter 5

Reputation and Alliance Formation

This chapter develops the general propositions from chapter 3 into a set of observable implications in the area of conflict. From these implications, I will derive a set of testable hypotheses that correspond to propositions 1, 2 and 4 and present empirical evidence in support of these hypotheses.

**Proposition 1** (Simplified from page 42): Under certain conditions, a holder that has entered an untested agreement will receive a reputational benefit and thus be more likely to enter into an agreement with a given observer than similar holders that have not entered such an agreement.

**Proposition 2** (Simplified from page 43): When actors are the same sort, then under certain conditions, the reputational benefits of entry endorsement will be stronger when observers and endorsers are similar and thus entry endorsement will make agreements between the holder and observer more likely when the observer is similar to the endorser.

**Proposition 4** (Simplified from page 50): Entry endorsement decays faster when an observer has been betrayed in the past and (if the observer has extensively been betrayed) entry endorsement produces an agreement between the holder and observer only if the entry endorsement occurred very recently.

The chapter is organized as follows. In the first section, I discuss the motivation for examining the conflict issue-area for evidence to test my theory. In this section, I also explain how and why focusing on alliance formation (the outcome of interest, or DV, in this chapter) is an adequate approach to testing a
theory about reputation. This first section concludes with hypotheses that test the theory from chapter 3. In the second section, I present a statistical analysis of alliance formation that supports my hypotheses. This section includes discussion of concept measurement, modeling strategy as well as model estimates and evaluation.

5.1 Where to look for evidence

This chapter tests the theory laid out in chapter 3 by examining international interactions in matters of conflict and security. There are a few different reasons that motivate testing the theory using evidence from the conflict issue area.

First, examining evidence from conflict and security helps to establish the general applicability of my theory. The extent to which my theory applies across issue areas is certainly not a given – demonstrating the applicability of my theory to economic issues does not guarantee equal applicability in the areas of conflict and security. In fact, scholars have put forth considerable effort to argue that factors enabling cooperation in the ‘low-politics’ arena of economics do not similarly apply to the ‘high-politics’ arena of conflict and security\(^1\). However, combining evidence from conflict and security along with evidence from economic interactions (chapter 4), helps to establish that my theory is not confined to a specific issue-area, but instead applies more generally to the assessment of reputation in international interactions.

Second, conflict/security issues tend to involve unlinked, similar-sort observers and therefore provide an ideal testing ground for proposition 2. As discussed more thoroughly in section 3.2.4, observers are unlinked when the utility derived from an observer’s interaction with a given holder is independent of the outcome of future interactions between the holder and other observers. In simpler terms, observers are unlinked when the early (endorsing) observer does not anticipate punishment from a betrayed observer later on. Also, as discussed in section 3.2.4, actors the same-sort, or a similar-sort when they are, essentially, the

\(^1\) See Mearsheimer, 1994.
same type of unit (e.g. persons, firms, states). The use of ‘sort’ here to avoid confusion with the terms ‘state’ or ‘type’ as used elsewhere (and for different purposes) in this document.

To see that matters of conflict and security tend to involve same-sort actors (at least as far as reputation is concerned), consider that matters of conflict and security are generally state-level matters, meaning that the ultimate decision making authority rests with leading state officials (i.e. the government). Accordingly, we can usually describe conflict/security interactions as interactions of states and other states. It follows, directly, that matters of conflict and security tend to involve actors of the same sort (states). Now, one might argue, based on Fearon\(^3\), and Putnam\(^4\) that domestic audiences matter and are observers of a different sort, but even if this is the case, domestic audiences are unlinked from entry endorsers, and lack any obvious way of directly punishing entry endorsers for making bad endorsements. Instead, domestic audiences will punish state leaders (i.e. the government) for failure/betrayal. It is a simple matter, then to build the state’s concern for domestic punishment into its utility function (Equation 3.2) allowing us to maintain focus on states with respect to the role of reputation in conflict/security.

To see that matters of conflict and security tend to involve unlinked actors, consider the previous paragraph and the following. In conflict/security settings a betrayed observer state seems manifestly more likely to punish the holder state that betrayed it instead of the earlier observer that endorsed the holder. Further, punishing a holder for betrayal is intuitively a more direct and responsive course of retribution than punishing an earlier (endorsing) observer. Consider, for example, a scenario in which France betrayed the UK in an alliance. The UK would most likely punish France directly, not, say, Poland for having allied with France at some earlier point in time. To the extent that we would expect Poland to be impacted at all, we might expect the UK to lower its view of Poland’s judgment. Generally though, and especially in consideration of the more immediate issues that might surround the formation of an alliance, we would expect Poland to treat this potential consequence (the depreciation of its status as a good judge of allies) as a trivial concern when forming new alliances.

\(^2\) The use of ‘sort’ here to avoid confusion with the terms ‘state’ or ‘type’ as used elsewhere (and for different purposes) in this document.
\(^3\) 1994.
\(^4\) 1988.
5.1.1 Observing reputation in conflict

This chapter focuses on patterns of alliance formation as the primary source of evidence to test the theory outlined in chapter 3. This may, initially seem odd – if the theory from chapter 3 yields propositions about reputation, why focus on alliance formation? The short answer is that alliance formation is a concrete, observable behavior that should reflect reputational considerations. The remainder of this section will elaborate on the difficulties of studying reputation in conflict and yield a justification for the focus on alliance formation as a suitable proxy for good reputation.

5.1.1.1 Ideal measurement of reputation

In an ideal world, the application of the theory from chapter 3 to the conflict/security issue area would be straightforward. We would simply observe and record a state’s reputation with an observer. Over time, we would record how and when the reputation changes. To test the theory we would then examine these changes in correspondence to changes in observer betrayal (proposition 4), entry into or membership in untested agreements (proposition 1), and similarity between the observer and prior, endorsing states (proposition 2).

Immediately, any effort to put this research plan into practice encounters problems. Recall from section 2.2 that reputation is a relationship between two actors, characterized largely by the observer’s assessment of the holder’s type. As a relationship, then, reputation is socially constructed and is not a natural, tangible phenomenon like trade or a war. This makes direct observation nigh impossible. At best, we can infer the existence and quality of this relationship from observable phenomena that would result from it.

In the most ideal settings, reputation is inferred from an observer’s report of it’s assessment of the holder’s type\(^5\). Unfortunately, and for obvious diplomatic reasons, states do not often publicly disseminate

\(^5\) These ideal settings do, in fact, occur somewhat frequently in certain settings. The entire business model for credit rating agencies, for example, is centered around reporting the agency’s assessment of a holder’s type.
their assessment of another state’s type. Even in cases where states do make such statements, they often do so with ulterior motives or for the purposes of soliciting public support. Consequently, it becomes difficult to differentiate the state’s genuine assessment of the other state’s type (i.e. it’s reputation) from political saber-rattling. Accordingly, public statements about a state’s type are unlikely to be a suitable way of examining reputation. Instead, to examine reputation, we will need to examine other, less proximate, phenomena that we expect to be heavily influenced by reputation.

5.1.1.2 Imperfectly observing reputation through proximate phenomena

While observer states are unlikely to (accurately) report a holder state’s reputation, we can make inferences about reputation from recorded interactions between observers and holders. The logic here is simple. We expect certain interactions to be influenced by a holder’s reputation for reliability (or, in the case of deterrence, credibility) with an observer. Thus, where we observe the interaction, we infer the holder to have a corresponding reputation for reliability.\(^6\)

Now, interpreting the outcome of scenarios/interactions as reflective of reputation sets forth a sensitive assumption and requires some commentary. A sensitive assumption is one that is necessary to maintain the validity of an argument. In this case, the assumption that I must assert, and upon which the validity of the following empirical analysis rests, is that the occurrence of certain interactions (namely alliances) between a holder state and an observer state is meaningfully reflective of the holder’s positive reputation with the observer state. The presence of this assumption imperils the meaningfulness of my empirical tests (and the support I claim for my theory) if it is also shown to be at odds with reality (i.e. if it can be shown to be a strong assumption). This would be the case if it could be shown that the interactions of interest (alliances) are wholly divorced from any systematic influence by reputation. I will return to this point further below, explaining why this is unlikely to be the case.

In the issue-area of conflict/security there are two obvious candidates for interactions that reflect

\(^6\) This logic is not novel and is implicit in any use of proxy measures/variables in empirical analysis.
reputation: threats and alliances. The early literature on reputation in conflict and security grew, largely, out of the scholarship on rational deterrence theory. This line of scholarship focused primarily on scenarios where state leaders felt compelled to follow through on costly, unpopular retaliatory threats. These scenarios posed an interesting conundrum: states would be otherwise unwilling to follow through on these threats if not for their own desire to appear willing to follow through. Empirical studies have examined the role of reputation in deterring (or encouraging) future attackers though some scholars are likely to object to the use of threats as a proxy for reputation.

Fundamentally, threats entail distinct problems that make them inadequate for this study. The use of threats to study reputation is, in itself, subject to methodological problems. As noted by Fearon, Gibler and others, data on threats and threat outcomes are subject to selection effects that will mislead our empirical analyses. The problems stem, largely, from the observation that states initiate crises selectively. They are less likely to initiate a crisis with states that will follow through on retaliatory threats. States that are more likely to follow through on retaliatory threats are thus less likely to appear in a crisis dataset and are also more likely to have a reputation for follow-through. In addition, observe that crises (and threats) are not always public and thus will not always be included in a crisis dataset. The extent to which threats are issued publicly likely corresponds to the probability of follow-through. Taken together, then, there are multiple, compounding selection effects in the data that will predispose an empirical analysis to understate the influence of reputation. These are noted in the literature on reputation and have likely motivated some scholars to move away from crises and threats when examining the role of reputation in the conflict/security issue-area.

Additionally, deterrence scenarios are inadequate for the specific goals of this research project. They are inadequate for two reasons. First, they are an imprecise fit for the theory as outlined in chapter 3. Most importantly, scenarios involving threats of retaliation yield ambiguous outcomes under conditions that

---

8 See Mercer, 1996; Press, 2005.
9 1997.
10 2008.
12 See Gibler, 2008.
are subtly different from more collaborative types of scenarios (such as the formation of an agreement for mutual gain) that are the focus of the theory presented in chapter 3. In the scenarios discussed in chapter 3, ambiguous outcomes occur when a holder’s commitment to an agreement goes untested by the occurrence of an independent, external event\textsuperscript{13}. In the context of deterrence, however, ambiguity would occur when a holder makes a retaliatory threat and the challenger (an observer) backs down. The conditions for ambiguity are thus not independent nor external to the interaction between the holder and the observer (challenger). Instead the occurrence of ambiguity is wholly endogenous to the behavior of the observer. This is not to say that the theory cannot apply to deterrence scenarios, but application is not straightforward and would require some adjustments to the underlying formal model\textsuperscript{14}. Put differently, I would still expect concepts of entry endorsement, betrayal and similarity to apply to the assessment of reputation in deterrence scenarios. But the model outlined in section 3.2 would require adjustment and, consequently, the observable implications would be different from those outlined in chapter 3.

The second reason that deterrence scenarios are unsuitable is because they represent a type of interaction that has no obvious analog in the economic issue area. That is, the scholarship on reputation and international political economy is primarily focused on scenarios in which states (and other actors), attempt to coordinate their actions for mutual gain (i.e. cooperate). Scenarios (and corresponding studies) analogous to deterrence do exist in economic settings\textsuperscript{15}, but they are either apolitical in nature, lack an international dimension, or otherwise are relegated strictly to the interests of economists. As one of the goals of this study is to establish the general applicability of the theory presented in chapter 3, it is desirable that we be able to compare, to some extent, the scenario studied in the economic issue-area to the one studied in the conflict issue area. Our ability to make this comparison is already strained, but we are better positioned to make a comparison if the scenarios studied in the economic and conflict issue areas are (at least structurally) similar.

In contrast to deterrence scenarios, alliance formation is more suitable for testing the theory from

\textsuperscript{13} Recall that this is, essentially, an event that occurs randomly, or, minimally, an event that requires action by a third party force/actor to occur. An example of this type of event might be a war or military dispute that compels a holder to come to the aid of its allies.

\textsuperscript{14} Though, importantly, it would not require adjustments to the reputation assessment mechanism outlined in equation (3.3) or equation (3.4).

\textsuperscript{15} The obvious example here would be market entry-deterrence type games, of which the chain store problem is a notable and somewhat relevant example, see Selten (1978)
5.1.2 Alliance formation and reputation

Alliance formation is an interaction that provides a suitable testing ground for examining reputation and testing the theory from chapter 3. Unlike crises data, alliance data is less prone to selection effects. Further, unlike deterrence scenarios, alliance formation (as a scenario), is essentially in-line with the broad international interaction model outlined in section 3.2, and there exist sufficiently similar\textsuperscript{16} scenarios in the economic issue-area. I will discuss each of these, briefly, in turn.

Alliance formation data seems less likely to be influenced by selection effects. Though not always the case, alliances are usually written down and available as public record\textsuperscript{17}. The broader selection effect that impacts the crisis data (a correlation between reputation and crisis onset, and thus, inclusion in the data) is also less of an issue, though this is attributable to other aspects of the research-design, namely the examination of all (politically relevant) dyad-years\textsuperscript{18}.

The formation of an alliance fits nicely into the model presented in section 3.2. Barring the nuances of alliance design and negotiation, the formation of an alliance is essentially a decision of each state to accept the other as its ally or reject it. This type of decision is present in the model from section 3.2, though it is further simplified, focusing on the decision of only one state\textsuperscript{19}. Intuitively, we would expect the observer to accept an alliance with a holder state if the holder’s reputation for reliability was high enough offset the relative benefit\textsuperscript{20} (to the observer) of avoiding betrayal and fighting alone. And it is precisely this intuition

\textsuperscript{16}The key points of similarity are that both the economic and conflict/security scenarios involve the \textit{ex ante} assessment of type, that observers and (at least some types of) holders would like to coordinate policy to achieve mutual gain

\textsuperscript{17}See Gibler, 2008.

\textsuperscript{18}This is certainly an acceptable approach when examining deterrence and crisis data, though in the context of deterrence, the selection process for crises involvement must be accounted for theoretically. Fittingly, scholars have developed the concept of ‘general deterrence’ (See Danilovic, 2001; Quackenbush, 2010), which is impacted by reputation but, as one might expect, is not easily adaptable to the model in section 3.2, nor does it find an obvious analog in the economic issue area.

\textsuperscript{19}It is possible to extend the model to account for both states symmetrically. Doing this would entail allowing each state to be both a holder and an observer. Both states would evaluate the reputation of the other and decide whether to accept or reject the other. This complicates the model considerably, and makes explanation of the model more convoluted than it already is. Its general interpretation is the same and the observable implications that result are largely consistent with those already presented in chapter 3

\textsuperscript{20}That is, relative specifically to the net benefit of fighting with the potential support of an ally (the holder).
that is outlined (albeit more abstractly) in section 3.2, specifically equation (3.2).

We likewise see congruity between alliance formation and the expanded model presented in section 3.2.1. We expect observer states to estimate the reliability of a potential ally (a holder state) by assessing the holder’s reputation for reliability\(^{21}\). This reputation is based on the outcome of the holder’s prior alliances. It should be plain that the assessment of reputation in this way corresponds directly to the role of reputation \((\pi)\) as discussed in section 3.2.1. Further, membership in untested alliances – where a holder is never called upon to defend an ally – corresponds to an ambiguous outcome. Entry into an untested alliance constitutes entry endorsement and the prior observers (endorsers) are the holder’s allies in the untested alliance.

5.1.3 Theoretical Expectations and Hypotheses

Given that alliance formation provides suitable testing grounds for the theory presented in chapter 3, and that the formation of an alliance can be represented by the model presented in section 3.2, we can form testable hypotheses that correspond to the propositions outlined in section 3.2.

5.1.3.1 Hypotheses about entry-endorsement and entry-recency

Recall from proposition 1, that we expect the entry of a holder into an agreement with an observer to improve its reputation for reliability, even if the holder’s willingness to uphold the agreement is untested (i.e. the outcome of the agreement is ambiguous). Since we assume that reputation informs a holder’s ability to enter into future alliances, we can infer a state’s reputation from it’s likelihood to enter into an alliance. Combined with proposition 1, we can expect that, all else equal, a state’s mere entry into an alliance (at some point in the past) improves its reputation and thus improves its likelihood of forming a new alliance later on. Moreover, we can expect that a state’s recent entry into an alliance, even if untested, improves its reputation more than a similar state who has not recently entered into an alliance.

\(^{21}\) See Crescenzi et al., 2012; Gibler, 2008; Miller, 2003.
**H 6:** States that have recently entered into untested alliances are more likely to enter into new alliances than states that have not recently entered into untested alliances.

Note that I do not expect to find unqualified support for this hypothesis since reputations eventually erode to a point of uncertainty. Instead, I anticipate finding support for this hypothesis only beyond some critical threshold at which the holder’s last tested alliance occurred sufficiently long ago and its most recent entry-endorsement occurred (entry into an untested alliance) occurred sufficiently recently.

Additionally, recall from section 3.2.1.1 that we expect the influence of observable behavior of a holder (i.e. demonstrations of compliance, entry into new agreements, etc) to change over time. Specifically, we expect that more recent entry endorsement will be more influential on reputation than less recent entry. Accordingly:

**H 7:** States that have more recently entered into an untested alliance will be more likely to form a new alliance than states that have less recently entered into an untested alliance.

### 5.1.3.2 Hypotheses about betrayal

As discussed in section 3.2.1.1, we expect the holder’s type to be dynamic, changing over time. The changes in a holder’s type over time are modeled as the outcome of a Markov process (see Equation 3.3). We expect the parameters of this function to be a product of the observer’s past experiences. When the observer is betrayed, it expects more frequent type-changes. This amounts to an increased rate of decay. In terms of alliance formation, we expect that a state that has been betrayed in its past alliances will discount entry-endorsement (i.e. the entry of a holder into an untested alliance) severely as the entry-endorsement becomes less and less recent. At the same time, we expect a state that has been betrayed less frequently in past alliances will continue to value entry endorsement for longer periods of time. Put more directly, we expect betrayed observers (when evaluating the reputation of a holder) only to consider the entry-endorsement of a

---

22 Note that recency implies recency relative to the last observed unambiguous alliance outcome (i.e. membership in a tested alliance). This is further discussed below.
holder if it is very recent, while we expect unbetrayed observers to value entry-endorsement even if it is less recent. This yields the following expectations about alliance formation:

**H 8:** States (observers) who have been in many alliances that ended poorly are less likely to form an alliance with a holder state that has entered an untested alliance unless the holder state has entered that alliance very recently.

Alternatively: Entry endorsement’s effect on new alliance formation is moderated by the extent to which an observer state has been betrayed in previous alliances.

### 5.1.3.3 Hypotheses about endorsement quality

Lastly, we derive a hypothesis about the similarity of observers and entry-endorsers. Recall that section 3.2.4 provides us with specific expectations about how unlinked actors estimate the reputation of a holder state on the basis of entry-endorsement. The expectation is that an observer state will be more certain about the value of an entry-endorsement when it is more similar to the entry endorsers (i.e. prior observers). As applies to alliance formation, ‘similarity’ applies most intuitively with respect to strategic interest (particularly with regard to alliances).

**H 9:** Holder states who have entered into an untested alliance are more likely to form new alliances with states that had similar strategic interests as the holder’s former alliance partners (entry endorsers).

Alternatively: the influence of entry endorsement on alliance formation is moderated by the similarity in strategic interests between the potential ally state (the observer) and the endorsers. Greater similarity between the observer and the endorsers makes the influence of entry endorsement on alliance formation stronger.
5.2 Examining Alliance Formation: Design and Methods

An analysis of alliance formation that tests the hypotheses outlined above is not entirely straightforward. It requires careful attention to nuances of research design and measurement, the finer points of which I outline here.

5.2.1 Unit of Analysis

The unit of analysis in this analysis will be the directed-dyad-year. Immediately, readers may object to this unit of analysis. Recent trends in the study of IOs have emphasized that a k-adic unit of analysis is more appropriate when analyzing multilateral events.

K-adic analysis serves to accommodate theoretical considerations that do not apply to this study. From the perspective of theory, there are important features of an interaction that are omitted when considering dyadic interactions in place of genuinely k-adic interactions. Poast illustrates this well when he observes that Belgium and Turkey would probably not be allies if not for their desire to be allies with the US through NATO. However, the assessment of reputation, as presented in chapter 3 is inherently dyadic and to the extent that the actions of other parties are relevant (e.g. entry endorsers, betrayal by former allies), they are represented in the analysis as measures attached to states in the dyad (see further discussion of this below).

More generally, k-adic analysis attempts to prevent over-representation of a single event. Using a directed-dyad-year unit of analysis, a single multilateral alliance formation (with $k$ members) will appear in the data multiple times ($k(k - 1)$ times). Thus, using a dyadic unit of analysis, alliance formation will be over-represented in the data.

The over-representation of alliance formation is not a problem for this analysis. This is the case for two reasons. First, the over-representation of alliance formation reflects the holder’s favorable reputation

---

23 See Poast (2010) for a discussion of the merits of this approach. See Poast (2012) and Chiba, Johnson, and Leeds (Forthcoming) for recent examples of this approach in use.

with each of the members of the alliance. Simply put, when a holder state attempts entry into an alliance, it must have a sufficiently favorable reputation with most (usually all) members of that alliance. Representing a multi-lateral alliance formation multiple times thereby corresponds (and represents) the notion that the holder state has a favorable reputation with each of the members of the alliance.

Second, the failure of multi-lateral alliance formation (or entry) biases estimates against supportive findings. While entry into a multi-lateral alliance can be taken as an indication of the holder’s good reputation with (essentially) all alliance members, rejection from a multi-lateral alliance indicates the holder’s disfavorable reputation with between one and all of the alliance members. A holder state that is rejected from a multi-lateral alliance will not register as forming an alliance with any members of the multi-lateral alliance. Thus, even if the holder had favorable reputations with all but one of the members of the alliance, those favorable reputations will not be indicated in the data. This implies that, using directed-dyad-years, there will be some amount of measurement bias (i.e. systematic measurement error) in the use of alliance formation as a proxy for reputation. Good reputations will be systematically under-represented. This, if anything, biases our estimates downward and reduces the likelihood of statistically significant coefficient estimates that support the hypotheses described above.

### 5.2.2 Population of Cases

The ambition of this analysis is to present findings that generalize to the relationship among potential allies at a given point in time. Strictly speaking, the population of cases considered in this analysis is the set of politically relevant dyad-years from 1815 to 2008. The dyads are generated by the program **Eugene**, and follow rules set out by Maoz and Russett. The decision to restrict the analysis to this specific population

---

25 For a more detailed explanation, consider that the formation of a multilateral alliance will require all members to accept one another as allies and that addition of a member often requires the unanimous acceptance by all current members. Viewed this way, the entry of a holder state into a multi-lateral alliance is essentially equivalent to the interaction described in section 3.2 (and illustrated in Figure 3.2) occurring in parallel between the holder and each of the observer states (i.e. other members of the alliance). The formation of an alliance suggests that, in each of these interactions, the observer state chose to accept, and thus found the holder to have a sufficiently reliable reputation. It is thus appropriate to represent the formation or entry into an alliance multiple times since it is indicative of the number of times that the holder’s reputation was assessed and found to be sufficient.

26 See Bennett and Stam, 2000.

27 1993. The level of contiguity is set to its most inclusive – states may be separated by 400 miles of water.
of cases is motivated by theoretical and methodological concerns. I will discuss each of these briefly, in turn.

There are many pairs of states that simply are not theoretically relevant. The interaction presented in chapter 3 presumes that actors have some potential interest in forming an agreement. For our purposes here, then, we would like to focus on cases in which potential alliance formation is at least minimally likely. Though not a perfect solution, we can move towards this goal by restricting the population of cases to pairs of states that have at least an opportunity to form an alliance.

Methodologically, restricting our population to the set of politically relevant dyads limits problems of zero inflation and helps to accommodate the rarity of alliance formation, thereby avoiding type-2 error. The inclusion of pairs of states that would not have an opportunity to consider an alliance with one another represents a source of zero-inflation in the data. That is, it represents a process by which alliances fail to form (i.e. lack of opportunity) that is inherently different from the the substantively relevant process by which alliances fail to form (i.e. failure of alliance negotiations). This makes alliance formation increasingly rare in the data and reduces the causal leverage afforded by the analysis. Restriction of the population of cases to only politically relevant dyads limits the extent of the zero-inflating process (though it does not eliminate it).

5.2.3 Measurement: Alliance Formation

The phenomenon of interest in this study is alliance formation. I use entry into an alliance as my measure of alliance formation. The data for this measure comes from the ATOP project.

The alliance formation measure does not capture entry into non-aggression pacts. The rationale behind this omission is that non-aggression pacts are fundamentally different from security agreements, I would expect observers to attach systematically different utilities to the benefits of such an agreement. Observers

\[29\] See Most and Starr, 1989.

\[29\] More formally, and in the context of the game represented in section 3.2, we expect that observers would anticipate some non-zero utility from an alliance with a given holder. Where this is not the case, the reputation threshold \((\pi^*)\) is non-sensical (either undefined or negative).

\[30\] See King and Zeng, 2001.

\[31\] See Leeds, 2005.
would set systematically different reputation thresholds \( (\pi^*_T) \) for entry into a non-aggression agreement. Accordingly, entry into a non-aggression pact would represent a level of holder reputation that is systematically different from the level of reputation indicated by entry into a security alliance. Generally, we would expect entry into security agreements to carry high reputation requirements and for entry into non-aggression pacts to carry lower reputation requirements. Focusing on entry into security agreements thereby represents a \textbf{harder} test for my theory, making it \textbf{more} difficult, if anything to find statistically significant support for the hypotheses outlined above.

\[ 5.2.4 \quad \text{Measurement: Untested Alliance Membership (Entry Endorsement)} \]

Measures of entry endorsement\textsuperscript{32} are centered primarily around recency of alliance entry. As discussed in section 3.2.3, and in discussion of hypothesis 7 and hypothesis 6, the observable effect of entry-endorsement should be qualified, hinging critically on the recency of a holder’s last entry into an untested alliance (i.e. the recency of the holder’s last entry-endorsement). But the discussion that follows proposition 1 indicates that recency is important here in two ways: (1) with respect to the time between the last test of a holder’s reliability and the time of last alliance entry and (2) with respect to the time between the last test and the year in question.

To represent both features of recency, I construct a measure called \textbf{ENTRY RECENCY}. This measure is constructed as:

\[
\text{ENTRY RECENCY} = \frac{\text{time of last entry} - \text{time of last test}}{\text{observation year} - \text{time of last test}}
\]

Examples of this measure are illustrated in Figure 5.1. This figure groups four timelines into two groups, according to entry recency. As should be clear from this figure, the level of recency is determined by the distance between the last tested agreement and the entry into a new security agreement relative to the

\textsuperscript{32} If not clear at this point, since entry endorsement is the mechanism by which membership in untested alliances influences holder’s reputation for reliability with an observer, the terms ‘entry endorsement’ and ‘entry into untested alliance’ are functionally equivalent, and I use them here and throughout somewhat interchangeably.
total temporal distance between the last tested agreement and the present.

Where the term ‘test’ should be taken here and throughout to mean an opportunity for a state to
demonstrate its reliability (i.e. when one of the state’s allies encounters conflict). And where ‘time of last
entry’ is the time of the state’s most recent entry into an alliance.

Readers should note, given the construction of this measure, that **ENTRY RECENCY** can take on positive
and negative values. Negative values of **ENTRY RECENCY** indicate that a state’s most recent alliance outcomes
are unambiguous – that the state has had an opportunity to demonstrate its reliability since it’s most recent
entry into an alliance. Positive values of **ENTRY RECENCY** indicate a holder’s most recent alliance outcome
as ambiguous – that entry endorsement is the most recent indication of a holder’s type. Essentially, positive
values of entry endorsement answer the question ‘how much more recent is the state’s last entry compared
to its last test?’ Where increasing positive values indicate comparatively increasing recency.

Since **ENTRY RECENCY** takes on two distinct meanings, depending on if it is positive or negative, I also
employ a dichotomous measure called **AMBIGUOUS**. This measure is scored ‘1’ if **ENTRY RECENCY** is greater
than zero, and 0 if **ENTRY RECENCY** is less than, or equal to zero. Effectively, this measure captures whether
or not the most recently observable behavior of a state was a test (unambiguous) – an opportunity for the
holder to demonstrate its reliability – or an entry endorsement (ambiguous) – an entry into an alliance with
no intervening test of the holder’s commitment before the year in question. Thus, by interacting these two
measures (**ENTRY RECENCY** and **AMBIGUOUS**), I can distinguish the effect of entry endorsement on alliance
formation when entry endorsement is the most recently observable behavior of a holder state.\(^{33}\)

The process for calculating **ENTRY RECENCY** (and consequently, **AMBIGUOUS**) is somewhat complicated.
It is essentially two separate measures, one indicating the time of a state’s most recent entry into an alliance
and the other indicating the last time a state’s ally encountered conflict (the last test). These two measures
are determined primarily through a series of cross-references between the ATOP and COW\(^{34}\) datasets.

\(^{33}\) Alternatively stated: when the holder has experienced ambiguous alliance outcomes
\(^{34}\) See Sarkees and Wayman, 2010.
5.2.5 Measurement: Observer-endorser similarity

To test hypothesis 9, we require a measure of the similarity in strategic interests that informs us of the similarity between an observer and the holder’s recent entry-endorser (i.e. states with which the holder has most recently formed and alliance). Signorino and Ritter\(^3\) already provide a suitable basis for assessing this through their ‘S’ measure of state similarity. I use this measure, as applied to alliance portfolios and weighted by national capability scores\(^3\) to generate \textit{ENDORSER} S.

The construction of \textit{ENDORSER} S is, like entry-recency, the product of a series of cross-references between ATOP and other (intermediately constructed) datasets. For any given dyad-year, I collect the allies of the holder state in its most recent alliance (these are the entry endorsers). The \textit{ENDORSER} S is constructed as an average of the ‘S’ score between the observer and the the endorsers. Though, by this point obvious, it is thus crucial to note: \textit{ENDORSER} S is not an indication of the similarity between the members of the dyad, but is instead similarity between the observer and the holder’s allies.

5.2.6 Measurement: Betrayal

To measure the extent to which an observer has been betrayed (which is crucial for testing hypothesis 8), I construct a measurement \textit{BETRAYAL}. This measure comes from an examination of the state’s membership in alliances prior to the year of a given observation. In addition to agreements that compel positive action, I also allow neutrality and non-aggression agreements to be considered in the calculation of the betrayal measurement. Fundamentally, \textit{BETRAYAL} is the number of times an observer has potentially been betrayed through membership in a failed alliance\(^3\), expressed as a proportion of the state’s total number of alliances since 1815.

---

\(^3\) Signorino and Ritter 1999.
\(^3\) Specifically, the calculated values for ‘S’ come from \textit{EUGene} (See Bennett and Stam, 2000). These scores, notably, do not include UN voting similarity. This is acceptable since the primary use of ‘S’ here is to indicate similarity in strategic interests qua alliance policy. Further analysis may include other dimensions to ‘S’, but I do not expect them to substantively alter the findings presented below.
\(^3\) by ‘failed alliance’, I generally mean an alliance that ended due to violation of the terms of the alliance, or other, similar conditions that a state would consider abysmal.
The measure BETRAYAL is constructed by reference primarily to the ATOP member\textsuperscript{38} dataset. This dataset contains a measure, TERMCAUS that categorizes the reasons a state has left an alliance. These are ordered such that higher numerical categories correspond to increasingly hostile or acrimonious reasons for alliance termination. Levels of TERMCAUS above three indicate reasons for alliance termination that include termination due to policy disputes, war among members of the allies and external wars that violate the provisions of the agreement. The observer’s BETRAYAL measure increases with each membership to an alliance that, during the period of the observer’s membership, ended with TERMCAUS greater than three.

5.2.7 Modeling Strategy

Since the dependent variable in this study is dichotomous, CLRM type models are inappropriate. Instead, to test the hypotheses listed above, I model alliance formation as a GLM, with interactions among the measures listed above, and structured as follows:

\[
\text{Alliance Formation}_{ijt} \sim B\left(\frac{1}{1 - e^{-XB + ZC + \epsilon}}\right)
\]

(5.1)

Where \(X\) is a vector of the key independent variables (and their interactions), and \(B\) is a vector of corresponding coefficients. And where \(Z\) is a vector of relevant control variables and \(C\) is a vector of corresponding coefficients.

The vector \(X\) (and corresponding \(B\)) is specified in two different ways, producing two slightly different models (both structured as in Equation (5.1)). The first model focuses on interactions between BETRAYAL, ENTRY RECENCY, and AMBIGUOUS. It includes ENDORSER \(S\) as a standalone, independent variable that is not interacted with any other variables in the model.

\textsuperscript{38} See Leeds, 2005.
Limited ENDORSER S interaction (Model 1)

\[
X = \begin{bmatrix}
\text{AMBIGUOUS}_{it} \\
\text{ENTRY RECENTY}_{it} \\
\text{AMBIGUOUS}_{it} \times \text{ENTRY RECENTY}_{it} \\
\text{ENDORSER} S_{ijt} \\
\text{ENDORSER} S_{ijt} \times \text{AMBIGUOUS}_{it} \\
\text{BETRAYAL}_{jt} \\
\text{BETRAYAL}_{jt} \times \text{AMBIGUOUS}_{it} \\
\text{BETRAYAL}_{jt} \times \text{ENTRY RECENTY}_{it} \\
\text{BETRAYAL}_{jt} \times \text{AMBIGUOUS}_{it} \times \text{ENTRY RECENTY}_{it}
\end{bmatrix}
\]

The rationale behind the Model 1 specification is that we do not necessarily expect the influence of ENDORSER S to change as a function of the the recency of entry endorsement, but only as a function of whether the holder’s most recently entered alliance was untested (i.e. yielded an ambiguous outcome).

It is, however, not unreasonable to suspect that ENDORSER S changes with recency as well. To accommodate this, avoid misspecification\(^{39}\), and for the sake of robustness, I provide a second model, with \(X\) specified:

\(^{39}\) This is particularly called for in McCullagh and Nelder (1989, chapter 4)
This model is noticeably more complex. Ultimately, both models yield considerably similar results.

### 5.2.8 Expectations and Conditions for Failure

The hypotheses outlined in section 5.1.3 lead to specific expectations for the model estimates. As with any statistical analysis, our goal here is to infer support for our hypotheses through a lack of evidence to the contrary. Accordingly, this section establishes the ways that model estimates might fail to support the hypotheses listed above.

Hypothesis 6 pertains to the reputational benefits of entry endorsement when alliance outcomes are ambiguous. Model estimates fail to show support for hypothesis 6 if the total effect of ambiguity on the probability of alliance formation is negative at every level of entry recency. Such a finding would indicate that entry endorsement does not have a positive effect on the likelihood of alliance formation. By extension, it would suggest that that mere entry into an alliance does not benefit reputation, that proposition 1 does
not apply to matters of alliance formation, and that the theory presented in chapter 3 does not describe the role of reputation in alliance formation or broader issues of conflict and security.

Hypothesis 7 pertains more specifically to the relationship between entry endorsement and recency. Model estimates fail to show support for hypothesis 7 if (the total effect of) more recent instances of entry endorsement (when outcomes are ambiguous) impose a flat (i.e. 0, or null) or negative effect on the likelihood of alliance formation. If this were the case, it would suggest that entry-endorsement does not influence reputation as anticipated and/or that the influence of entry-endorsement does not occur through reputation. This would suggest that neither proposition 1 nor the reputational mechanics as outlined in Figure 2.3 apply to alliances as expected. As alliance formation is expected, more than other matters in conflict/security to be influenced by reputation, a finding of the sort described here would suggest an overall lack of applicability of the theory to any phenomena in conflict/security.

Hypothesis 8 is about the influence of betrayal on reputation assessment. For this hypothesis, the simplest condition for failure occurs if model estimates suggest that increases in the frequency of observer betrayal yield to identical or greater likelihoods of alliance formation, regardless of levels entry recency or ambiguity. Further, this hypothesis fails to find support if a specific relationship between ENDORSER S and BETRAYAL is not observed when AMBIGUOUS =1. Put intuitively: increases in BETRAYAL must indicate increases in the rate of decay of the effect of entry. This should manifest as an increase in the threshold at which increases in entry recency produce statistically and substantively meaningful increases on alliance formation, and in increase in the steepness of the response function of entry recency.

Lastly, hypothesis 9 pertains to the relationship between entry-endorsement and the similarity between endorsers and observers. Model 1 estimates fail to show support for hypothesis 9 if increases in ENDORSER S produce a null or negative effect on alliance formation when AMBIGUOUS =1. In Model 2, estimates fail to show support for hypothesis 9 under generally the same conditions. Additionally, we might also expect to see the effect of ENDORSER S on alliance formation increase with increases in ENTRY RECENCY and when  

---

40 As a technical matter, the expectation here is that increases in the second-order partial derivative of the model with respect to entry recency when ambiguity is 1
Note, however, that failure to find evidence for this trend is not, in itself, reason to reject hypothesis 9.

### 5.3 Findings

Model estimates are provided in Table 5.1 and Table 5.2. Because of the number of terms in the model, only relevant independent variable estimates are shown. It is important to note that, as discussed in section 4.3, tabularized model estimates are uninterpretable in any meaningful way. This is due to the number of interaction terms in the model and the fact that the marginal effect of any given term involved in an interaction is conditioned on the **entirety** of the covariate profile. Since each of the hypotheses described above is tied to an interaction, none of them can be evaluated from tabular results alone. These tables are presented entirely to illustrate the breadth of models estimated and the corresponding consistency of model estimates.

Estimates were generated using three different methods for each model. A standard set of maximum-likelihood estimates were generated using the entire population of politically relevant states from 1815 to 2008. These are represented in the first column of each table. To account for the fact that alliance formation is a rare event (occurring in about 1% of observations in the dataset used for this analysis), two alternative methods were used to generate estimates. The second column of each table reports estimates generated and corrected using rare events logistic regression as prescribed in King and Zeng. In both models, these estimates are, generally, slightly lower than the standard, ML estimates, but are otherwise similar.

The third column of each table report estimates from a choice-based sample. A choice-based sample is one in which sampling is stratified on levels of the dependent variable. All observations of alliance formation are selected along with a random sample of observations that did not form alliances. The size of the non-alliance observations was five times the size of the sample of alliance-forming observations. For both

---

41 Estimates generated using Zelig (2009)
43 See ibid., pp. 698-700.
models, these estimates are generally the lowest for each coefficient, however their sign and significance remain consistent with the other two sets of estimates. It is, perhaps, notable that these estimates also produce the lowest percentage of correctly classified cases (PPC)\textsuperscript{44}.

\footnote{It is also notable that these estimates also produce the largest reduction in error rate. This is, likely attributable to the reduction in relative rarity of alliance formation in the model.}
Figure 5.1: Conceptual examples of low and high levels of entry recency

Low Entry Recency
(Weak Influence of Entry Endorsement):

Security Agreement Entry
(Entry Endorsement)

Distant Past
Security Agreement

Last Tested
Security Agreement

Present

High Entry Recency
(Strong Influence of Entry Endorsement):

Security Agreement Entry
(Entry Endorsement)

Distant Past
Security Agreement

Last Tested
Security Agreement

Present
<table>
<thead>
<tr>
<th></th>
<th>(Full Sample)</th>
<th>(Full Sample)</th>
<th>(Selective Sample)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>logistic</td>
<td>rare events</td>
<td>logistic</td>
</tr>
<tr>
<td>ENTRY RECENCY</td>
<td>0.084*</td>
<td>0.079*</td>
<td>0.075*</td>
</tr>
<tr>
<td></td>
<td>(0.031)</td>
<td>(0.031)</td>
<td>(0.031)</td>
</tr>
<tr>
<td>AMBIGUOUS</td>
<td>−4.366*</td>
<td>−4.348*</td>
<td>−2.974*</td>
</tr>
<tr>
<td></td>
<td>(0.301)</td>
<td>(0.301)</td>
<td>(0.272)</td>
</tr>
<tr>
<td>AMBIGUOUS:ENTRY RECENCY</td>
<td>5.971*</td>
<td>5.966*</td>
<td>4.574*</td>
</tr>
<tr>
<td></td>
<td>(0.270)</td>
<td>(0.270)</td>
<td>(0.236)</td>
</tr>
<tr>
<td>ENDORSER S</td>
<td>0.071</td>
<td>0.072</td>
<td>0.101</td>
</tr>
<tr>
<td></td>
<td>(0.150)</td>
<td>(0.150)</td>
<td>(0.161)</td>
</tr>
<tr>
<td>ENDORSER S:AMBIGUOUS</td>
<td>1.612*</td>
<td>1.610*</td>
<td>1.477*</td>
</tr>
<tr>
<td></td>
<td>(0.166)</td>
<td>(0.166)</td>
<td>(0.184)</td>
</tr>
<tr>
<td>BETRAYAL</td>
<td>−0.727</td>
<td>−0.656</td>
<td>−0.689</td>
</tr>
<tr>
<td></td>
<td>(0.622)</td>
<td>(0.622)</td>
<td>(0.638)</td>
</tr>
<tr>
<td>BETRAYAL:ENTRY RECENCY</td>
<td>0.255</td>
<td>0.249</td>
<td>0.318</td>
</tr>
<tr>
<td></td>
<td>(0.208)</td>
<td>(0.208)</td>
<td>(0.218)</td>
</tr>
<tr>
<td>BETRAYAL:AMBIGUOUS</td>
<td>−3.168*</td>
<td>−3.118</td>
<td>−3.663*</td>
</tr>
<tr>
<td></td>
<td>(1.614)</td>
<td>(1.614)</td>
<td>(1.527)</td>
</tr>
<tr>
<td>BETRAYAL:AMBIGUOUS:ENTRY RECENCY</td>
<td>3.569*</td>
<td>3.458*</td>
<td>3.986*</td>
</tr>
<tr>
<td></td>
<td>(1.558)</td>
<td>(1.558)</td>
<td>(1.489)</td>
</tr>
<tr>
<td>Observations</td>
<td>127,798</td>
<td>127,798</td>
<td>9,569</td>
</tr>
<tr>
<td>PPC</td>
<td>0.985</td>
<td>0.985</td>
<td>0.867</td>
</tr>
</tbody>
</table>

*p<0.05

Note: Additional coefficients are suppressed here to conserve space. Model estimates included controls for the following: economic wealth (GDP), time dependence (splines), major power status, contiguity, security level (De Mesquita 1983; 1985), material capabilities (CINC), regime type (Polity), and measures of reputation based on prior (observable) behavior from Crescenzi et al. (2012).
Table 5.2: Alliance Formation Estimates. Full ENDORSER S Interaction (Model 2)

<table>
<thead>
<tr>
<th></th>
<th>(Full Sample)</th>
<th>(Full Sample)</th>
<th>(Selective Sample)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>logistic</td>
<td>rare events</td>
<td>logistic</td>
</tr>
<tr>
<td>ENTRY RECENCY</td>
<td>0.249*</td>
<td>0.243*</td>
<td>0.255*</td>
</tr>
<tr>
<td></td>
<td>(0.064)</td>
<td>(0.064)</td>
<td>(0.070)</td>
</tr>
<tr>
<td>AMBIGUOUS</td>
<td>-2.979*</td>
<td>-2.936*</td>
<td>-2.640*</td>
</tr>
<tr>
<td></td>
<td>(0.392)</td>
<td>(0.392)</td>
<td>(0.397)</td>
</tr>
<tr>
<td>AMBIGUOUS:ENTRY RECENCY</td>
<td>4.031*</td>
<td>3.997*</td>
<td>3.698*</td>
</tr>
<tr>
<td></td>
<td>(0.384)</td>
<td>(0.384)</td>
<td>(0.391)</td>
</tr>
<tr>
<td>BETRAYAL</td>
<td>-1.072</td>
<td>-1.032</td>
<td>-1.063</td>
</tr>
<tr>
<td></td>
<td>(0.640)</td>
<td>(0.640)</td>
<td>(0.657)</td>
</tr>
<tr>
<td>BETRAYAL:AMBIGUOUS</td>
<td>-3.862*</td>
<td>-3.800*</td>
<td>-3.756*</td>
</tr>
<tr>
<td></td>
<td>(1.676)</td>
<td>(1.676)</td>
<td>(1.575)</td>
</tr>
<tr>
<td>BETRAYAL:ENTRY RECENCY</td>
<td>0.051</td>
<td>0.024</td>
<td>0.084</td>
</tr>
<tr>
<td></td>
<td>(0.199)</td>
<td>(0.199)</td>
<td>(0.209)</td>
</tr>
<tr>
<td>BETRAYAL:AMBIGUOUS:ENTRY RECENCY</td>
<td>4.860*</td>
<td>4.789*</td>
<td>4.734*</td>
</tr>
<tr>
<td></td>
<td>(1.615)</td>
<td>(1.615)</td>
<td>(1.534)</td>
</tr>
<tr>
<td>ENDORSER S</td>
<td>-0.309</td>
<td>-0.314</td>
<td>-0.301</td>
</tr>
<tr>
<td></td>
<td>(0.177)</td>
<td>(0.177)</td>
<td>(0.192)</td>
</tr>
<tr>
<td>ENDORSER S:AMBIGUOUS</td>
<td>-0.534</td>
<td>-0.561</td>
<td>0.930</td>
</tr>
<tr>
<td></td>
<td>(0.486)</td>
<td>(0.486)</td>
<td>(0.483)</td>
</tr>
<tr>
<td>ENDORSER S:ENTRY RECENCY</td>
<td>-0.206*</td>
<td>-0.204*</td>
<td>-0.215*</td>
</tr>
<tr>
<td></td>
<td>(0.065)</td>
<td>(0.065)</td>
<td>(0.070)</td>
</tr>
<tr>
<td>ENDORSER S:AMBIGUOUS:ENTRY RECENCY</td>
<td>2.897*</td>
<td>2.927*</td>
<td>1.255*</td>
</tr>
<tr>
<td></td>
<td>(0.487)</td>
<td>(0.487)</td>
<td>(0.486)</td>
</tr>
</tbody>
</table>

Observations 127,798 127,798 9,569
PPC 0.985 0.985 0.867

*p<0.05

Note: Additional coefficients are suppressed here to conserve space. Model estimates included controls for the following: economic wealth (GDP), time dependence (splines), major power status, contiguity, security level (De Mesquita 1983:1985), material capabilities (CINC), regime type (Polity), and measures of reputation based on prior (observable) behavior from Crescenzi et al. (2012).
5.3.1 Hypothesis Evaluation

I will evaluate the hypotheses presented above graphically. This approach is ideal since evaluation of the hypotheses outlined above requires interpretation of complicated interactions in the model. Moreover, it is difficult to present marginal effects (or even predicted values) in a non-graphical way that is both intuitive to the reader and adequate for evaluating the hypotheses listed above. This difficulty arises largely through the fact that Model 1 and Model 2 are non-linear models. As noted by Ai and Norton\textsuperscript{45}, when involved in an in interaction, the marginal effect of a variable in a non-linear model is not calculated the same way that it is in a linear model\textsuperscript{46}. Importantly, the size, direction and significance of a variable’s marginal effect are contingent on the values of the component terms of the interaction along with the values of every other variable in the model\textsuperscript{47}.

The figures here are generated using simulated quantities of interest. Ordinarily, simulations are used to generate predictions over a range of values of one or two covariates while the remainder of covariates are held constant (usually at their mean). This approach cannot be taken here for reasons just discussed. That is, since the effect of a variable on alliance formation is contingent on the values of the other variables in the model, we cannot simply hold control variables constant across simulations. Instead, I employ a slightly different approach as discussed in Hamer and Ozan Kalkan\textsuperscript{48}. Essentially, I generate a set of simulated coefficients and use them to predict the probability of alliance formation for every observation in the data, holding the variables of interest constant. I then record the average of these predicted values, increment the variable(s) of interest and repeat the process. Completing this for every value in a specified range of the independent variable(s) of interest constitutes one set of simulations.

To be precise, the simulated quantity of interest is the average predicted probability of alliance formation. I will refer to this quantity throughout this section as, simply the QI, or, as it’s interpretation is

\textsuperscript{45} 2003.

\textsuperscript{46} This is commonly computed as a partial derivative. In a model $y = \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$, the marginal effect of $x_1$ is: $\frac{\partial(y)}{\partial(x_1)} = \beta_1 + \beta_{12} x_2$. In a non-linear model, however, the marginal effect is calculated differently, essentially as a result of the chain-rule. Consider a non-linear model (with a mean function $g()$): $y = g(\beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2)$. The marginal effect is $\frac{\partial(y)}{\partial(x_1)} = g'(\beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2) (\beta_1 + \beta_{12} x_2)$.

\textsuperscript{47} Note also that this prohibits effective evaluation of the hypotheses from reference to Table 5.1 and Table 5.2 alone.

\textsuperscript{48} 2013.
functionally similar, the predicted probability of alliance formation. The figures generated here are each the result of 1000 sets of simulated QI. Lines indicate the average value of the QI across simulations and for a variable value corresponding to a point on the X axis. The ribbons indicate the interval between the .975 percentile estimate and the .025 percentile estimate at each value of the x axis. This is to be taken as an approximation for a 95% confidence interval.

To evaluate hypothesis 6 and hypothesis 7, consider Figure 5.2 and Figure 5.3. In both of these figures, the x axis represents a levels of ENTRY RECENCY while the y axis represents the probability of alliance formation. Both figures contain two lines apiece. One line indicates simulated QI when AMBIGUOUS is held constant at 0, the other corresponds to simulated QI when AMBIGUOUS is held constant at 1.

Keen readers will note that for each line, certain values of ENTRY RECENCY seem troubling or contradictory. For example, ENTRY RECENCY >0 do not make much sense when AMBIGUOUS is held constant at 0, since AMBIGUOUS is determined by ENTRY RECENCY. Considering this, we might expect the confidence intervals for values of AMBIGUOUS =0 to extend towards infinity as move from left to right along the AMBIGUOUS =0 line, since there are no observations in the data for which both AMBIGUOUS =0 and ENTRY RECENCY >0. Instead, we get steady confidence intervals throughout. This is a function of the way the QI are simulated. That is, since the QI are generated from the entire dataset, other covariates contribute to the generated QI, even though the values of these independent variables (i.e. AMBIGUOUS =0 and ENTRY RECENCY >0) do not occur simultaneously in the data. This allows us to generate an average QI with a sensible confidence interval. In essence, this allows us to generate QI for impossible combinations of covariates, or, put more directly, it allows us to approximate counter-factual conditions.

As anticipated, predicted alliance formation is greater when AMBIGUOUS is held constant at 1, but only when ENTRY RECENCY is positive, and high (between about .7 and 1). This suggests that states whose most recent membership in an alliance was untested (and hence ambiguous) benefit simply from entering the alliance (if the entry occurred recently). Recall that ENTRY RECENCY is relative to the time of last entry and the last time the holder was in a tested alliance. Accordingly, levels of ENTRY RECENCY do not correspond
to fixed or absolute periods of time. Generally, however, an ENTRY RECENCY of .7 corresponds to an alliance entry occurring between 4 and 6 years prior to the year of a given observation.

Both Figure 5.2 and Figure 5.3 demonstrate qualified support for hypothesis 6. As discussed in section 5.1.3, estimates only fail to support hypothesis 6 if ambiguity produces a negative effect across all possible covariate profiles. These figures, however, illustrate support for hypothesis 6 at every point for which the lines labeled AMBIGUOUS =1 are higher than the lines labeled AMBIGUOUS =0. Thus, it would appear that, in at least some instances, mere entry into an (untested) alliance makes future alliance entry more likely (suggesting a benefit to reputation).

Additionally, Figure 5.2 and Figure 5.3 demonstrate greater support for hypothesis 7. The benefits of membership in an untested alliance on future alliance formation are strongest immediately after alliance entry and fade quickly thereafter. This is indicated by the steep decline in predicted alliance formation as one moves from ENTRY RECENCY =1 to ENTRY RECENCY =.7, along the AMBIGUOUS =1 line. At levels below ENTRY RECENCY =.7, a holder state is more likely to form new alliances if it has had an ally that has experienced conflict (and thus had an opportunity to demonstrate its reliability – indicated by the AMBIGUOUS =0 line).
Figure 5.2: Alliance Formation on Entry Recency at levels of Ambiguity (Model 1)
Figure 5.3: Alliance Formation on Entry Recency at levels of Ambiguity (Model 2)
To evaluate hypothesis 8, consider Figure 5.4 and Figure 5.5. Like with Figure 5.2 and Figure 5.3, these figures are essentially the same, except Figure 5.4 was generated from Model 1 estimates while Figure 5.5 was generated using Model 2 estimates. These figures each contain three lines, corresponding to three different values of BETRAYAL. Points on each line indicate the probability of alliance formation at a corresponding value of ENTRY RECENCY (x axis) and for the level of BETRAYAL indicated by the line.

Crucially, note that increases in BETRAYAL generally correspond to decreases in predicted alliance formation. This is seen by moving from a line indicating a low level of BETRAYAL (such as the line BETRAYAL =0) to a higher level (such as BETRAYAL =.8) while remaining at the same point along the x axis. This suggests, at least, minimal support for hypothesis 8 since at no point does the probability of alliance formation appear to increase in BETRAYAL.

Perhaps more importantly, note the specific nature of the effect of increasing betrayal on the curvature of the relationship between ENTRY RECENCY and the probability of alliance formation. The simulated QI for each value of BETRAYAL appears to converge at the extremes of entry recency at 0 and near .06. At the same time, increasing levels of BETRAYAL create a pronounced effect in simulated QI through the middle range of ENTRY RECENCY. As a result, increases in BETRAYAL alter the steepness (or curvature) of the relationship between ENTRY RECENCY and alliance formation. Accordingly, as BETRAYAL increases, so too does the threshold at which ENTRY RECENCY (and thus entry endorsement) begins to have a meaningful, non-zero effect. Focusing on Figure 5.5 (though Figure 5.4 has a similar interpretation) When BETRAYAL =0, this threshold occurs as low as ENTRY RECENCY =.06. When, on the other hand, BETRAYAL =.8, ENTRY RECENCY does not begin to have a meaningful effect on the probability of alliance formation until a threshold of nearly ENTRY RECENCY =.8.

This is precisely the type of relationship suggested in section 3.2.7, and expected in section 5.1.3. It

49 The values of BETRAYAL that are illustrated here are deliberately few in number. inclusion of additional levels of BETRAYAL creates overlap in the figure, which makes it unreadable. Nevertheless, an overlay of all values of BETRAYAL (as a gradient, or contour map) confirms that the trend displayed in this figure is consistent.

50 Since these figures illustrate simulated QI for which AMBIGUOUS =1, the lower boundary of entry recency is, effectively, 0. As with Figure 5.2 and Figure 5.3, QI can be generated for ENTRY RECENCY <0, but simulated QI converge and remain at 0 for all levels of BETRAYAL and at all levels of ENTRY RECENCY <0. Accordingly, they are omitted from these figures.

51 Though not shown in the illustration, this is the threshold at which simulated QI, generated while holding AMBIGUOUS =1, exceed the counter-factual QI generated when AMBIGUOUS =0
indicates that observers that have been increasingly betrayed anticipate faster transitions in the holder’s state (or type) and thus grow skeptical of all but the most recent entry endorsements. Put differently, these figures suggest that a holder’s reputation with an observer decays faster when (1) the holder’s recent alliance behavior is ambiguous (i.e. occurring primarily through entry endorsement – entry into untested alliances) AND (2) the observer has been frequently betrayed in previous alliances.
Figure 5.4: Alliance Formation on Entry Recency at levels of Betrayal (Model 1)

Note: AMBIGUOUS held constant at 1
Figure 5.5: Alliance Formation on Entry Recency at levels of Betrayal (Model2)

Note: AMBIGUOUS held constant at 1
Finally, we evaluate hypothesis 9 through consideration of Figure 5.6 and Figure 5.7, based on estimates from Model 1 and Model 2 respectively. These figures are substantially different as they are derived from substantially different model specifications.

Figure 5.6 shows simulated QI across ENDORSER S when AMBIGUOUS =1 and when AMBIGUOUS =0. Unlike Figure 5.2 and Figure 5.3, all points along the line AMBIGUOUS =0 are plausible since AMBIGUOUS is not dependent on ENDORSER S. When AMBIGUOUS =0, changes in ENDORSER S appear to have no influence on the probability of alliance formation. This is consistent with our expectations since, when a holder is a member of a tested alliance, it has an opportunity to clearly and unequivocally demonstrate its reliability, thereby reducing the need for future observers to rely on entry endorsement (thus mitigating the influence of similarity between observers and entry endorsers).

In Figure 5.6, increases in ENDORSER S yield higher probabilities of alliance formation when AMBIGUOUS =1. This, again, is consistent with our expectations and supports hypothesis 9. It suggests that the effect of entry endorsement (i.e. membership in an untested alliance) are greater when the observer is more similar to the states that endorsed a holder.

Further, Figure 5.6 is compelling because it suggests that entry endorsement shifts from benefit to curse precisely at the point at which endorsers and observers are likely to shift from friends to enemies. That is, we see that simulated QI for AMBIGUOUS =1 become significantly greater than simulated QI for AMBIGUOUS =0 near ENDORSER S ≥0 and that simulated QI for AMBIGUOUS =1 are less than or equal to simulated QI for AMBIGUOUS =0 when ENDORSER S ≤0. This shift in the value of entry endorsement across values of ENDORSER S sits nicely with our theory.

Moving beyond Model 1, Figure 5.6 provides additional support for hypothesis 9. This figure shows the predicted probability of alliance formation across ENDORSER S and at differing levels of ENTRY REGENCY. The fundamental relationship between ENDORSER S and alliance formation remains, though it strongest when entry endorsement is most recent (ENTRY REGENCY =.9). The relationship appears weaker at lower levels of ENTRY REGENCY, suggesting that the recency of entry endorsement moderates the relationship between
ENDORSED S and alliance formation. This suggests that the influence of similarity between endorsers and observers erodes over time, similar to the decay of entry endorsement itself.
Figure 5.6: Alliance Formation on Endorser S at levels of Ambiguity (Model 1)
Figure 5.7: Alliance Formation on Endorser S at levels of Entry Recency (Model 2)

Note: AMBIGUOUS held constant at 1
5.4 Summary

Broadly speaking, we find support for the theory presented in chapter 3 as it applies to security agreements and a state’s reputation for reliability. Particularly, we find support for proposition 1, proposition 2, and proposition 4. The findings here are clear in their demonstration of the influence of entry endorsement on state reputation. The evidence provided here is even more stark in nature than the findings from chapter 4.

In conclusion, then, this analysis provides broad support for the hypotheses outlined in section 5.1.3. This, in turn implies some applicability of proposition 1, proposition 4, and proposition 2 to alliance formation and perhaps to matters of conflict and security more broadly. Combined with the findings from chapter 4, we see the influence of entry endorsement in the two main issue areas that encompass international relations: political economy and conflict. Further, when combined with the findings from chapter 4, we examined both proposition 2 and proposition 3, finding them to apply in different issue-areas as anticipated. This means that, with the conclusion of this chapter, we have conducted at least some empirical test of each of the propositions laid out in chapter 3; for each of these we have found empirical support as expected.
Chapter 6

Conclusion

This dissertation has sought to answer a simple question: how does membership in an untested agreement influence a state’s reputation? If extant models were correct, then we should have expected membership in an untested agreement to produce an overall negative influence on a state’s reputation.

Instead, I have argued in chapter 3 that untested membership can actually be beneficial. This is because entry into an agreement, whether or not it is tested, tells us something about a state. Namely, it tells us about how the state is viewed by its partners. By allowing a state to enter an agreement, its partners signal their expectations that it is sufficiently likely that the state will comply with the agreement – that the state has an adequately high reputation for reliability. We have called this concept entry-endorsement.

In addition to the argument that entry-endorsement influences state reputations, my theory leads to some additional, novel expectations about the factors influencing reputation. I argue that specific qualities about a state's partners influence its reputation, primarily through the mechanism of entry-endorsement. Further, I argue that the relevant qualities are subject to change depending on context. In circumstances such as political economy, where the observers of reputation are sub-state firms/actors and the endorsers are states, I submit that endorser vulnerability to punishment is a crucial determinant of the quality of entry-endorsement. On the other hand, in issues such as international conflict/security, where observers and endorsers are both likely to be states, I argue that the quality of entry-endorsement is determined by the similarity in strategic interests between the observer and endorser. Another novelty of the argument I
present is the explicit focus on the experiences of the observer in influencing reputation. Where observers have been more extensively betrayed, they are more likely to be skeptical of entry-endorsement unless it has occurred very recently.

In the area of international political economy, we examined the role of untested membership as it applies to international investment decisions. Specifically, we examined a type of investment that was highly likely to be informed by a host state's reputation for reliably upholding investor protections: greenfield investment. We found that entry into a BIT that had not been invoked as the basis for arbitration would make future greenfield investment more likely and, as distinctly anticipated by our theory, that the influence of entry into a BIT eroded over time. Further, we found that the vulnerability of an endorser to punishment for making bad endorsements influences the impact of endorsement by that state. Specifically, we found that sensitivity to punishment by special interests – indicated by the importance of foreign investment assets to the economy – makes new BITs (i.e. entry-endorsement) more impactful for holder (host) state reputations. We did not find evidence that endorser-state sensitivity to electoral pressure influenced the quality of entry-endorsement. We also found that the extensiveness of betrayal by investors in the investor (observer) state produced decreasing benefits from entry-endorsement as those endorsements grew older. This finding was consistent with our expectations, though it produced the additional, unexpected finding that very recent entry-endorsement produces a greater reputational benefit if an observer has been recently betrayed.

In chapter 5 we examined the role of untested membership as it applied to international conflict/security. Specifically, we focused on security agreements. We found that entry into a security agreement that was not tested by conflict produced a noticeable effect on the ability of a state to enter into a new agreement later on (thus suggesting a reputational benefit). As expected, the benefit of entering an untested security agreement declined over time. Additionally, as expected when the observers and endorsers are the same sort, we found that entry into an untested agreement produces a stronger reputational benefit when the observer and endorser have similar strategic interests. We also found that the extent to which an observer has been betrayed in the past influences the speed at which the benefit of entry-endorsement
erodes. Generally, the findings with regard to security agreements was more stark than for investment.

The examination of both investment and security agreements provide tests of each of the propositions laid out in chapter 3; this would not be possible by examining any single issue. The findings presented here indicate the overall applicability of each of the propositions to international relations in one form or another and thereby suggest the applicability of the theory to explaining international reputation. Furthermore, by examining phenomena from both economic and conflict/security issue-areas, we establish that the scope of the theory is not confined to any one issue-area.

6.1 Implications of this study

This study has focused, somewhat narrowly on a specific question, but its findings are likely to have broader implications for the study of reputation and the study of international relations. I will focus on what I see as the three most meaningful implications of the findings presented here.

6.1.1 Implications for our understanding of Reputation

Obviously, the central implication of this document is that the determinants of reputation are more nuanced than previously thought, and extend beyond a simple record of a state’s prior behavior. This document affirms and extends recent arguments by Gray¹, stating that the attributes of a state’s partners influence its reputation. It continues to present two additional factors that should be considered in future research on reputation: membership in untested agreements and the role of the observer’s prior experiences.

If the findings presented here are accurate, the role of these new factors demands that additional research on reputation change in some fundamental ways, primarily by focusing more on reputation as a relationship instead of reputation as an attribute of a singular actor. Previous studies of reputation have tended to focus on the aspects of the holder when explaining reputation. The relevance of other actors such

¹ 2013.
as endorsers and observers demand further consideration of reputation as constituted by contributions of various actors, and not merely as sole attribute of the holder. In practice, a focus on the relational aspects of reputation places greater empirical demands on future testing. Indeed, in the course of conducting this study, I was confronted with the difficulty of finding or assembling data that reflected the relationship between endorsers, holders and observers. It is possible that these very difficulties have motivated prior research to focus exclusively on the holder’s attributes. Since the behavioral revolution, however, data has continued to increase in fidelity and quality, and more recent developments such as the growing prevalence of massive data stores on the internet, have only accelerated this process. As time wears on, therefore, studies of reputation with a focus on reputation as a relationship should become more tractable.

6.1.2 Implications for ‘cheap talk’

This dissertation carries additional implications for the role of ‘cheap talk’ in international relations. Cheap talk\(^2\) is essentially the concept of communication that is relatively costless to send, non-binding and unverifiable (note the parallel to ambiguity). Scholars in IR have seized on this concept, with realists particularly placing emphasis on the notion that cheap talk should not influence strategic outcomes in many forms of international interaction. Further, scholarship contends that signing onto an agreement is cheap talk unless it carries substantial cost and/or substantial consequences for non-compliance (i.e. is dem facto binding). As such, scholars have wondered whether entry into these types of agreements can carry any actual influence over the course of international interactions.

Certainly, we would expect entry into an agreement that is unlikely to be tested – such as a security agreement that is likely to deter conflict – to qualify as cheap talk. To see this, consider that if a holder considers the likelihood of conflict to be low, then it will be unlikely to be put in a position where it faces punishment for non-compliance – thereby limiting the negative consequences of the agreement. At the same time, the most serious costs associated with such an agreement (direct contribution to an ally’s defense)

\(^2\) See Crawford and Sobel (1982) for the originating paper. Farrell (1987) is largely responsible for the phrase itself, as well as presentation of the concept in a more approachable fashion.
are only incurred in the event of an attack, and are not due at the time of entry. And yet, if the findings presented here are correct, we see that this very form of cheap talk should have an impact on the payoffs that a holder expects to receive. Namely entry into an agreement improves its reputation through entry endorsement.

The chief implication for international relations scholarship, then, is that cheap talk may very well influence strategic interaction, and do so within the paradigm of rationalism. This is important as the bulk of the theoretical work that has been done to show that cheap talk matters has occurred within the constructivist paradigm\textsuperscript{3} and does not necessarily explain how cheap talk can lead to cooperation in scenarios that allow for gains from betrayal. Moreover, this is important since the bulk of rationalist-oriented work in international relations has treated cheap talk as functionally meaningless.

\subsection*{6.1.3 Implications for research on the frequency of agreement formation}

Lastly, the findings presented here carry implications for the frequency with which we expect to see states engage in international agreement formation. Though not a novel insight, the empirical evidence presented explicitly seeks evidence of reputation decay over time, and in the absence of new information\textsuperscript{4}. Because reputations appear to erode over time, and since mere entry into new agreements provides a reputational benefit, we might expect to see states enter agreements with considerable frequency, if for no other reason than to maintain their good reputation. This is especially likely to be the case where tests of an agreement are costly and/or tend to indicate non-compliance and where there are considerable, low-cost opportunities to enter new agreements.

In practice, this implication may be relevant to explaining the growth of BITs and other, similar agreements in international political economy. To see why this is the case, consider that a test of membership

\textsuperscript{3} For prominent examples, see work by Risse, Finnemore and Sikkink, particularly, Risse (1999) and Finnemore and Sikkink (2001). Sartori (2002) is notable standout here, providing a rationalist explanation through which states develop a reputation for honest, cheap-talk, communication in order to preserve their ability to bluff when needed.

\textsuperscript{4} Crescenzi et al. (2012) and Gibler (2008) are recent examples of models that account for decay, but it is notable that both fold the influence of decay into larger, composite measures, thus preventing the ability to examine the isolated influence of decay. This study provides clearer evidence for the role of decay, particularly through the influence of the ENTRY RECENCY term.
in a BIT can be the bringing of a complaint under that agreement. Now, this type of test is more likely than not to negatively influence a state’s reputation, since the state needs to violate (or appear to violate the agreement in order to provoke a test). Under such circumstances, it is possible that states, because they are compliant, avoid testing and attempt to quickly settle any potential claims before they reach the arbitration stage. Then, to prevent decay from eroding their reputation, states enter new agreements, leveraging the benefits of entry-endorsement to maintain their good reputation. This certainly does not account for the entirety new BITs, but it may help to account for some of the increase in new BIT formation.

6.2 Avenues for future research on reputation

There are two main avenues for future research that are unexplored in this dissertation, but follow directly from the argument provided here.

First, the model provided in chapter 3 suggests novel implications for the rate and method by which state reputations recover from bad behavior. Prior research has indicated that reputations can only be repaired by taking on painful action. In Tomz’s book, for example, a state regains reputation by taking on punitive loan terms and repaying them. This model, however, suggests that reputation might be repaired through less costly measures. The markov nature of the model suggests that reputations decay to perfect uncertainty over time – i.e. a .5 probability of compliance. Thus if a state has a particularly bad reputation (i.e. it is expected that there is zero probability of compliance), then, after a long-enough period of time, its reputation will return to perfect uncertainty. A state can improve its reputation further if it can manage to enter additional, untested agreements. This suggests that the path to reputation repair might actually be rather cheap.

Second, the argument presented here presents additional questions about how and when states might behave within untested agreements. A compelling, substantive, application of these questions involves the behavior of states in untested security agreements. Consider that states in a security agreement face a high cost to complying with their agreement obligations and enjoy reputational benefits from untested membership
while an agreement goes untested. We might expect, therefore, that states will be more likely to work to
deescalate an crisis between an ally and a non-agreement member (thereby preserving the agreement as
untested while avoiding the costs of contributing to the ally’s defense). On the other hand, because the
benefits of entry-endorsement erode over time, states may choose, after a period of time in an untested
agreement, to actively stoke a crisis (a small conflict being an opportunity to demonstrate compliance and
maintain the state’s reputation).

Lastly, this dissertation points toward, but does not answer questions regarding negotiations over
agreements in light of the benefits of untested membership. This essentially envisions untested membership
as a potential outcome of the enforcement phase of a two-phase negotiation. The potential payoffs to
untested membership may influence the scope of an agreement. For example, it may motivate states to
negotiate shorter-term agreements if they expect them to go untested, since a shorter agreement will afford
members more frequent opportunities to re-negotiate agreements and thus renew the benefits of entry-
endorsement. Alternatively, it may motivate states to reduce the breadth of conditions that trigger tests
(i.e. compel behavior that indicates compliance), thereby making untested membership more likely.

---

Bibliography


Leeds, Brett Ashley (2005). **Alliance treaty obligations and provisions.** Dataset. Rice University, Department of Political Science, Houston.


Poast, Paul (2010). “(Mis) using dyadic data to analyze multilateral events”. Political Analysis, mpq024.


Appendix A

Theory Chapter Appendix

A.1 Extended discussion of probability of state transition over time

Equation (3.4) indicates that $b$ infers $\pi_s'$ from observing $b'$ behavior. This is demonstrable through straight-forward application of Bayes Rule, and described as follows.

Recall that $\pi = pr(y_T = 1)$, and that the use of reputation to inform the assessment of $\pi$ comes from the application of a decay process and an interpretation of $\pi_t = pr(y_t = 1)$. Under perfect observability, the value of $\pi_t = pr(y_t = 1|a_t^h)$. This is calculated by bayesian updating. Since $a_t^h$ determines $y_t$ by construction, then $\pi_t$ separates perfectly such that $\pi_t = 1$ if $a_t^h =$ comply and $\pi_t = 0$ if $a_t^h =$ renege. When $a^h$ is unambiguous, $b$ updates in a similar fashion with regard $h$'s entry into an agreement with $b'$, calculating $pr(y_s = 1|a_s^{b'})$ as\(^1\) follows.

Begin by assuming $pr(y_s = 1) \geq \pi_s'$. In other words, assume that, in deciding $a_s^{b'}$, $b'$ has determined that $\pi_s \geq \pi_s'$. As prescribed by the endorser’s best response function($\sigma^{b'}$), $a_s^{b'} =$ accept, regardless of $h$’s

\(^1\) Note that I denote $pr(y_s = 1|a_s^{b'})$ as somewhat interchangeably with $\pi_s$. The notation is used to be specific about the precise inputs to $b$’s assessment. These can vary, depending on the scenario under discussion. For example, if $h$’s most recent agreement ended with an observable outcome, at time $s$, then $h$ would update based on that observation, expressed as $pr(y_s|a_s^h)$. I use the notation $\pi_s$ to reference $b$’s assessment of $h$’s type at $s$ as derived from the most appropriate information. My use of the more explicit notation is purposeful and intended to improve the clarity of the math as I present it here.
actual state. In other words: \( pr(a^b_s = \text{accept}|y_s) = 1 \). It follows:

\[
pr(y_s = 1|\text{accept}) = \frac{pr(\text{accept}|y_s) \ast pr(y_s = 1)}{pr(\text{accept})} \geq \frac{1}{\pi_s^{s'}}
\]

and

\[
E[pr(y_s = 1|\text{accept})] \in (\pi_s^{s'}, 1]
\]

To reaffirm this, consider the contrary. Assume that \( pr(y_s = 1) = \pi_s \leq \pi_s^{s'} \). Under equilibrium, \( \sigma^{b'} \), \( a^b_s = \text{reject} \), regardless of \( h \)'s actual state. Thus, if \( pr(y_s = 1) = \pi_s \leq \pi_s^{s'} \), then \( pr(a^b_s = \text{accept}|y_s) = 0 \).

Accordingly:

\[
pr(s_h = 1|\text{accept}) = \frac{pr(\text{accept}|s_h) \ast pr(y_s = 1)}{pr(\text{accept})} = \frac{0 \pi_s^{s'}}{1} = 0 \quad \text{(A.1)}
\]

From equations (3.1) and (3.2), it follows that \( \pi_s^{s'} > 0 \), strictly. Further, from equation (A.1) above, \( a^b_s = \text{accept} \) is not in the support of \( pr(y_s|a^b_s) \). In equilibrium, this implies a contradiction since, if \( \pi_s \leq \pi_s^{s'} \), and \( a^b_s = \text{accept} \) is observed, then \( \pi_s = 0 \), and \( \sigma^{b'} \neq \text{accept} \). Therefore, in equilibrium, \( a^b_s = \text{accept} \) cannot be observed if \( \pi_s \leq \pi_s^{s'} \).

### A.1.1 Reputational costs of rejection

A separate matter is largely not considered in this document: what occurs if \( b' \) rejects at \( s \). In other words what is \( pr(y_s = 1|a^b_s = \text{reject}) \). In the context of the model laid out in section 3.2 and expanded in section 3.2.2, this is quantity is not as easily or precisely calculated. Generally, the updated assessment of \( \pi_s \) if \( a^b_s = \text{reject} \) is on the range \( [0, \pi_s^{s'}] \). As a practical matter, however, it is important to note that rejected agreements are not always widely publicized. Further, any effort to estimate or measures of rejection will likely underestimate its prevalence since a prospective partner \( b \) might reject an agreement at any time, up until the point of acceptance. Lastly, rejected parties have incentives to deliberately attempt to hide or obfuscate failed attempts at agreements. Accordingly, and to maintain the simplicity of the model, assume
that rejected agreements are either unnoticed or uninformative to b’s assessment of \( \pi \). Accordingly:

\[
pr(y_s = 1|a_s^{b'} = \text{reject}) = pr(y_s = 1|a_s^b)
\]

which implies:

\[
E[m(\cdot)|\pi_t, \alpha, t] = \pi_T
\]

and

\[
E[m(\cdot)|\pi_t, \alpha, t, a_s^{b'} = \text{reject}] = \pi_T
\]

and thus

\[
\pi_T \perp a_s^{b'} = \text{reject}
\]

\[\text{---}\]

2 A further extension of this model may wish to attach some probability that a negotiation between \( h \) and \( b' \) fails spectacularly, negatively impacting \( b' \)’s assessment of \( \pi_t \). Such an extension requires significant modification to the model. Specifically, the structure would need to be altered to allow \( h \) to engage more fully in strategic decision making.