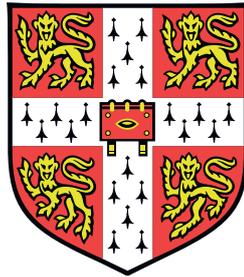# Contract Design for Collaborative Response to Service Disruptions

Marc Christiaan Jansen
Downing College

This dissertation is submitted for the degree of Doctor of Philosophy

December 2016

Judge Business School
University of Cambridge

Preface

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Acknowledgements and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text.

The research described in this dissertation was performed at the Judge Business School at the University of Cambridge between October 2012 and December 2016. This dissertation does not exceed 80,000 words.

Marc C. Jansen
Cambridge, UK
December 2016

Acknowledgements

First and foremost my thanks go out to my two PhD supervisors at the Judge Business School: Prof. Daniel Ralph and Dr. Nektarios Oraiopoulos. With great patience and attention to my progress as a budding researcher they enabled me to develop the skills and persistence necessary to craft the papers that make up this dissertation. Alongside my research they also gave me the opportunity to find and nurture my passion for teaching. Many thanks to Dr. Niyazi Taneri at National University of Singapore. Four very productive and exciting months spent during my intermission at Singapore University of Technology and Design allowed me do foundational work on the software and experiments behind Chapter 3 in this dissertation. I am grateful also to Dr. Jane Davies for guiding me through my MPhil program and inspiring me to take on this challenge.

The first chapter is based on a paper written in collaboration with Prof. Ralph and Dr. Oraiopoulos. The second chapter is based on a paper written in collaboration with Dr. Oraiopoulos and Dr. Taneri. Beyond collaborating on the research that led to two of the chapters in this dissertation, all three have been instrumental in helping me build and improve my work. Their passion for research and drive for real world impact have fueled my pursuit of the topics central to this dissertation.

Many people at the Judge Business School have shaped my thinking and inspired me to go the extra mile the past five years, starting the moment I got off the train in Cambridge. I had the pleasure of sharing most of my PhD journey with Michael Freeman, my office mate, fellow conference-traveller and sounding board for new ideas. Thank you for all your support and advice. The same holds for the rest of the Operations Group members: thank you all for letting me be a part of such an open and inspiring work environment.

Beyond the business school, my time in Cambridge was influenced by the most intelligent and driven people I have met in my life. I thank my fellow MCR members and housemates at Downing College for all the great moments. I would specifically like to express my gratitude to Phil Sitte, whom I met at Downing and with whom I shared a house for two brilliant years. Special thanks also to Nancy Yoo, who has been by my side supporting me throughout the final months of completing this dissertation.

Lastly, I owe it to my parents and my brother to have come to this point. Each in their own way they have shown me the value of hard work, allowing me to take on challenges of my own choosing — despite the challenges life has thrown at us. After all I have had the chance to see and experience, it is all too easy to lose touch with reality. Thank you, truly, for keeping me grounded and making me who I am today.

Abstract

This dissertation studies firms' strategic interactions in anticipation of random service disruption following technology failure. In particular it is aimed at understanding how contracting decisions between a vendor and one or multiple clients affect the firms' subsequent decisions to ensure disruption response and recovery are managed as efficiently as possible. This dissertation consists of three studies that were written as standalone papers seeking to contribute to the literature on contract design and technology management in operations management. Together, the three studies justify the importance of structuring the right incentives to mitigate disruption risks.

In the first study we contribute to this literature by means of an analytical model which we use to examine how a client and vendor should balance investments in response capacity when both parties' efforts are critical in resolving disruption and each may have different risk preferences. Responding to information technology (IT) system failure often requires a collaborative approach in which both the client and the vendor need to invest in response capacity. By investing more in response capacity, the client might make the vendor's response capacity more effective in the system restoration stage. Yet, in doing so, the client also encourages free-riding by the vendor. To understand how a client should balance the need to support the vendor while setting the right incentives for the vendor to invest, we develop an analytical model that combines the key characteristics of value co-creation (i.e., complementarity between the firms' investments in response capacity) with standard maintenance contract practices (i.e., penalty contracts that penalize the vendor for system downtime). We study the difference in the client's expected utility between the observable effort case (in case of low system complexity) and the non-observable effort case (in case of high system complexity). We refer to this difference as the cost of complexity. This study presents two key findings. Firstly, we show that the cost of complexity to client is decreasing in the risk aversion of vendor but increasing in her own risk aversion. Secondly, we find that the effect of risk aversion on the average system downtime is diametrically opposite depending on whether or not the client's investment is observable.

In the second study we further examine the context of the first study through a controlled experiment. We examine how differences in risk aversion and access to information on a contracting partner's risk preferences interact in affecting contracting and investment decisions between the client and vendor. We design an experiment in which subjects take the role of the client and make both contracting and investment decisions in order to control the vendor's incentives to invest and jointly minimize disruption costs. We implement an innovative pre-test that measures subjects' risk aversion and allows us to manipulate the difference in risk aversion between the subject and the automated vendor in the main experiment. We show that subjects deviate from theory in predictable ways: subjects appear to follow decision heuristics leading to over-investment and over-penalization. In addition, we find providing information on vendor risk aversion can be counter-effective in aiding subject decision making, reinforcing reliance on inefficient heuristics through 'cognitive overload'.

In the third study we model the effect of contract design on a provider's response capacity allocation in a setting where multiple clients may be disrupted and available response capacity is limited. Increasingly large client pools have come to depend on continuous availability and security of equipment from a single specialist provider. Faulty equipment can be a direct cause of disruption to the client and in some cases cause life-threatening circumstances to consumers or patients downstream. At the onset of disruption, the scale of the disruption is typically unknown. We examine how contracting decisions between a technology provider and multiple clients can enable efficient allocation of response capacity under incomplete and imperfect information on the scale of a disruption. We demonstrate that by setting disclosure rewards, the provider can create a powerful incentive for the clients to put in investigation effort and communicate truthfully, in order to facilitate response capacity allocation. However, competition for limited emergency resources may lead clients to deliberately under-report to the provider. Even when it is feasible to implement the optimal contract terms, miscommunication and delay of response can turn out to be optimal under decentralized decision making, echoing real world observations surrounding response to medical equipment failures.

Exposure of this work

Conferences

Jansen, M.C., N. Oraiopoulos, D. Ralph, Enabling Collaborative Response to IT Service Disruption under Risk Aversion. *POMS Annual Conference*, Atlanta, GA, USA. May 2014.

Jansen, M.C., N. Oraiopoulos, D. Ralph, Enabling Collaborative Response to IT Service Disruption under Risk Aversion. *INFORMS Annual Meeting*, San Francisco, CA, USA. November 2014.

Jansen, M.C., N. Oraiopoulos, D. Ralph, Enabling Collaborative Response to IT Service Disruption under Risk Aversion. *MSOM Conference*, Toronto, Canada. June 2015.

Jansen, M.C., N. Oraiopoulos, D. Ralph, Enabling Collaborative Response to IT Service Disruption under Risk Aversion. *INFORMS Annual Meeting*, Philadelphia, PA, USA. November 2015.

Jansen, M.C., Managing Escalations: Equipment Failure and Response Capacity Allocation. *INFORMS Annual Meeting*, Philadelphia, PA, USA. November 2015.

Jansen, M.C., Managing Escalations: Equipment Failure and Response Capacity Allocation. *INFORMS Annual Meeting*, Nashville, TN, USA. November 2016.

Jansen, M.C., Risk Preferences and Joint Disruption Mitigation: Experimental Evidence. *INFORMS Annual Meeting*, Nashville, TN, USA. November 2016.

Seminars

Jansen, M.C., Risk Preferences and Joint Disruption Mitigation: Experimental Evidence. *Nanyang Technological University*, Singapore. March 2016.

Jansen, M.C., Risk Preferences and Joint Disruption Mitigation: Experimental Evidence. *City University of Hong Kong*, Hong Kong. April 2016.

*Presenting author highlighted in **bold** font.*

Contents

Chapter 1

General Introduction

1.1 Motivation

The choice of topic for this dissertation was motivated by shifts in the risk landscape many companies face, driven by increased outsourcing (Apte and Mason 1995, Van Mieghem 1999, Benjaafar et al. 2007) and servitization[1] (Sawhney2004, Cohen et al. 2006, Kastalli and Van Looy 2013). We study disruption risk mitigation through the lens of service contracting, focusing on joint decision making by client(s) and a vendor in contexts where a client has outsourced maintenance or operation of advanced technology or equipment to a vendor. The chapters in this dissertation are motivated by cases involving potential service disruption as a result of information technology (IT) or medical equipment failure. In these contexts the technology is often mission critical[2] to the client's day-to-day service operations and the nature of the technology is such that multi-sourcing or emergency sourcing are not possible in the event of disruption (Kim et al. 2010).

In recent years, various firms in consulting and insurance, as well as independent international organizations like the World Economic Forum have produced reports detailing the changing risk landscape for supply and service chains. In 2009, PriceWaterhouseCoopers (PwC) issued a note on risk management in response to the alarming number of bankruptcies and general financial distress in the wake of the financial crisis. The note highlighted the damage supply and service disruptions do to 'brand and bottom line', with data showing share prices dropping up to nine percentage points and increased volatility for up to two years after a disruption. These damages come on top of the direct damages as a result from delayed or discontinued operations. PwC state that priority should be with identifying and nurturing critical vendor relationships and making a transition from a traditional crisis response attitude to a early identification and corrective action in the event of disruption (Nally and Pittman 2009). Results from a survey conducted by Zurich Insurance Group and DHL, the Business Continuity Institute in 2011 highlight the source and frequency of business interruption. Of over five hundred respondents, 85% indicated having recorded at least one disruption in the year preceding the survey. As much as 35% of the recorded disruptions were related to outsourcing (Glendon 2011). A 2013 report by the World

---

[1]Defined as the process of adding value to core corporate offering through inclusion of services (Vandermerwe and Rada 1988)

[2]Defined as 'extremely important or necessary for a company, activity, etc. to operate successfully', in *Cambridge Dictionary* Online, from http://dictionary.cambridge.org/dictionary/english/mission-critical

Economic Forum and Accenture echoes this with survey results showing 80% of respondent firms are concerned about the resilience of their business (Bhatia et al. 2013). The report by Chacon et al. (2012) highlights the dangers in that lie with the common reliance on crucial business process inputs like availability of data, and extensive subcontracting as some of the major vulnerabilities. Dependence on IT systems provided by external parties is mentioned as a particular issue in contingent business interruption (Galey 2002). Finally (Dubey and Wagle 2007, p. 11) highlight that as technology has come to integrate the tools for multiple business processes, "complaints frequently involve incidents that are mission critical, thus demanding immediate attention".

Academics in operations management (OM) have recognized these issues, resulting in a growing body of work considering how decisions made in advance of disruptions, including contracting decisions, help mitigate disruption risks (see e.g. Tang (2006) and Sodhi et al. (2012) and references therein). Whereas most of these works consider disruptions in traditional supply chain contexts, attention is increasingly given to service disruptions. In an interview with CIO Magazine in 2006, Prof. Sheffi at MIT said with reference to recent IT disruptions (Patton 2006):

"We've learned that the fates of companies and government agencies are sealed before the disaster hits. Organizations that get ready perform well; those that don't prepare don't do well. (…) There is something in the DNA of resilient companies that is missing from those that falter and suffer, [which] goes beyond just redundancy."

This dissertation contributes to this body of work by addressing a number of themes which we address in the next section.

1.2   Central Themes

The study of mitigating random disruptions has generated a large body of work in OM. Although contracting and risk management as a broader field of study is well-developed, the OM papers that form the core of the supporting literature for this dissertation were mainly published after the turn of the century. Some of the fundamental concepts can, however, be traced back to earlier works. Take for instance the paper by Shavell (1984), which illustrates the trade-off between investment in accident prevention and the expected cost of an accident $c(x, h)$ with the simple function $c(x, h) = x + p(x)h$, where x is the level of care taken, $p(x)$ is likelihood of an accident given the level of care and h is the harm caused by the accident. Since these early models a stream of work in OM has sought to address how disruptions should be managed in varying contexts, seeking to make an impact on real decisions.

One part of this stream focuses on demand risks and developing improved inventory management policies and capacity allocation decisions by means of contracts (see for instance Tomlin (2003), Bernstein and Federgruen (2005) and references therein). Alongside this stream another stream

developed, considering management of supply risks and addressing how the impact of disruptions may be mitigated, particularly in manufacturing settings (e.g. Tang (2006)). Broadly, this literature can be classified according to how reliability of supply is modeled (Wang et al. 2010). Random capacity refers to the case in which the realized supply is the minimum of the order quantity and the realization of a stochastic capacity level, independent of the order quantity. Hence in this case the realized supply can never be greater than the order quantity. In the case of random yield the realized supply is a random fraction of the order quantity. Generally, this fraction is allowed to be greater than one without an explicit capacity limit, meaning realized supply can in this case be greater than the order quantity. Random *disruption* can be interpreted as a special case of random yield in which a supplier's realized supply is either all (i.e. a proportion of 1) when the supplier is not disrupted, or nothing (i.e. a proportion of 0) when the supplier is disrupted. A central tenet of this literature is that contracts ultimately determine the incentives for firms to address and manage risks and are thus an important instrument to manage these disruption risks.

This dissertation relates contracting decisions to investments in response capacity and information gathering in mitigation of random disruption risks. We consider random disruption through a service lens: when mission critical technology is down, so is the service process it supports. Throughout the dissertation we may use 'technology' and 'system' interchangeably, in all cases referring to hardware and software facilitating firm level operations and decision making.

Response, interdependence and collaboration

The OM literature generally views mitigating disruption as a combination of two strategies: steps can be made to lower the likelihood of a disruption occurring (*preventive* capacity) and improving the mean-time-to-repair (*response* capacity) (Jain et al. 2013, Kim and Tomlin 2013). This dissertation concentrates on the latter, studying how timely and effective response can be facilitated by the right contract design. Without an accurate and timely response, most major disruptions cannot be contained even with a lot of slack built into the system. The right alert and response system can make the difference between a controlled loss of operations and long term loss of business.

A disruption in a supply or service network is harder to contain if multiple parties are simultaneously affected. Tomlin and Wang (2005) incorporate product demand correlations and their effect on the choice of supply chain design. Fewer works have addressed correlated supply disruptions, with some papers like Tomlin (2009) explicitly taking supplier yields as independent random variables. When it comes to major disruptions it is unlikely that this independence holds up. Particularly as advanced technology increasingly forms the backbone of daily operations and many firms and organizations may depend on a common technology, one disruption may have a widespread impact. Accounting for potential correlation between disruptions affects which mitigation investment strategy is optimal. A paper by Bhattacharya et al. (2012) provides

another perspective by comparing single-sourcing and multi-sourcing strategies for information services projects when vendor and client efforts are interdependent and there is a disconnect between outcomes and the metric used to generate the right incentives. Kim and Tomlin (2013) study a context in which technological systems risk failure if one or more subsystems fail and joint failure of multiple subsystem is possible. Results following from their model indicate firms' tendency to overinvest in response capacity, yet underinvest in prevention, particularly when joint failure is likely.

Incentives and contingent contracts

When response to disruption depends on the efforts of multiple parties in a supply or service chain, achieving effective response becomes a matter of aligning incentives. In turn incentives are determined by the parties' respective objectives and manipulated through contract design (Narayanan and Raman 2004). For these contracts to be implementable, the relevant information needs to be not just *observable*, but also *verifiable* to a court. Verifiable information is information that "can be certified or authenticated once disclosed" (Bolton and Dewatripont 2005, p. 171). In case "performance cannot be verified by a court, contracts that are contingent on performance [...] cannot be made, as the courts will be unable to enforce them" (Tirole 1988, p. 38). For a contingent contract to be implementable, the contingency must be verifiable, yet the decisions or actions giving rise to that contingency need not be, e.g. capacity investments may be observable but not verifiable (Cachon and Lariviere 2001). From the perspective of incentives, contracts are feasible as long as *incentive compatibility* and *participation* (or *individual rationality*) constraints are met (Laffont and Martimort 2001). Here, risk preferences are an important consideration: risk seeking and risk averse decision makers can make very different decisions under the same contract terms.

Contingent contracts is the class of contracts that include payments or transfers that are conditional on a contingency, i.e. the occurrence of a certain outcome, or a particular set of criteria being met (Bazerman and Gillespie 1998). As maintenance services have increasingly replaced traditional sale and implementation of technology as a core business, *performance-based* contracts (PBCs) have replaced the simpler fixed price and time-and-materials contracts. PBCs are a special form of contingent contract where a vendor is paid as long as the technology provided is operational. The implications of PBCs for service provision and quality have been examined in various settings, including aerospace and defense (Kim et al. 2007), software outsourcing (Dey et al. 2010), healthcare (Jiang et al. 2012) and advertising (Dellarocas 2012). Whatever the setting, improved vendor performance is the objective and has been observed empirically (e.g. in Guajardo et al. (2012)).

Information, communication and learning

In general, the role of information in contracting problems can be classified into multiple categories. Knowledge or observability of information about certain variables, events or payoffs might not be the same for all players, i.e. there may be *asymmetric* or *incomplete* information. The former refers to contexts in which a party "knows or observes something that other players do not observe" (Watson 2008, p. 290). The latter refers to when "some player do not know the payoffs of the others" (Fudenberg and Tirole 1991, p. 209). In game theoretical terms, *imperfect* information means there is "at least one contingency in which the player on the move does not know exactly where he is in the tree", i.e. he does not know the result of a previous event (Watson 2008, p. 162). Harsanyi (1967) suggested that by modeling 'moves of nature' resulting in a player's privately observed 'type' determining his payoffs, a game of incomplete information can be reinterpreted as a game of *imperfect* information on nature's moves (Fudenberg and Tirole 1991, p. 209). Players' access to information can be affected through communication and learning, which can affect contract design (Laffont and Martimort 2001).

When it comes to communication, evidence exists that managers exhibit systemic bias in the disruptions they report (Maggi and Rodriguez-Clare 1995, Schmidt and Raman 2012). Schmidt and Raman (2012) distinguish between internal and external disruptions and study the differences in likelihood to impact shareholder value, which affects the extent to which they are communicated accurately. Errors in decision making are typically only revealed when something goes wrong, mostly in case disruption. As it may be beneficial to misreport disruptions in certain circumstances, communication in the face of disruption becomes a strategic decision. Firms with information on an impending or realized disruption thus do not always possess full knowledge of the situation, or may have the incentive to hide or misrepresent information. On the receiving end, parties might therefore question the possibility of false alarm upon receiving warnings. The completeness, accuracy and credibility of the information has to be weighed against the need for expedited response, as it may only be feasible to control disruption damages in a limited time frame.

Information sharing is one way for firms to learn about potential risks and direct their mitigation efforts. A relevant stream of work considers information sharing in settings subject to demand risk. It is commonly assumed that the downstream firm has most insight into what demand will be like; upstream firms are typically less informed. Several works study how competition in a supply network affects communication of demand forecasts (see e.g. Ziv (1993), Anand and Goyal (2009)). Analogously, in light of supply risks, firms may not be equally informed on exposure to certain risks, the gravity of existing problems, or even possible solutions. Considering asymmetric information on supply risks, upstream firms (i.e. suppliers, vendors) are typically considered to be more informed (see e.g. Yang et al. (2009), Tomlin (2009)). In Chapter 4 we employ a different assumption: in a technology management setting clients may be more informed on the state of the technology and potential reasons for non-performance.

A different mode of learning can be characterized as 'learning by doing'. In a setting at risk of disruption, learning by doing equates to investigating red flags and confirming the nature of the problem and, if necessary, formulating a response strategy. Tomlin (2009) investigates the effect of learning in the face of future disruptions along the supply chain, while assuming no delay in response to the first signs of a disruption. Tomlin (2006) does include a measure of supply chain response time, but does not allow for improvement over time through experience.

1.3   Summary of Results and Contributions

Chapters 2 through 4 were written as standalone papers and can be read as such. Nonetheless, all three chapters respectively address one or more of the central themes addressed in the preceding sections, forming a single dissertation on contract design for collaborative response to service disruptions.

In Chapter 2 we examine how a client and vendor should balance investments in response capacity when both parties' efforts are critical in resolving disruption and each may have different risk preferences. By means of an analytical model we study the performance of a penalty contract in aligning incentives for investment in response capacity. We contribute to the literature by showing how, in this context, differences in risk preferences can result in either under- or over-investment, depending on which party is the more risk averse party.

In Chapter 3 we build on the findings in Chapter 2 and further examine the performance of a penalty contract in aligning incentives for investment in response capacity through a series of controlled experiments. We contribute to the literature by testing how differences in risk preferences between the stakeholders in a contract interact with stakeholders' access to intelligence on a contracting partner's risk profile and together impact contracting decisions and performance.

In Chapter 4 we expand on the bilateral contracting settings studied in Chapter 2 and Chapter 3. We model the effect of contract design on a provider's response capacity allocation in a setting where multiple clients may be disrupted, available response capacity is limited and information on the nature of disruptions is both imperfect and incomplete. We contribute to the literature by demonstrating disclosure rewards are a powerful incentive for truthful communication to improve response capacity allocation between multiple clients subject to disruption.

References

Anand KS, Goyal M (2009) Strategic information management under leakage in a supply chain. *Management Science* 55(3):438–452.

Apte UM, Mason RO (1995) Global disaggregation of information-intensive services. *Management Science* 41(7):1250–1262.

Bazerman MH, Gillespie JJ (1998) Betting on the future: The virtues of contingent contracts. *Harvard Business Review* 77(5):155–60.

Benjaafar S, Elahi E, Donohue KL (2007) Outsourcing via service competition. *Management Science* 53(2):241–259.

Bernstein F, Federgruen A (2005) Decentralized supply chains with competing retailers under demand uncertainty. *Management Science* 51(1):18–29.

Bhatia G, Lane C, Wain A (2013) Building resilience in supply chains. Technical report, World Economic Forum.

Bhattacharya S, Gupta A, Hasija S (2012) Single sourcing versus multisourcing: The role of effort interdependence, metric-outcome misalignment, and incentive design, working paper.

Bolton P, Dewatripont M (2005) *Contract theory* (MIT press).

Cachon GP, Lariviere MA (2001) Contracting to assure supply: How to share demand forecasts in a supply chain. *Management Science* 47(5):629–646.

Chacon N, Doherty S, Hayashi C, Green R, Lever I (2012) New models for addressing supply chain and transport risk. Technical report, World Economic Forum.

Cohen MA, Agrawal N, Agrawal V (2006) Winning in the aftermarket. *Harvard Business Review* 84(5):129.

Dellarocas C (2012) Double marginalization in performance-based advertising: Implications and solutions. *Management Science* 58(6):1178–1195.

Dey D, Fan M, Zhang C (2010) Design and analysis of contracts for software outsourcing. *Information Systems Research* 21(1):93–114.

Dubey A, Wagle D (2007) Delivering software as a service. *The McKinsey Quarterly* 6(2007):2007.

Fudenberg D, Tirole J (1991) *Game theory*, volume 393.

Galey G (2002) Contingent business interruption and other special covers. Technical report, SwissRe.

Glendon L (2011) Supply Chain Resilience 2011. Technical Report November, BCI, Zurich and DHL.

Guajardo JA, Cohen MA, Kim SH, Netessine S (2012) Impact of performance-based contracting on product reliability: An empirical analysis. *Management Science* 58(5):961–979.

Harsanyi JC (1967) Games with incomplete information played by Bayesian players part I. The basic model. *Management Science* 14(3):159–182.

Jain N, Hasija S, Popescu DG (2013) Optimal contracts for outsourcing of repair and restoration services. *Operations Research* 61(6):1295–1311.

Jiang H, Pang Z, Savin S (2012) Performance-based contracts for outpatient medical services. *Manufacturing & Service Operations Management* 14(4):654–669.

Kastalli IV, Van Looy B (2013) Servitization: Disentangling the impact of service business model innovation on manufacturing firm performance. *Journal of Operations Management* 31(4):169–180.

Kim SH, Cohen MA, Netessine S (2007) Performance contracting in after-sales service supply chains. *Management Science* 53(12):1843–1858.

Kim SH, Cohen MA, Netessine S, Veeraraghavan S (2010) Contracting for infrequent restoration and recovery of mission-critical systems. *Management Science* 56(9):1551–1567.

Kim SH, Tomlin B (2013) Guilt by association: Strategic failure prevention and recovery capacity investments. *Management Science* 59(7):1631–1649.

Laffont JJ, Martimort D (2001) *The theory of incentives: The principal-agent model* (Princeton university press).

Maggi G, Rodriguez-Clare A (1995) Costly distortion of information in agency problems. *The RAND Journal of Economics* 26(4):675–689.

Nally D, Pittman D (2009) 10 Minutes on supply chain risk management. Technical report, PricewaterhouseCoopers.

Narayanan V, Raman A (2004) Aligning incentives in supply chains. *Harvard Business Review* 82(11):94–102.

Patton S (2006) Logistics expert Yossi Sheffi talks disaster recovery. CIO.com, URL `http://www.cio.com/article/2447457/disaster-recovery/logistics-expert-yossi-sheffi-talks-disaster-recovery.html`, retrieved 18/12/2016.

Schmidt W, Raman A (2012) When supply chain disruptions matter.

Shavell S (1984) A model of the optimal use of liability and safety regulation. *The Rand Journal of Economics* 15(2):271–280.

Sodhi MS, Son BG, Tang CS (2012) Researchers' perspectives on supply chain risk management. *Production and Operations Management* 21(1):1–13.

Tang CS (2006) Perspectives in supply chain risk management. *International Journal of Production Economics* 103(2):451–488.

Tirole J (1988) *The theory of industrial organization* (MIT press).

Tomlin B (2003) Capacity investments in supply chains: Sharing the gain rather than sharing the pain. *Manufacturing & Service Operations Management* 5(4):317–333.

Tomlin B (2006) On the value of mitigation and contingency strategies for managing supply chain disruption risks. *Management Science* 52(5):639–657.

Tomlin B (2009) Impact of supply learning when suppliers are unreliable. *Manufacturing & Service Operations Management* 11(2):192–209.

Tomlin B, Wang Y (2005) On the value of mix flexibility and dual sourcing in unreliable newsvendor networks. *Manufacturing & Service Operations Management* 7(1):37–57.

Van Mieghem JA (1999) Coordinating investment, production, and subcontracting. *Management Science* 45(7):954–971.

Vandermerwe S, Rada J (1988) Servitization of business: Adding value by adding services. *European Management Journal* 6(4):314–324.

Wang Y, Gilland W, Tomlin B (2010) Mitigating supply risk: Dual sourcing or process improvement? *Manufacturing & Service Operations Management* 12(3):489–510.

Watson J (2008) *Strategy: An introduction to game theory* (W.W. Norton & Company), second edition.

Yang Z, Aydin G, Babich V, Beil DR (2009) Supply disruptions, asymmetric information, and a backup production option. *Management Science* 55(2):192–209.

Ziv A (1993) Information sharing in oligopoly: The truth-telling problem. *The RAND Journal of Economics* 455–465.

Chapter 2

Enabling Collaborative Response to IT Service Disruptions under Risk Aversion

2.1 Introduction

To protect their place in today's competitive landscape businesses have come to depend on mission critical technology that supports their operations. Specifically information technology (IT) systems have become mission critical as they are increasingly relied upon to maintain and improve business performance (Fitoussi and Gurbaxani 2012). Maintaining continuity of mission critical systems has become a major concern (Kim et al. 2010, Kim and Tomlin 2013), complicated by the fact that IT activities are increasingly outsourced, meaning that risk management has come to rely heavily on effective contract design (Gopal et al. 2003, Gopal and Sivaramakrishnan 2008, Chen and Bharadwaj 2009). Moreover, IT downtime can inflict serious costs through various channels, including delays to or loss of the information flows necessary to maintain revenue flows.

Examples of costly IT disruptions abound across industries. In the oil and gas industry, exploration activities simultaneously require data coming from sensors close to the action and access to previously recorded data. Losing such facilities may mean losing as much as $1 million per hour (Forbes 2013). In the airline industry, a two-hour IT glitch at United Airlines in 2012 resulted in 257 delayed flights and 10 cancelled flights (The Telegraph 2012). In financial services, a single server failure in early 2014 affected 7,000 Lloyds Banking Group ATMs throughout the UK (Sky News 2014).

The frequent occurrence of such major disruptions indicates that no matter how hard firms try to anticipate them, the risk of disruption can never be fully eliminated. For that reason, a number of recent studies from both practitioners (see e.g. FM Global (2010), CA Technologies (2011)) and academics (e.g Jain et al. (2013) and references therein) have emphasized the importance of risk mitigation through building sufficient response capacity. Such investments in response capacity by the vendor are often incentivized by the use of penalty contracts imposed by the client. Yet, in settings of complex system integration between the client and vendor, the client's investments in response capacity might be as important as the vendor's.

Consider the case of telecommunications firm Bharti Airtel which outsourced its IT service to IBM (Martinez-Jerez and Narayanan (2007), cited in Bhattacharya et al. (2014)). To keep up with network expansion, Bharti Airtel planned to take a new strategic course and let IBM take responsibility for the hardware and software elements of the IT architecture. At the same time, Bharti was to maintain the technological responsibilities on the telecom infrastructure. Through previous acquisitions of telecom operators, Bharti's IT architecture had become a complicated system of inherited sub-systems integrated with its telecom infrastructure. As a result, in the design of the system there were strong interdependencies between the efforts of the client (Bharti Airtel) and the vendor (IBM).

Such interdependencies between the client and the vendor are also critical in the event of a service disruption. For instance, a vendor's efforts to restore the functionality of a complex IT system are unlikely to be effective unless the client provides sufficient access, information and personnel to conduct a thorough response in the event of system failure. As such, disruption response needs to be a collaborative process, in which both the client and the vendor contribute resources to facilitate system recovery. Importantly, the more complex the interface of technologies between the client and vendor, the more difficult it becomes to specify and monitor the role of the client in facilitating system recovery. For example, given the highly complex and integrated nature of the Bharti Airtel IT systems, it was unclear what would be a sufficient level of the support provided by Bharti Airtel.

Examples can also be found of outsourcing settings where integration and collaboration are essential, yet the client's contribution is more readily observable. Take for instance the case of Gothaer Systems outsourcing the collection systems of its parent company Gothaer Insurance Group to SAP (Basten et al. 2014). Similar to Bharti, Gothaer grew its business through a series of mergers and acquisitions, resulting in a patchwork of IT systems to manage collections which was due for replacement by a single new system. Whereas Bharti and IBM's new system required complex integration between the respective parties IT and telecom systems, Gothaer looked to replace its systems with a standard insurance solution offered by SAP. As a result, despite the typical integration difficulties, the lines between Gothaer Systems and SAP technology remained visible. The sharp contrast between the Bharti and the Gothaer cases echoes a point made by Susarla et al. (2010), whose empirical work highlights that fewer interdependencies in the client-vendor technological interface are associated with higher observability of efforts.

The fact that the client has to contribute resources that facilitate the vendor's response and at the same time make the vendor liable for that response, typically through penalty contracts (see Chan et al. (2014) for examples in medical equipment maintenance services) introduces an interesting trade-off: On the one hand, by contributing more resources the client makes the vendor's resources more effective. On the other hand, more resources from the client might give rise to free-riding and discourage the vendor from investing his own resources. The goal of our research is to understand how the client should balance these two opposing forces: Supporting the vendor and at the same time keeping the vendor liable through an appropriate contract.

Specifically, we develop an analytical model to study the effect of two critical parameters of the environment in which the parties operate. First, in some settings, the client's contribution to the response capacity might be limited to rather tangible resources that can be easily observed and verified (e.g. as in the Gothaer and SAP case). In other settings, the client's resource contribution may require in-depth know-how that cannot be easily assessed or verified (e.g. as in the Bharti and IBM case). In the former setting, the client's investment in response capacity is contractible, while in the latter, it is not. In general, a highly complex and integrated system is likely to be associated with an environment where the client's efforts are less observable. On the contrary, when the vendor and the client's system are connected through a more standardized interface, the client's efforts are more observable. Second, we examine how the relative risk attitude of the client and the vendor affect investments in the collaborative response process. Typically, cash flow volatility is more distressing for small- and medium-sized clients and vendors than for larger ones. However, recent empirical studies show that even large companies can exhibit risk aversion in contract design (Ning et al. 2014).

Our analysis sheds light on how the interaction effect of these two parameters (observability of resources and risk attitude) jointly determine the efficiency of the investments in collaborative response. We capture the latter through two performance metrics: i) The cost of complexity which we define as the difference in the client's expected utility between the observable effort case (low complexity) and the non-observable effort case (high complexity); and ii) the total expected system uptime.

We make the following three contributions to the literature on mitigating disruption risk in IT outsourcing (ITO). First, we find that the cost of complexity diminishes as the vendor becomes more risk averse. This counter-intuitive result occurs because when the client's effort is observable, it acts as a risk transfer mechanism to the vendor. As the vendor becomes more risk averse, this mechanism becomes less effective, and as such, the difference between the observable and non-observable cases decreases. Second, we show that the exact opposite effect takes place when the client's risk aversion increases. Hence, in that case, the cost of complexity increases because complexity hampers the ability of the client to transfer risk. Finally, we show that the effect of higher risk aversion on the system uptime is contingent on the level of the system complexity: Under low complexity (observable effort) it decreases in vendor risk aversion, but is invariant in client risk aversion. On the contrary, under high complexity (non-observable effort) it increases in both vendor and client risk aversion. In summary, we note that increasing uptime, which sounds positive, is actually a signal of inefficiency due to complexity.

2.2   Literature Review

Firms choose to outsource the implementation and maintenance of their IT to augment their productivity, leverage knowledge outside the firm, and improve core competencies (DiRomauldo

and Gurbaxani 1998), objectives that are of an increasingly strategic nature (Susarla 2012). Such outsourcing relationships are typically governed by contracts that fall within three categories: fixed-price contracts, cost-plus or time and materials contracts, and performance-based contracts (PBCs) (Chen and Bharadwaj 2009, Dey et al. 2010). Provided the IT is sufficiently critical to the client's operations, IT contracts will typically contain provision for and metrics related to disruption risks. Chen and Bharadwaj examine a sample of ITO contracts in light of contract provisions along four dimensions, including monitoring provisions. Monitoring provisions may cover benchmarking of vendor's performance against prespecified standards to ensure compliance and adherence to guidelines such as disaster recovery plans in case of failures. These provisions may cover risks that are potentially outside of the vendor's control. Vendors and clients share these risks through a combination of fixed price and performance-based contract terms (e.g. penalties), while the vendor may be protected by a guaranteed payment (fixed fee) independent of performance (Fitoussi and Gurbaxani 2012). Susarla and Barua (2011) underscore that inappropriate contract offers by the client pose serious financial risks to IT vendors. Dey et al. highlight that the client's business needs and industry standards should inform the choice of product performance metric. They note that in case of outsourcing mission critical IT systems, which we study in this chapter, metrics should cover system downtime.

IT contracts typically require significant integration and coordination of actions (Bapna et al. 2010), resulting in potential agency costs when mitigating system failure risks (Anderson and Dekker 2005, Wu et al. 2012). In settings where coproduction is necessary and agency costs are likely to be high, PBCs are the most effective way of aligning the incentives across the firms (Roels et al. 2010). Empirical evidence of the efficacy of PBC's has been demonstrated in the context of product reliability of commercial aircraft engines (Guajardo et al. 2012). The importance and widespread use of PBCs in IT service outsourcing are discussed extensively in Bhattacharya et al. (2014) and the references therein. Following this analytical and empirical support for PBC's, we focus our analysis on the performance of a simple PBC in the context of mission critical ITO.

The role of contracts in facilitating joint efforts has been studied in the context of IT contracting, from software development (Chen and Bharadwaj 2009, Dey et al. 2010, Ceccagnoli et al. 2012) to cyber infrastructure (August and Tunca 2011, Lee et al. 2013) and IT product improvement (Bhattacharya et al. 2014). Chan et al. (2014) provide empirical evidence of the importance of contract design in facilitating joint production of value between the client and vendor in restoring complex technology. The papers by August and Tunca (2011) and Bhattacharya et al. (2014) are closest to our study. In the context of joint efforts to minimize disruptions through cyber attacks, August and Tunca (2011) examine and compare different liability policies. Among their findings is that imposing partial patching cost liability is preferable to full cost liability from a welfare optimization perspective. In their model, the vendor and user can invest to prevent disruption. Instead, we study a context in which the vendor and client invest in response capacity, which shortens the expected disruption length but does not prevent disruptions from happening.

Bhattacharya et al. (2014) explore the context of joint product improvement of a client and a customer support center where a double moral hazard problem arises as improvement efforts by the latter may hurt their revenues as service requests are lowered, yet client efforts are required for improvement as well. Their paper studies the efficacy of linear gain-share contracts contingent on realized volume of service requests; we work with a linear penalty contract contingent on realized downtime of an IT service.

In our model, we focus on investments in response capacity to improve time to repair for a mission critical IT system. Kim and Tomlin (2013) examine the trade-off between prevention and mitigation and show that firms in a decentralized setting tend to overinvest in response capacity. Closer to our work is Jain et al. (2013), who analyze after-sales contracting in a double moral hazard setting and point out that PBCs expose the service provider to financial distress, which may lead to suboptimal investment decisions as a result of risk aversion. In their model, the client's investment affects the interfailure time and the vendor's investment affects the time to repair. Kim et al. (2007) and Kim et al. (2010) also study the role of risk aversion in the context of investment in response capacity. In contrast to Jain et al. (2013), we are interested in a setting where both the client and vendor need to dedicate resources to respond to disruption and their joint resources are complementary in reducing the length of the disruption (without affecting the interfailure time). To our knowledge, our model is the first to consider the extent to which risk preferences of the client and vendor as well as the collaborative nature and observability of investments in system availability should be accounted for in performance-based ITO contracting.

2.3 Model Assumptions

In this section we describe the characteristics of our model in two subsections. First, we specify the assumptions regarding the disruption process and the response capacity investments by the client and vendor. Second, we specify the contract terms and the payoff functions for the client and vendor along with the sequence of events of the contracting game.

2.3.1 Disruption Process and Response Capacity Decisions

We consider a setting in which a client has outsourced the maintenance of a mission critical IT system (hereafter 'the system') to a vendor. The high degree of asset specificity tied to customized solutions prevents the client from multi-sourcing or resorting to alternative options (Chen and Bharadwaj 2009). The system is assumed to be a mission critical system such that the client's entire revenue flow stops any time the system is in a state of disruption. Formally, the client's average revenues, denoted by R, are proportional to the uptime of the system over the contracting period (Kim et al. 2010).

Disruptions to the system arrive following a Poisson process with rate λ. Also, let N be the associated random variable representing the number of failures per period. The failure rate is common knowledge, exogenous and state-independent (Kim and Tomlin 2013). The disruption length D_i of each disruption $i \in [1, N]$ follows an exponential distribution that is stochastically dependent on both parties' investments through the rate parameter μ, where $\mu > 1$. As in Kim and Tomlin (2013), we assume $1/\mu \ll 1/\lambda$, such that no disruptions overlap.

A key characteristic of our model is the complementarity between the vendor's and client's efforts. To capture this complementarity, we use the Cobb-Douglas production function (Bhattacharyya and Lafontaine 1995, Roels et al. 2010). Specifically, we assume that the joint response capacity is given by $\mu = \bar{\mu} x^\alpha y^\beta$, where $\bar{\mu}$ is a scaling factor, x denotes the investment by the client, and y denotes the investment by the vendor $(x, y > 0)$. In the non-observable case, the investments made by the client and vendor are assumed to be non-verifiable and, therefore, non-contractible, which introduces double-sided moral hazard to the problem. In the observable case we assume that the client's efforts are contractible. The exponents α and β reflect the client's and vendor's input elasticity to lowering the expected disruption length, respectively. For instance, a higher α for the client means that her investments are more effective at reducing the expected disruption length. To make sure neither party's investment is trivial and the combined investments have diminishing returns we assume $\alpha, \beta \geq 0$ and $\alpha + \beta \leq 1$ such that there are decreasing returns on investment in response capacity.

Following this, the expected cumulative downtime (hereafter 'expected system downtime') is equal to:

$$\mathbb{E}\left[\sum_{i=1}^{N} D_i | x, y\right] = \frac{\lambda}{\mu} = \frac{\lambda}{\bar{\mu} x^\alpha y^\beta}. \tag{2.1}$$

The cost of investment is assumed to be linear in the level of investment (Corbett et al. 2005, Roels et al. 2010, Kim and Tomlin 2013). The client's and vendor's marginal investment costs are c_c and c_v, respectively. Throughout this chapter the indices $\{c, v\}$ refer to the client and vendor, respectively. Different investment costs could stem from a difference in aptitude in responding to system downtime: Putting the right people and technology in place to respond to disruption is less costly for the party that has more experience in dealing with IT disruptions.

2.3.2 Contract Structure and Sequence of Events

At the outset, the client offers the vendor a take-it-or-leave-it contract. The length of each disruption is observable and verifiable by both parties, and as such, contracts contingent on disruption length are enforceable. Let F be the fixed payment and p, the penalty attributed per unit of expected cumulative disruption time. Then, the transfer payment from the client to the vendor is $T(F, p) = F - p \cdot \sum_{i=1}^{N} D_i$, with an expected value $\mathbb{E}[T(F, p)|x, y] = F - p\lambda/(\bar{\mu} x^\alpha y^\beta)$.

As such, the expected profit functions for the client and the vendor, respectively, are as follows:

$$\mathbb{E}[\Pi_c] = R\left(1 - \frac{\lambda}{\bar{\mu}x^\alpha y^\beta}\right) + \frac{p\lambda}{\bar{\mu}x^\alpha y^\beta} - c_c x - F, \tag{2.2}$$

$$\mathbb{E}[\Pi_v] = F - \frac{p\lambda}{\bar{\mu}x^\alpha y^\beta} - c_v y. \tag{2.3}$$

In the non-observable case (NO), the sequence of events is as follows. In the first stage, the client offers the vendor a contract that outlines the fixed fee, F, and the penalty, p. If the vendor accepts the contract, the client and vendor simultaneously invest in response capacity in the second stage. In the final stage, disruptions and payoffs materialize. In the observable case (O), in addition to the contract payment terms above, the vendor can observe the investment the client makes in response capacity. Formally, we structure the observable case such that, the client's investment is set in the first stage rather than the second stage (as in the non-observable case) and comes with the contract offer. The sequence of events for both contract structures is displayed in Figure 4.1 and notation used in the model is summarized in Table 2.1.

Figure 2.1 Sequence of Events
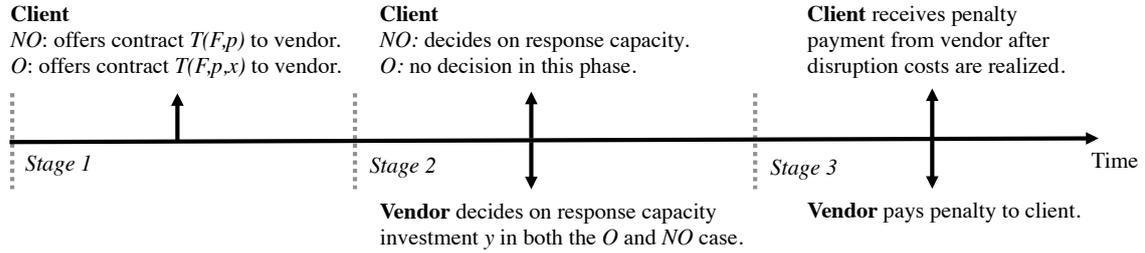
Table 2.1 Summary of notation

R	Client's revenues over an undisrupted contracting period
λ	Disruption arrival rate (Poisson)
x	Client's investment in response capacity
y	Vendor's investment in response capacity
$\bar{\mu}$	Production function scale factor (baseline response capacity)
α	Elasticity of expected disruption length to the client's investment
β	Elasticity of expected disruption length to the vendor's investment
F	Fixed fee
p	Downtime penalty
c_c	Marginal cost of response capacity investment to the client
c_v	Marginal cost of response capacity investment to the vendor

2.4 Analysis

In this section we characterize the optimal contract offered by the client and the equilibrium investments in response capacity (hereafter 'investments') and discuss the implications for the average system uptime and the client's expected utility. We begin with the case where both the client and vendor are risk-neutral. In this setting, we benchmark the equilibrium investment levels of each case (non-observable and observable) against the first-best investment levels that optimize the joint payoff of the two firms (Section 2.4.2). We begin by examining the effect of vendor risk aversion (Section 2.4.3) and move on to examine client risk aversion (Section 2.4.4).

2.4.1 First-best under Risk Neutrality

The first-best levels of investment can be found by maximizing the joint payoff of the client and the vendor. Formally:

$$(\mathcal{D}^{FB}) \qquad \Pi^{FB} = \max_{x,y \geq 0} R \left(1 - \frac{\lambda}{\bar{\mu} x^{\alpha} y^{\beta}} \right) - c_c x - c_v y. \tag{2.4}$$

After solving the first order condition and some algebraic manipulation, the first-best investment levels are:

$$x^{FB} = \left(\frac{\alpha c_v}{\beta c_c} \right)^{\frac{\beta}{\sigma}} \left(\frac{R\alpha\lambda}{\bar{\mu} c_c} \right)^{\frac{1}{\sigma}} \qquad \text{and} \tag{2.5}$$

$$y^{FB} = \left(\frac{\beta c_c}{\alpha c_v} \right)^{\frac{\alpha}{\sigma}} \left(\frac{R\beta\lambda}{\bar{\mu} c_v} \right)^{\frac{1}{\sigma}}, \tag{2.6}$$

where $\sigma = \alpha + \beta + 1$. Using these findings, it is easy to show that:

$$\frac{x^{FB}}{y^{FB}} = \frac{\alpha c_v}{\beta c_c}. \tag{2.7}$$

2.4.2 Contracting under Risk Neutrality

In the non-observable case (NO), the client's investment in response capacity is not observable; only the contract terms are set in the first stage. In the second stage, once the vendor agrees to the contract, both parties simultaneously set their levels of investment. The simultaneous and non-verifiable nature of the investments gives rise to double-sided moral hazard. At the start of the game, the client takes this into account and offers a contract that solves the following

optimization problem:

$$(\mathcal{D}^{NO}) \quad \max_{F,p} \quad R\left(1 - \frac{\lambda}{\bar{\mu}(x^*)^{\alpha}(y^*)^{\beta}}\right) + \frac{p\lambda}{\bar{\mu}(x^*)^{\alpha}(y^*)^{\beta}} - c_c x^* - F,$$

$$\text{s.t.} \quad F - p\frac{\lambda}{\bar{\mu}(x^*)^{\alpha}(y^*)^{\beta}} - c_v y^* \geq k, \tag{IR}$$

$$x^* = \arg\max_{x \geq 0} R\left(1 - \frac{\lambda}{\bar{\mu}(x)^{\alpha}(y^*)^{\beta}}\right) + \frac{p\lambda}{\bar{\mu}(x)^{\alpha}(y^*)^{\beta}} - c_c x - F, \tag{IC1}$$

$$y^* = \arg\max_{y \geq 0} F - \frac{p\lambda}{\bar{\mu}(x^*)^{\alpha}(y)^{\beta}} - c_v y, \tag{IC2}$$

where x^* and y^* denote the equilibrium investments and IR and IC stand for individual rationality and incentive compatibility, respectively. This formulation is analogous to the problem statement used in Bhattacharya et al. (2014), adapted to the assumptions in Section 2.3.2 of this chapter. Proposition 1 characterizes the solution to $\mathcal{D}^{NO}$. All proofs are found in Appendix 2.A.

Proposition 1 (OPTIMAL CONTRACT UNDER NON-OBSERVABILITY AND RISK NEUTRALITY).
(i) In the non-observable case, given the optimal penalty p^{NO} set in Stage 1, the second-stage equilibrium investments are:

$$x^{NO} = \left(\frac{\alpha\lambda(R - p^{NO})}{\bar{\mu}c_c}\right)^{\frac{1+\beta}{\sigma}} \left(\frac{p^{NO}\beta\lambda}{\bar{\mu}c_v}\right)^{\frac{-\beta}{\sigma}}, \qquad y^{NO} = \left(\frac{\alpha\lambda(R - p^{NO})}{\bar{\mu}c_c}\right)^{\frac{-\alpha}{\sigma}} \left(\frac{p^{NO}\beta\lambda}{\bar{\mu}c_v}\right)^{\frac{\alpha+1}{\sigma}},$$

where x^{NO} decreases in p^{NO} and y^{NO} increases in p^{NO}. The optimal penalty satisfies the following optimality condition:

$$\frac{R\lambda(\beta R - (\alpha + \beta)p^{NO})}{\bar{\mu}\sigma p^{NO}(R - p^{NO})(x^{NO})^{\alpha}(y^{NO})^{\beta}} - \frac{c_v(\beta p^{NO}(R - p^{NO}) + \alpha((p^{NO})^2 + (2\beta - 1)p^{NO}R - \beta R^2))}{\beta\sigma(p^{NO})^2(R - p^{NO})y^{NO}} = 0$$

, with $0 < p^{NO} < R$ and where the optimal fixed fee is set such that the IR constraint is binding.

(ii) Let $\alpha = \beta = 0.5$, $c_c = c_v$ and $\bar{\mu} = 1$. Given these symmetry conditions, we find $p^{NO} = R/2$ and $x^{NO} = y^{NO} = x^{FB}/\sqrt{2} = y^{FB}/\sqrt{2}$.

Proposition 1 part (i) shows that x^{NO} and y^{NO} are such that when neither player's actions are verifiable, it is never optimal for the client to transfer all the risk to the vendor by setting $p = R$. Non-observability, and by extension complexity of system integration, therefore necessitates risk sharing: The client never fully transfers the potential cost of downtime to the vendor. In this case, the best the client can do is to use the penalty as a lever to distribute investments according to the respective marginal return on investments for herself and the vendor. In the case of symmetry, it is therefore optimal to share the potential cost of downtime equally by setting $p^{NO} = R/2$.

Part (ii) shows that given symmetric firms and non-observable investment by the client, both the vendor and client underinvest with respect to the first-best by a factor $1/\sqrt{2}$ and produce a second-best outcome. This result is in line with the general finding in Bhattacharyya and

Lafontaine (1995) that a linear penalty yields second-best results under double moral hazard. Next, we examine the case in which the client's investment in response capacity is observable at the outset of the contracting game (Stage 1).

In the observable case (O), only the vendor invests in response capacity in the second stage. In Stage 1, the client offers a contract that specifies the terms F and p as well as her investment x. The client's optimization problem in the observable case ($\mathcal{D}^O$) can be stated as follows:

$$(\mathcal{D}^O) \qquad \max_{x \geq 0, F, p} \qquad R\left(1 - \frac{\lambda}{\bar{\mu} x^\alpha (y^*)^\beta}\right) + \frac{p\lambda}{\bar{\mu} x^\alpha (y^*)^\beta} - c_c x - F,$$

$$\text{s.t.} \qquad F - p\frac{\lambda}{\bar{\mu} x^\alpha (y^*)^\beta} - c_v y^* \geq k, \qquad (IR)$$

$$y^* = \arg\max_{y \geq 0} F - \frac{p\lambda}{\bar{\mu} x^\alpha y^\beta} - c_v y, \qquad (IC)$$

where y^* denotes the equilibrium investment by the vendor in the Stage 2 problem. This problem formulation differs from $\mathcal{D}^{NO}$ in that x is a variable in the upper level of the problem, removing the need for ($IC2$). Proposition 2 gives the solution to $\mathcal{D}^O$.

Proposition 2 (Optimal contract under observability and risk neutrality). *In the observable case, the client sets the optimal penalty $p^O = R$ and fixed fee F such that the IR constraint is binding and the optimal contract guarantees first-best investment levels, i.e. $y^O = y^{FB}$ and $x^O = x^{FB}$.*

Proposition 2 shows that the client can always induce a first-best investment by the vendor by combining a maximum penalty $p^O = R$ with an observable first-best investment by herself. Thus, if the client's investment is observable, the underinvestment described in Proposition 1 is prevented and efficiency is restored.

This result can be understood by noting that when facing a penalty equal to R, the risk-neutral vendor fully internalizes the potential cost of disruption. As such, the vendor acts as if the client's loss of revenue were his own. Although both parties still need to invest, the client's observable investment in response capacity reduces the double-sided moral hazard problem to a one-sided one. Given a first-best investment by the client in Stage 1, the vendor is incentivized to make a first-best investment in Stage 2. In short, as a result of observability, the client reaps all the benefits while fully transferring her exposure to the cost of downtime to the vendor.

This finding is similar to the main result in Demski and Sappington (1991). In this model, both the principal and the agent exert efforts towards improving the profitability of the business. The contract offered by the principal requires the agent to purchase the principal's business, if the agent's investment is low enough, and the business is likely to underperform. The threat of buying a potentially loss-making business prevents the agent from shirking, and instead, the agent invests efficiently. Anticipating this, the principal also invests efficiently and preempts the transfer of ownership, much like the first-best investment made by the client in our model.

In general, throughout our analysis the observable case always dominates the non-observable case from the perspective of the client. That is, the observable case always leads to higher expected profits for the client as it puts her in a Stackelberg leader position and reduces the moral hazard issue from double-sided to single-sided. This, however, does not imply that the first-best efforts are maintained. As we show in the following subsection, when the vendor is risk-averse, observability may not exclusively lead to efficient investments in response capacity.

2.4.3 Contracting under Vendor Risk Aversion

A critical issue with PBCs is that they expose vendors to increased financial risks by linking uncertain outcomes to payment terms (Kim et al. 2010, Jain et al. 2013). Such exposure is particularly important for small- and medium-sized vendors as it can lead to financial distress and, potentially, bankruptcy. Moreover, risk aversion can also be exhibited by large corporations (Ning et al. 2014). To understand the implications of these considerations, we now extend our analysis to allow for vendor risk aversion (VRA); that is, we characterize the optimal contracts and equilibrium investment levels as functions of the vendor's level of risk aversion for both the non-observable and observable cases.

We model risk aversion by capturing the disutility of such volatile returns on the vendor's payoffs: The higher the VRA, the higher the utility loss for the vendor. For tractability reasons we apply the mean-standard deviation (MSD) framework which has the same directional effects on incentives as the mean-variance framework (Pratt 1964, Levy and Markowitz 1979, Samuelson 1986, Van Mieghem 2007).

Let SD[·] be the standard deviation operator and define $P(p,x,y) \equiv p \cdot \sum_{i=1}^{N}[D_i|x,y]$ to reflect the penalty paid over the cumulative downtime by the vendor and let γ_v be the vendor's risk aversion parameter. The parameter γ_v ($\gamma_v > 0$) is a measure of the degree of the risk aversion: The higher γ_v, the more uncertainty there is with regard to the payoff factors into the utility.

Then, we can define the vendor's risk aversion utility function as follows, where we use the subscript vra to refer to the case of vendor risk aversion:

$$\mathbb{U}_{vra}(y) \equiv F - \mathbb{E}[P(p,x,y)] - c_v y - \frac{\gamma_v}{2}\text{SD}[P(p,x,y)], \tag{2.8}$$

where $P(p,x,y)$ contains a compound Poisson random variable.

Lemma 1 (Variance of the compound Poisson process, (Ross 2003)). *The variance of the compound Poisson random variable can be computed: $VAR(X(t)) = \lambda\mathbb{E}[VAR(D_1) + \mathbb{E}[D_1]^2]$. Given $P(p,x,y)$, we find $SD[P(p,x,y)] = p\sqrt{2\lambda}/\bar{\mu}x^\alpha y^\beta$.*

From Lemma 1 it follows that:

$$\mathbb{U}_{vra}(y) = F - p\frac{\lambda}{\bar{\mu}x^\alpha y^\beta} - c_v y - \frac{\gamma_v p\sqrt{2\lambda}}{2\bar{\mu}x^\alpha y^\beta}. \tag{2.9}$$

Also note that the vendor's individual rationality constraint becomes $\mathbb{U}_{vra}(y) \geq k$. Proposition 3 characterizes the solution to the client's optimization problem under VRA in the non-observable case.

Proposition 3 (Optimal contract under non-observability and VRA). *Define* $\Phi \equiv (2\lambda + \gamma_v\sqrt{2\lambda})/(2\lambda)$. *Note that* $\Phi \geq 1$. *Given the optimal penalty* p_{vra}^{NC} *set in Stage 1, the second-stage equilibrium investments are:*

$$x_{vra}^{NO} = x^{NO}(p_{vra}^{NO})\Phi^{\frac{-\beta}{\sigma}}, \qquad y_{vra}^{NO} = y^{NO}(p_{vra}^{NO})\Phi^{\frac{1+\alpha}{\sigma}},$$

where $x^{NO}(\cdot)$ *and* $y^{NO}(\cdot)$ *are the second-stage equilibrium investments in the risk-neutral case (Proposition 1). The optimal penalty* p_{vra}^{NO} *satisfies the following optimality condition:*

$$\frac{\beta c_c \hat{x} - (1+\alpha)c_v \hat{y}}{\sigma \hat{p}} + \frac{(1+\beta)c_c \hat{x} - \alpha c_v \hat{y}}{\sigma(R-\hat{p})} + \frac{R\lambda}{\sigma\bar{\mu}\hat{x}^\alpha\hat{y}^\beta}\left(\frac{\beta}{\hat{p}} - \frac{\alpha}{R-\hat{p}}\right)$$
$$- \frac{(1+\alpha)\gamma_v\sqrt{2\lambda}}{2\sigma\bar{\mu}\hat{x}^\alpha\hat{y}^\beta} - \frac{\alpha\hat{p}\gamma_v\sqrt{2\lambda}}{2(R-\hat{p})\bar{\mu}\hat{x}^\alpha\hat{y}^\beta} = 0, \qquad (2.10)$$

where we use $\hat{p} \equiv p_{vra}^{NO}$, $\hat{x} \equiv x_{vra}^{NO}$, *and* $\hat{y} \equiv y_{vra}^{NO}$ *as a shorthand. Moreover, under symmetry, i.e.,* $\alpha = \beta = 0.5$, $c_c = c_v$ *and* $\bar{\mu} = 1$, *the penalty,* p_{vra}^{NO}, *and the vendor's effort,* y_{vra}^{NO}, *decrease in* γ_v, *while the client effort's,* x_{vra}^{NO}, *increases in* γ_v.

One might expect that a more risk-averse vendor would invest more in response capacity. Yet, Proposition 3 shows that the opposite is true: As γ_v increases, the vendor's investment decreases while the client's investment increases. To see why this happens, note that in our model the investment levels are determined by two effects. The direct effect follows from the vendor's incentive to limit exposure to downtime by investing more in response capacity. This effect is captured through the parameter Φ. The indirect effect follows from the role of the penalty p that the client sets in the contract. As γ_v increases and the vendor becomes more risk-averse, the client has to lower the penalty p and increase the fixed fee F in order to ensure that the vendor's participation constraint (IR) is met. At optimality, the vendor's participation constraint is binding. Importantly, the indirect effect (of the lower penalty) outweighs the direct effect (of the vendor's exposure to downtime), and as a result, a risk-averse vendor underinvests compared to a risk-neutral vendor, i.e. $y_{vra}^{NO} < y^{NO}$ and y_{vra}^{NO} decreases in γ_v. This result is consistent with Jain et al. (2013), who examine the effect of financial distress on the efficiency of linear contracts. They show that as the vendor's financial distress threshold is reduced and bankruptcy becomes more likely, the vendor invests less in response capacity and the performance of the linear contract deteriorates.

At the same time, the client overinvests with respect to the risk-neutral case $x_{vra}^{NO} > x^{NO}$ and x_{vra}^{NO} increases in $\gamma_v > 0$. In essence, the client prefers to invest in response capacity herself rather than incentivize the vendor (through the penalty p) to do so. The latter would require an even higher fixed fee, which the client would find suboptimal to offer. Next, we examine the role of observability of the client's investment when facing a risk-averse vendor.

Proposition 4 (OPTIMAL CONTRACT UNDER OBSERVABILITY AND VRA). *Define* $\eta \equiv (2\lambda + \gamma_v \sqrt{2\lambda})/(2\lambda + (1 + 1/\beta)\gamma_v \sqrt{\lambda(2 + \lambda)})$ *and note that* $\eta \le 1$. *Given the optimal penalty* p^O_{vra} *and optimal client investment* x^O_{vra}, *the vendor's second-stage equilibrium investment is* $y^O_{vra} = y^O \eta^{\frac{1+\alpha}{\sigma}} < y^O$, *where* y^O *is the vendor's investment under risk neutrality (Proposition 2). The client's optimal investment and penalty, set in Stage 1, are* $x^O_{vra} = x^O \eta^{\frac{-\beta}{\sigma}} > x^O$ *and* $p^O_{vra} = 2\beta\lambda R/(2\beta\lambda + (1 + \beta)\gamma_v \sqrt{2\lambda})$, *respectively. Moreover,* x^O_{vra} *increases in* γ_v, *while* p^O_{vra} *and* y^O_{vra} *decrease in* γ_v.

Proposition 4 states that, given VRA, observability does not exclusively lead to the efficient (first-best) investments that we described in the risk-neutral situation examined in Proposition 2. Recall that in Proposition 2, efficiency was achieved in two steps: The client knows her investment is observable and incentivizes the vendor to also invest efficiently by fully transferring the risk to him. The latter is effective when facing a risk-neutral vendor but much less so when facing a risk-averse vendor. In fact, as VRA increases, the penalty p^O_{vra} decreases, indicating that the vendor bears less of the risk. This is optimal for the client because doing otherwise, i.e. transferring more risk to a risk-averse vendor, would require a prohibitively high fixed fee to meet the vendor's participation constraint.

Figure 2.2 illustrates the key results of Propositions 3 and 4. The first plot shows how penalties in the non-observable and observable cases decrease in VRA. The second plot shows the client's and vendor's investments as VRA increases from $\gamma_v = 0$ to $\gamma_v = 2$ in increments of $\Delta = 0.1$. In all figures, the remaining parameters are set to $R = 100$, $\bar{\mu} = 5$, $\lambda = 2$, $k = 0$, $c_c = 1$, $c_v = 1$, $\alpha = 0.5$, $\beta = 0.5$.

So far we have discussed the effect of VRA on the optimal contract and the vendor's and client's equilibrium investments. We now examine the effect of VRA on effect of observability on our key performance metrics: average system uptime (Corollary 1) and the client's expected utility (Corollary 2). To ensure tractability, we present the results for the case of symmetric firms, i.e. $\alpha = \beta = 0.5$ and $c_c = c_v$, below.

Corollary 1 (EXPECTED SYSTEM UPTIME UNDER VRA). *For* $\alpha = \beta = 0.5$ *and* $c_c = c_v$, *under observability (non-observability), the expected system uptime decreases (increases) in the vendor's risk aversion.*

Corollary 1 states that the effect of higher VRA on the average system uptime depends critically on whether the client's investment in response capacity is observable. This is illustrated in the top panel of Figure 2.3, in which we plot the expected system uptime for the observable and non-observable strategies. In the observable case, the expected system uptime decreases in VRA, while it increases in the non-observable case. First, recall that the average system uptime in our model increases in the client's and vendor's investments in response capacity. Also recall that higher VRA leads to higher client investment and lower vendor investment in both the observable (Proposition 4) and non-observable (Proposition 3) cases. In the observable case, however, the

Figure 2.2 Effect of increasing VRA on the optimal penalty and investments
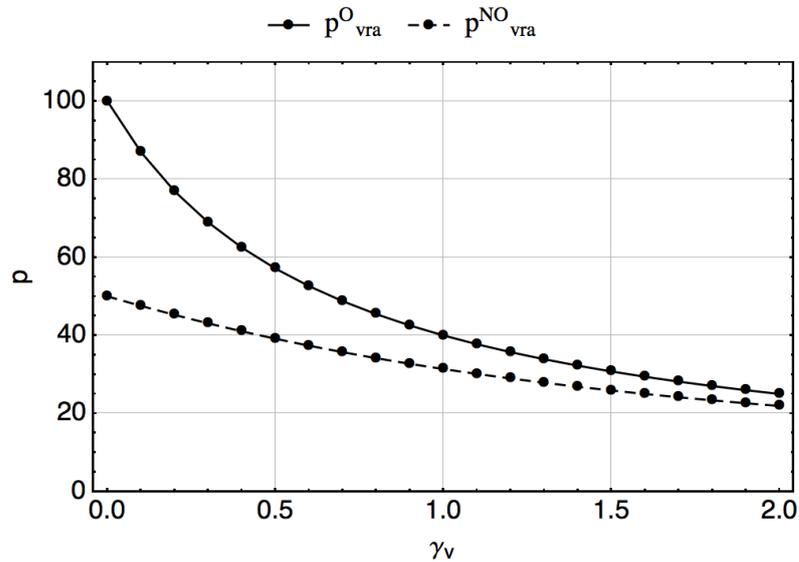
(i) effect of VRA on the optimal penalty
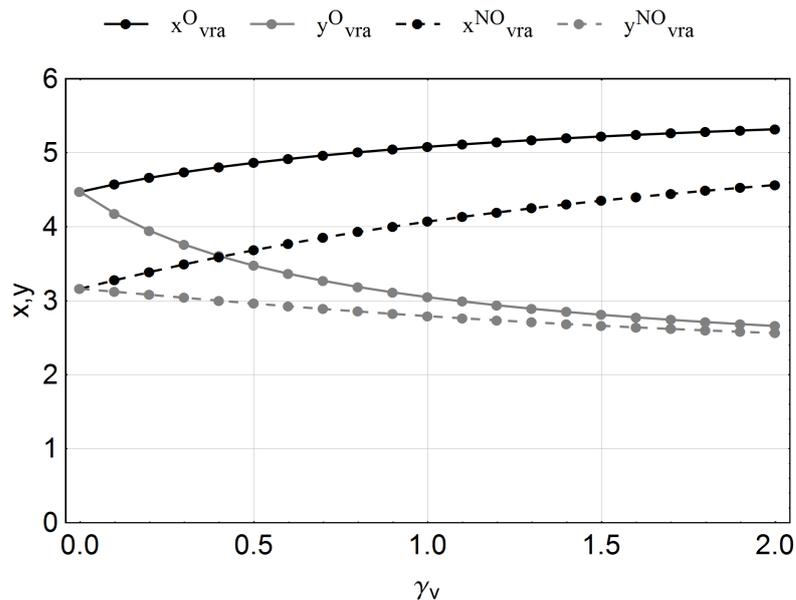
(ii) effect of VRA on the equilibrium investments

drop in the vendor's investment outweighs the increase in the client's investment. As such, the joint response capacity decreases and the average uptime decreases. In the non-observable case, the reverse is true and the average uptime increases.

As VRA increases, the client has to decide whether to overcompensate for the vendor's underinvestment or accept a longer system downtime. In the observable case, the vendor bears most of the cost of downtime through the relatively high penalty. From the client's perspective, compensating for the vendor's underinvestment is not worthwhile. Instead, the client trades off longer downtime (since the vendor is still primarily liable for it) against lower investment in response

capacity. By contrast, in the non-observable case, the penalty for the vendor is much lower and, therefore, the client bears an increasing share of the cost of downtime (as VRA increases). This finding is reflected in an empirical study by Susarla et al. (2010), who demonstrate that for settings with lower interdependency (hence better observability) of efforts, incentives should be more high-powered (hence shifting more risk to the vendor). As such, it is in the client's interest to overcompensate for the vendor's underinvestment and increase the joint response capacity. This is why higher VRA leads to longer uptime in the case of non-observability but longer downtime in the case of observability. Lastly, note that for high VRA, the penalties for the observable and non-observable cases converge and the effect of higher VRA on the average system uptime is marginal.

Corollary 2 (CLIENT'S EXPECTED UTILITY UNDER VRA). *For $\alpha = \beta = 0.5$ and $c_c = c_v$, the difference between the client's expected utility in the observable and non-observable cases decreases as VRA increases.*

The key managerial implication of Corollary 2 is that a lack of observability of the client's investment results in both lower expected uptime and utility for the client. As such there is a cost of complexity, linking complexity of system integration to system performance through the efficacy of collaborative response. One might reason that a observability of resources by the client would be particularly effective in the case of highly risk-averse vendors, who are particularly concerned about the volatility of their payoff. Yet, according to Corollary 2, the cost of complexity for the client diminishes as the vendor becomes more risk-averse. To see why this is the case, recall from Proposition 2 that the observability of the client's investment induces efficient investments by transferring risk from the client to the vendor. As VRA increases, the risk premium that the client needs to pay to the vendor (through the up-front fixed fee) becomes prohibitively expensive and the client instead prefers to bear more of the risk herself. This leads to underinvestment by the vendor and, in turn, to lower expected profits for the client. The bottom panel of Figure 2.3 plots the client's expected profits with respect to VRA (using the same parameter settings as in Figure 2.2.).

Collectively, our results on the effect of VRA indicate that lack of observability of investment is not necessarily an impediment to collaboration, particularly when those resources are not observable to more risk-averse vendors, which are presumably small- and medium-sized businesses.

Figure 2.3 Effect of increasing VRA on the client's average system uptime and expected utility
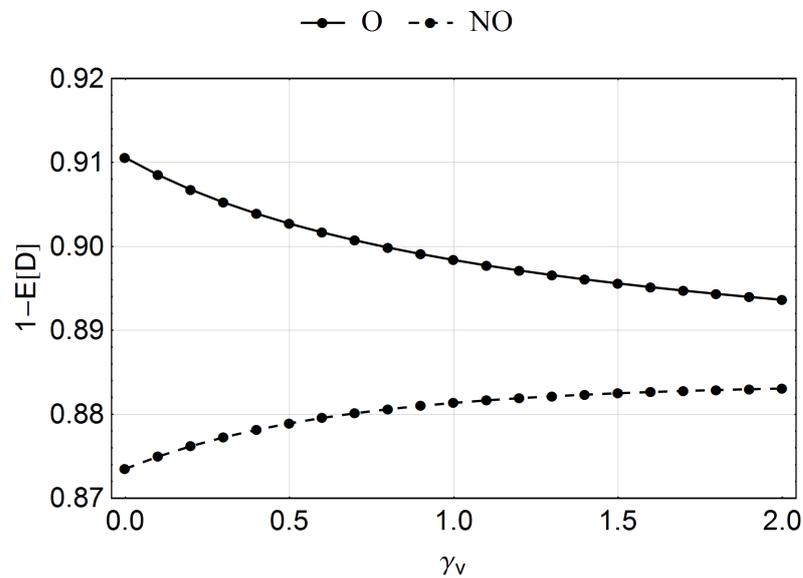
(i) effect of VRA on average system uptime
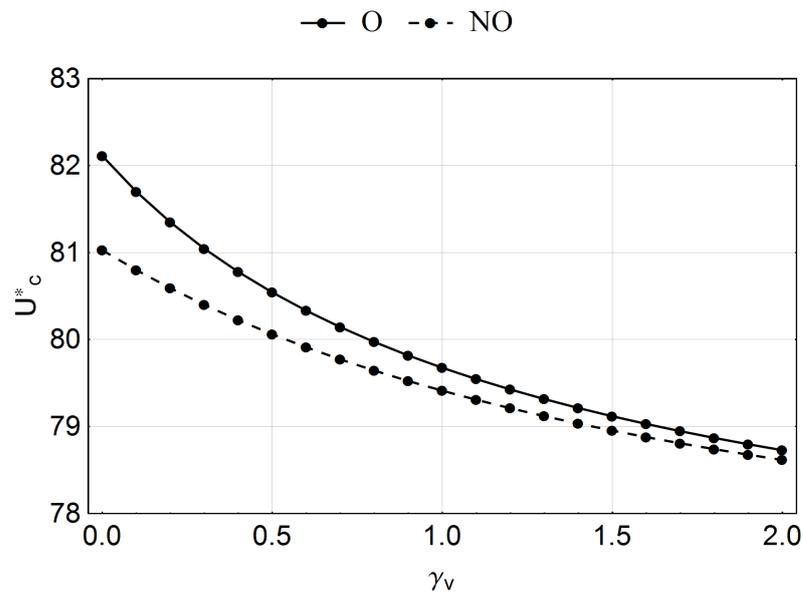
(ii) effect of VRA on the client's expected utility

2.4.4   Contracting under Client Risk Aversion

In many settings it is the client rather than the vendor who is risk-averse. In this section, we study the role of client risk aversion (CRA) on the optimal contract and the investments in response capacity. In line with the previous section, we account for CRA through the MSD framework. Specifically, we define the client's risk aversion utility function as follows, using the index *cra* to refer to the case of client risk aversion:

$$\mathbb{U}_{cra}(F,p,x) \equiv R\left(1 - \frac{\lambda}{\bar{\mu}x^\alpha y^\beta}\right) - c_c x + \frac{p\lambda}{\bar{\mu}x^\alpha y^\beta} - \frac{\gamma_c\sqrt{(R-p)^2}\sqrt{2\lambda}}{2\bar{\mu}x^\alpha y^\beta} - F. \qquad (2.11)$$

In the above formulation, note that the client experiences variance due to the uncertain downtime, which results in loss of revenue R, but this effect is offset by the penalty, p, paid to her by the vendor. The client thus experiences disutility proportional to *effective* standard deviation: $\sqrt{(R-p)^2}\mathrm{SD}\left[\sum_{i=1}^n D_i | x,y\right] \equiv |R-p| \cdot \mathrm{SD}\left[\sum_{i=1}^n D_i | x,y\right]$. The client's optimization problem is to maximize the above utility subject to the vendor's participation constraint and incentive compatability. In the remainder of this section, to disentangle the effect of CRA, we assume a risk-neutral vendor (i.e. $\gamma_v = 0$). Proposition 5 and Proposition 6 present our results for the cases of non-observability and observability, respectively, under CRA.

Proposition 5 (Optimal contract under non-observability and CRA). *Define $\Phi' \equiv (2\lambda + \gamma_c\sqrt{2\lambda})/(2\lambda)$ and note that $\Phi' \geq 1$. Given the optimal penalty p_{cra}^{NO} set in Stage 1, the second-stage equilibrium investments are:*

$$x_{cra}^{NO} = x^{NO}(p_{cra}^{NO})\Phi'^{\frac{1+\beta}{\sigma}}, \qquad y_{cra}^{NO} = y^{NO}(p_{cra}^{NO})\Phi'^{\frac{-\alpha}{\sigma}},$$

where $x^{NO}(\cdot)$ and $y^{NO}(\cdot)$ are the second-stage equilibrium investments in the risk-neutral case (Proposition 1). The optimal penalty satisfies the following optimality condition:

$$\frac{\beta c_c\hat{x} - (1+\alpha)c_v\hat{y}}{\sigma\hat{p}} + \frac{(1+\beta)c_c\hat{x} - \alpha c_v\hat{y}}{\sigma(R-\hat{p})} + \frac{R\lambda}{\sigma\bar{\mu}\hat{x}^\alpha\hat{y}^\beta}\left(\frac{\beta}{\hat{p}} - \frac{\alpha}{R-\hat{p}}\right)$$

$$+ \frac{(1+\beta)\gamma_c\sqrt{2\lambda}}{2\sigma\bar{\mu}\hat{x}^\alpha\hat{y}^\beta} + \frac{\beta\sqrt{(R-\hat{p})^2}\gamma_c\sqrt{2\lambda}}{2\sigma\hat{p}\bar{\mu}\hat{x}^\alpha\hat{y}^\beta} = 0, \quad (2.12)$$

where we use $\hat{p} \equiv p_{cra}^{NO}$, $\hat{x} \equiv x_{cra}^{NO}$, and $\hat{y} \equiv y_{cra}^{NO}$ as a shorthand. Moreover, under symmetry, i.e., $\alpha = \beta = 0.5$, $c_c = c_v$ and $\bar{\mu} = 1$, the penalty, p_{cra}^{NO}, and the vendor's effort, y_{cra}^{NO}, increases in γ_c, while the client's effort, x_{cra}^{NO}, decreases in γ_c.

Proposition 5 illustrates how in the non-observable case, the role of CRA is diametrically opposite to the role of VRA (discussed in Proposition 3). In particular, the inefficiency in the investment levels is now expressed by the client underinvesting and the vendor overinvesting relative to the investments made under risk neutrality. As we see in the top panel of Figure 2.4, as CRA increases, so does the penalty p. By increasing the penalty, the client can offset her exposure to variability by creating an incentive for the vendor to increase his investment. As a result, as γ_c increases, the client's investment decreases and the vendor's increases (bottom panel of Figure 2.4). Next, we examine the effect of CRA in the observable case.

Proposition 6 (Optimal contract under observability and CRA). *In the observable case, the client sets the optimal penalty $p_{cra}^O = R$, which completely offsets the client's exposure to variance and induces first-best investments, i.e. $y_{cra}^O = y^{FB}$ and $x_{cra}^O = x^{FB}$.*

Proposition 6 shows that if the client's investment is observable, CRA is completely inconsequential under vendor risk neutrality; i.e. both firms' investment levels and expected utilities

Figure 2.4 Effect of increasing CRA on the optimal penalty and equilibrium investments
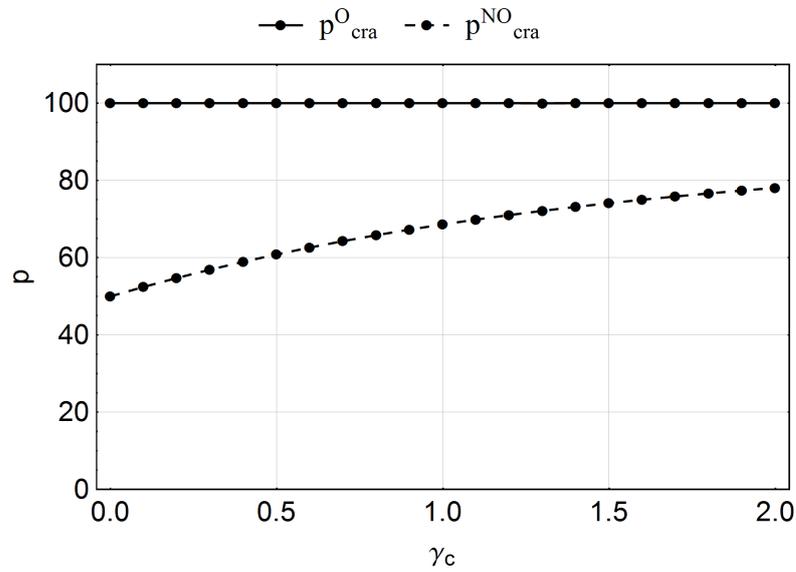
(i) effect of CRA on the optimal penalty
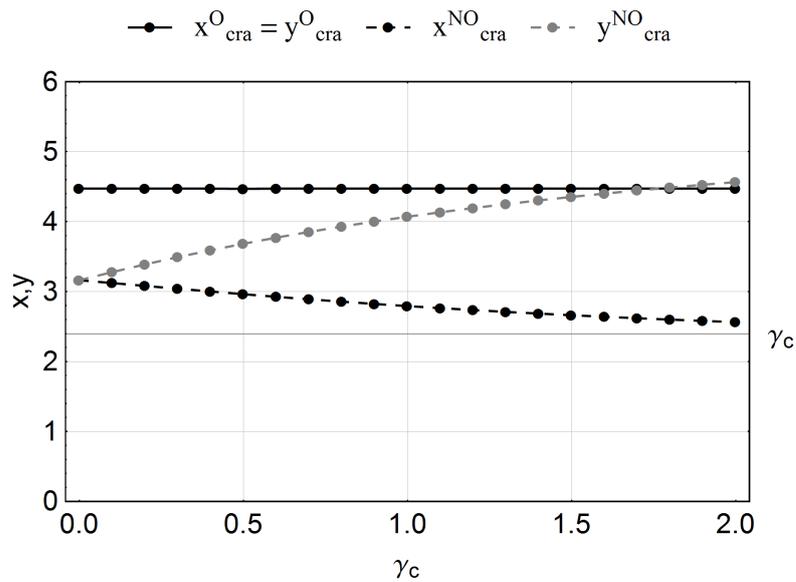
(ii) effect of CRA on the equilibrium investments

are identical to the case of observability under risk-neutral firms. Recall from the discussion following Proposition 2 that by setting $p^O = R$, the client is able to incentivize the vendor to invest efficiently. Moreover, in this case, by setting $p^O_{cra} = R$, the client is able to offset her exposure to volatility fully, i.e. reduce her effective standard deviation to zero. As a result, the investment decisions are invariant in the client's risk aversion. These findings are also illustrated in Figure 2.4.

We now examine the effect of CRA on the average system uptime and the client's expected utility. As in Section 4.2, we present the results for the case of symmetric firms, i.e. $\alpha = \beta = 0.5$ and $c_c = c_v$.

Corollary 3 (EXPECTED UPTIME AND CLIENT'S EXPECTED UTILITY UNDER CRA). *For $\alpha = \beta = 0.5$ and $c_c = c_v$, the difference between the expected system uptime in the observable and non-observable cases decreases as the client's risk aversion increases. However, the difference between the client's expected utilities increases as CRA increases.*

The first part of Corollary 3 shows that in the case of non-observability, even though the client's investment decreases in her risk aversion, the increase in the vendor's investment overcompensates for it, and as a result, the joint response capacity increases. This, leads in turn to higher expected system uptime. Yet, in the observable case, the average system uptime is invariant in CRA. As a result, the cost of complexity on the expected system uptime decreases. The effect is illustrated in the top panel of Figure 2.5 and is in line with the result of Corollary 1.

The second part of Corollary 3 stands in sharp contrast to the corresponding result in Corollary 2. In the non-observable case, higher CRA leads to lower expected utility for the client. Again, in the observable case, the client's expected profits are invariant in her level of risk aversion. As such, higher CRA leads to a widening of the performance gap between the observable and non-observable cases (bottom panel of Figure 2.5). This comparison has an important managerial implication: It is particularly beneficial for a highly risk-averse client to operate in an environment where investments are observable (i.e. limited complexity of system integration) but this becomes fairly inconsequential when the vendor is highly risk-averse.

2.5   Robustness Tests

To demonstrate the generalizability of our main findings in previous section, we present a number results of numerical robustness tests we performed. Firstly, we show that our key directional results regarding the effect of VRA and CRA remain valid under the case where both the client and the vendor are risk averse. Secondly, we relax the assumption of symmetry between the vendor and client in terms of their return on investment in response capacity and show that our results still hold.

2.5.1   Vendor and Client Risk Aversion

So far we have discussed only the effect of one-sided risk aversion, i.e. either vendor risk aversion (VRA) or client risk aversion (CRA). In this section we conduct a numerical analysis to examine the robustness of our main findings when both the client and the vendor are risk-averse. As before, we use γ_v for VRA and γ_c for CRA. When $\gamma_v > \gamma_c$ we say the vendor is relatively risk-averse and when $\gamma_c > \gamma_v$ we say the client is relatively risk-averse.

Figure 2.5 Effect of increasing CRA on the client's expected utility and average system uptime
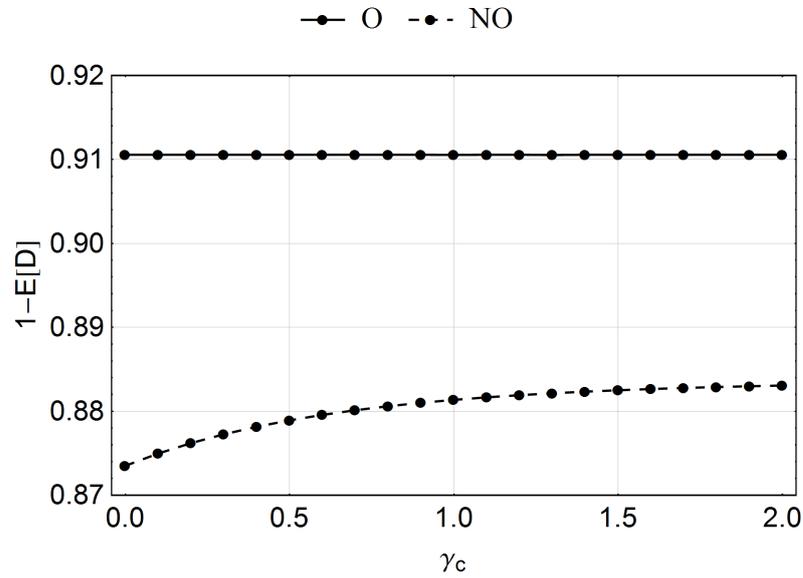
(i) effect of CRA on average system uptime
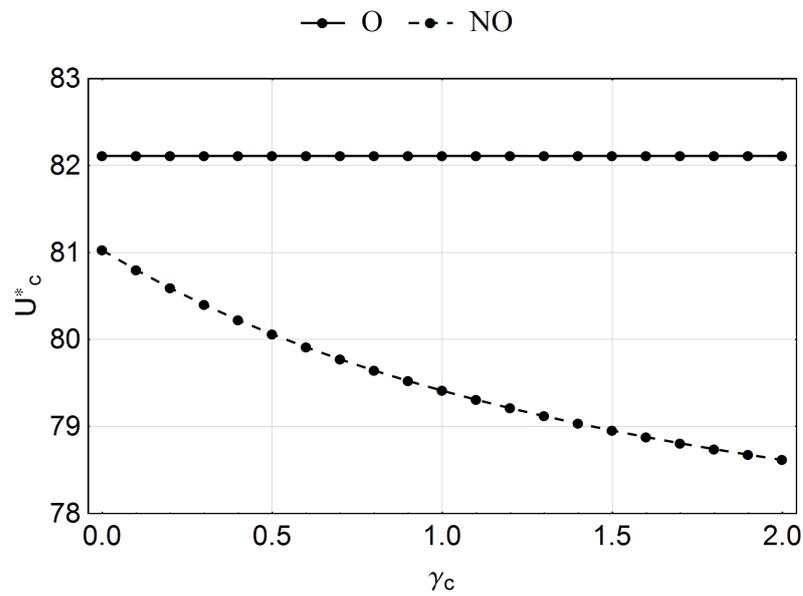
(ii) effect of CRA on the client's expected utility

Figure 2.6, shows the effect of VRA for different levels of CRA, for both the observable and non-observable cases. Conversely, Figure 2.7 shows the effect of CRA for different levels of VRA, for both the observable and non-observable cases. To facilitate comparison, all graphs are plotted using the same parameter values that we used in our earlier analyses, that is, $\alpha = \beta = 0.5$, $c_c = c_v = 1, R = 100, \bar{\mu} = 5, \lambda = 3, k = 0$. As such, the curves in Figure 2.6 for which $\gamma_c = 0$ are the same as in panel (ii) in Figure 2.3. Similarly, the curves in Figure 2.7 for which $\gamma_v = 0$ are the same as in panel (ii) in Figure 2.4.

Figure 2.6 The effect of VRA for different levels of CRA
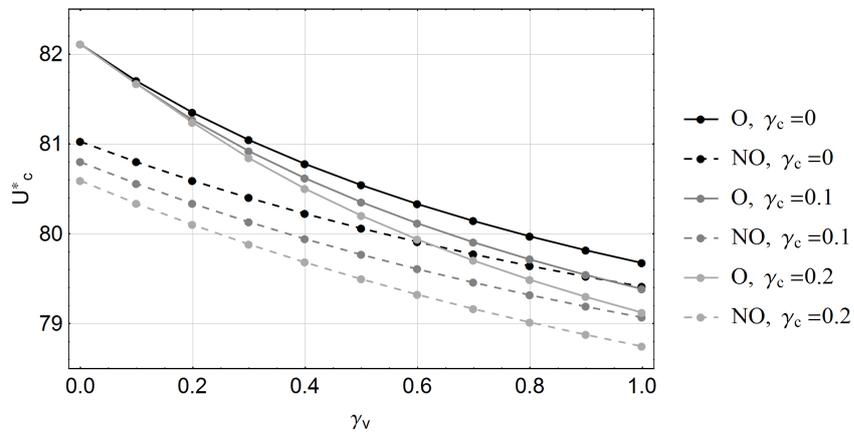
There are two key observations from this analysis. First, for any level of CRA, the directional effect of VRA in Figure 2.6 remains similar to Figure 2.3. That is, as VRA increases, the client's expected utilities for both the observable and non-observable cases decrease, and the difference between them decreases as well. Similarly, for any level of VRA, the directional effect of CRA in Figure 2.7 remains similar to the right-hand panel in Figure 2.5. That is, the performance difference between the observable and non-observable cases decreases in CRA. The effects of vendor and CRA to the rest of our key variables (e.g., the penalty and the equilibrium efforts) are similarly consistent with our earlier analysis. We omit this figures for brevity. Altogether, these results confirm the robustness of our main findings.

Second, when both firms are risk-averse, the client can no longer transfer all the risk to the vendor but instead she needs to balance the shielding effect of the penalty with regard to her own risk aversion with the risk premium that she needs to pay to vendor. This leads to having both parties carrying some of the risk, and a corresponding inefficiency in their investment levels.

Figure 2.7 The effect of CRA for different levels of VRA
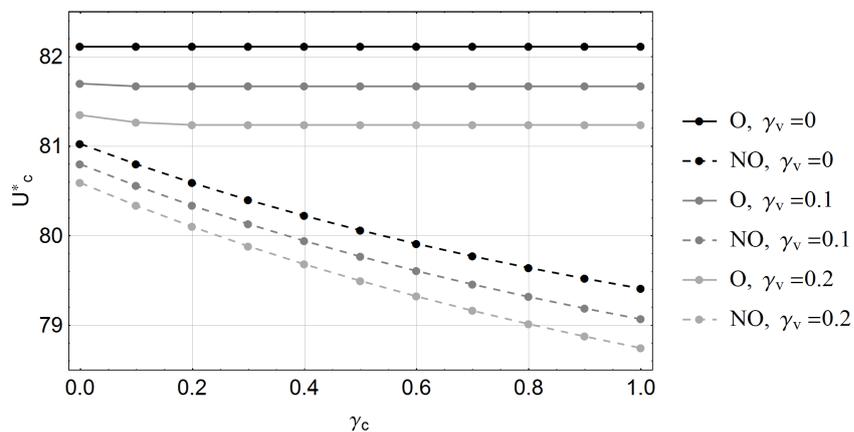
2.5.2   Asymmetric Firms

In the interest of tractability, our analytical results are derived under symmetry between the client and vendor with regard to their return on investment (α and β, respectively, for the client and vendor) as well as their cost of investment (c_c and c_v). In this section, we present a numerical analysis that shows the robustness of our key findings when we allow for asymmetries between the two firms. Specifically, Figure 2.8 below illustrates the effect of VRA on the client's utility (left-hand panel) and the equilibrium investments (right-hand panel) for $\beta > \alpha$ ($\beta = 0.5$ and $\alpha = 0.3$). By the same token, Figure 2.9 plots the same metrics for $\alpha > \beta$ ($\alpha = 0.5$ and $\beta = 0.3$).

As we would expect, a higher β leads to higher investment levels by the vendor, and higher α leads to higher investment levels by the client. Otherwise, all the directional effects of VRA on the client's expected utility and equilibrium investments for asymmetric firms ($\alpha \neq \beta$) remain the same as in the case of symmetric firms ($\alpha = \beta$). The same results hold for asymmetric investment costs ($c_c \neq c_v$), and therefore not presented here for the sake of brevity.

Figure 2.8 Effect of VRA ($\beta > \alpha$)
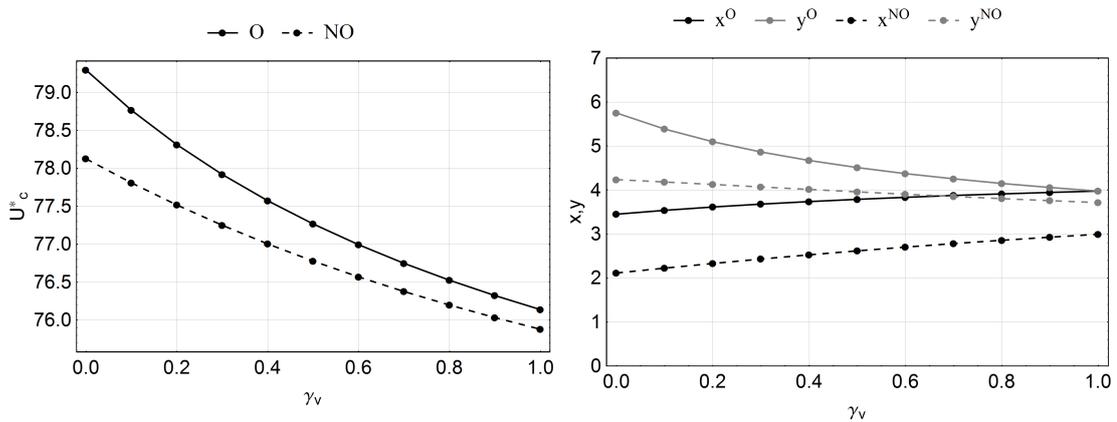
Figure 2.9 Effect of VRA ($\alpha > \beta$)
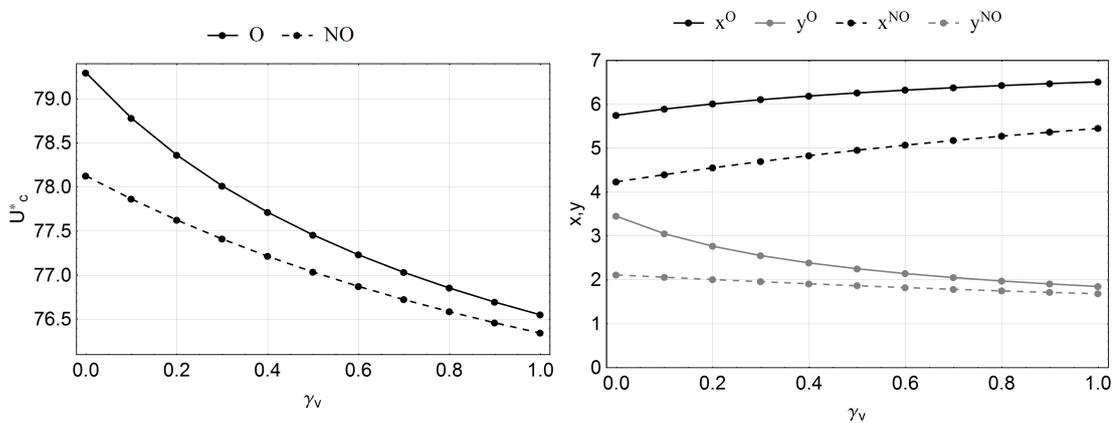
Lastly, we examine the effect of CRA under asymmetric firms. Figure 2.10 and Figure 2.11 below plot the client's expected utility (left-hand panel) and the equilibrium investments (right-hand

panel) for $\beta > \alpha$ and $\alpha > \beta$, respectively. Again, we observe that while the absolute magnitude of the investment levels depends on α and β, all the directional results of the effect of CRA are consistent with our earlier analysis (e.g., see Figure 2.5, panel (b)).

Figure 2.10 Effect of CRA ($\beta > \alpha$)
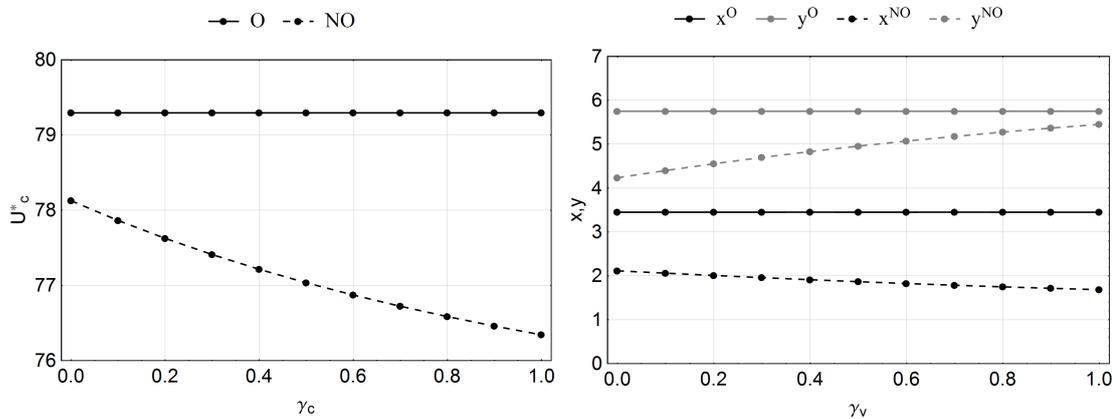
Figure 2.11 Effect of CRA ($\alpha > \beta$)

(i)   (ii)
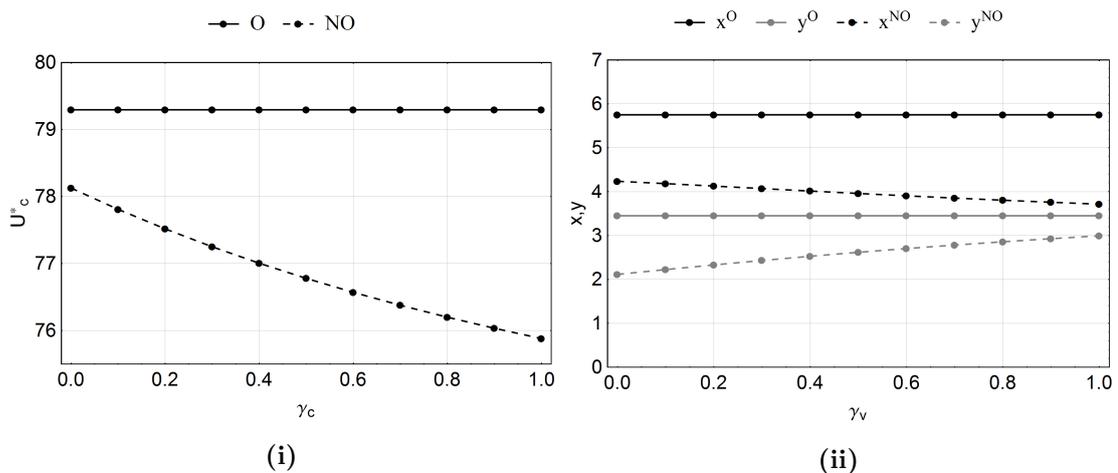
2.6 Conclusions

Responding to mission critical IT system failures often requires a collaborative approach in which both the vendor and client need to invest in response capacity. In such a setting, the client needs to balance two opposing forces. By investing more resources, the client makes the vendor's resources more effective when restoring operations during an outage. At the same time, by providing more resources, the client encourages free-riding by the vendor. To better understand how a client should balance these forces, we develop a novel model that combines the key characteristics of value co-creation (i.e. complementarity between the firms' investments in response capacity) with standard maintenance contract practices (i.e. penalty contracts that penalize the vendor for system downtime). We study the difference in the client's expected

utility between the observable effort case (low complexity) and the non-observable effort case (high complexity). We refer to this difference as the cost of complexity. Our analysis has the following managerial implications.

First, as vendor risk aversion (VRA) increases, the cost of complexity to the client decreases. When facing a risk-neutral vendor, and the clients' effort is observable, the client can perfectly disentangle her investment from the vendor's. By doing so, the client reduces the double-sided moral hazard to single-sided moral hazard. For the latter, an appropriately designed performance-based contract can achieve efficient levels of investments. In essence, the client's resources do not only provide support to the vendor, but also act as a risk transfer mechanism. As VRA increases, it becomes increasingly expensive for the client to transfer risk to the vendor through such a contract. As such, the client prefers to keep more risk herself and the advantage of observability becomes less relevant. A key implication of this result is that the cost of complexity is lower when the client is working with small- and medium-sized vendors, who are likely to be more risk-averse.

Second, we show that the cost of complexity to the client increases as the client becomes more risk-averse. In fact, in the observable case, the client's expected utility is invariant in her level of risk aversion as all the risk can be effectively transferred to the vendor. In the non-observable case, however, as client risk aversion (CRA) increases, it becomes more imperative for the client to transfer risk to the vendor. The only means of doing so is to increase the penalty fee that the vendor incurs. However, this risk transfer is costly to the client as it requires the client to pay higher up-front fees to the vendor. This result illustrates that it is highly risk-averse clients who are set to benefit the most from operating in an environment where their own investment in response capacity is observable.

Taken together, the previous results also suggest that when the client is highly risk-averse, she has a lot to gain by making her efforts more observable (e.g., by investing in monitoring mechanisms or processes and systems that make her efforts more transparent). On the contrary, when the client is working with a highly risk-averse vendor, investing in increasing her observability of efforts will not have a substantial impact on her profits. Thus, the relative risk preferences of the two firms play a critical role in the extent to which asymmetric information affects the efficiency of the collaborative process.

Third, we show that higher VRA leads to higher system downtime in the observable case, but lower system downtime in the non-observable case. As VRA increases, the vendor invests less and the client has to decide whether to overcompensate for the vendor's weaker efforts or accept a longer system downtime. In the observable case, the vendor incurs most of the downtime cost due to the high penalty fees. As such, the client trades off longer downtime against lower investment in response capacity. By contrast, in the non-observable case, the client cannot transfer most of the risk to the vendor, and as VRA increases, the cost of downtime to the client increases as well. This is why, when the client's investment in response capacity is not

observable, she prefers to overcompensate for the vendor's lower efforts by investing even more herself, and the combined effect is lower expected system downtime. Notably, higher uptime, which sounds positive, is actually a signal of inefficiency due to higher complexity associated with the non-observable case. In short, the effect of risk aversion on the average system downtime is diametrically opposite depending on whether or not the client's investment is observable.

These three main findings do not come without limitations. In particular, concerns may be raised surrounding a number of the core assumptions to the model which may not find reflection in reality. Firstly, we built our model on the assumption of both a one-shot contract and static investment in response capacity. These are assumptions made elsewhere in the literature. For example, Kim et al. (2007) give an example of the US Department of Defense offering one-shot contracts on an annual basis despite the lifetime of technology involved. In addition, preparedness to respond to disruption is likely to require a sizeable upfront and non-reversible investment, in turn meaning dynamic adjustments of the investment levels may not be feasible. Nonetheless, in the context of IT outsourcing subject to both frequent disruption as well as significant technological advances, both investment and contract terms may be more reactionary to disruptions. Moreover, we assume whether or not the client's investment is observable and verifiable, and hence the system complexity, is given exogenously. If the technology is indeed critical to the client, then there is reason for the client to decide on the architecture with the dependence on the vendor's efforts in disruption response in mind. In other words, ensuring verifiability of her resource contribution may be a client-side decision that is endogenous to the contracting game. Broadly, inefficient decisions following from differences in risk aversion are likely to persist even under dynamic decision making and endogenous verifiability of investments may only be worthwhile so long as acting on it does not present immediate costs that are higher than discounted expected cost of future disruptions under limited verifiability of investments.

To conclude, our work shows that improving the efficiency of a collaborative response requires a clear understanding of two key parameters: the nature of the investments in response capacity that are required (whether or not the client's investment can be disentangled from the vendor's); and the risk attitudes of each firm. Our analysis highlights that the implications of these two parameters cannot be considered in isolation and instead must be jointly considered.

Appendix

2.A   Proofs

Proof of Proposition 1. Part (i) We first state the expected profit functions for the client and the vendor:

$$\mathbb{E}[\Pi_c] = R\left(1 - \frac{\lambda}{\bar{\mu}x^\alpha y^\beta}\right) + \frac{p\lambda}{\bar{\mu}x^\alpha y^\beta} - c_c x - F, \qquad \mathbb{E}[\Pi_v] = F - \frac{p\lambda}{\bar{\mu}x^\alpha y^\beta} - c_v y. \tag{2.13}$$

We find the solution to this problem through backward induction, starting with a solution to the Nash subgame in stage 2. The firms' best response functions are:

$$x(y) = \left(\frac{\alpha\lambda(R-p)}{\bar{\mu}c_c y^\beta}\right)^{\frac{1}{\alpha+1}}, \qquad y(x) = \left(\frac{p\beta\lambda}{\bar{\mu}c_v x^\alpha}\right)^{\frac{1}{\beta+1}}. \tag{2.14}$$

From the second-order derivatives on the expected profit function for both the client we can verify both independently have a unique solution for any decision of the other player, as long as $0 < p < R$:

$$\frac{\partial^2\mathbb{E}[\Pi_v]}{\partial y^2} = \frac{-(1+\beta)\beta p\lambda}{\bar{\mu}y^{2+\beta}x^\alpha} < 0, \qquad \frac{\partial^2\mathbb{E}[\Pi_c]}{\partial x^2} = \frac{-\alpha(1+\alpha)(R-p)\lambda}{\bar{\mu}x^{2+\alpha}y^\beta} < 0. \tag{2.15}$$

Simultaneously solving the FOCs gives the second stage equilibrium investments, y^N and x^{NO}:

$$y^{NO} = \left(\frac{\alpha\lambda(R-p)}{\bar{\mu}c_c}\right)^{\frac{-\alpha}{\sigma}} \left(\frac{p\beta\lambda}{\bar{\mu}c_v}\right)^{\frac{\alpha+1}{\sigma}}, \qquad x^{NO} = \left(\frac{\alpha\lambda(R-p)}{\bar{\mu}c_c}\right)^{\frac{\beta+1}{\sigma}} \left(\frac{p\beta\lambda}{\bar{\mu}c_v}\right)^{\frac{-\beta}{\sigma}}, \tag{2.16}$$

where $\sigma = 1 + \alpha + \beta$. From the above we can see that x^{NO} decreases in p and y^{NO} increases in p.

At optimality, the vendor's individual rationality constraint is always binding, so the client sets the fixed fee $F^{NO} = \frac{p^{NO}\lambda}{\bar{\mu}(y^{NO})^\beta(x^{NO})^\alpha} + c_v y^{NO} + k$. Subsituting F^{NO}, x^{NO} and y^{NO} into the client's expected profit function gives an updated objective function for the stage one problem:

$$\mathbb{E}[\Pi_c] = R\left(1 - \frac{\lambda}{\bar{\mu}(x^{NO})^\alpha(y^{NO})^\beta}\right) - c_c x^{NO} - c_v y^{NO}. \tag{2.17}$$

Taking the derivative with respect to p gives the optimality condition for p^{NO}

$$\frac{R\lambda(\beta R - (\alpha+\beta)p^{NO})}{\bar{\mu}\sigma p^{NO}(R-p^{NO})(x^{NO})^\alpha(y^{NO})^\beta} - \frac{c_v(\beta p^{NO}(R-p^{NO}) + \alpha((p^{NO})^2 + (2\beta-1)p^{NO}R - \beta R^2)}{\beta\sigma(p^{NO})^2(R-p^{NO})y^{NO}} = 0. \tag{2.18}$$

This optimality condition does not have a closed-form solution, however, numerical analysis shows that there exists a unique p that maximizes the client's expected profit. We omit the results of this analysis for the sake of brevity.

Part (ii) Let $\alpha = \beta = 0.5$, $c_v = c_c = c$ and $\bar{\mu} = 1$. First note that the total investment cost is minimized for $p = R/2$. By part (i), the equilibrium investments x^{NO} and y^{NO} are unique and given $c_v = c_c = c$, the total investment cost at optimality is $C^{NO}(p) \equiv c(x^{NO}(p) + y^{NO}(p))$ with:

$$\frac{\partial C^{NO}}{\partial p} = \frac{c_v R(R - 2p)\left(\frac{p\lambda}{c\bar{\mu}}\right)^{3/4}}{4\sqrt{2}p^2(p - R)\left(\frac{(R-p)\lambda}{c\bar{\mu}}\right)^{1/4}}, \tag{2.19}$$

$$\frac{\partial^2 C^{NO}}{\partial p^2} = \frac{c_v R(12p^2 - 12pR + 5R^2)\left(\frac{p\lambda}{c\bar{\mu}}\right)^{3/4}}{16\sqrt{2}p^3(p - R)^2\left(\frac{(R-p)\lambda}{c\bar{\mu}}\right)^{1/4}} > 0. \tag{2.20}$$

The first order derivative given by Equation 2.19 is zero at $p = R/2$. As the second order derivative in Equation 2.20 is positive everywhere, C^{NO} is strictly convex and cost is minimized for $p = R/2$.

Next we show the product of the investments is maximized for $p = R/2$. Let $H^{NO} \equiv x^{NO}(p)y^{NO}(p)$. Then,

$$\frac{\partial H^{NO}}{\partial p} = \frac{(2p - R)\sqrt{x^{NO}y^{NO}}}{4p(p - R)}, \tag{2.21}$$

$$\frac{\partial^2 H^{NO}}{\partial p^2} = -\frac{(2p^2 - 2pR + 3R^2/2)\sqrt{x^{NO}y^{NO}}}{8p^2(p - R)^2} < 0. \tag{2.22}$$

The first order derivative given by Equation 2.21 is zero at $p = R/2$. As the second order derivative in Equation 2.22 is negative everywhere, H^{NO} is strictly concave and it is maximized for $p = R/2$. By maximizing H^{NO}, the expected uptime $1 - \lambda/(\bar{\mu}\sqrt{x^{NO}y^{NO}})$ is maximized. Setting $p = R/2$ thus simultaneously maximizes H^{NO} and minimizes C^{NO}, which means $\mathbb{E}[\Pi_c]$ in Equation 2.17 is maximized. Under symmetry we have $x^{FB} = y^{FB} = \left(\frac{R\lambda}{2c}\right)^{\frac{1}{2}}$, thus $x^{NO} = x^{FB}/\sqrt{2} < x^{FB}$ and equivalently $y^{NO} = y^{FB}/\sqrt{2} < y^{FB}$. $\square$

Proof of Proposition 2. The expected profit for the vendor in case of a linear penalty contract is:

$$\mathbb{E}[\Pi_v] = F - \frac{p\lambda}{\bar{\mu}x^\alpha y^\beta} - c_v y. \tag{2.23}$$

We first solve the FOC for y, which gives the vendor's best response function:

$$y(x) = \left(\frac{p\beta\lambda}{\bar{\mu}c_v x^\alpha}\right)^{\frac{1}{1+\beta}}. \tag{2.24}$$

Showing that the second order derivative for the vendor expected profit function is negative indicates the solution for y in this stage is unique given $p > 0$ and $x > 0$:

$$\frac{\partial^2 \mathbb{E}[\Pi_v]}{\partial y^2} = \frac{-(1 + \beta)\beta p\lambda}{\bar{\mu}y^{2+\beta}x^\alpha} < 0. \tag{2.25}$$

At optimality the vendor's IR constraint is always binding, the client sets the fixed fee $F^O = \frac{p^O \lambda}{\bar{\mu}(y^O)^\beta (x^O)^\alpha} + c_v y^O + k$. Substituting F^O and $y(x)$ into the client's expected profit function and taking the partial derivative with respect to x produces:

$$\frac{\partial \mathbb{E}[\Pi_c]}{\partial x} = \frac{\alpha \lambda (\beta p + R)}{(1+\beta)\bar{\mu} x^{1+a}} \left(\frac{c_v \bar{\mu} x^\alpha}{\beta p \lambda} \right)^{\frac{\beta}{1+\beta}} - c_c. \tag{2.26}$$

Note that for $p = R$, we can solve for x to find:

$$x^O = \left(\frac{c_v \alpha}{\beta c_c} \right)^{\frac{\beta}{\sigma}} \left(\frac{\alpha \lambda R}{c_c \bar{\mu}} \right)^{\frac{1}{\sigma}}, \tag{2.27}$$

retrieving the first best investment by the client, i.e. $x^O = x^{FB}$. Here again $\sigma = 1 + \alpha + \beta$. Given that the best response function Equation 3.7 is the same as found when solving $\mathcal{D}^{FB}$, the first best investment level is retrieved for y^O as well, such that

$$y^O = \left(\frac{\beta c_c}{\alpha c_v} \right)^{\frac{\alpha}{\sigma}} \left(\frac{R \beta \lambda}{\bar{\mu} c_v} \right)^{\frac{1}{\sigma}}. \tag{2.28}$$

Given that the client can extract all rents by adjusting the fixed fee in the contract and knowing that the first best investments yield the best possible system profit, it follows that setting $p = R$ is indeed optimal. Moreover, we find that as all parameters are non-negative, $p \leq R$ and $0 < \beta < 1$ it holds that:

$$\frac{\partial^2 \Pi_c}{\partial p^2} = \frac{c_v(\beta(p - 2R) - R)}{(1+\beta)^2 p^3 \left(\frac{c_v \bar{\mu} x^\alpha}{\beta p \lambda} \right)^{\frac{1}{1+\beta}}} < 0, \tag{2.29}$$

$$\frac{\partial^2 \Pi_c}{\partial x^2} = \frac{-\alpha \sigma \lambda (\beta p + R) \left(\frac{c_v \bar{\mu} x^\alpha}{\beta p \lambda} \right)^{\frac{\beta}{1+\beta}}}{(1+\beta)^2 \bar{\mu} x^{2+\alpha}} < 0, \tag{2.30}$$

$$\frac{\partial^2 \Pi_c}{\partial p \partial x} = \frac{\alpha c_v (p - R)}{(1+\beta)^2 p^2 x \left(\frac{c_v \bar{\mu} x^\alpha}{\beta p \lambda} \right)^{\frac{1}{1+\beta}}} < 0. \tag{2.31}$$

The determinant of the Hessian matrix is:

$$\begin{vmatrix} \frac{\partial^2 \Pi_c}{\partial p^2} & \frac{\partial^2 \Pi_c}{\partial p \partial x} \\ \frac{\partial^2 \Pi_c}{\partial p \partial x} & \frac{\partial^2 \Pi_c}{\partial x^2} \end{vmatrix} = -\frac{\alpha \sigma c_v^2 (\beta(p - 2R) - R)(\beta p + R)}{\beta(1+\beta)^4 p^4 x^2 \left(\frac{c_v \bar{\mu} x^\alpha}{\beta p \lambda} \right)^{\frac{2}{1+\beta}}} - \frac{\alpha^2 c_v^2 (p - R)^2}{(1+\beta)^4 p^4 x^2 \left(\frac{c_v \bar{\mu} x^\alpha}{\beta p \lambda} \right)^{\frac{2}{1+\beta}}}. \tag{2.32}$$

As $p \leq R$, both terms on the RHS are positive. Then, as $0 < \alpha, \beta \leq 1$ and $(\beta(p - 2R) - R)(\beta p + R) = R^2 + \beta^2(2R^2 + 2pR - p^2) > (p - R)^2$, the second term on the RHS is smaller than the first term, such that the determinant is positive. Together with the fact that $\frac{\partial^2 \Pi_c}{\partial p^2}$ is negative, this means the Hessian matrix is negative-definite, which proves joint concavity in x and p. $\square$

Proof of Lemma 1. The following is based on elements of Chapter 5 in (Ross 2003) and adapted to the assumptions in our model. A stochastic process $\{N(t), t \geq 0\}$ is called a counting process if $N(t)$ represents the total number of events that occur by time t. Particularly, the counting process must satisfy:

1. $N(t) \leq 0$

2. $N(t)$ is integer valued

3. If $s < t$, then $N(s) \leq N(t)$

4. for $s < t$, $N(t) - N(s)$ equals the number of events in the interval $(s, t]$.

The Poisson process is a special case of a counting process which adheres to the following properties. A counting process $\{N(t), t \geq 0\}$ is Poisson with rate λ, $\lambda > 0$, if:

1. $N(0) = 0$

2. The increments of the process are independent, i.e. the number of events in an interval is independent of the interval.

3. The number of events in the interval is distributed with mean λt.

This means, $\forall s, t \geq 0$, it holds that $\mathbb{P}\{N(t+s) - N(s) = n\} = e^{-\lambda t} \frac{(\lambda t)^n}{n!}$, where $n = 0, 1, \ldots$. In words, the interarrival time is found to be distributed i.i.d. with an exponential distribution. To find the characteristics of the Compound Poisson process that is central to the model in this chapter, we need a number of additional assumptions. The stochastic process $\{X(t), t \geq 0\}$ is a Compound Poisson process if:

$$X(t) = \sum_{i=1}^{N(t)} Y_i, \quad t \geq 0, \tag{2.33}$$

where $\{N(t), t \geq 0\}$ is a Poisson process and $\{Y_i, i \geq 1\}$ is a family of i.i.d. random variables also independent of the Poisson process. Specifically, $X(t)$ is here called the Compound Poisson random variable, which has moments like any other random variable. The mean and variance of a Compound Poisson random variable are:

$$\mathbb{E}[X(t)] = \lambda t \mathbb{E}[Y_i] = \mathbb{E}[N]\mathbb{E}[Y_i], \tag{2.34}$$

$$\text{VAR}(X(t)) = \lambda t \mathbb{E}[(Y_i)^2]$$

$$= \lambda t \mathbb{E}[\text{VAR}[Y_i] + \mathbb{E}[Y_i]^2], \tag{2.35}$$

which holds for any i as all Y_i are i.i.d. To find Equation 2.35, we require computation of variance by conditioning. We know $\text{VAR}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$. Then, note:

$$\mathbb{E}[\text{VAR}(X|Y)] = \mathbb{E}[\mathbb{E}[X^2|Y] - (\mathbb{E}[X|Y])^2]$$

$$= \mathbb{E}[\mathbb{E}[X^2|Y]] - \mathbb{E}[(\mathbb{E}[X|Y])^2]$$

$$= \mathbb{E}[X^2] - \mathbb{E}[(\mathbb{E}[X|Y])^2], \tag{2.36}$$

where the second equality holds as, by definition, $\mathbb{E}[X] = \sum_y \mathbb{E}[X|Y = y]\mathbb{P}\{Y = y\}$. Moreover, we need:

$$\text{VAR}(\mathbb{E}[X|Y]) = \mathbb{E}[(\mathbb{E}[X|Y])^2] - (\mathbb{E}[\mathbb{E}[X|Y]])^2$$

$$= \mathbb{E}[(\mathbb{E}[X|Y])^2] - (\mathbb{E}[X])^2, \tag{2.37}$$

such that:

$$\mathbb{E}[\text{VAR}(X|Y)] + \text{VAR}(\mathbb{E}[X|Y]) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \text{VAR}(X). \tag{2.38}$$

Now for the Compound Poisson random variable $S = \sum_{i=1}^{N} X_i$, where $X_i \sim F(\mu, \sigma^2)$ and the length of the interval is normalized to 1, we have $\mathbb{E}[S|N=n] = n\mu$ and:

$$\begin{aligned}
\text{VAR}(S|N=n) &= \text{VAR}\left(\sum_{i=1}^{N} X_i | N = N\right) \\
&= \text{VAR}\left(\sum_{i=1}^{n} X_i | N = n\right) \\
&= \text{VAR}(nX_i) \\
&= n\sigma^2.
\end{aligned} \tag{2.39}$$

Combining the results above we can compute:

$$\begin{aligned}
\text{VAR}(S) &= \mathbb{E}[\text{VAR}(S|N)] + \text{VAR}(\mathbb{E}[S|N]) \\
&= \mathbb{E}[N\sigma^2] + \text{VAR}(N\mu) \\
&= \sigma^2 \mathbb{E}[N] + \mu^2 \text{VAR}(N).
\end{aligned} \tag{2.40}$$

In the context of after-sales disruptions we can model the arrival rate of failures using a Poisson process, with $\mathbb{E}[N] = \lambda$. The length of each individual disruption can be modelled using an exponential disruption, with $\mathbb{E}[X] = \mu^{-1}$ and $\text{VAR}(X) = \mu^{-2}$. Combining the two characteristics we find a Compound Poisson random variable S, which, given Equation 2.40, has the following variance:

$$\text{VAR}(S) = \lambda(\mu^{-2} + \mu^{-2}) = \frac{2\lambda}{\mu^2}. \tag{2.41}$$

$\square$

Proof of Proposition 3. We first state the expected utility for the client and vendor:

$$\mathbb{E}[\mathbb{U}_c] = R\left(1 - \frac{\lambda}{\bar{\mu}x^\alpha y^\beta}\right) - c_c x + \frac{p\lambda}{\bar{\mu}x^\alpha y^\beta} - F, \qquad \mathbb{E}[\mathbb{U}_v] = F - \frac{p\lambda}{\bar{\mu}x^\alpha y^\beta} - c_v y - \frac{p\gamma_v \sqrt{2\lambda}}{2\bar{\mu}x^a y^b}. \tag{2.42}$$

Next we find the best response functions given the penalty set in stage one:

$$x(y|p) = \left(\frac{\alpha\lambda(R-p)}{\bar{\mu}c_c y^\beta}\right)^{\frac{1}{1+\alpha}}, \qquad y(x|p) = \left(\frac{\beta p(2\lambda + \gamma_v\sqrt{2\lambda})}{2\bar{\mu}c_v x^\alpha}\right)^{\frac{1}{1+\beta}}. \tag{2.43}$$

Substituting $x(y|p)$ into $y(x|p)$, we find the vendor's optimal investment as a function of p:

$$y_{vra}^{NO}(p) \equiv \left(\frac{\beta p(2\lambda + \gamma_v\sqrt{2\lambda})}{2\bar{\mu}c_v\left(\frac{\alpha\lambda(R-p)}{\bar{\mu}c_c y^\beta}\right)^{\frac{\alpha}{1+\alpha}}}\right)^{\frac{1}{1+\beta}} = y^{NO}(p)\Phi^{\frac{1+\alpha}{\sigma}}, \tag{2.44}$$

where $\Phi \equiv \frac{2\lambda + \gamma_v\sqrt{2\lambda}}{2\lambda}$ and the finding for y^{NO} in Proposition 1. Similarly we can find: $x_{vra}^{NO}(p) = x^{NO}(p)\Phi^{\frac{-\beta}{\sigma}}$.

At optimality, the IR constraint of the vendor is binding and accounts for the the second stage equilibrium investments, i.e. $F_{vra}^{NO} = \frac{p_{vra}^{NO}(2\lambda + \gamma_v\sqrt{2\lambda})}{2\bar{\mu}(x_{vra}^{NO})^\alpha(y_{vra}^{NO})^\beta} + c_v y_{vra}^{NO} + k$. Substituting F_{vra}^{NO}, y_{vra}^{NO} and x_{vra}^{NO} into the client's expected utility function gives the client's objective function for the stage one problem:

$$\mathbb{E}[\mathbb{U}_c] = R\left(1 - \frac{\lambda}{\bar{\mu}(x_{vra}^{NO})^\alpha(y_{vra}^{NO})^\beta}\right) - c_c x_{vra}^{NO} - c_v y_{vra}^{NO} - \frac{\gamma_v p\sqrt{2\lambda}}{2\bar{\mu}(x_{vra}^{NO})^\alpha(y_{vra}^{NO})^\beta}. \quad (2.45)$$

Solving the FOC, using shorthand $\hat{p} \equiv p_{vra}^{NO}$, $\hat{x} \equiv x_{vra}^{NO}$ and $\hat{y} \equiv y_{vra}^{NO}$, and finally simplifying gives the optimality condition for p_{vra}^N:

$$\frac{\beta c_c \hat{x} - (1+\alpha)c_v \hat{y}}{\sigma\hat{p}} + \frac{(1+\beta)c_c\hat{x} - \alpha c_v\hat{y}}{\sigma(R-\hat{p})} + \frac{R\lambda}{\sigma\bar{\mu}\hat{x}^\alpha\hat{y}^\beta}\left(\frac{\beta}{\hat{p}} - \frac{\alpha}{R-\hat{p}}\right)$$
$$- \frac{(1+\alpha)\gamma_v\sqrt{2\lambda}}{2\sigma\bar{\mu}\hat{x}^\alpha\hat{y}^\beta} - \frac{\alpha\hat{p}\gamma_v\sqrt{2\lambda}}{2(R-\hat{p})\bar{\mu}\hat{x}^\alpha\hat{y}^\beta} = 0. \quad (2.46)$$

As before, this optimality condition does not have a closed-form solution, however, numerical analysis again shows that there exists a unique $\hat{p}$ that maximizes the client's expected profit. First we show that $p_{vra}^{NO}(\gamma_v)$ decreases in γ_v for $\alpha = \beta = 0.5$ and for γ_v near zero. Let $F(\gamma_v, \hat{p})$ be the LHS of Equation 2.46. By the implicit function theorem (IFT):

$$\frac{dp_{vra}^{NO}(0)}{d\gamma_v} = -\frac{\frac{\partial F(0,\hat{p})}{\partial\gamma_v}}{\frac{\partial F(0,\hat{p})}{\partial p}} < 0. \quad (2.47)$$

To see this holds, first note that $\frac{\partial x_{vra}^{NO}}{\partial\gamma_v} < 0$ and $\frac{\partial y_{vra}^{NO}}{\partial\gamma_v} > 0$, such that the partial derivative in γ_v of the first two terms of $F(\gamma_v, \hat{p})$ is negative. Next, taking the partial derivative in γ_v of the remaining three terms gives:

$$-\frac{\sqrt{2}\beta^2 R^2\lambda + \hat{p}R((1+\alpha)^2\gamma_v\sqrt{\lambda} + \sqrt{2}\lambda(\sigma + \alpha + \alpha^2 - \beta^2)) - \hat{p}^2((1+\alpha)\gamma_v\sqrt{\lambda} + \sqrt{2}\lambda\sigma)}{\sigma^2\hat{p}(R-\hat{p})\bar{\mu}(\sqrt{2}\gamma_v + 2\sqrt{\lambda})\hat{x}^\alpha\hat{y}^\beta}. \quad (2.48)$$

For $\alpha = \beta = 0.5$ and $\gamma_v = 0$ we find the numerator of this derivative is $\frac{R^2 + 10\hat{p}R - 8\hat{p}^2\lambda}{2\sqrt{2}} > 0$ as $\hat{p} < R$, which means that Equation 2.48 is negative. Moreover, from the concavity in p of the risk neutral client's objective function we know $\frac{\partial F(0,\hat{p})}{\partial p} < 0$. From Equation 2.47 it then follows that p_{vra}^{NO} is decreasing in γ_v for $\alpha = \beta$ and γ_v near zero.

We now show that p_{vra}^{NO} also decreases in γ_v for $\gamma_v > 0$. From Equation 2.48 it can be shown that, given $\alpha = \beta$, the derivative is negative for any γ_v. This means the numerator on the RHS of Equation 2.47 is never zero and, therefore, $\frac{dp_{vra}^{NO}}{d\gamma_v}$ does not have a stationary point. Without a stationary point on the derivative of p_{vra}^{NO} in γ_v, p_{vra}^{NO} cannot be decreasing-increasing in γ_v. Taken together, p_{vra}^{NO} monotonically decreases in γ_v. Finally, as vendor risk aversion increases and the client lowers the penalty to the vendor, the vendor exerts less effort (both the cost of downtime to him and disutility from γ_v decrease), while the client exerts more effort. □

Proof of Proposition 4. The first order condition for the vendor is:

$$\frac{\partial\mathbb{E}[\mathbb{U}_v]}{\partial y} = \frac{\beta p\lambda}{\bar{\mu}x^\alpha y^{1+\beta}} + \frac{\beta p\gamma_v\sqrt{2\lambda}}{2\bar{\mu}x^\alpha y^{1+\beta}} - c_v = 0. \quad (2.49)$$

Solving this equation for y gives the vendor's best response function to the x and p:

$$y(x|p) = \left(\frac{\beta p \left(2\lambda + \gamma_v \sqrt{2\lambda} \right)}{2\bar{\mu} c_v x^\alpha} \right)^{\frac{1}{1+\beta}}. \tag{2.50}$$

This best response function gives a unique solution for y as:

$$\frac{\partial^2 \mathbb{E}[\mathbb{U}_c]}{\partial y^2} = -\frac{p\beta(1+\beta)(2\lambda + \gamma_v \sqrt{2\lambda})}{2\bar{\mu} x^\alpha y^{2+\beta}} < 0. \tag{2.51}$$

At optimality, the vendor's IR constraint is binding, such that $F^O = \frac{p\lambda}{\bar{\mu} x^\alpha y^\beta} + c_v y + \frac{p\gamma_v \sqrt{2\lambda}}{2\bar{\mu} x^\alpha y^b} + k$, where we can let $k = 0$ without loss of generality. Similar to Proposition 2, we maximize for both x and p, by simultaneously solving $\frac{\partial \mathbb{E}[\mathbb{U}_c]}{\partial x} = 0$ and $\frac{\partial \mathbb{E}[\mathbb{U}_c]}{\partial p} = 0$. We proceed by finding the FOC in p:

$$\frac{\partial \mathbb{E}[\mathbb{U}_c]}{\partial p} = -\frac{2\beta(p - R)\lambda + (1 + b)p\gamma_v \sqrt{2\lambda}}{x^\alpha y^\beta (1 + \beta) p\bar{\mu}} = 0. \tag{2.52}$$

Solving this equation gives:

$$p_{vra}^O = \frac{2\beta R\lambda}{2\beta\lambda + (1 + \beta)\gamma_v \sqrt{2\lambda}}, \tag{2.53}$$

which is decreasing in γ_v at a decreasing rate. To confirm that this penalty p_{vra}^O indeed maximizes client utility, we find:

$$\frac{\partial^2 \mathbb{E}[\mathbb{U}_c]}{\partial p^2} = \frac{\beta g(p)}{2^{1/(1+\beta)}(1 + \beta)^2 p^2 \bar{\mu} y^{\beta/(1+\beta)}} < 0, \tag{2.54}$$

where $g(p) = 2\beta p + (1 + \beta)p\gamma\sqrt{2\lambda} - (1 + 2\beta)2R\lambda$. Observe that $g(p)$ is increasing in p and that $g(p) = 0$ for $\bar{p} = \frac{2(R\lambda + 2\beta R\lambda)}{2\beta\lambda + (1+\beta)\gamma_v \sqrt{2\lambda}}$. Clearly, $\bar{p} > p_{vra}^O$, therefore $g(p_{vra}^O) < 0$ and $\frac{\partial^2 \mathbb{E}[\mathbb{U}_c]}{\partial p^2}\big|_{p=p_{vra}^O} < 0$. With this result we can now solve the client's FOC with $p = p_{vra}^O$ to find the equilibrium investment by the client, x_{vra}^O:

$$x_{vra}^O = \left(\frac{\alpha R\lambda}{\bar{\mu}} \right)^{\frac{1+\beta}{\sigma}} \left(\frac{\beta R\lambda\eta}{c_v \bar{\mu}} \right)^{\frac{-\beta}{\sigma}} = x^O \left(\frac{1}{\eta} \right)^{\frac{\beta}{\sigma}}, \tag{2.55}$$

where $\eta \equiv \frac{2\lambda + \gamma_v \sqrt{2\lambda}}{2\lambda + (1+1/\beta)\gamma_v \sqrt{2\lambda}}$. The result for x_{vra}^C is unique as $\frac{\partial}{\partial x} \frac{x^{\frac{\alpha\beta}{1+\beta}}}{x^{1+\alpha}} < 0$, therefore $\frac{\partial^2 \mathbb{E}[\mathbb{U}_c]}{\partial x^2}\big|_{p=p_{vra}^O} < 0$. As $\eta < 1$, $x_{vra}^O > x^O$. Finally, we find:

$$y_{vra}^O = y(x_{vra}^O, p_{vra}^O) = \left(\frac{\beta c_c}{\alpha c_v} \right)^{\frac{\alpha}{\sigma}} \left(\frac{R\beta\lambda}{\bar{\mu} c_v} \right)^{\frac{1}{\sigma}} \eta^{\frac{1+\alpha}{\sigma}} = y^O \eta^{\frac{1+\alpha}{\sigma}}, \tag{2.56}$$

where we can see that as $\eta < 1$, $y_{vra}^O < y^O$. $\qquad\square$

Proof of Corollary 1. From the proof of Propositions 1 and 3, the product of equilibrium investments under symmetry is: $x_{vra}^{NO}(p)y_{vra}^{NO}(p) = \Phi^{\frac{1}{\sigma}}(x^{NO}(p)y^{NO}(p))$. Taking the derivative of this function in γ_v gives $(x^{NO}(p)y^{NO}(p))(\Phi^{\frac{1}{\sigma}})' + \Phi^{\frac{1}{\sigma}}(x^{NO}(p)y^{NO}(p))'$. Note that x^{NO} and y^{NO} only depend on γ_v through p. The first term of the derivative is non-negative for any $\gamma_v > 0$ as $(x^{NO}y^{NO}) \geq 0$ and $(\Phi^{\frac{1}{\sigma}})' = \frac{1}{\sigma\sqrt{2\lambda}} \left(\frac{2\lambda + \gamma_v \sqrt{2\lambda}}{2\lambda} \right)^{\frac{1}{\sigma}-1} > 0$. For γ_v near zero, p_{vra}^{NO} is close to $R/2$ and therefore $(x^{NO}(p)y^{NO}(p))' = 0$. As such, the second term of the derivative is zero and

the sum of the two terms is positive. Also note that the lowest value for the derivative occurs when γ_v goes to infinity. For high γ_v, from Proposition 3, $(p_{vra}^{NO})'$ goes to zero and therefore, $(x^{NO}(p)y^{NO}(p))' = 0$. Moreover, for high γ_v, $(\Phi^{\frac{1}{\sigma}})'$ goes to zero. Therefore, the total derivative of $x_{vra}^{NO}(p)y_{vra}^{NO}(p)$ with respect to γ_v goes to zero. Taken together, this means $x_{vra}^{NO}(p)y_{vra}^{NO}(p)$ increases in γ_v.

We now show that $x_{vra}^O y_{vra}^O$ decreases in γ_v. Proposition 4 shows that $x_{vra}^O = x^O \left(\frac{1}{\eta}\right)^{\frac{\beta}{\sigma}}$ and $y_{vra}^O = y^O \eta^{\frac{1+\alpha}{\sigma}}$, such that when $\alpha = \beta = 0.5$ and $c_c = c_v = c$, the total response capacity in the observable case and under vendor risk aversion is:

$$\mu_{vra}^O(\gamma_v) = \bar{\mu}(x_{vra}^O y_{vra}^O)^{1/2} = \bar{\mu}(x^O y^O)^{1/2} \eta^{1/4}. \tag{2.57}$$

As $\eta < 1$ and decreasing in γ_v, we find that in the observable case and under vendor risk aversion, total response capacity is decreasing in γ_v. Thus, as γ_v increases, joint investment in the non-observable case increases, whereas the joint investment in the observable case decreases. Therefore, expected uptime decreases in vendor risk aversion when the client's investment is observable and expected uptime increases in vendor risk aversion when it is not. $\qquad\square$

Proof of Corollary 2. Let $\alpha = \beta = 0.5$ and $c_c = c_v = c$. From Corollary 1 we know that as vendor risk aversion increases, the difference in the joint capacity between the observable and non-observable case decreases in γ_v. From the client's expected utility functions in the observable and non-observable case we can see that what remains to be shown in order for the difference of those functions to decrease in γ_v is that the difference in the sum of the investments also decreases, i.e. $\frac{\partial}{\partial \gamma_v}\left((x_{vra}^O + y_{vra}^O) - (x_{vra}^{NO} + y_{vra}^{NO})\right) < 0$. First, we find:

$$\frac{\partial}{\partial \gamma_v}(x_{vra}^O + y_{vra}^O) = \frac{-2\sqrt{2}\lambda}{2(\sqrt{2}\gamma_v + 2\sqrt{\lambda})\left(1 - \gamma_v/((3/2)\gamma_v + \sqrt{2\lambda}/2)\right)^{\frac{1}{4}}\left(\sqrt{2}(3/2)\gamma_v + \sqrt{\lambda}\right)^2} < 0. \tag{2.58}$$

Next, we show that $\frac{\partial}{\partial \gamma_v}(x_{vra}^{NO} + y_{vra}^{NO}) > 0$. Let $x' \equiv \frac{\partial x_{vra}^{NO}}{\partial \gamma_v}$ and let $y' \equiv \frac{\partial y_{vra}^{NO}}{\partial \gamma_v}$. Assume the opposite is true, i.e. $0 < x' < -y'$. From Proposition 3 we know $0 < y < x$ for $\gamma_v > 0$. Then we have $0 < x'y < -y'x \Rightarrow xy' + x'y < 0$, which is false as by Corollary 1 we have $(xy)' = xy' + x'y > 0$. We thus have $\frac{\partial}{\partial \gamma_v}(x_{vra}^O + y_{vra}^O) < 0$ and $\frac{\partial}{\partial \gamma_v}(x_{vra}^{NO} + y_{vra}^{NO}) > 0$, which means $\frac{\partial}{\partial \gamma_v}\left((x_{vra}^O + y_{vra}^O) - (x_{vra}^{NO} + y_{vra}^{NO})\right) < 0$. $\qquad\square$

Proof of Proposition 5. We first state the expected utility for the client and vendor:

$$\mathbb{E}[\mathbb{U}_c] = R\left(1 - \frac{\lambda}{\bar{\mu}x^\alpha y^\beta}\right) - c_c x + \frac{p\lambda}{\bar{\mu}x^\alpha y^\beta} - \frac{\gamma_c\sqrt{(R-p)^2}\sqrt{2\lambda}}{2\bar{\mu}x^a y^b} - F, \qquad \mathbb{E}[\mathbb{U}_v] = F - \frac{p\lambda}{\bar{\mu}x^\alpha y^\beta} - c_v y. \tag{2.59}$$

Given these expected utility functions, the best response functions for the client and vendor are:

$$x(y|p) = \left(\frac{(R-p)\alpha(2\lambda + \gamma_c\sqrt{2\lambda})}{2\bar{\mu}c_x y^\beta}\right)^{\frac{1}{1+\alpha}}, \qquad y(x|p) = \left(\frac{\beta p\lambda}{\bar{\mu}c_v x^\alpha}\right)^{\frac{1}{1+\beta}}. \tag{2.60}$$

Cross-substituting the best response functions, solving for x and y and defining $\Phi' = \left(\frac{2\lambda + \gamma_c\sqrt{2\lambda}}{2\lambda}\right)$, we find:

$$x_{cra}^{NO} = x^{NO}(p)\Phi'^{\frac{1+\beta}{\sigma}}, \qquad y_{cra}^{NO} = y^{NO}(p)\left(\Phi'\right)^{\frac{-\alpha}{\sigma}}. \tag{2.61}$$

At optimality, the IR constraint of the vendor is binding and accounts for the the second stage equilibrium investments, i.e. $F_{cra}^{NO} = \frac{p_{cra}^{NO}\lambda}{\bar{\mu}(x_{cra}^{NO})^\alpha(y_{cra}^{NO})^\beta} + c_v y_{cra}^{NO} + k$. Substituting F_{cra}^{NO}, x_{cra}^{NO} and x_{cra}^{NO} into the client's expected profit function gives the client's objective function for the stage one problem:

$$\mathbb{E}[\mathbb{U}_c] = R\left(1 - \frac{\lambda}{\bar{\mu}(x_{cra}^{NO})^\alpha(y_{cra}^{NO})^\beta}\right) - c_c x_{cra}^{NO} - c_v y_{cra}^{NO} - \frac{\gamma_c\sqrt{(R-p)^2}\sqrt{2\lambda}}{2\bar{\mu}(x_{cra}^{NO})^\alpha(y_{cra}^{NO})^\beta} \tag{2.62}$$

. Using shorthand $\hat{p} \equiv p_{cra}^{NO}$, $\hat{x} \equiv x_{cra}^{NO}$ and $\hat{y} \equiv y_{cra}^{NO}$, p_{cra}^{NO} the FOC gives the optimality condition for p_{cra}^{NO}:

$$\frac{\beta c_c \hat{x} - (1+\alpha)c_v\hat{y}}{\sigma\hat{p}} + \frac{(1+\beta)c_c\hat{x} - \alpha c_v\hat{y}}{\sigma(R-\hat{p})} + \frac{R\lambda}{\sigma\bar{\mu}\hat{x}^\alpha\hat{y}^\beta}\left(\frac{\beta}{\hat{p}} - \frac{\alpha}{R-\hat{p}}\right)$$
$$+ \frac{(1+\beta)\gamma_c\sqrt{2\lambda}}{2\sigma\bar{\mu}\hat{x}^\alpha\hat{y}^\beta} + \frac{\beta\sqrt{(R-\hat{p})^2}\gamma_c\sqrt{2\lambda}}{2\sigma\hat{p}\bar{\mu}\hat{x}^\alpha\hat{y}^\beta} = 0. \tag{2.63}$$

As before, this optimality condition does not have a closed-form solution, however, numerical analysis shows that there exists a unique $\hat{p}$ that maximizes the client's expected profit. We now show that $p_{cra}^{NO}(\gamma_c)$ decreases in γ_c for $\alpha = \beta$ and γ_c near zero. Let $F(\gamma_c, \hat{p})$ be the LHS of Equation 2.63. By the IFT:

$$\frac{\partial p_{cra}^{NO}(0)}{\partial\gamma_c} = -\frac{\frac{\partial F(0,\hat{p})}{\partial\gamma_c}}{\frac{\partial F(0,\hat{p})}{\partial p}} > 0. \tag{2.64}$$

To see this holds, first note that $\frac{\partial x_{cra}^{NO}}{\partial\gamma_v} > 0$ and $\frac{\partial y_{cra}^{NO}}{\partial\gamma_v} < 0$, such that the partial derivative in γ_c of the first two terms in Equation 2.63 is positive. Next, taking the partial derivative in γ_c of the remaining three terms gives:

$$-\frac{1}{\vartheta}\left(\sqrt{2}\alpha\beta\sqrt{(\hat{p}-R)^2}(R^2 - \hat{p}R)\lambda - \sqrt{2}\alpha^2\sqrt{(\hat{p}-R)^2}\hat{p}R\lambda\right)$$
$$+ \frac{1}{\vartheta}\left((1+\beta)(\hat{p}-R)^2(\hat{p}+\beta R)\gamma_c\sqrt{\lambda} - \sqrt{2}\sigma(\hat{p}-R)^2(\hat{p}+\beta R)\lambda\right). \tag{2.65}$$

For $\alpha = \beta = 0.5$ and $\gamma_v = 0$ we find the numerator of this derivative is $\frac{R^3\lambda}{\sqrt{2}} < 0$, which implies Equation 2.65 is positive. From the concavity in p of the risk neutral client's objective function we know $\frac{\partial F(0,\hat{p})}{\partial p} < 0$. It follows that p_{cra}^{NO} is increasing in γ_v for $\alpha = \beta = 0.5$ and $\gamma_v \to 0$. Moreover, for $\alpha = \beta = 0.5$ and $\hat{p} < R$, the numerator Equation 2.65 becomes $-\frac{R-\hat{p}}{2\sqrt{2}}(3R^2(4\gamma_c + \lambda) + 6\hat{p}R(2\gamma_c + \lambda) - 8p^2(3\gamma_c + \lambda)) < 0$. This means the numerator on the RHS of Equation 2.47 is never zero and implies that p_{cra}^{NO} monotonically increases in γ_c when $\alpha = \beta = 0.5$. Finally, as client risk aversion increases and the client increases the penalty, risk is transferred from the client to the vendor. As a result, the vendor exerts more effort, while the client exerts less effort. $\square$

Proof of Proposition 6. As the vendor is risk neutral, the vendor's best response function $y(x|p)$ is the same as in Proposition 2, i.e. $y(x|p) = \left(\frac{\beta p\lambda}{\bar{\mu}c_v x^\alpha}\right)^{\frac{1}{1+\beta}}$. Similar to prior proofs, the client can

guarantee the vendor's IR constraint by setting $F_{cra}^O = \frac{p\lambda}{\bar{\mu}(x_{cra}^O)^\alpha (y_{cra}^O)^\beta} + c_v y_{cra}^O + k$. Substituting these functions in to the client's expected profit function we can see that if the client sets $p = R$, she effectively cancels out her disutility term. Following this we solve the client's FOC in x and rewrite to retrieve the equilibrium investment by the client:

$$x_{cra}^O = \left(\frac{\alpha c_v}{\beta c_c} \right)^{\frac{\beta}{\sigma}} \left(\frac{\alpha R\lambda}{\bar{\mu} c_c} \right)^{\frac{1}{\sigma}} = x^O. \tag{2.66}$$

Substituting the optimal penalty and client's equilibrium investment into the vendor's best response function and simplifying, gives:

$$y_{cra}^O = \left(\frac{\beta c_c}{\alpha c_v} \right)^{\frac{\alpha}{\sigma}} \left(\frac{\beta R\lambda}{\bar{\mu} c_v} \right)^{\frac{1}{\sigma}} = y^O. \tag{2.67}$$

The solutions for x_{cra}^O and y_{cra}^O are the same as the first-best investments found in the observable case and under risk neutrality are as such not a function of γ_c. $\qquad\square$

Proof of Corollary 3. The proof follows the same steps as the proof of Corollary 1. In this case, the product $x_{cra}^{NO} y_{cra}^{NO}$ increases in γ_c, because the penalty now increases in client risk aversion, $\frac{\partial p_{cra}^{NO}}{\partial \gamma_c} > 0$, and at the same time, $p_{cra}^{NO} > \frac{R}{2}$, and therefore, $\frac{\partial (x^{NO} y^{NO})}{\partial p_{cra}^{NO}} < 0$. Thus, as γ_c increases, inefficiency in the non-observable case increases, whereas the first-best result is maintained with observability. As a result, the difference between the client's expected utility with and without observability increases under client risk aversion. $\qquad\square$

References

Anderson SW, Dekker HC (2005) Management control for market transactions: The relation between transaction characteristics, incomplete contract design, and subsequent performance. *Management Science* 51(12):1734–1752.

August T, Tunca TI (2011) Who should be responsible for software security? A comparative analysis of liability policies in network environments. *Management Science* 57(5):934–959.

Bapna R, Barua A, Mani D, Mehra A (2010) Research commentary-cooperation, coordination, and governance in multisourcing: An agenda for analytical and empirical research. *Information Systems Research* 21(4):785–795.

Basten D, Joosten D, Mellis W, Wallmueller C (2014) Keep IT simple: The challenge of interlaced IT architecture at Gothaer Systems. *Journal of Information Technology Teaching Cases* 4(1):34–40.

Bhattacharya S, Gupta A, Hasija S (2014) Joint product improvement by client and customer support center: The role of gain-share contracts in coordination. *Information Systems Research* 25(1):137–151.

Bhattacharyya S, Lafontaine F (1995) Double-sided moral hazard and the nature of share contracts. *RAND Journal of Economics* 26:761–781.

CA Technologies (2011) The Avoidable Cost of Downtime: The impact of IT downtime on employee productivity. URL `http://www3.ca.com/~/media/files/articles/avoidable_cost_of_downtime_part_2_ita.aspx`, [retrieved 11/12/2013].

Ceccagnoli M, Forman C, Huang P, Wu D (2012) Co-creation of value in a platform ecosystem: The case of enterprise software. *MIS Quarterly* 36(1):263–290.

Chan T, De Vericourt F, Besbes O (2014) Contracting in Medical Equipment Maintenance Services: An Empirical Investigation, working Paper, ESMT.

Chen Y, Bharadwaj A (2009) An empirical analysis of contract structures in IT outsourcing. *Information Systems Research* 20(4):484–506.

Corbett CJ, DeCroix GA, Ha AY (2005) Optimal shared-savings contracts in supply chains: Linear contracts and double moral hazard. *European Journal of Operational Research* 163(3):653–667.

Demski JS, Sappington DE (1991) Resolving double moral hazard problems with buyout agreements. *The RAND Journal of Economics* 22(2):232–240.

Dey D, Fan M, Zhang C (2010) Design and analysis of contracts for software outsourcing. *Information Systems Research* 21(1):93–114.

DiRomauldo A, Gurbaxani V (1998) Strategic intent for IT outsourcing. *MIT Sloan Management Review* 39(4):67–80.

Fitoussi D, Gurbaxani V (2012) IT outsourcing contracts and performance measurement. *Information Systems Research* 23(1):129–143.

FM Global (2010) Flirting with natural Disasters - why companies risk it all? URL `http://www.fmg-dev.co.uk/wp-content/uploads/2012/04/Flirting-with-Natural-`

Disasters-Why-Companies-Risk-It-All_Flirting_with_natural_disasters.pdf, retrieved 31/08/2016.

Forbes (2013) How expensive is your it downtime? how about $1m/hour? URL http://www.forbes.com/sites/netapp/2013/10/07/whats-it-downtime-cost/, retrieved 11/12/2013.

Gopal A, Sivaramakrishnan K (2008) Research note-on vendor preferences for contract types in offshore software projects: The case of fixed price vs. time and materials contracts. *Information Systems Research* 19(2):202–220.

Gopal A, Sivaramakrishnan K, Krishnan MS, Mukhopadhyay T (2003) Contracts in offshore software development: An empirical analysis. *Management Science* 49(12):1671–1683.

Guajardo JA, Cohen MA, Kim SH, Netessine S (2012) Impact of performance-based contracting on product reliability: An empirical analysis. *Management Science* 58(5):961–979.

Jain N, Hasija S, Popescu DG (2013) Optimal contracts for outsourcing of repair and restoration services. *Operations Research* 61(6):1295–1311.

Kim SH, Cohen MA, Netessine S (2007) Performance contracting in after-sales service supply chains. *Management Science* 53(12):1843–1858.

Kim SH, Cohen MA, Netessine S, Veeraraghavan S (2010) Contracting for infrequent restoration and recovery of mission-critical systems. *Management Science* 56(9):1551–1567.

Kim SH, Tomlin B (2013) Guilt by association: Strategic failure prevention and recovery capacity investments. *Management Science* 59(7):1631–1649.

Lee CH, Geng X, Raghunathan S (2013) Contracting information security in the presence of double moral hazard. *Information Systems Research* 24(2):295–311.

Levy H, Markowitz HM (1979) Approximating expected utility by a function of mean and variance. *The American Economic Review* 308–317.

Martinez-Jerez FA, Narayanan V (2007) Strategic Outsourcing at Bharti Airtel Limited. *Harvard Business School Case* 9:107–003.

Ning J, Babich V, Handley J, Keppo J (2014) Risk-aversion and b2b contracting under asymmetric information: Evidence from managed print services, working Paper, Weatherhead School of Management.

Pratt JW (1964) Risk aversion in the small and in the large. *Econometrica: Journal of the Econometric Society* 122–136.

Roels G, Karmarkar US, Carr S (2010) Contracting for collaborative services. *Management Science* 56(5):849–863.

Ross SM (2003) *Introduction to probability models* (Academic press).

Samuelson W (1986) Bidding for contracts. *Management Science* 32(12):1533–1550.

Sky News (2014) Lloyds banking group sorry over IT glitch. URL http://news.sky.com/story/1201359/lloyds-banking-group-sorry-over-it-glitch, retrieved 10/06/2016.

Susarla A (2012) Contractual flexibility, rent seeking, and renegotiation design: An empirical analysis of information technology outsourcing contracts. *Management Science* 58(7):1388–1407.

Susarla A, Barua A (2011) Contracting efficiency and new firm survival in markets enabled by information technology. *Information Systems Research* 22(2):306–324.

Susarla A, Barua A, Whinston AB (2010) Multitask agency, modular architecture, and task disaggregation in saas. *Journal of Management Information Systems* 26(4):87–118.

The Telegraph (2012) The ten biggest computer failures of 2012. URL `http://www.telegraph.co.uk/technology/picture-galleries/9743652/The-ten-biggest-computer-failures-of-2012.html?frame=2427040`, retrieved 08/03/2013.

Van Mieghem JA (2007) Risk mitigation in newsvendor networks: Resource diversification, flexibility, sharing, and hedging. *Management Science* 53(8):1269–1288.

Wu D, Ding M, Hitt LM (2012) IT implementation contract design: Analytical and experimental investigation of IT value, learning, and contract structure. *Information Systems Research* 24(3):787–801.

Chapter 3

Risk Preferences and Joint Disruption Risk Mitigation: An Experimental Study

3.1 Introduction

In the past two decades manufacturers across different sectors have undergone a shift towards including maintenance services in their vendor offering. This shift followed demand from clients looking to outsource care of high-tech equipment supporting their business operations (Sawhney et al. 2004). The contract type generally used in this setting is the performance-based contract (PBC). Whether the vendor takes responsibility for part or all of the potential downtime under such contracts, collaboration between the client and vendor is necessary to achieve system performance over time Chan et al. (2016). The primary reason for this is that the client has access to equipment and the vendor has the knowledge of the technical details of the equipment. Resolving disruption requires both parties to dedicate response capacity (Kim and Tomlin 2013). Achieving effective collaboration in service environments may be difficult because of moral hazard and asymmetric information (Roels et al. 2010). In this chapter, we test whether it is also hindered by behavioral factors. Specifically, we investigate to what extent differences in the risk preferences of the two parties and the information available to the client on these differences impact their joint ability to mitigate downtime risk.

A well-known example of a PBC is Roll-Royce offering 'Power by the Hour' contracts to clients, linking compensation to availability in hours flown of their engines (Kim et al. 2007). The concept has found its way across to other industries, with companies like Hitachi and Caterpillar offering similar contracts (Visnjic et al. 2013). One context in which collaboration is increasingly critical is that of IT in the financial sector. In February 2015 Deutsche Bank signed a 10-year, multi-billion dollar deal with Hewlett Packard to secure maintenance and modernization of the bank's IT system as well as data storage. Deutsche Bank meanwhile will retain responsibility for application development and data security. Partnering in this fashion, banks hope to obtain scale and access to new technologies. Maintaining uptime is a core concern throughout the industry (Shotter et al. 2015). Key in these examples is that the collaboration efforts are exerted over an extended horizon and that stakeholders' objectives and incentives overlap, but may not be

entirely aligned. With this study, we aim to understand how clients outsourcing mission critical technology balance positive incentives (complementary investment by the client) with negative incentives (penalties to the vendor per unit of downtime) to ensure effective response to service disruptions.

While existing experimental research primarily focuses on the direct link between partners' decisions and performance, we focus on the element of collaboration that creates feedback between decisions and leads to both direct and indirect effects on performance. We consider a simple setting with a single client (she) and a single vendor (he) in which both parties contract to invest in mitigating the cost of IT system ('the system') downtime over the contracting horizon. We examine the effect of differences in risk aversion on the contracting and investment decisions. Moreover, we study how this effect is moderated by the decision maker's access to information on the risk aversion of the other party. To test for these effects we ran controlled experiments with subjects acting as the client and deciding on a performance-based contract parameter and her own investment level to influence the vendor's investment decision and optimize system performance.

The work in this chapter distinguishes itself from the existing experimental contracting literature by making two main contributions. First, we introduce a two-stage approach by measuring all subjects on a multi-dimensional risk aversion scale. Using this measurement we calibrated a sequence of computerized vendors, each with a different level of risk aversion, allowing us to manipulate the difference in risk aversion between the subject and the automated vendor in the main experiment. Second, we investigate how differences in risk preferences between the client and the vendor interact with the provision of cognitive feedback on the vendor's risk profile, which provides insight on whether and when it may be beneficial to provide information and could lead to actionable corrections to real life client contracting decisions when collaboration is essential.

3.2   Literature Review

This literature review covers three streams of work. We start our literature review by describing key works on disruption risk and maintenance service contracting in operations management (OM). Next we present an overview of the relevant experimental papers studying contracting problems within OM. Finally we take a closer look at how risk aversion has been measured and featured in experimental settings, mostly within behavioral economics and psychology.

3.2.1   Maintenance Service Contracting

In concert with the incidence of high-profile disruptions, the growth of maintenance service outsourcing has generated a stream of literature concerned with contract design for business

continuity. This stream of work considers misaligned incentives as an important factor driving downtime frequency and length. Performance-based contracts (PBCs) are known to remedy moral hazard issues, although in double moral hazard situations, first-best (FB) results may not be attained (see e.g. Bhattacharyya and Lafontaine (1995) and Roels et al. (2010)). Most papers in this stream of work consider the implications of PBCs in different contexts. Kim et al. (2007) consider model a setting in aerospace and defence with one principal facing uptime targets and outsourcing after sales services to multiple suppliers. They find the optimal balance between performance-based and cost-sharing elements in the contract offered by the principal, specifically for a context in which the suppliers' actions are unobservable and channel members are risk averse. Kim et al. (2010) reconsider the power of PBCs in the light of managing mission critical system uptime where the supplier can exert effort to minimize downtime after a disruption. They find the efficacy of the contract is dependent on characteristics of the failure process and how these are accounted for in the contract, demonstrating implementing PBCs may be costly when disruptions are very infrequent. Kim and Tomlin (2013) continue the line of work to show that decentralized decision making may lead to overinvestment in response capacity and underinvestment in prevention, as the former allows free-riding on other firms' investments in case of joint failure.

Jain et al. (2013) are to our knowledge the first to consider joint decisions between a vendor and client in a maintenance service setting, where both decisions are unobservable and affect system availability. They highlight the link between PBCs and induced financial distress, showing that in their context non-linear PBCs are best in overcoming both double moral hazard and vendor risk aversion. Two recent papers in this stream of work re-evaluate the differences between traditional resource based contracts and PBCs, considering trade-offs between spare parts inventory and investment in product reliability. Although inefficiencies arise in case of both contracts, PBCs provide better incentives for suppliers to invest in product reliability (Kim et al. 2015), but are also useful as a quality signaling mechanism to suppliers (Bakshi et al. 2015).

Theory thus suggests designing the appropriate performance-based contract should help to align incentives and improve maintenance services, aside from drawbacks under some conditions. The role of contract design in performance of maintenance services has also been observed empirically in aviation (Guajardo et al. 2012) and health care equipment (Chan et al. 2016). At the same time it is widely acknowledged that, although theoretically superior, complex contracts can be prohibitively difficult to implement (Kalkanci et al. 2011). We add to the maintenance service contracting literature by explicitly testing human decision making performance in the context of maintenance service contracting by means of a controlled experiment.

3.2.2 Experiments in Operations Management

There is a growing body of work validating and extending the analytical work within OM in experimental settings. A paper by Schweitzer and Cachon (2000) forms one of the cornerstones

to the experimental literature around contracting and supply decisions. In their experiment, subjects are asked to make inventory orders in a newsvendor setting across profit and demand conditions. They find subjects deviate from theoretically optimal decisions and find their observations are best explained with ex-post inventory error reduction, anchoring and insufficient adjustment bias. Their work inspired a line of work dedicated to examining behavior in a newsvendor decision making context, finding evidence for system neglect in making demand forecasts (Kremer et al. 2011) and the efficacy and complementarity of task decomposition and decision support systems in order decisions (Lee and Siemsen 2013). Beyond the classic newsvendor context, decisions are equally found to generally be sub-optimal to some degree in other decision making contexts. This can be the result of bounded rationality, biases and heuristics and social preferences (see e.g. Katok and Wu (2009)). Because of the nature of our experiments (human vs. computer only) we do not make inferences about social preferences (Katok and Wu 2009, Kalkanci et al. 2011). With the help of information and learning across interactions, decision makers' performance can improve (see e.g. Bolton and Katok (2008)). Using data from surgery decisions, KC et al. (2013) find that decisions makers tend to learn from their own successes and others' failures, in line with predictions from attribution theory.

Katok and Wu (2009) are the first to consider decision making biases in comparing experimental results on supply chain coordination under different risk sharing contract types. The main finding in their paper is that theoretical improvements of supply chain performance from a sub-optimal contract to an optimal contract do not come out as strong in the laboratory setting. Additionally they find mathematically equivalent contracts actually yield different outcomes in an experimental setting. The authors attribute this finding to loss aversion, but note that the effect dissipates with subjects' experience. The design of our study is such that we can directly examine the effect of risk aversion on contracting decisions as well as how this effect is moderated by subjects' learning and access to relevant information.

Overall the experimental evidence in OM supports the general finding that individual decisions by subject typically show deviations from the theoretically predicted decisions, which can be explained by a combination of well-known biases. Across repetitions and on aggregate levels, subjects do, however, perform closer to theoretical results. With this chapter we contribute to the growing body of experimental papers in OM by examining deviations from theoretically optimal decision adjustments under shifts in contracting conditions and available feedback on these shifts.

3.2.3 Risk Aversion in Experiments

Friedman and Sunder state that "many models assume that economic agents can be classified by their attitudes toward risk", where risk aversion are typically regarded as an intrinsic trait to the economic agent. (Friedman and Sunder 1994, Ch. 4.2). Broadly, the experimental literature is split between three approaches to examine the role of uncertainty and risk aversion in decision-

making. A first approach considers assumptions on risk aversion ex-post in fitting models to experimental data. A second approach controls risk aversion directly through design of the experiment, for instance by designing the payoff structure to induce risk neutral behavior. A third approach assigns experimental subjects to treatments conditional on pre-tested measures of risk aversion. In this case, risk aversion is generally considered in an ordinal sense within the experimental sample. This chapter takes this third approach to calibrate computerized vendors with different levels of risk aversion based on the subjects' risk preferences, allowing us to directly examine the effect of (differences between) players' risk aversion on joint disruption mitigation.

A multitude of approaches to measure subjects' risk aversion exists, though generally all approaches fall into either an implied quantitative category or a self-reported qualitative category. Examples of methods in the first category often involve a sequence of choices from a menu of lotteries, referred to as the Multiple Price List ('MPL') method (Charness et al. 2013). Tying decisions made to real payoffs to the subjects makes this method incentive compatible over methods that depend on self-rating. A well-known version of the MPL method is the one developed by Holt et al. (2002), referred to as the Holt-Laury ('HL') measure. The original HL measure has ten pairs of two-outcome gambles, with the probabilities of the better (worse) outcomes in each gamble increasing (decreasing) from the first to the last pair, while possible outcomes are kept constant across pairs. One of the gambles in each pair has a wider spread of outcomes, such that there should be one point at which a subject is willing to switch from the safer to the more daring gamble. This switching point implies the subject's risk aversion level and can be quantified using a simple utility function of the form $u(x) = x^{1-r}$, where is x is wealth and r the coefficient of constant relative risk aversion (CCRA) (Charness et al. 2013).

An early method in the second category was developed by Jackson et al. (1972), who built a multitrait-multimethod measure of risk-taking behavior encompassing financial risk, as well as physical, social and ethical risk. A more recent example in the same category is the Domain Specific Risk Taking ('DOSPERT') scale developed by Weber et al. (2002) and later fine-tuned in Blais and Weber (2006). The DOSPERT scale employs self-reported 7-point scale items on perception of risk as well as propensity to undertake risky action. Similar to the method in Jackson et al. (1972), the DOSPERT scale covers different dimensions of risk taking: investment, gambling, health and safety, recreation, ethics and social. They find these factors are not consistently correlated, highlighting the need to account for the decision context. Closest to our work among the papers that employ DOSPERT is the work by Bapna et al. (2010), who use the DOSPERT to study the effect of risk aversion on bidding strategies in auctions.

Specifically concerning financial risk aversion the literature commonly assumes subject risk aversion to be both stationary and scale-free, following a utility function characterized by constant absolute risk aversion (CARA) as in Pratt (1964). Still, evidence deviating from these assumptions exists. Thaler et al. (1997) show that myopic decision makers gravitate towards lower risk taking over time. Holt et al. (2002) find that the effect of risk aversion is actually scale-

dependent, but only when real money is at stake. Moreover, subjects are found to display differences in risk aversion across different elicitation methods (see Anderson and Mellor (2009) and references therein).

Another factor that is found to interact with risk aversion is the information available in the decision making context. Roth and Malouf (1979) show knowledge of other subjects' (potential) payoff affects outcomes of bargaining over respective probabilities of each subject winning his/her payoff. Ang and Schwarz (1985) study sequences of trades under different information regimes. They compare trades in an experimental market comprised of risk averse investors and one comprised of less risk averse investors. The information regimes are distinct by whether a subset of traders has (im)perfect information on future market states and whether this is common knowledge to all traders. Interestingly, less risk averse traders appear less affected by false information regimes, i.e. where there is a false suspicion of superior information being present among some traders. Thaler et al. (1997) also relate information provision to risky financial decisions. They induced myopia by designing treatments on the basis of decision and performance information provision frequency across an investment horizon. Those subjects endowed with more flexibility to adjust decisions and more information on performance generally performed worse.

The role of risk aversion in experimental settings has been studied extensively outside of OM. Although risk aversion can be measured with a number of well-documented measures, subjects may display different preferences depending the measure and decision making context. Our study adds to this literature by combining multiple risk aversion measures, confirming consistency across the measures and calibrating experimental conditions in our main experiment on the basis of one of these measures.

3.3 Theoretical Model

This chapter builds on a simplified version of the model in Chapter 2. In this chapter we studied outsourcing of mission critical system maintenance by a client to a specialized vendor through a contracting game. This game is set in a Principal-Agent framework in which the combined investments in response capacity by the principal and agent affect the expected cumulative downtime. In this section, we first describe the optimal solutions to the model and then turn to the simulation technique used to integrate the model into the software we developed for our experiment.

3.3.1 Optimal Decisions

Cumulative downtime $\sum_{i=1}^{N} D_i$ follows a compound Poisson process, where disruption frequency follows a Poisson distribution with rate λ and disruption length follows an exponential distribu-

tion with rate μ. To capture joint investment and its complementarity, we define μ as a function of the respective investments by the client x and the vendor y using a Cobb-Douglas production function: $\mu = \bar{\mu} x^\alpha y^\beta$, where we assume $\bar{\mu}, \alpha, \beta > 0$. The parameter $\bar{\mu}$ is a scale parameter that reflects the baseline capacity for both firms to combine resources to respond to disruptions; α and β respectively reflect the client's and vendor's return on investment in response capacity in terms of reducing the expected disruption length. This gives the expected cumulative downtime:

$$\mathbb{E}\left[\sum_{i=1}^{N} D_i | x, y\right] = \frac{\lambda}{\bar{\mu} x^\alpha y^\beta}. \tag{3.1}$$

For simplicity, a number of assumption are made: 1) the client earns a return at a rate R over the contracting horizon as long as the system is up; 2) the cost of investment is linear in the size of the investment, i.e. $c_c x$ for the client and $c_v y$ for the vendor; 3) the contract the client can offer to the vendor is restricted to a simple linear penalty contract, where p is the penalty per unit of downtime and F is a fixed fee. The sequence of events is such that the client moves first and decides on her level of investment in response capacity and the penalty, offering a take-it-or-leave-it contract to the vendor. The vendor observes the client's decisions and upon accepting the contract, decides on his level of investment. Given the decisions made, cumulative downtime materializes and determines payoffs for the client and vendor. This can be cast as the client's optimization problem, assuming risk neutrality ($\mathcal{N}$) for both the client and vendor:

$$(\mathcal{N}) \quad \max_{x \geq 0, F, p} \quad R\left(1 - \frac{\lambda}{\bar{\mu} x^\alpha (y^*)^\beta}\right) + \frac{p\lambda}{\bar{\mu} x^\alpha (y^*)^\beta} - c_c x - F,$$

$$\text{s.t.} \quad F - p\frac{\lambda}{\bar{\mu} x^\alpha (y^*)^\beta} - c_v y^* \geq k, \tag{IR}$$

$$y^* = \arg\max_{y \geq 0} F - \frac{p\lambda}{\bar{\mu} x^\alpha y^\beta} - c_v y, \tag{IC}$$

where the first constraint (IR) is the individual rationality constraint setting the minimum expected payoff for the vendor to accept the contract and the second constraint (IC) ensures the vendor maximizes his objective function.

To depart from the risk neutrality assumption, in Chapter 2 we applied a mean-standard deviation framework in which the client and vendor may experience disutility in proportion to the standard deviation of their respective payoffs. Let γ_v be the vendor's level of risk aversion and γ_c the client's level of risk aversion. The client's optimization problem when both parties are risk averse to some degree ($\mathcal{A}$) becomes:

$$(\mathcal{A}) \quad \max_{x \geq 0, F, p} \quad R\left(1 - \frac{\lambda}{\bar{\mu} x^\alpha y^\beta}\right) - c_c x + \frac{p\lambda}{\bar{\mu} x^\alpha y^\beta} - \frac{\gamma_c \sqrt{(R-p)^2}\sqrt{2\lambda}}{2\bar{\mu} x^\alpha y^\beta} - F,$$

$$\text{s.t.} \quad F - p\frac{\lambda}{\bar{\mu} x^\alpha (y^*)^\beta} - c_v y^* - \frac{\gamma_v p\sqrt{2\lambda}}{2\bar{\mu} x^\alpha (y^*)^\beta} \geq k, \tag{IR}$$

$$y^* = \arg\max_{y \geq 0} F - \frac{p\lambda}{\bar{\mu} x^\alpha y^\beta} - c_v y - \frac{\gamma_v p\sqrt{2\lambda}}{2\bar{\mu} x^\alpha y^\beta}. \tag{IC}$$

For the purpose of our experiments, we model and automate the response of the vendor as triggered by contracting and investment decisions of the client, whose role is played by the

subjects in our experiments. Given the order of events, the best response by vendor is fully characterized by the solution to the first order condition of the vendor:

$$y(x,p) = \left(\frac{\beta p \left(2\lambda + \gamma_v \sqrt{2\lambda} \right)}{2 \bar{\mu} c_v x^{\alpha}} \right)^{\frac{1}{1+\beta}}. \tag{3.2}$$

When both the client and vendor are risk neutral, i.e. in case of problem $\mathcal{N}$, we can find closed form solutions for the optimal penalty and investment decision by the client. This is the result of Proposition 2 in Chapter 2, an adjusted version of which we reproduce below in Proposition 7. An abridged version of the proof can be found in Appendix 3.A of this chapter.

Proposition 7 (Optimal decisions under risk neutrality). *The client solves $\mathcal{N}$ by setting the optimal penalty $p^N = R$ and fixed fee F such that the IR constraint is binding. Under this contract it is optimal for both client and vendor to set their respective investments equal to the first-best (FB) investment levels, i.e. $y^N = y^{FB}$ and $x^N = x^{FB}$, where:*

$$x^{FB} = \left(\frac{\alpha c_v}{\beta c_c} \right)^{\frac{\beta}{1+\alpha+\beta}} \left(\frac{R\alpha\lambda}{\bar{\mu} c_c} \right)^{\frac{1}{1+\alpha+\beta}} \qquad and \qquad y^{FB} = \left(\frac{\beta c_c}{\alpha c_v} \right)^{\frac{\alpha}{1+\alpha+\beta}} \left(\frac{R\beta\lambda}{\bar{\mu} c_v} \right)^{\frac{1}{1+\alpha+\beta}}.$$

All model parameters except for γ_c and γ_v were set to the same values for each subject and across treatments. To keep the computation example in the instructions as simple as possible, while maintaining reasonable expected outputs given the context, we set $R = 100, \lambda = 2, \mu = 5, \alpha = \beta = 1/2$ and $c_v = c_c = 1$. This gives the optimal benchmark penalty $p^N = 100$ and benchmark investments under risk neutrality: $x^N = y^N = 2\sqrt{5} \approx 4.47$.

Closed-form solutions to $\mathcal{A}$ cannot be found in all instances, particularly when neither the client nor the vendor can be assumed to be perfectly risk neutral, which are of interest to us in this study. Using *Wolfram Mathematica* we numerically find the optimal penalty and investment decisions for the client for combinations $\{\gamma_c, \gamma_v\}$ in case neither γ_c nor γ_v are equal to zero. An overview of all possible optimal decisions and associated expected client profit figures, conditional on the respective risk aversion of the client and vendor are given in Table 3.B.1 in Appendix 3.B. These benchmarks form the basis for the directional predictions in the first of our hypotheses in the next section. In the rest of the chapter we follow the convention of identifying subjects with a positive γ_c as 'risk averse' and identifying subjects with a negative γ_c as 'risk seeking'. Moreover we consider a vendor to be 'more risk averse' than the subject when $\gamma_v - \gamma_c > 0$.

3.3.2 Simulating Random Disruptions

To simulate the failure process conditional on decisions made by the subjects in our experiment we use the Inverse Transform Method (ITM) (Sigman 2010). Let $F(x), x \in \mathbb{R}$ be any continuous cumulative distribution function and note that $F : \mathbb{R} \to [0,1]$ is non-negative and non-decreasing. Furthermore let $F^{-1}(y)$ be the inverse function such that $F^{-1}(y) = \min\{x : F(x) = y\}, y \in [0,1]$, which gives $F^{-1}(F(x)) = x$. ITM uses a uniform random variable $U \sim unif(0,1)$ to generate a

random variable $X \sim F$, given that by setting $X = F^{-1}(U)$ it must hold that $P(X \leq x) = F(x)$. For this to work, a closed form expression for the inverse function of F is necessary to map the randomly generated values of the uniform distribution to values of X with the right distribution.

The theoretical model behind our experiment builds on an exponential distribution for disruption length and a Poisson distribution for the number of disruptions within the contracting horizon. Fortunately both are easy to simulate using the ITM. In case of the exponential distribution, $F(x) = 1 - e^{-\mu x}$, $x \leq 0$ with μ as the rate parameter and $F^{-1}(y) = -(1/\mu)\text{Ln}(1 - y)$. Using values generated by U as the input gives $X = -(1/\mu)\text{Ln}(U)$ as $1 - U$ is also uniformly distributed on $[0, 1]$. Specific to our model, the vendor's and client's investment decisions affect the rate parameter as $\mu = \bar{\mu}x^{\alpha}y^{\beta}$. Every round in the experiment, the software takes the subject's investment decision $\hat{x}$ and penalty decision $\hat{p}$, generates the best-response by the vendor $\hat{y}$ and then uses $\hat{x}, \hat{p}$, parameters α, β and a random vector drawn from a uniform distribution to generate a vector of random disruption lengths.

Now we still need to simulate the number of disruptions in each round. To do so, let $\{N(t) : t \leq 0\}$ be the counting process of a Poisson process at rate λ, where the number of events in each period follows a Poisson distribution with rate λ. The disruption interarrival times X_i are i.i.d. with an exponential distribution also with rate λ, such that the time between the first and n^{th} disruption equals $t_n = X_1 + X_2 + \ldots + X_n$. To include the count of the first disruption, let $Y = N(1)+1 = \min\{n \leq 1 : X_1+\ldots+X_n > 1\}$. Again using ITM to sample from an exponential distribution we find, $Y = \min\{n \leq 1 : -(1/\lambda)\text{Ln}(U_1) + \ldots - (1/\lambda)\text{Ln}(U_n) > 1\} = \min\{n \leq 1 : U_1 \cdot U_2 \ldots \cdot U_n < e^{-\lambda}\}$. Multiplying sample values from an exponential distribution until the product is less than $e^{-\lambda}$ gives Y, which by subtracting 1 gives the number of disruptions in any given period. Repeating this generates a vector of random disruption frequencies per round we can use in the experiment. Finally we sum across the first n random disruption lengths in each round to find the simulated cumulative downtime.

3.4 Hypotheses and Metrics

The aim of this study is to improve understanding of the interaction between risk aversion and information impacting service contracting between a client and vendor in a service chain setting. Specifically we're interested in the impact within the context of coproduction in services. In the following we justify the hypothesis and relevant metrics for our study.

3.4.1 Hypotheses

We have two sets of hypotheses for this study, respectively related to differences in risk aversion between the client and vendor and the client's access to information about the vendor. The first set of hypotheses is in line with an established body of work in economics recognizing the

role of risk aversion as a driver behind contracting decisions (see e.g. (Laffont and Martimort 2009, p. 166)), which has recently seen empirical support from work in OM, for instance in the context of biopharmaceutical alliances (Taneri and De Meyer 2016). In this study we are interested in the differences in risk aversion on maintenance service contracting decisions. We define Hypotheses 1(a) and (b) based on findings in Chapter 2. From Table 3.B.1 in Appendix 3.B we can observe that given the level of subject risk aversion (γ_c), it is optimal to increase the investment and simultaneously decrease the penalty when facing a vendor with a higher level of risk aversion (γ_v goes from $\gamma_c + 0$ to $\gamma_c + 0.5$ or from $\gamma_c + 0.5$ to $\gamma_c + 1$). Therefore under increasing vendor risk aversion, we define our first hypothesis as follows.

Hypothesis 1. (THEORETICAL ADJUSTMENTS UNDER INCREASING VENDOR RISK AVERSION) *(a) Subjects facing a vendor with a higher level of risk aversion will invest proportionally more toward the joint output and (b) Subjects facing a vendor with a higher level of risk aversion will set a lower penalty.*

The second set of hypotheses examines the role of information about the risk aversion of the vendor on the decision process of the client. Decision makers generally have difficulty making unaided decisions in complex environments characterized by uncertainty and systematic changes in the decision context (Balzer et al. 1989, Sterman 1989). Providing feedback on performance outcomes seems an obvious remedy. However, evidence points towards cognitive feedback being the better alternative, particularly in the form of task information and to some extent functional validity information (Balzer et al. 1989). Task information helps the decision maker to understand the decision environment by making relations (e.g. trade-offs) in the decision environment more explicit. Functional validity information provides support to improve calibrating decisions to the environment. Both help to improve performance through better knowledge (matching decisions to the context) and control (making consistent decisions). Various papers in OM relate to these findings. Sengupta and Abdel-Hamid (1993) show subjects managing simulated software development projects adjust better to the dynamic context given cognitive feedback, rather than outcome feedback. Bolton and Katok (2008) find that providing performance feedback is not sufficient on its own to improve newsvendor decisions making performance. A paper by Lurie and Swaminathan (2009) find that more frequent feedback may be degrading to decision maker performance if the feedback information comes from a noisy source. Their results show more frequent feedback made decision makers more focused on recent feedback and less selective in their information acquisition.

Feedback characteristics as well as the decision environment thus affect how decision makers use information, which may both improve and degrade performance. Particularly of interest to this study is the finding by Ang and Schwarz (1985) that there is a relation between subjects' risk aversion and the role of feedback on decision makers' performance. In the context of our study, under-appreciating differences in risk aversion may result in inefficient decisions. When given the right information on the vendor's risk aversion, we therefore hypothesize clients may

do better tailoring their investment and contracting decisions to the context. Additionally, we hypothesize this information is more helpful, the larger the differences in risk aversion.

Hypothesis 2. (Feedback on Vendor Risk Aversion) *(a) With information on the risk aversion level of the vendor, subjects make more efficient decisions. (b) Information has a stronger impact on improved decision making for larger differences in risk aversion between the subject and the vendor.*

3.4.2 Metrics for the Analysis

In light of Hypotheses 1(a) and (b) we directly analyze subjects' penalty and investment decisions across rounds and treatments. We analyze subject behavior in light of Hypotheses 2(a) and (b) by means of a number of additional, computed, metrics:

1) Deviation from first-best profit: Similar to Bolton and Katok (2008), we are interested in the financial impact of decisions subjects make across treatments. To ensure we can compare performance across treatments, we normalize the metric by subtracting observed expected profits in each round from the first best profit level. We track expected profit rather than realized profit to focus on what subjects perceive when they submit choice and avoid letting random disturbances to profit affect the comparisons.

2) Expected uptime: Subjects see both expected profit and expected uptime in their game window, prior to submitting round decisions. It is possible that subjects choose to optimize uptime at the expense of expected profit rather than following the objective to optimize total profit over time. Therefore we are also interested in analyzing between- and within-subject differences in optimal uptime.

3.5 Experimental Design

The experiments conducted for this study consisted of two parts: Part I and Part II. In Part I subjects completed a pre-test consisting of a three-part assessment of their risk aversion. In Part II subjects played a contracting game taking the role of the client and interacting with a sequence of computerized vendors. To address our first research hypothesis, all subjects were exposed to a within-subject ('vendor risk aversion') treatment sequentially exposing each subject to three vendors with different levels risk aversion. In particular, we focus on the case where the vendor is more risk averse than the client. Each vendor's risk preference was pegged to the subject's risk preference as measured in Part I. Each subject faced one vendor with an equal risk preference, i.e. $\gamma_c = \gamma_v$; and two more risk aversion vendors, such that $\gamma_c = \gamma_v + 0.5$ and $\gamma_c = \gamma_v + 1$. We denote these differences by $\Delta \in \{0, 0.5, 1\}$. Using this approach, differences in risk preference were controlled regardless of the respective actual subject risk aversion. To our knowledge, this method of configuring treatments on the basis of subject risk aversion has not be

applied elsewhere in the behavioral operations management literature. A note should be made that Bapna et al. (2010) deliberately test for the risk profiles of the subjects in their experimental study in a post-game test, so as to prevent affecting subjects' behavior. However, in our case this would prevent us from studying decision adjustments while manipulating differences in risk aversion.

To address our second research hypothesis subjects were randomly allocated to one of two between-subject ('information') treatments. In the 'informed' treatment, subjects were informed that, between the subject and the vendor, the vendor was more risk averse and, in case of the second (third) vendor in the sequence, whether the second (third) vendor was more or less risk averse that the previous vendor(s). This additional type of information can be classified as cognitive feedback (Sengupta and Abdel-Hamid 1993). This information was given to the subjects through pop-up windows. A sample of these message prompts can be seen in Appendix 3.D. Subjects in the 'non-informed' treatment saw prompts without this information. Further details on the design of pre-test and the main part of the experiment are given in the following.

3.5.1 Part I: Measuring Subject Risk Aversion

The assessment in Part I consisted of a quantitative measurement of risk aversion as developed by Holt et al. (2002) (HL) as well as an adaptation of the HL measure appropriate for our theoretical model ('HLa'), in that order. Part I additionally consisted of a qualitative measurement of risk aversion as developed by Blais and Weber (2006) (DOSPERT). We used the HLa measure to compute each subject's level of risk aversion, γ_c. The HL and DOSPERT measures were used to compute three additional risk preference measures used to validate the HLa measure, which we report in Section 3.5.4 below.

The HLa measure works by letting subjects make a series of choices between the outcome of a gamble and a deterministic outcome, respectively presented in two columns in the experiment window. Let the uncertain outcomes in the gamble be $\underline{b}$ (low outcome) and $\overline{b}$ (high outcome), each with a 50% probability and let $\mathbf{c} = \{c_1, c_2, ... c_n\}$ be a vector of certain outcomes. For each decision, the subject evaluated whether she prefers $b = (\underline{b} + \overline{b})/2$ or c_i. In our experiments we ordered $\mathbf{c}$ in decreasing value, such that $b = (\underline{b} + \overline{b})/2 < c_i$ for $i \in [1, m)$ and $b = (\underline{b} + \overline{b})/2 > c_i$ for $i \in [m, n]$. We set $\underline{b} = 2.95$, $\overline{b} = 8.95$ and $\mathbf{c} = \{8, 7.5, ... , 3.5\}$. Rational (risk neutral) subjects should therefore prefer the safe outcome for the first $m-1$ decisions and 'switch' to the gamble for the remaining $n-m$ decisions. However, subjects may choose differently: switching at $\tilde{m} < m$ indicates a degree of risk seeking and switching at $\tilde{m} > m$ indicates a degree of risk aversion. Assuming a mean-standard deviation framework, each subject will have some indifference point at which she evaluates $b - \gamma_c \sigma = \hat{c}$, where $c_{\tilde{m}-1} < \hat{c} < c_{\tilde{m}}$ and σ is the standard deviation of the gamble. To arrive at an approximate value for γ_c for each subject, we therefore evaluated $(b - c_{\tilde{m}-1})/\sigma$ and $(b - c_{\tilde{m}})/\sigma$ and took the average of the two values.

Part I was the same for all subjects, regardless of allocation to treatments. The HL measure and the HLa measure were presented on separate windows. In both cases subjects were forced to complete all choices. However, subjects were allowed to make multiple switches between gambles in the HL case and multiple switches between certain outcomes and gambles in the HLa case. As such switches are inconsistent from a utility point of view, this allowed removal of subjects from the experiment data should such inconsistencies have occured. In the remainder of Part I, subjects completed the DOSPERT in blocks of questions spread over six consecutive windows with the same interface. After each set of ten questions, subjects proceeded by clicking the 'save and continue' button. After submitting the final ten answers, the subject proceeded to Part II upon entering a correct password. The password was announced once all subjects finished Part I, so as to coordinate subjects' start times for Part II and limit disturbance by subjects who completed the experiment before others.

3.5.2   Part II: The Contracting Game

In Part II, subjects took the role of the client and interacted with the three automated vendors in a random order. We denote the order in which the subject interacted with the vendor by permutations of $\{1, 2, 3\}$, where 1 stands for the vendor with $\gamma_c = \gamma_v$; 2 stands for the vendor with $\gamma_c = \gamma_v + 0.5$; and 3 stands for the vendor with $\gamma_c = \gamma_v + 1$. Randomizing the order counterbalances the within-subject treatment. Subjects interacted with each of the three vendors for five warm-up rounds and thirty 'live' rounds. No explicit time constraints were imposed. The vendors were programmed to act in line with model in Section 4.3 and thereby focuses the experiment on identification of the effects of risk aversion as well as decision biases (Katok and Wu 2009, Bapna et al. 2010, Kalkanci et al. 2011). In each round, subjects were asked to decide on an investment in response capacity as well as a downtime penalty set to the vendor, affecting payoffs as described in Section 4.3. Subjects were instructed to decide on the two variables in each round in order to maximize their total profit.

Subjects were presented with a new user interface for Part II of the experiment. This interface was split into two panels. The left-hand panel functioned as a 'what-if' decision support tool for subjects to test decisions before submitting them. The right-hand panel showed a performance log. Respectively, the information presented in the two panels can be classified as feedforward and outcome feedback (Sengupta and Abdel-Hamid 1993). The user interface in Part II remained the same across treatments, except for the fields 'vendor risk profile' and 'investment sequence', which were empty for non-informed subjects. The computation of expected and realized results varied across treatments, though this was automated and not observable to the subjects. At the end of each round, the decisions and realized profits and other performance indicators were recorded in the performance log and remained visible as feedback for the duration of the treatment. Upon completing the final round of the contracting game the software asked all subjects to fill out an exit survey with on demographic and study-related information.

In total 34 participants took part in the experiment, across four sessions and each subject participating in only one session. Because of the uncertainty in the participant recruiting process some experimental conditions were allocated to more than others, resulting in an unbalanced design after data collection. Therefore, to achieve a balanced design across the levels for both factors, 24 participants were subsampled at random from the data to form our final data set for analysis.[1] The treatments and sample sizes are summarized in Table 3.1. We developed the software used for the experiment in *Visual Basic*. We ran trial experiments to test Part I and II for comprehension and possible software errors on 60 participants before the main experiments. Images of the user interface for both parts of the experiment can be found in Appendix 3.D.

Table 3.1 Experimental Design and Sample Sizes

| | Vendor Order | | | | | | |
Info	(123)	(132)	(213)	(231)	(312)	(321)	Total
Informed	2*	2*	2*	2	2	2	12
Non-informed	2*	2*	2	2	2	2*	12
Total	4	4	4	4	4	4	24

Notes. * Randomly re-sampled experimental conditions.

3.5.3   Experimental Protocol

Each subject participated in a single session that covered Part I and Part II and took up to 90 minutes. At the start of each session all subjects were asked to read a detailed instruction sheet. Subsequently, the experiment and software were again explained verbally by the experimenter, and subjects' questions were answered. All subjects completed Part I and Part II using software pre-installed on identical computers in a computer laboratory. The experiments were planned to coincide with the university's class-free time, to control for convenience sampling. All treatments were represented across different testing days. Upon arrival at the venue for the experiment, subjects were randomly allocated to a computer terminal. The software was set-up to trigger a particular experimental condition when given a number from a list randomly assigned across the terminals. This ensured the experiment was single-blind and prevented subjects from choosing seats close to familiar other subjects or gaining information on their assigned experimental condition. Finally, measures were taken to prevent participants from being able to observe information provided and decisions made at other terminals. Subjects were paid S\$5 for participating in the game and an additional performance-based payment conditional on their achieved profit. To prevent wealth effects, payments were made on the basis of accumulated profits converted from a fictitious in-game currency to Singapore Dollars at a pre-determined conversion rate (Friedman and Sunder 1994, p. 49).

---

[1]The main analysis was repeated for the full sample and two additional stratified subsamples. These results are presented in Section 3.7.

3.5.4   Validity of the Risk Preference Measure

As described in Section 3.5.1 of this chapter we used the adjusted Holt-Laury measure (HLa) in Part I of the experiment to compute each subject's risk preference, γ_c, and assign each subject to treatments on the basis of each subject's γ_c. To validate the HLa measure, we included two other tests for risk aversion in Part I. We included the original HL measure following the scoring procedure in Charness et al. (2013) as well as the DOSPERT measure as developed in Blais and Weber (2006). A valid HLa measure should produce risk preference score that are in line with the scores produced by the other two measures insofar as that there is a significant correlation between the measures and the correlation has the correct sign. As additional confirmation of validity, we report the median split agreement. We ran the validation tests after the two trial experiments before running our final experiments. Here we present the validation results on the final sample data. Note that for these tests $n = 23$ after listwise deletion because of missing values on DOSPERT data for one subject. Missing data on the DOSPERT section of the pre-test does not impact the game data and has therefore no impact on the analysis outside of this section. Furthermore we note that no inconsistent switches on the HL and HLa occurred in the sample. Figure 3.1 shows the distribution of HLa values, which shows a concentration of subjects around the risk neutral mark, with some subjects displaying a risk averse preferences (positive γ_c) and some displaying more risk seeking preferences (negative γ_c).

Figure 3.1 HLa Distribution
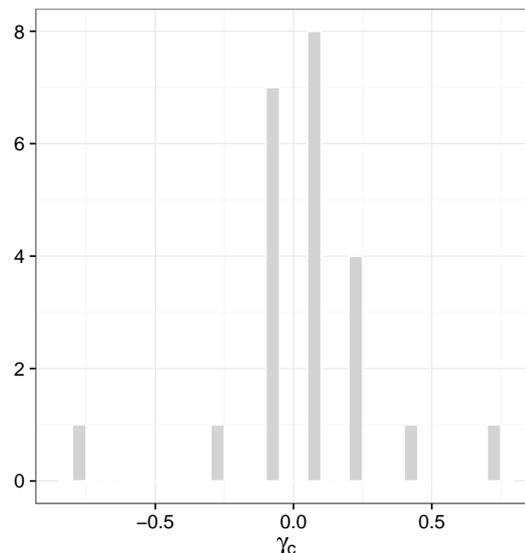
The original Holt-Laury measure produces a measure of constant relative risk aversion (the r-value) for each subject. Our HLa measure instead produces a measure of constant absolute risk aversion, hence the assumptions underlying each measure are different. Moreover, recall that as subjects choose between a set of guaranteed outcomes and a set of gambles in case of the HLa measure, rather than between two sets of gambles in the original HL measure. This also

means subjects may find it easier to indicate truthful preferences in the HLa case. Nevertheless we would expect those subjects identified as risk averse (seeking) in one measure to equally be identified as such by the other measure.

The DOSPERT scale can be used to produce various alternative risk preference measures. Taking a simplified version of the approach in Blais and Weber (2006), we can regress each subject's score across all thirty questions pertaining to the likelihood of undertaking a certain action on the subject's score across the same questions pertaining to the perception of riskiness. This gives a β coefficient for each subject, where a negative β indicates a risk averse preference (i.e. generally unwilling to undertake actions perceived as risky) and positive β indicates a risk seeking preference. We would thus expect subjects with a positive γ_c to give answers on the DOSPERT scale that yield a negative β. An alternative approach used in Bapna et al. (2010) restricts the focus to the subset of DOSPERT items related to financial risk taking and perception. Following this approach, we calculate each subject's risk taking and perception scores and multiply the risk taking score by 1 if the risk perception score is above the sample mean risk perception score and -1 otherwise. We denote the resulting score with s. As such subjects high (low) and positive (negative) s across the financial items are considered strongly (weakly) risk averse (seeking). We would therefore expect that subjects with a high γ_c generally have a high s also and vice versa.

Figure 3.2 shows the distributions of the alternative measures described above (r, β, and s). Considering the differences in measurement and computation of each of the measures, naturally the distributions have no close resemblance. For each of the measures we do see that both the risk averse and risk seeking domain are represented in the sample (where in the case of s this is by construction). For r, like with γ_c, we observe a concentration around the risk neutral mark; for β we observe that subjects are generally more risk averse.

Figure 3.2 Alternative Risk Preference Distributions

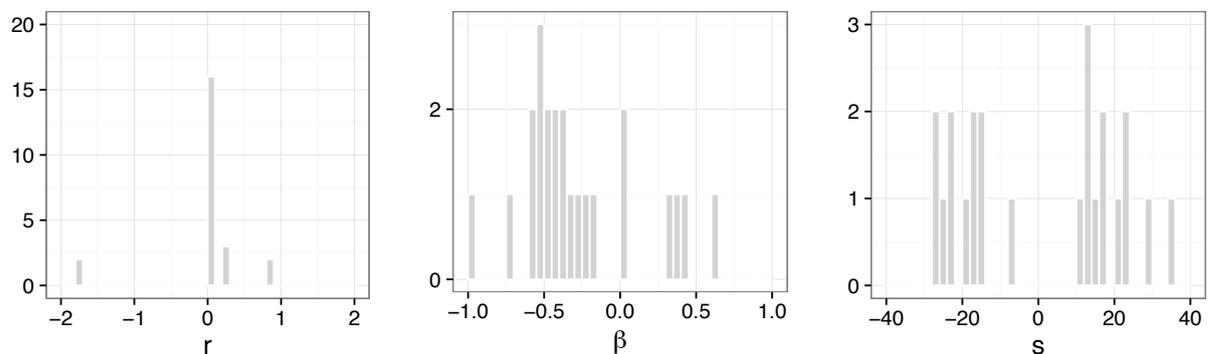Figure 3.D.2 shows the scatterplots between γ_c and the other three risk preference measures. The scatterplots overlayed with a simple linear regression line (thick black lines) and 95% confidence bands as well as the median values for each of the measures (thin black lines). From the regression lines we can cautiously confirm our expectations with regards to the agreement between γ_c and

the other three risk preference measures. More formally, however, only between γ_c and r we find a statistically significant relation ($p < 0.05$) with the correct sign.

The median split agreement gives a weaker notion of agreement between the measures. Between γ_c and r and γ_c and s, the median split agreement is equal to the number of observations in sample that are on or above (below) the median on both measures. Between γ_c and β, the median split agreement is equal to the number of observations in the sample that are on or above (below) the median of the former measure and on or below (above) the median on the latter measure. We find 22 out of 23 observations are consistent between γ_c and r, 16 out of 23 observations are consistent between γ_c and β and 14 out of 23 observations are consistent between γ_c and s. Overall we can thus conclude that there is a reasonable correspondence between the HLa measure and the other measures and γ_c is a good proxy for subject level risk preference.

Figure 3.3 Risk Preference Scatterplots
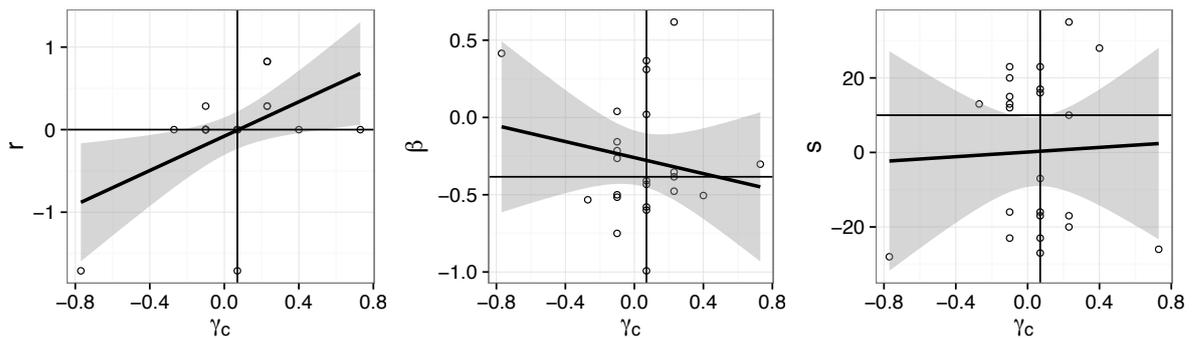
3.6 Results

In this section we first analyze the results with respect to Hypotheses 1(a) and (b), before turning to analyzing the results with respect to Hypotheses 2(a) and (b). The analysis in each section is done by means of the metrics highlighted before in Section 3.4.2. The optimal benchmarks in the table in Appendix 3.B are used as a reference values for the observed decisions in this section. We refer to these results as 'benchmarks' for brevity.

3.6.1 Hypothesis 1 Results: Theoretical Predictions

Following Section 3.4, we study subject (client) decisions on a within-subject basis to determine the effect of the differences in risk aversion between the subject and automated vendor on subject investment and penalty decisions. For this part of the analysis, we pool together the informed and non-informed subjects. We exclude data from the warm-up rounds to ensure we capture

only decisions the subjects were compensated for during live rounds. We analyze both mean and median decisions to prevent extreme decisions from causing misleading results.

Figure 3.1 shows the median investment and penalty decisions computed across thirty rounds for all subjects conditional on Δ being equal to 0, 0.5 or 1. Note that here we do not consider the order in which the subjects interacts with different vendors. Recall that when both parties are risk neutral, the theoretically optimal client investment is approximately 4.7 and the optimal penalty is 100. Also recall that each subject's risk aversion forms the reference point for the difference in risk aversion.

The lower two panels in the plot indicate that, as predicted, subjects set lower penalties when the vendor is comparatively more risk averse. This observation holds for both mean and median penalties set. A note should be made that when $\Delta = 0$, the mean penalty is noticeably lower than the median penalty, meaning some subjects set comparatively low penalties. The upper two panels, suggest little adjustment in the investments is made to accommodate for different vendor risk aversion levels, in line with the benchmarks. However, a noticeable difference between the mean and median results indicate some subjects make relatively high investments. Judging from the figure, all four plots show that when the difference in risk aversion is zero, subjects set their investments above the risk neutral benchmark, and set penalties under the risk neutral benchmark. This reflects that vendors are more risk averse than the subjects by design. However, it may also reflect that the sample has a risk averse slant, as reported in Section 3.5.4.

We used Wilcoxon Signed-Rank tests (Siegel 1957) to determine whether there is statistical support for the observations from Figure 3.1. Table 3.1 reports the results for non-parametric pairwise comparisons between decisions made by the same subjects facing vendors with different risk aversion. We see that there while there is no consistent support support for Hypothesis 1(a), there is strong support for Hypothesis 1(b). These results indicate subjects tend to set both higher and lower investments when facing a more risk averse vendor instead of only investing more as theory would predict.

Table 3.1 Within-subject comparison of investments and penalties

Δ	Investments[a]			Penalties[a]		
	0	0.5	1	0	0.5	1
Hypothesis	$x < x'$	$x' < x''$	$x < x''$	$p > p'$	$p' > p''$	$p > p''$
Mean	$*$	$-$	$-$	$* * *$	$* * *$	$* * *$
Median	$-$	$-$	$-$	$* * *$	$* * *$	$* * *$

Notes. [a] Significance levels for Wilcoxon Signed-Rank tests.
[b] $x, x', x''(p, p', p'')$ respectively indicate investments (penalties)
under $\Delta \in \{0, 0.5, 1\}$.
$-n.s.$, $^*p < 0.1$,$^{**}p < 0.05$,$^{***}p < 0.01$.

Figure 3.1 Subject mean and median investments and penalties
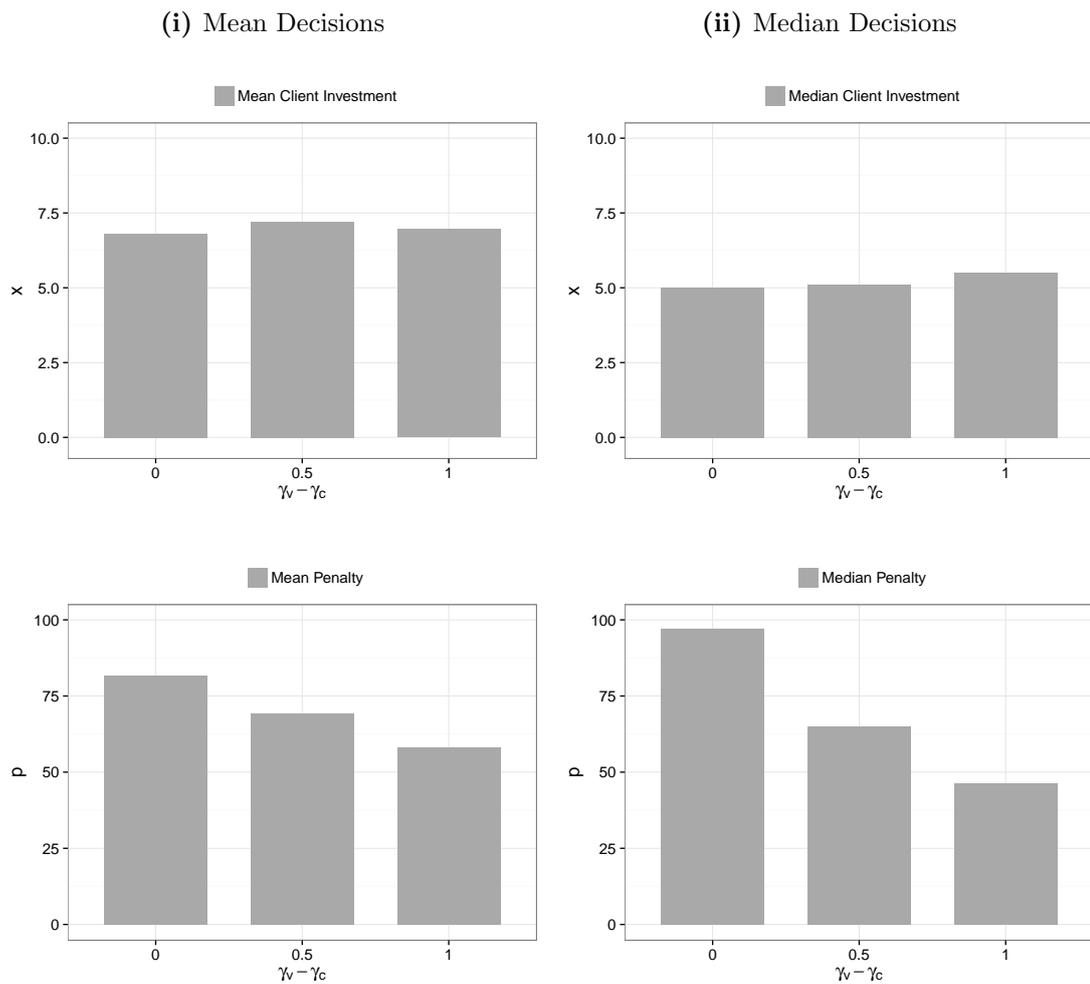
(i) Mean Decisions   **(ii)** Median Decisions

Notes. Mean and median investments and penalties for the plots in this figure were taken across the thirty live rounds where Δ was equal to 0, 0.5 or 1 for all subjects regardless of the order of the vendors.

We can make similar pairwise comparisons between the conditionally optimal benchmarks and the observed mean and median decisions. Table 3.2 reports the results for these comparisons and provides strong evidence for over-investment by subjects across the within-subject conditions. In case of penalty decisions, there is strong evidence for over-penalization by subjects when difference in risk preferences are non-zero (i.e. $\Delta = 0.5, 1$). Interestingly, when the subject and vendor are equally risk averse there is strong evidence for *under*-penalization. Based on the results in Table 3.1 and Table 3.2 we can make two observations regarding decision biases. Firstly, subjects tend to set and hold on to an inefficiently high investment level. Secondly, subjects tend to set and hold on to a penalty that is too high when interacting with more risk averse vendors and too low in case the vendor is equally risk averse.

Table 3.2 Within-subject comparison of investments and penalties

| | Investments[a] | | | Penalties[a] | | |
Δ	0	0.5	1	0	0.5	1
Hypothesis	$x > x^*$	$x > x^*$	$x > x^*$	$p < p^*$	$p > p^*$	$p > p^*$
Mean	$* * *$	$* * *$	$* * *$	$* * *$	$* * *$	$* * *$
Median	$* * *$	$* *$	$* * *$	$* * *$	$* * *$	$* * *$

Notes. [a] Significance levels for Wilcoxon Signed-Rank tests.
[b] $x^*(p^*)$ indicates a conditionally optimal investment (penalty) subject to γ_c and γ_c.
$-n.s.$, $^*p < 0.1$,$^{**}p < 0.05$,$^{***}p < 0.01$.

3.6.2 Hypothesis 2 Results: Vendor Risk Preference Information

For this part of the analysis, we consider whether information on the vendor's risk aversion helps subjects make more efficient decisions. Recall that the information treatment prompted subjects with information with regards to the vendor's risk aversion relative to the subject's risk aversion as well as the directional difference in risk aversion between one vendor and the next. We analyze efficiency of decision making on the basis of two outcome metrics: expected profit and expected uptime. As before, we exclude data from the warm-up rounds.

Table 3.3 summarizes the results for between-subjects tests with regards to expected profit and expected uptime. We ran Two-Way ANOVA tests to determine whether the informed group results are significantly different from the non-informed group as well as to test for the interaction between Δ and the information treatment. These tests were run with the game-round as the observation level. Hence, any significant results indicate differences between groups across all rounds. Additionally, we ran the Kruskal-Wallis test (in this case equivalent to a Mann-Whitney U test). Although the Kruskal-Wallis test is restricted to a single factor, it compares medians between groups and does not require the assumption of normality for the dependent variable.

From the ANOVA results we observe that with expected profit as the dependent variable, there is a significant difference between informed and non-informed subjects. However we find neither a significant within-subject effect of difference in risk aversion, nor do we find a significant interaction effect between the information treatment and difference in risk aversion. We find contrasting results with expected uptime as the dependent variable. In this case we find a significant effect of difference in risk aversion as well as a significant interaction effect between the information treatment and difference in risk aversion. The results of Kruskal-Wallis tests further support the ANOVA results for expected profit, but speak against the ANOVA results for expected uptime. Taken together the results of the two tests suggest there is a difference in performance caused by the information treatment, although it is manifested differently for expected profit than for expected uptime.
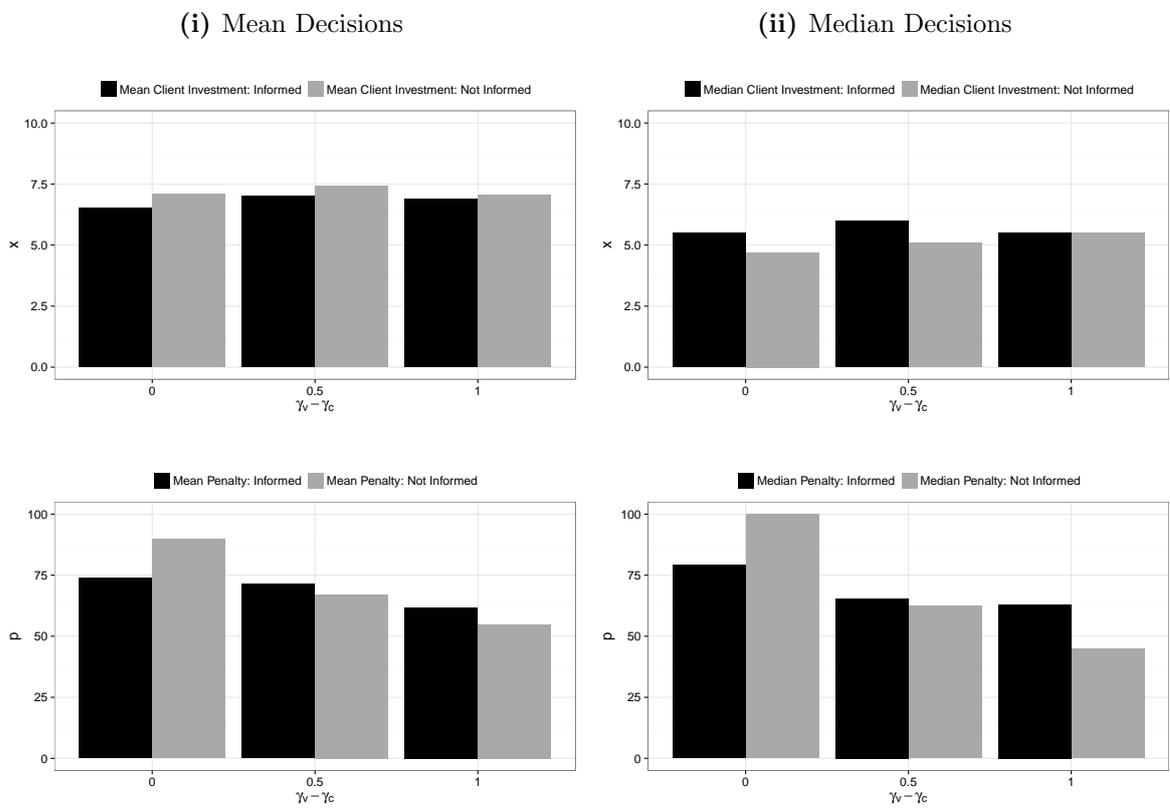
Table 3.3 Between-subject comparisons of observed outcomes

	Two-way ANOVA (F)		Kruskal-Wallis (χ^2)	
	Expected Profit	Expected Uptime	Expected Profit	Expected Uptime
INFO (yes=1)	3.91**	0.04	87.49***	4.66**
Δ	0.16	6.69***		
INFO$\times\Delta$	0.77	50.17***		

Notes. $^{*}p < 0.1,^{**}p < 0.05,^{***}p < 0.01$

To determine what may be driving differences in performance, Figure 3.2 splits out the plots in Figure 3.1 across the informed and non-informed subjects, both for mean and median investments and penalties. We can observe that the differences between mean and median decisions persist across the two treatments, although the difference is stronger for the subjects who were not informed. Said differently, high investments are not influenced as much by extreme decisions among informed subjects. With regards to the penalties we can see informed subjects tend to set lower penalties than the non-informed subjects when $\Delta = 0$, yet set higher penalties when $\Delta = 0.5$ or $\Delta = 1$. This suggests the decision-bias observations made in Section 3.6.1 are stronger for subjects in the informed treatment. However, testing whether there are indeed significant differences between the two treatment groups with respect to the observed decisions we find only the penalties set in case of $\Delta = 0$ are significantly different. Table 3.4 presents these results. On the basis of 30-round mean or median decisions we can thus not conclude decision biases are stronger for the informed subjects.

Figure 3.2 Subject mean and median investments and penalties

(i) Mean Decisions

(ii) Median Decisions

Notes. Mean and median investments and penalties for the plots in this figure were taken across the thirty live rounds where Δ was equal to 0, 0.5 or 1 for all subjects regardless of the order of the vendors.

Table 3.4 Between-subject comparison of investments and penalties

Δ	Investments[a]			Penalties[a]		
	0	0.5	1	0	0.5	1
Hypothesis	$x^I \neq x^{NI}$	$x^I \neq x^{NI}$	$x^I \neq x^{NI}$	$p^I \neq p^{NI}$	$p^I \neq p^{NI}$	$p^I \neq p^{NI}$
Mean	–	–	–	**	–	–
Median	–	–	–	*	–	–

Notes. [a] Significance levels for Mann-Whitney U tests.

[b] $x^I, x^{NI}(p^i, p^{NI})$ respectively indicate investments (penalties) under the informed (I) and not informed (NI) treatments.

$-n.s., {}^*p < 0.1, {}^{**}p < 0.05, {}^{***}p < 0.01.$

Our analysis so far does not consider how decisions may change as subjects progress through the experiment and learn from repeated interactions. Focusing on subject performance with regards to expected profits, we test for the role of time in conjunction with information with a regression. The full regression equation for our model is:

$$\ln(\Pi^{FB} - \Pi^{Obs}) = intercept + \beta_1 \times INFO + \beta_2 \times \Delta_{0.5} + \beta_3 \times \Delta_1 + \beta_4 \times ROUND$$
$$+ \beta_5 \times [\Delta_{0.5} \times INFO] + \beta_6 \times [\Delta_1 \times INFO]$$
$$+ \beta_7 \times [ROUND \times INFO] + \boldsymbol{\gamma}^T \boldsymbol{Z} + \epsilon, \tag{3.3}$$

where the dependent variable is the different between the first best expected profit (Π^{FB}) and the observed expected profit (Π^{Obs}) and $\boldsymbol{Z}$ is a control vector. We take the logarithm of the dependent variable to correct for the fact that it's highly positively skewed. The variable $INFO$ is a binary variable with the non-informed treatment as the reference category. The variables $\Delta_{0.5}$ and Δ_1 are levels of a factor variable for the difference in risk aversion, with Δ_0 as the reference category. The variable $ROUND$ is a variable capturing progression through the experiment in terms of number of rounds completed. Finally, the control vector includes categorical variables for subject nationality and study major as well as the order of interaction with the vendors. For all regressions we excluded data from the warm-up rounds.

Table 3.5 gives the regression estimates for three regressions, where from the first to the second regression and from the second to the third the effects of learning and the controls as well as interaction effects are sequentially introduced. Broadly, we find strong evidence against Hypothesis 2a: information shows to be counterproductive across all three model specifications. More precisely, from the regression coefficients for $INFO$ we see that informed subjects' decisions lead to a significantly larger gap between the first best expected profit and observed expected profits. In line with expectation, coefficients for $\Delta_{0.5}$ and Δ_1 are positive and significant, meaning subjects generally perform worse when differences in risk aversion are larger. The interaction effects between the difference in risk aversion and the information treatment seem give to support Hypothesis 2b. The signs of the interaction effects across the three regressions are correct. However, given that information appears counterproductive in terms of the direct

effect, differences in risk aversion offset the deleterious effect of information rather than add to the hypothesized benefit of information. Considering the effect of time, we observe that subjects don't do significantly better in later rounds. There is a significant interaction with $INFO$, suggesting that the counterproductive effect of information is also offset in later rounds. However, the effect size is negligible compared to the effect of $INFO$.

Table 3.5 Regressions Coefficients for equation Equation 3.3

DV:	$\ln(\mathbb{E}[\Pi_{FB} - \Pi_{Obs}]$		
$INFO$ (yes=1)	1.69***	2.25**	1.94***
	(0.14)	(0.19)	(0.20)
$\Delta_{0.5}$	1.80***	1.80***	1.80***
	(0.14)	(0.14)	(0.12)
Δ_1	2.06***	2.06***	2.06***
	(0.14)	(0.14)	(0.12)
$\Delta_{0.5} \times INFO$	−1.56***	−1.56***	−1.56***
	(0.19)	(0.19)	(0.17)
$\Delta_1 \times INFO$	−1.31***	−1.31***	−1.31***
	(0.19)	(0.19)	(0.17)
$ROUND$		0.00	0.00
		(0.00)	(0.00)
$ROUND \times INFO$		−0.01***	−0.01***
		(0.00)	(0.00)
$Controls$	No	No	Yes
R^2	0.15	0.16	0.35
Log-likelihood	-4374	-4362	-4083
No. of obs. (groups)		2160(24)	

Notes. The group variable is individual subject.

SE's are reported in parentheses.

$^*p < 0.1, ^{**}p < 0.05, ^{***}p < 0.01.$

3.7 Robustness Tests

To ensure the main findings in the previous section are not dependent on the subsample used, we performed a series of robustness tests. We repeated the analysis on median decisions as presented in Tables 3.1, 3.2 and the full regression model results as presented in 3.5 for two additional random 24-subject subsamples (denoted RS1 and RS2) and the full sample of 34 subjects (denoted FS). The subsamples were produced by stratifying the full sample into the experimental conditions and sampling 2 subjects without replacement in each overpopulated experimental condition in the full sample. The results in Table 3.1 and Table 3.2 show the main results are consistent across the two additional subsamples and the full sample. Note that there

are minor differences in the significance for the within subject comparisons for the observed and optimal decisions (lower half of Table 3.1). Also note that, although there are no considerable changes in sign and significance of the main effects between the full regression results in Table 3.5 and Table 3.2, effect sizes are generally smaller for the results produced using the additional subsamples as well as the full sample. Overall we can conclude that our main findings are consistent across the samples.

Table 3.1 Within-subject comparison (median decisions) for additional stratified subsamples

Δ	Investments[a]			Penalties[a]		
	0	0.5	1	0	0.5	1
Hypothesis[b]	$x < x'$	$x' < x''$	$x < x''$	$p > p'$	$p' > p''$	$p > p''$
RS1	−	−	−	∗∗∗	∗∗∗	∗∗∗
RS2	−	−	−	∗∗∗	∗∗∗	∗∗∗
FS	−	−	−	∗∗∗	∗∗∗	∗∗∗
Hypothesis[c]	$x > x^*$	$x > x^*$	$x > x^*$	$p < p^*$	$p > p^*$	$p > p^*$
RS1	∗∗∗	∗∗∗	∗∗	∗∗∗	∗∗	∗∗
RS2	∗∗∗	∗∗	∗∗	∗∗∗	∗∗∗	∗∗∗
FS	∗∗∗	∗∗∗	∗∗∗	∗∗∗	∗∗∗	∗∗∗

Notes. [a] Significance levels for Wilcoxon Signed-Rank tests.

[b] $x, x', x''(p, p', p'')$ respectively indicate investments (penalties) under $\Delta \in \{0, 0.5, 1\}$.

[c] $x^*(p^*)$ indicates a conditionally optimal investment (penalty) subject to γ_c and γ_c.

$-n.s.,$ $^*p < 0.1,$ $^{**}p < 0.05,$ $^{***}p < 0.01$.

3.8 Discussion and Conclusion

Over the duration of an outsourcing relationship, a client and vendor may encounter problems only to be overcome through effective collaboration, whether it be problems arising through product design flaws or failing information technology systems. Contracts are the prime conduit of any such relationship, but leave room for behavioral factors by either party to be in the way of performance. We hypothesize differences in risk aversion between the client and vendor are at the heart of misaligned incentives and, hence, misaligned decisions in anticipation of future problems. We also hypothesize information on the vendor's risk aversion may offer the client some remedy to this misalignment. We test these hypotheses through a controlled experiment in which subjects play the role of the client in a contracting game with a computerized vendor. Our study makes a contribution to the experimental literature on contracting by implementing a novel approach to measure and control for subject risk aversion and provides empirical justification of the importance of considering risk aversion in contract decision making.

Table 3.2 Regressions results for additional stratified subsamples

DV:	$\ln(\mathbb{E}[\Pi_{FB} - \Pi_{Obs}]$		
Sample	RS1	RS2	FS
INFO (yes=1)	1.02***	0.68***	0.69***
	(0.16)	(0.19)	(0.15)
$\Delta_{0.5}$	1.57***	1.47***	1.31***
	(0.09)	(0.12)	(0.10)
Δ_1	1.66***	1.51***	1.39***
	(0.09)	(0.12)	(0.09)
$\Delta_{0.5} \times$ INFO	−1.50***	−1.14***	−0.98***
	(0.13)	(0.17)	(0.15)
$\Delta_1 \times$ INFO	−0.89***	−0.37**	−0.21***
	(0.13)	(0.17)	(0.15)
ROUND	-0.00	−0.00*	0.00
	(0.00)	(0.00)	(0.00)
ROUND $\times$ INFO	−0.01***	−0.00	−0.01***
	(0.00)	(0.00)	(0.00)
Controls	Yes	Yes	Yes
R^2	0.61	0.37	0.29
Log-likelihood	-3452	-4062	-5881
No. of obs. (groups)	2160(24)		3060(34)

Notes. The group variable is individual subject.

SE's are reported in parentheses.

$^*p < 0.1, ^{**}p < 0.05, ^{***}p < 0.01.$

First, regarding subjects' decisions under differences in risk aversion we only find support for Hypothesis 1(b): subjects set lower penalty when facing a more risk averse vendor. Subjects tend to set both higher and lower investments when facing a more risk averse vendor instead of only investing more as theory would predict. Comparing decisions with the conditionally optimal benchmarks we arrive at two observations that highlight possible heuristic decision biases. Firstly, subjects tend to set and hold on to an inefficiently high investment level. Secondly, subjects tend to set and hold on to a penalty that is too high when interacting with more risk averse vendors and too low in case the vendor is equally risk averse. This provides further evidence that the anchoring and insufficient adjustment bias as observed by Schweitzer and Cachon (2000) and other papers in behavioral operations management extend beyond newsvendor decision contexts.

Rather than explaining decisions through heuristics leading to insufficient adjustment, an alternative explanation could be offered, relating to heuristics leading to interior choice preference. As a complement to the well-known central tendency bias, there is experimental evidence from various contexts that subjects tend favor interior decisions over extreme decisions in their de-

cision space, even if extreme decisions are theoretically optimal. For instance, in the case of incumbent bidding against an entrant in a contract-renewal auction context, Wan et al. (2012) find that subjects bid more aggressively than theory predicts because they either avoid undercutting the competition or avoid boycotting the auction by reverting to the pre-auction contract price when the respective strategies are optimal. Particular to our context, this may have lead subjects to avoid setting extreme penalties (i.e. a lower (higher) penalty for a more (less) risk averse vendor), despite the fact that they are theoretically optimal.

Finally, we find strong evidence against Hypothesis 2a. Surprisingly, information on the vendor's risk aversion appears to have counterproductive effects on subject's performance in the experiment. Informed subjects displayed a wider the gap between first best expected profit and observed expected profit. Although it appears the information treatment has a reinforcing effect on the heuristic decision biases observed, there is not sufficient support for this statement. Evidence does appear to support Hypothesis 2b. However, we find that under larger differences in risk aversion, informed subjects actually only end up offsetting inefficient decisions resulting from the information treatment.

Together with the lack of improvement over time between the two treatment, it appears subjects internalize additional information in the form of cognitive feedback to set decision heuristics which they do not adjust sufficiently to changing conditions. However, considered in a different light, this might be a strong indication of limitations to the design of our experiment. In the first place risk aversion is a difficult concept for subjects to reflect on. Even though our approach of using a multiple price list arrives at an implied risk preference, it is likely subjects were either not prepared to make sensible decisions in the pre-test or they were aware of the 'rational' solution and made their decisions accordingly if perceiving this as a desirable outcome. Next the task of deciding on contract parameters in the main experiment is sufficiently complex for it to be likely subjects simply used the decision support in the game window as to search for a satisfactory local optimum. Moreover, under the fixed parameters, the differences in vendor risk aversion from one computerized vendor to the next may not have lead to sufficiently noticeable differences in expected outcomes for subjects to respond to. Finally, the way we decided to operationalize the information treatment as information windows given during the game, rather than additional training upfront, raises the question whether the meaning and potential value of the information indeed came across to the subjects.

Combined, given the complexity of the pre-test, main experiment and information provision stand in the way of the current findings ultimately being convincing. Future work on this chapter will firstly require revision of the experiment to make the context and decisions much easier to grasp, i.e. a relatable context for joint problem solving with only one degree of freedom. Should different information treatments remain a part of future experiments, more care must also be taken to ascertain differences in informed states actually materialize as a result of the treatments. Secondly, the link between subject level risk preferences and subsequent contracting decisions is perhaps too tenuous to factor into the analysis as it has. Instead, risk preferences may

have been inferred directly from subjects' repeated decisions, then linked to subject performance throughout the experiment. Lastly, as a central objective of this chapter is to study collaboration under differences in risk preferences, a potentially larger contribution to the literature would come from removing the computerized vendor and designing a human-to-human experiment instead.

Acknowledging our work has notable limitations, we caution we posit two key managerial implications. The first is that in adjusting to different contracting partners, managers responsible for contract design and implication should be wary of anchoring on previous decisions, particularly when conditions may have shifted. The second is that investing effort in understanding the vendor's risk aversion may make contracting decisions all the more difficult. Particular when it comes to understanding how the vendor's preferences relate to her own, a possible 'cognitive overload' leads to over-reliance on existing heuristics where adjustments are needed to correct for changing conditions. In conclusion, as outsourcing of services and technology become more commonplace and are often paired with contingent contracts, care must be taken to let contracting decisions in advance of possible disruptions facilitate response rather create inadvertent hurdles when disruption does hit.

Appendix

3.A Proofs

Proof of Proposition 1 (Abridged). To complete this proof we first establish the first-best solution, which solves the contracting game as if both investment decisions x, y are made centrally by a single decision maker, without an intervening penalty contract. Formally, this problem can be stated as:

$$(\mathcal{C}) \qquad \Pi^{FB} = \max_{x,y \geq 0} R\left(1 - \frac{\lambda}{\bar{\mu} x^\alpha y^\beta}\right) - c_c x - c_v y. \tag{3.4}$$

The first-best levels of investment can be found by maximizing the payoff over x and y simultaneously. Solving the first order condition and doing some algebraic manipulation, we find the the first-best investment levels:

$$x^{FB} = \left(\frac{\alpha c_v}{\beta c_c}\right)^{\frac{\beta}{\sigma}} \left(\frac{R\alpha\lambda}{\bar{\mu} c_c}\right)^{\frac{1}{\sigma}} \qquad \text{and} \qquad y^{FB} = \left(\frac{\beta c_c}{\alpha c_v}\right)^{\frac{\alpha}{\sigma}} \left(\frac{R\beta\lambda}{\bar{\mu} c_v}\right)^{\frac{1}{\sigma}}, \tag{3.5}$$

where $\sigma = \alpha + \beta + 1$. Now we can turn to solving the contracting game $\mathcal{N}$. The expected profit for the vendor in case of a linear penalty contract is:

$$\mathbb{E}[\Pi_v] = F - \frac{p\lambda}{\bar{\mu} x^\alpha y^\beta} - c_v y. \tag{3.6}$$

We first solve the FOC for y, which gives the vendor's best response function:

$$y(x) = \left(\frac{p\beta\lambda}{\bar{\mu} c_v x^\alpha}\right)^{\frac{1}{1+\beta}}. \tag{3.7}$$

Showing that the second order derivative for the vendor expected profit function is negative indicates the solution for y in this stage is unique given $p > 0$ and $x > 0$:

$$\frac{\partial^2 \mathbb{E}[\Pi_v]}{\partial y^2} = \frac{-(1+\beta)\beta p\lambda}{\bar{\mu} y^{2+\beta} x^\alpha} < 0. \tag{3.8}$$

At optimality the vendor's IR constraint is always binding, the client sets the fixed fee $F^N = \frac{p^N \lambda}{\bar{\mu}(y^N)^\beta (x^N)^\alpha} + c_v y^N + k$. Substituting F^N and $y(x)$ into the client's expected profit function and taking the partial derivative with respect to x produces:

$$\frac{\partial \mathbb{E}[\Pi_c]}{\partial x} = \frac{\alpha\lambda(\beta p + R)}{(1+\beta)\bar{\mu} x^{1+a}} \left(\frac{c_v \bar{\mu} x^\alpha}{\beta p \lambda}\right)^{\frac{\beta}{1+\beta}} - c_c. \tag{3.9}$$

Note that for $p = R$, we can solve for x to find:

$$x^N = \left(\frac{c_v\alpha}{\beta c_c}\right)^{\frac{\beta}{\sigma}} \left(\frac{\alpha\lambda R}{c_c\bar{\mu}}\right)^{\frac{1}{\sigma}}, \tag{3.10}$$

retrieving the first best investment by the client, i.e. $x^N = x^{FB}$. Here again $\sigma = 1 + \alpha + \beta$. Given that the best response function Equation 3.7 can be shown to be the same as found when solving $\mathcal{C}$, the first best investment level is retrieved for y^N as well, such that

$$y^N = \left(\frac{\beta c_c}{\alpha c_v}\right)^{\frac{\alpha}{\sigma}} \left(\frac{R\beta\lambda}{\bar{\mu}c_v}\right)^{\frac{1}{\sigma}}. \tag{3.11}$$

Given that the client can extract all rents by adjusting the fixed fee in the contract and knowing that the first best investments yield the best possible system profit, it follows that setting $p = R$ is indeed optimal. For a proof on uniqueness of the solution we refer to the full proof of Proposition 2 in Chapter 2. $\qquad\square$

3.B Theoretical Benchmarks

Table 3.B.1 Optimal Decisions and Conditional Client Expected Profit

γ_c	γ_v	x^*	y^*	p^*	$\mathbb{E}[\Pi_c^*]$
-0.77	-0.77	3.51	3.51	100.00	81.58
-0.77	-0.27	3.93	2.49	45.02	80.79
-0.77	0.23	4.12	2.17	29.08	80.33
-0.60	-0.60	3.74	3.74	100.00	81.83
-0.60	-0.10	4.16	2.72	48.28	81.23
-0.60	0.40	4.35	2.37	31.82	80.83
-0.43	-0.43	3.96	3.96	100.00	81.98
-0.43	0.07	4.37	2.95	51.14	81.54
-0.43	0.57	4.58	2.57	34.35	81.19
-0.27	-0.27	4.16	4.16	100.00	82.06
-0.27	0.23	4.56	3.15	53.56	81.74
-0.27	0.73	4.77	2.75	36.57	81.44
-0.10	-0.10	4.36	4.36	100.00	82.11
-0.10	0.40	4.76	3.36	55.88	81.88
-0.10	0.90	4.97	2.94	38.78	81.63
0.00	**0.00**	**4.47**	**4.47**	**100.00**	**82.11**
0.00	0.50	4.86	3.47	57.14	81.93
0.00	1.00	5.08	3.05	40.00	81.71
0.07	0.07	4.55	4.55	100.00	82.11
0.07	0.57	4.94	3.56	57.98	81.96
0.07	1.07	5.16	3.12	40.83	81.75
0.23	0.23	4.72	4.72	100.00	82.08
0.23	0.73	5.11	3.74	59.79	82.00
0.23	1.23	5.33	3.29	42.64	81.83
0.40	0.40	4.90	4.90	100.00	82.04
0.40	0.90	5.28	3.92	61.54	82.01
0.40	1.40	5.50	3.46	44.44	81.87
0.73	0.73	5.22	5.22	100.00	81.89
0.73	1.23	5.59	4.27	64.54	81.95
0.73	1.73	5.82	3.79	47.64	81.87
0.90	0.90	5.39	5.39	100.00	81.80
0.90	1.40	5.74	4.44	65.91	81.90
0.90	1.90	5.97	3.95	49.15	81.84

Notes. The solutions in columns 3-5 were obtained by solving problem $\mathcal{A}$ in Section 4.3 using γ_v and γ_c in columns 1-2. The optimal decisions in were then substituted back into the client's profit function under risk neutrality found through Proposition 7 in order to compare outcomes under differences in risk aversion. The FB solution is highlighted in **bold**.

3.C   Instructions

This is an experiment about decision-making in a business context. No prior knowledge or experience is necessary to take part in this experiment. The instructions below tell you all you should know. First it is important that from this moment onwards you focus on your own station: please do not create a disruption by talking, laughing or making noises. The investigator can ask you to leave the room without warning. If this happens you will only receive the participation fee. If you complete the experiment, your total payoff will depend on your performance.

Game Scenario

In this experiment we simulate a setting in which a CLIENT (e.g. a large bank) has outsourced the maintenance of an IT system to a VENDOR. The system is subject to random disruptions. The IT system is assumed to be a mission critical system such that the CLIENT's entire revenue flow stops any time the system is in a state of disruption. Both the CLIENT and the VENDOR must make a response capacity INVESTMENT to make sure the time-to-repair after any disruption is minimized. Moreover, the outsourcing relation between the CLIENT and the VENDOR is governed by a contract with a PENALTY rate for downtime. Alongside her investment, the CLIENT sets this rate for the VENDOR to pay in case of downtime. The CLIENT's decisions impact what the VENDOR decides to invest. This then impacts the payoffs as the investments by the CLIENT and VENDOR are complementary and jointly lower time-to-repair. The CLIENT and VENDOR therefore each need to find the balance between the cost of investing in response capacity and the potential cost of downtime in attempt to maximize their respective profits.

Experimental Procedure

The experiment consists of two main parts. In PART I you will be asked to complete several sequences of questions spread across eight sections. The section is indicated at the top left of the window. You will need to complete all questions in each section to advance to the next section. Click 'Save and Continue' when you have completed the section and you are happy with your answers. There is a 'Clear' button in each section which resets your answers, which you can use in case you want to restart the section. Completing all eight sections is required to advance to PART II and obtain your payoff for today's experiment.

In the first section of PART I you are asked to choose one of two gambles in ten hypothetical situations. In the second section you are asked to choose between a gamble and a value you would receive for sure in another ten hypothetical situations. Each gamble has an expected value equal to the weighted average of the two outcomes. For instance, a gamble that says "50% of \$5, 50% of \$10" has a 50% probability of resulting in a payoff of \$5 and a 50% probability of resulting in a

payoff of $10. The expected value of the gamble is $50\% * \$5 + 50\% * \$10 = 0.5 * \$5 + 0.5 * \$10 = \$7.5$. You are not expected to calculate these expected values, but a quick mental calculation to get an idea of the expected value of a gamble may help inform your choice. Keep in mind the gamble never pays out the expected value: there is a chance of obtaining the high outcome, but also a risk of obtaining the low outcome. Sections three until five each describe ten separate activities or behaviors for which you are asked to indicate how likely it is that you would engage in the activity or behavior. Sections six until eight present the same situation, but instead ask you to rate the situations in terms of how risky you perceive them to be.

In PART II you will play the role of a CLIENT in an investment game in the abovementioned service chain setting with one VENDOR. The VENDOR is automated and acts in his own interest. As the CLIENT you must decide on the PENALTY and your INVESTMENT in response capacity in each round. You will play the game for 30 periods across a number of rounds with different conditions to which you will be randomly assigned. In each period you will have the chance to try out different penalty and investment decisions using the relevant scroll bars and see expected results following your decisions before you submit them. Do this in the 'Try Before You Decide' panel in the game window. Do not click on the 'submit' button or hit enter until you are sure. After submitting your decisions, you will be able to observe the vendor's response and realized results in the 'Performance Log' panel in the game window. At the start of the game you will have five warm-up periods to get used to the user interface. Your performance in the warm-up rounds is not recorded and will not count towards your payoff. Your expected profit in each period is calculated for you using the following formula:

$$\mathbb{E}[\text{profit}] = R \left(1 - \frac{\lambda}{\bar{\mu} x^\alpha y^\beta}\right) + \frac{p\lambda}{\bar{\mu} x^\alpha y^\beta} - c_c x - F, \tag{3.12}$$

where in the first term R is the revenue per period when no disruptions occur, λ is the expected frequency of disruptions per period, $1/(\bar{\mu} x^\alpha y^\beta)$ is the expected length for each disruption. Specifically, μ, α and β are parameters that together with the CLIENT's (your) INVESTMENT x and the INVESTMENT by the VENDOR y produce a joint response capacity. Furthermore, p is the penalty you set, c_x is the cost per unit of investment and F is a fixed fee that is part of contract with the VENDOR. For the game, we set $R = 100$, $\bar{\mu} = 5$, $\lambda = 2$, $\alpha = \beta = 0.5$, $c_x = 1$ and F is determined automatically to meet some constraints of the game. Because $\alpha = \beta$ it is assumed both the CLIENT and VENDOR have the same return on investment in terms of reducing expected downtime. Realized profits will be different from the expected profit as in each round a random number of disruptions and a random length for each disruption that occurs is generated. For instance, say you choose to set your investment x at 4 and set the penalty p at 40. Now let's say the VENDOR decides to invest 1 and to meet the game constraints, F is set at 10. Then your expected profit is:

$$\mathbb{E}[\text{profit}] = 100 * \left(1 - \frac{2}{5 * \sqrt{4}\sqrt{1}}\right) + \frac{40 * 2}{5 * \sqrt{4}\sqrt{1}} - 1 * 1 - 10 = 77. \tag{3.13}$$

In expectation you face two disruptions each period (as $\lambda = 2$) and each disruption is expected to last one-tenth of the period (as $1/(5 * \sqrt{4}\sqrt{1}) = 1/10$), such that the expected cumulative

downtime is two-tenths of the period. As the rate of disruptions and the length of each disruption are random, however, the realized number of disruptions may be different (e.g. 1, or 3) and the length of each disruption may be different from 1/10 as well (e.g. 1/20, or 1/4). As a result the cumulative downtime may be more or less than the expected 2/10, resulting in a different realized profit for the period. As mentioned before, the realized results are recorded in the 'Performance Log' panel.

Your Payoff

Your payoff consists of two parts. You will receive a SGD 5 participation fee and additional payment conditional on your performance. Your performance-based payoff will be based on your cumulative profits in PART II of the experiment. The total CLIENT's earnings are added up and converted to Singapore Dollars at a rate of SGD 1 per 200 points. You can make up to SGD 20 including the participation fee. After you answer the questions appearing on your screen, your result will appear. **Please do not close your game window**. You will need to confirm your result with the investigators before writing it down on your payment paper with your bank information and completing the IDC claim form. The payment will be transferred to your bank account. Alternatively, you can request for a cheque to be issued. Please make sure that you write your name exactly as it appears on your bank account. In case you leave before completing the experiment, you will only receive the participation fee.

3.D   Screenshots

Subjects participated in the experiment using software developed in *Visual Basic* using *Microsoft Visual Studio Professional 2013*. The figures on the following pages show screenshots of the HL and HLa measures and a sample of questions from the DOSPERT measures in PART I of the experiment; the user interface for PART II of the experiment; and finally an example of the three information pop-ups seen by a subject in the information treatment and allocated to the vendors in order (213) upon starting the game and completing rounds 30 and 60.

Figure 3.D.1 PART I: HL Measure

Figure 3.D.2 PART I: HLa Measure

Figure 3.D.3 PART I: Sample of the DOSPERT Measure questions

Figure 3.D.4 PART II: Investment Game

Figure 3.D.5 PART II: Sample message prompts for an informed subject facing vendors in order (213)

(i) First message

Investment Game

You will interact with three vendors with varying degrees of risk aversion. Vendors with differing levels of risk aversion will respond differently to your contract and investment decisions. You will interact with each vendor for thirty consecutive rounds after five warm-up rounds. Relative to each vendor, you have a lower level of risk aversion.

OK

(ii) Second message

Investment Game

You will now interact with the second vendor. The second vendor is less risk averse than the first vendor, so he will respond differently to your decisions. Your level of risk aversion is still lower than the second vendor.

OK

(iii) Third message

Investment Game

You will now interact with the third vendor. The third vendor is more risk averse than either of the previous two vendors. Your level of risk aversion is still lower than the third vendor.

OK

References

Anderson LR, Mellor JM (2009) Are risk preferences stable? comparing an experimental measure with a validated survey-based measure. *Journal of Risk and Uncertainty* 39(2):137–160.

Ang JS, Schwarz T (1985) Risk a version and information structure: An experimental study of price variability in the securities markets. *The Journal of Finance* 40(3):825–844.

Bakshi N, Kim SH, Savva N (2015) Signaling new product reliability with after-sales service contracts. *Management Science (Articles in Advance)* 0(0):1–18.

Balzer WK, Doherty ME, O'Connor R (1989) Effects of cognitive feedback on performance. *Psychological Bulletin* 106(3):410.

Bapna R, Dellarocas C, Rice S (2010) Vertically differentiated simultaneous Vickrey auctions: theory and experimental evidence. *Management Science* 56(7):1074–1092.

Bhattacharyya S, Lafontaine F (1995) Double-sided moral hazard and the nature of share contracts. *RAND Journal of Economics* 26:761–781.

Blais AR, Weber EU (2006) A domain-specific risk-taking (dospert) scale for adult populations. *Judgment and Decision Making* 1(1).

Bolton GE, Katok E (2008) Learning by doing in the newsvendor problem: A laboratory investigation of the role of experience and feedback. *Manufacturing & Service Operations Management* 10(3):519–538.

Chan T, De Vericourt F, Besbes O (2016) Contracting in medical equipment maintenance services: an empirical investigation, working Paper, ESMT.

Charness G, Gneezy U, Imas A (2013) Experimental methods: Eliciting risk preferences. *Journal of Economic Behavior & Organization* 87:43–51.

Friedman D, Sunder S (1994) *Experimental methods: A primer for economists* (Cambridge university press).

Guajardo JA, Cohen MA, Kim SH, Netessine S (2012) Impact of performance-based contracting on product reliability: An empirical analysis. *Management Science* 58(5):961–979.

Holt CA, Laury SK, et al. (2002) Risk aversion and incentive effects. *American Economic Review* 92(5):1644–1655.

Jackson DN, Hourany L, Vidmar NJ (1972) A four-dimensional interpretation of risk taking. *Journal of Personality* 40(3):483–501.

Jain N, Hasija S, Popescu DG (2013) Optimal contracts for outsourcing of repair and restoration services. *Operations Research* 61(6):1295–1311.

Kalkanci B, Chen KY, Erhun F (2011) Contract complexity and performance under asymmetric demand information: An experimental evaluation. *Management Science* 57(4):689–704.

Katok E, Wu DY (2009) Contracting in supply chains: A laboratory investigation. *Management Science* 55(12):1953–1968.

KC D, Staats BR, Gino F (2013) Learning from my success and from others' failure: Evidence from minimally invasive cardiac surgery. *Management Science* 59(11):2435–2449.

Kim SH, Cohen MA, Netessine S (2007) Performance contracting in after-sales service supply chains. *Management Science* 53(12):1843–1858.

Kim SH, Cohen MA, Netessine S (2015) Reliability or inventory? an analysis of performance-based contracts for product support services .

Kim SH, Cohen MA, Netessine S, Veeraraghavan S (2010) Contracting for infrequent restoration and recovery of mission-critical systems. *Management Science* 56(9):1551–1567.

Kim SH, Tomlin B (2013) Guilt by association: Strategic failure prevention and recovery capacity investments. *Management Science* 59(7):1631–1649.

Kremer M, Moritz B, Siemsen E (2011) Demand forecasting behavior: System neglect and change detection. *Management Science* 57(10):1827–1843.

Laffont JJ, Martimort D (2009) *The theory of incentives: the principal-agent model* (Princeton university press).

Lee YS, Siemsen E (2013) Task decomposition and newsvendor decision making. *Available at SSRN 2232073* .

Lurie NH, Swaminathan JM (2009) Is timely information always better? the effect of feedback frequency on decision making. *Organizational Behavior and Human Decision Processes* 108(2):315–329.

Pratt JW (1964) Risk aversion in the small and in the large. *Econometrica: Journal of the Econometric Society* 122–136.

Roels G, Karmarkar US, Carr S (2010) Contracting for collaborative services. *Management Science* 56(5):849–863.

Roth AE, Malouf MW (1979) Game-theoretic models and the role of information in bargaining. *Psychological review* 86(6):574.

Sawhney M, Balasubramanian S, Krishnan VV (2004) Creating growth with services. *MIT Sloan Management Review* 45(2):34.

Schweitzer ME, Cachon GP (2000) Decision bias in the newsvendor problem with a known demand distribution: Experimental evidence. *Management Science* 46(3):404–420.

Sengupta K, Abdel-Hamid TK (1993) Alternative conceptions of feedback in dynamic decision environments: an experimental investigation. *Management Science* 39(4):411–428.

Shotter J, Noonan L, Davies S (2015) Deutsche Strikes Deal with HP on IT System. URL `http://www.ft.com/cms/s/0/c9e98f82bc3811e4b6ec00144feab7de.html`, retrieved 18/03/2016.

Siegel S (1957) Nonparametric statistics. *The American Statistician* 11(3):13–19.

Sigman K (2010) Inverse transform method. URL `http://www.columbia.edu/~ks20/4404-Sigman/4404-Notes-ITM.pdf`, retrieved 22/02/2016.

Sterman JD (1989) Modeling managerial behavior: Misperceptions of feedback in a dynamic decision making experiment. *Management science* 35(3):321–339.

Taneri N, De Meyer A (2016) Contract theory: Implications for biopharmaceutical alliance structure and performance. *Available at SSRN 2269381* .

Thaler RH, Tversky A, Kahneman D, Schwartz A (1997) The effect of myopia and loss aversion on risk taking: An experimental test. *The Quarterly Journal of Economics* 647–661.

Visnjic I, Taija T, Neely A (2013) When innovation follows promise. Technical report, Cambridge Service Alliance (University of Cambridge).

Weber EU, Blais AR, Betz NE (2002) A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. *Journal of Behavioral Decision Making* 15(4):263–290.

Chapter 4

Managing Escalations: Equipment Failure and Response Capacity Allocation

4.1   Introduction

Complex technologies increasingly support the day-to-day operations of firms and organizations across many sectors: airlines depend on direct access to booking systems, the financial markets require live and secure financial data streams and hospitals need their system of medical equipment to be safe, ready and available. While technology helps to improve operating efficiency, they also increase exposure to disruption risk when the maintenance is outsourced. This exposure is further amplified when multiple clients contract with a single provider to implement and service their technology. Motivated by past disruptions in the medical devices sector, the goal of this study is to understand how a provider should manage disruption risk across a clients base.

A report by the McKinsey Center for Government (2013) notes that while US medical device sales grew at a steady clip of 9% between 2001 and 2009, the number of patient adverse events linked to medical devices grew at nearly twice that rate. Sheffi (2005) highlights a series of such adverse events related to Baxter International, a medical equipment company facing blame for 53 deaths in the United States and across Europe linked to their dialysis equipment in 2001 (Hammonds 2002). First indications of a problem came to light in Madrid mid-August when four elderly patients died within hours of receiving routine dialysis. It was not until another string of more unusual post-dialysis deaths occurred in Valencia a week later that red flags were raised. Matters escalated to a full inquiry when another 21 patients died under similar circumstances in Croatia. A worldwide product recall of the dialysis filters followed when two patient casualties also occurred in the US. Yet as Baxter's equipment was not the only common factor between the reported incidents, determining whether the dialysis filters were indeed the root cause could not readily be confirmed (Young 2001a, Young 2001b). A few weeks into the investigation it was found that traces of a fluid called perfluorohydrocarbon that is used for testing leaks in the filters during production with a subsidiary had been left behind on a particular series of filters. Initially, the conclusion of the inquiry was that this fluid may have lead to pulmonary embolisms with the patients (Young 2001c). Later, in early 2002, new evidence pointed to a possible poisonous

by-product of production with the fluid leading to the deaths (Young 2002). Hard evidence for the true cause of the casualties was never found; still the company had to absorb $189M in costs as a result of the events and subsequent response (Hammonds 2002). Strikingly, the US Food and Drug Administration (FDA) did not announce their investigation into the causes of the disruption until November. This was after the company had already issued a worldwide product recall following their own investigation (FDA 2002), suggesting initial response may often be enacted by the provider – well before any regulatory and legal action.

In his discussion of the Baxter case, Sheffi underlines that a company like Baxter can neither afford responding to all red flags, nor can they afford to miss a defect that leads to preventable deaths. Baxter International has manufacturing and R&D facilities spread across the world and provides its products and maintenance services to client hospitals in 100 countries (Baxter 2001), meaning one small defect can have a major impact. Moreover, the complexity of casualties can come with significant noise, such that it is hard to discern the reality of any red flags. In situations like Baxter's, allocating response capacity to where it is truly needed is critical. Yet each hospital is primarily concerned with preventing further casualties and ensuring their faulty equipment is repaired or replaced as soon as possible, rather than focusing on investigating and reporting equipment failures. Recent reports by the FDA point toward frequent misreporting to manufacturers by hospitals despite federal regulations (Terhune 2016). In one instance, failure among hospitals to report casualties linked to contaminated medical scopes put 350 patients across 41 medical centers worldwide at risk.

As a result of outsourcing mission critical equipment, service providers and clients each have roles to play in response to disruption. While the provider possesses the technical expertise to resolve issues, the clients operate the technology on a day-to-day basis and are likely to be more informed on the nature of problems when they do arise (Chan et al. 2014). In light of the case involving Baxter International, one of the hospitals (i.e. the clients) was first to note the problems with the dialysis equipment and Baxter (i.e. the provider) had to decide on what action to undertake to control the damage. However, as the nature of disruptions is rarely understood fully at the onset, neither the clients nor the provider can be certain whether the disruption is 'local' (only affecting one hospital) or 'global' (affecting multiple or even all clients). What the provider and clients believe the scale of disruption is, will be primarily determined by the red flags they have observed (Sheffi 2005). Moreover, information held between the provider and clients is frequently fragmented as each firms' reception and interpretation of disruption intelligence are unobservable. This makes it difficult for the provider to respond effectively to a disruption possibly affecting multiple clients.

Adequate response to disruptions depends on the provider's ability to decide on and the deliver the necessary response measures after onset of disruptions. The ability to discern global disruptions from local problems in turn depends on the efforts made by the clients to identify whether response is necessary as well as incentives to communicate the potential extent of the disruption truthfully. Clients may call for support or initiate response measures after noticing a

malfunction in the equipment or reports of similar problems experienced elsewhere.[1] Placing a call for support either directly or via local authorities means the client escalates the problem to the provider to initiate a response. Investments in disruption prevention and response improvement alone thus do not guarantee that the resources are put to use in the most effective way. Misallocation of limited capacity to respond to disruption prevents faster resolution, thereby increasing the cost of disruption.

Generalizing from the case of Baxter International, three factors may result in suboptimal response to disruptions. Firstly, response capacity may be misallocated as a result of a tension between the need for the clients to be judicious in raising alarm with the provider, i.e. avoiding false alarms (false positive cost), and the need for the clients and provider to be proactive in response, i.e. avoiding potential negligence (false negative costs). Secondly, some equipment is easily sold to many clients, whereas the after-sales service is difficult to manage at scale (Cohen et al. 2006). This makes the available response capacity a scarce good and exacerbates the effect of misallocation through competition over these resources. Thirdly, the contract that facilitates the interactions between a provider and various clients may set the wrong incentives to initiate response calls when not accounting for hidden action and information (moral hazard and adverse selection).

In this chapter we examine the relation between the provider's response capacity allocation decision and the clients' response behavior given private information on the type of disruption. We focus on the provider's trade-off between 'broad' deployment of measures to respond to a potentially widespread problem, at the risk of wasting costly resources, and 'focused' deployment of measures to respond to a potentially isolated problem, at the risk of costly response delays in other locations. Specifically, we analyze how the provider should design a contract to manage allocation of response capacity among clients by ensuring clients invest in the ability to identify the potential ramifications of a disruption and report truthfully in calling for response.

The rest of the chapter is structured as follows. In Section 4.2 we give an overview of the relevant literature. In Section 4.3 we discuss the model setup and assumptions. In Section 4.4 we examine the first-best result under centralized decision making before moving on to find the results of the contracting game between the provider and clients. Finally, in Section 4.5 we conclude and discuss the managerial implications of our findings.

4.2 Literature Review

Existing literature in operations management (OM) has extensively discussed supply disruption risk management. Firms can control their exposure to disruptions resulting from outsourcing

---

[1] Alerts surrounding the Baxter equipment were raised through local Croatian authorities when the incidents in Croatia were linked to those in Spain, triggering Baxter to send its representatives to the affected site (BBC News 2001).

activities via a combination of investments in inventory, multi-sourcing, back-up suppliers, resource flexibility, procurement through PSPs, and access to spot markets (see e.g. Van Mieghem (1998), Tomlin (2006), Yang et al. (2009), Wang et al. (2010), Yang et al. (2012) and references therein). Tang (2006) provides a more a extensive review of earlier literature on supply chain risk management. In the context of after-sales product or system performance, various papers have examined the use of performance-based contracts to improve supplier inventory decisions (Kim et al. 2007, Bakshi et al. 2015) and decentralized investments in prevention (reliability) and 'response capacity' [2] (Kim et al. 2010, Jain et al. 2013, Kim and Tomlin 2013). These papers typically cast maintenance service contracting in a classical principal-agent framework where the client acts as the principal. However, the need for collaboration alters the nature of contracting (Roels et al. 2010), such that suppliers may take the role of principal. This is exemplified by medical equipment manufacturers offering contracts to control incentives of the equipment operators (e.g. hospitals) as well as their own (Chan et al. 2014). Although it is recognized that allocation of resources to emergency tasks is a major operational challenge (Angalakudati et al. 2014, Wex et al. 2014), little work in OM has addressed the possible detriment of tactical misuse of response capacity as a result of preceding strategic (contracting) decisions.

We study a two-level service system where clients (providing services to patients) are first to recognize a disruption and stand to benefit from triggering the provider's response (maintenance service). As such the interaction between the clients and the provider is similar to what is studied in the gatekeeping literature. This literature studies settings characterized by a gatekeeper who classifies incoming customers or patients and decides on providing the service herself or referring to a specialist. The referral decision is made complicated by the fact that the gatekeeper runs the risk of not providing the right service and the specialists' capacity is limited and costly (Argon and Ziya 2009). In the OM literature, Shumsky and Pinker (2003) are the first to capture this setting as a principal-agent problem where the gatekeeper has private information on the complexity of the service required as well as her ability to provide the right service. Their focus is on designing incentives for the agent to make system optimal referral decisions. A later paper by Lee et al. (2012) extend this by considering the possibility to outsource either the expert or gatekeeper. In our study the gatekeeper (client) similarly has private information on the service (disruption response) required and has a decision to make on escalating the problem to the specialist (provider). What is different in our case is that multiple clients compete for the response capacity of the provider, which may incentivize miscommunication, and both the provider and clients play a role in resolving the disruption.

We further depart from the classic assumptions in the gatekeeping literature when it comes to the clients' incentives to under-investigate and misreport disruption status for their own benefit. The literature on audit and hazard disclosure has studied these problems. Baiman et al. (2000) were among the first to model the the effect of contractual incentives on product quality in the presence of moral hazard. They consider both moral hazard on the part of the supplier and

---

[2]Also referred to as repair or recovery capacity in the literature.

buyer, respectively to do with product quality and quality appraisal decisions. Recent papers in this stream of work are generally concerned with regulators or buyers controlling exposure to environmental and social hazards through firms' or suppliers actions. Plambeck and Taylor (2015) examine how buyers should incentivize suppliers to exert responsibility effort rather than to exert effort in hiding information to pass audits. They find increasing audit frequency or penalties conditional on revealed audit evasion may backfire. Regulators' disclosure requirements may equally backfire by removing firms' incentives for voluntary disclosure as well as incentives to investigate potential impacts of hazards in the first place. Kim (2015) and Wang et al. (2016) both build a dynamic model to examine how a regulator's decision on penalties and inspection policy (random or periodic) and a producer's disclosure strategy (voluntary disclosure or risking detection) interact. Kim (2015) find that producers facing random inspection may find partial disclosure beneficial, omitting violations to keep producing in certain circumstances. Wang et al. (2016) instead find that the optimal inspection policy and resulting disclosure depend on options to provide rewards for disclosure as well as the post-investigation signal precision. In both papers, hazards occur randomly and without intent of the firm at risk. The exception is the paper by Babich and Tang (2012), who compare the efficacy of deferred payments and inspection in controlling supplier incentives for deliberate product adulteration. In our setting disruptions do not result from deliberate actions by the clients. However, clients can deliberately limit investigation effort and miscommunicate to the provider.

Two papers outside the abovementioned streams of work are close to this study because of their model set up and assumptions. A paper by Levitt and Snyder (1997) consider a principal-agent setting in which an agent has private information on the likelihood of success of a project. A key finding of the paper is that a principal should commit not to intervene on the basis of received information to efficiently incentivize the agent to exert high effort in project performance and report the project's chances truthfully before completion. In our setting, one of agents' decision to report a potential bad outcome early (i.e. the scale of the disruption) through a response call may in instead be undesirable if the information is not accurate. We study how the principal's strategy should control the escalation behavior of the agents through an appropriate contract design. Also relevant is the paper by Schlapp et al. (2015), who consider how incentives affect how project managers in a firm evaluate and report their project status in competitions over the firm's resources. One of their main findings is that the precision in the evaluation process determines whether incentives should be focused on bringing out status communication with the firms' aggregate performance in mind or self-serving evaluations to allow agents to concentrate on their added value. Another finding is that information asymmetry may ultimately lead to the firm underinvesting in information acquisition and spreading resources too thin. Rather than identifying the right projects to allocate resources to, our model is concerned with targeted allocation of resources to respond to equipment failure effectively.

4.3   Model Assumptions

Consider two clients $i \in \{1, 2\}$ who are both dependent on the same type of equipment for continuity of their operations: if the equipment malfunctions, operations are disrupted and the clients experience disruption costs that are proportional to the disruption length.[3] Upon arrival of a disruption, either both (a 'global' disruption) or just one of the clients is affected (a 'local' disruption). If affected, a client (she) has private and imperfect information on the nature of the disruption. We assume the two clients are ex ante symmetric. Note, however, that because of the stochastic nature of disruption, one clients is the first (and possibly only) to observe indications of the disruption. Without loss of generality we will refer to this client as client 1 or the 'first responder'. The system that the clients depend on is original equipment that has been implemented by a single provider p (he), who is responsible for maintenance servicing of the equipment for both the clients. The provider learns about the disruption nature through communication with client 1 and allocates response capacity on the basis of this communication. Over the course of the disruption, all players learn the true nature of the disruption.

In case equipment failure occurs, the clients resolve the problem through a combination of their own input and the provider's support. Each client has a local response capacity to respond to a disruption independently and is able to further resolve the problem faster with help of the provider's allocated response capacity. Moreover, the provider has a limited response capacity such that he cannot commit full response measures to both clients simultaneously. Therefore, client 1 must decide whether to raise alarm and, if raising alarm, whether to raise alarm for a local or global disruption. To inform this decision, client 1 must first exert investigation effort to improve her understanding of the nature of the disruption. Should client 2 be disrupted as well, her disruption costs add to the system disruption costs. It is therefore in the interest of the provider to allocate available response capacity to her as well.

The provider internalizes downtime costs proportional to the cumulative disruption length across both clients. The provider therefore faces a trade-off between rolling out a broad response, i.e. allocating response capacity to all possibly affected clients, and executing a focused response to client 1 first to learn about the nature of the disruption before rolling out further response measures. If client 1 does not have an incentive to exert effort in investigating the nature of disruption or might benefit from misrepresenting her information to the provider, the provider is at risk of misallocating response capacity. Knowing this, the provider aims to minimize the disruption costs net of contract payments by setting the right incentives for clients to investigate disruptions and report truthfully.

We study this context analytically through a three-stage resource allocation game, which has the following sequence of events (see Figure 4.1). In phase (i) the provider offers a take-it-or-

---

[3]We assume this to include all financial, operational and legal costs ensuing disruption. In reality, as exemplified in the Baxter case, these costs may not be immediately evident. However, clearly the longer a disruption of that nature is left unaddressed, the more costs will result over time.

leave-it contract to the clients. In phase (ii) a disruption potentially common to both clients, but of unknown nature arrives and client 1 decides on the investigation effort to exert to gain information on the scale of disruption through an imperfect private signal. Given the signal about the disruption client 1 has the choice to call the provider for support and, if making the call, decides between reporting a local disruption or global disruption[4] — upon which the provider commits the required response capacity to both clients. Once the provider has committed a share of response capacity to each client, the resources are committed for the duration of the game. In case client 1 reports a local disruption, the provider only allocates support to client 1 in this phase, otherwise both clients receive response capacity. In phase (iii) the true nature of the disruption is revealed through involvement of the provider - upon which the provider allocates the remaining committed response capacity to client 2 in case the disruption is of the global type and client 1 decided to report a local disruption in the previous phase. If client 1 did not send a message in phase (ii) the provider does not learn the true nature of the disruption.[5] In the following we first specify the information structure that is fundamental to the game. Next we formalize the response capacity allocation process and determine the players' respective payoff functions.

Figure 4.1 Sequence of Events
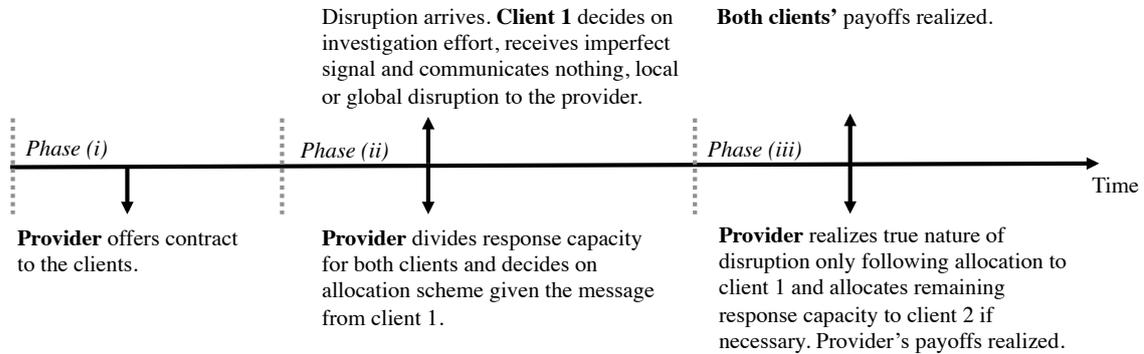
4.3.1 Information Structure and Message Space

We assume the disruption can be one of two types: a local disruption ($\Psi = l$) that affects only one of the two clients or a global disruption ($\Psi = g$) that affects both clients. In phase (ii), client 1 receives a random private signal $s \in \{l, g\}$ that gives an imperfect indication of the nature of the disruption. Analogous to the paper by Schlapp et al. (2015) and references therein, the

---

[4]Equipment or additional services may in some cases make continuous performance feedback available to the provider and clients (for example Philips Remote Services (Philips 2014)), although it is not yet widely implemented because of perceived risks (Paluch and Wünderlich 2016). In this case, whether and how a client escalates the problem in addition to the performance feedback is still a relevant decision. As the Baxter case illustrates, disruptions may arise despite seemingly normal equipment performance.

[5]We assume monitoring by the provider at the clients' premises is prohibitively costly, hence eliciting warnings from the client will depend on the incentive scheme (Levitt and Snyder 1997, Wu and Babich 2012).

informativeness of the signal received is characterized by signal precision q. Formally, $q \in [1/2, 1)$ is the probability that the signal reflects the true nature of the disruption. We assume signal precision is conditionally independent given the true type of the disruption (Wang et al. 2016), such that $Pr(s = g | \Psi = g) = Pr(s = l | \Psi = l) = q$.

To receive an informative signal with precision $q \in (1/2, 1)$, client 1 must exert a high investigation effort ($e = h$) in phase (ii). Otherwise, by exerting low effort ($e = l$) she remains uninformed, i.e. $q = 1/2$. Exerting high effort comes with a cost of k, whereas exerting low effort has zero cost to the client. As such the investigation effort cost function can be captured with $k \mathbb{1}_{\{e=h\}}$. The signal precision q that is feasible with high effort captures the client's ability to learn through investigation: the closer q is to 1, the more accurate the information gained through investigation effort is. Hence, the signal received could also be interpreted as the (imperfect) conclusion of the client's investigation. The achievable signal precision q after high investigation effort is common knowledge to both clients and the provider, but the client's decision whether to exert high investigation effort is not observable to the provider. We assume the provider does not receive a signal of his own and remains uninformed on the nature of the disruption unless he receives a message from client 1.

Given the signal, client 1 decides whether to call for response and, next, whether to report the signal truthfully to the provider by means of a message $m \in (l, g)$, i.e. whether to report a local disruption ($m = l$) if $s = l$ or report a global disruption ($m = g$) if $s = g$, or report a message in contradiction to the signal. In phase (iii) the true nature of the disruption is revealed as long as client 1 decided to communicate ($m \neq \emptyset$). As client 1 only decides on communication following a signal concerning an acknowledged disruption, we assume a 'pure' false alarm does not occur in our model: client 1 cannot raise alarm for a non-existing disruption. Client 1's alarm can, however, be false to the extent that it under- or overstates the gravity of the disruption.

4.3.2 Response Capacity Allocation

To respond to the disruption, each client has her own baseline response capacity μ_0 (Kim and Tomlin 2013) and receives a share of the provider's available response capacity μ_p on the basis of the message communicated by client 1. This increases her effective response capacity μ_i. The clients' baselines response capacities and the provider's available response capacity are common knowledge to the provider and both clients. In case client 1 decides not to call for support, she will not receive a share of the provider's total response capacity. We assume each client can receive at most $\bar{\mu}$ from the provider and, to avoid trivial cases, impose $\bar{\mu} < \mu_p < 2\bar{\mu}$. If client 1 reports a local disruption ($m = l$) in phase (ii), client 1 will receive $\bar{\mu}$ from the provider, such that $\mu_1(l) = \mu_0 + \bar{\mu}$. In this case $\mu_p - \bar{\mu}$ remains available and is committed to client 2, such that $\mu_2(l) = \mu_0 + \mu_p - \bar{\mu}$. Instead, if client 1 reports a global disruption in phase (ii), the provider splits the available response capacity equally, meaning $\mu_1(g) = \mu_2(g) = \mu_0 + \mu_p/2$. Note that, given $\bar{\mu} < \mu_p < 2\bar{\mu}$, it must mean that $\mu_p - \bar{\mu} < \frac{\mu_p}{2} < \bar{\mu}$. The implicit assumption here is that

the provider does not outright ignore a call for support, yet if client 1 does not call for response the provider will not be able to allocate response capacity in time meaning $\mu_1 = \mu_0$. Because of symmetry, this means in this situation also $\mu_2 = \mu_0$.

Additionally, we model a 'patching' benefit that results from sequenced allocation of response capacity. More precisely, allocating $\bar{\mu}$ to client 1 in phase (ii) and allocating the remaining response capacity $\mu_p - \bar{\mu}$ to the other client in phase (iii) boosts client 2's total response capacity by α. The argument for this benefit is that focusing response capacity on one client early may lead to solutions also applicable to the other client, similar to software patches.[6] Where q captures the learning ability of the client, α can be interpreted as capturing the learning ability of the provider, or alternatively the complexity of the disruption faced. The higher α is, the more applicable solutions found by dispatching response capacity to one client are to resolve issues with other clients. Note that if the disruption truly is global, broad response capacity is optimal from the provider's point of view, whereas focused response capacity allocation is optimal if the disruption turns out to be local. In case client 1 calls for focused allocation, but the disruption is in fact global, patching counteracts part of the cost following client 1's potential miscommunication, but cannot offset the damage fully.

4.3.3   Contract and Payoffs

The two types of disruption differ by what rate d_i^{Ψ} disruption costs build up each client as long as a disruption is not resolved. We assume the costs of disruption are proportional to the length of the disruption and that the cost rates are common knowledge to the provider and clients. A global disruption has a cost rate of $d_i^g > 0$ for each client, such that $d_1^g = d_2^g = d^g$ and a local disruption with client 1 has a cost rate of $d_1^l > 0$ and $d_2^l = 0$, where $d^g > d_1^l$ and $d^g < d_1^l$ are both possible scenarios.

In phase (i), the provider announces a pay-per-service contract with a 'disclosure' reward $T(f, r)$ to both clients, which sets a fee $f \geq 0$ for a client to receive response capacity from the provider and a disclosure reward $r \geq 0$ in case client 1's message coincides with the true nature of the disruption ex post, i.e. $m = \Psi$. The provider's contract parameter decisions are his levers to control incentives for client 1's escalation behavior. Empirical evidence suggests a pay-per-service contract results in higher service reliability and more efficient service provision when compared to full coverage plans (Chan et al. 2014). We therefore focus our analysis on the performance of a pay-per-service contract in combination with the disclosure reward. We assume that the provider cannot discriminate between clients and therefore offers the same f to each client. Recall client 1 is the first to be aware of the disruption. Client 1 weighs off the disruption

---

[6]Patching has been studied in the context of network security and user patching behavior, where installing patches prevents disruption by reducing network vulnerabilities (August and Tunca 2006). In our context, we interpret patching through a restorative lens: once a problem has been resolved for one client, a provider may apply the same or a similar solution to another client, speeding up resolution of the disruption.

cost reduction against the expected disclosure reward conditional on her signal as well as its precision. The balance will determine whether the client makes the call for response in the first place and, subsequently, what to report. The decisions by client 1 affect the response capacity allocation by the provider and, in turn, impact the disruption costs faced by client 2.

Let D_1 and D_2 respectively be the total costs experienced by client 1 and client 2 as a result of the disruption. Similar to Kim and Tomlin (2013) (and references therein) we assume the disruption length is exponentially distributed with rate parameter μ_i such that the expected disruption length is given by $1/\mu_i$. The parameter μ_i is interpreted as the response capacity: the more capacity there is to respond to a disruption, the shorter the expected disruption length. Given the assumptions in Section 4.3.2 the expected cost to client 2 evaluated in phase (ii) is therefore:

$$\mathbb{E}[D_2|m] = \frac{d_2^{\Psi}}{\mu_2(m) + \alpha \mathbb{1}_{m=l}}, \tag{4.1}$$

where the exponent Ψ is the verified nature of the disruption, $\mathbb{1}_{m=l}$ is an indicator variable that equals 1 if $m = l$ and 0 otherwise and α is the patching benefit generated in case response capacity is allocated to the two clients in sequence. We require $\mu_0 \geq 1$ such that $\mu_i(m) \geq 1$, meaning expected disruption costs are non-increasing in response capacity.

In phase (ii), client 1 can only form an expectation of the cost of disruption, which depends on the signal and the signal's quality. In turn, this depends on the investigation effort exerted in the preceding phase. The response capacity client 1 receives upon calling for support depends on the message m she decides to convey to the provider. The expected cost to client 1 evaluated in phase (ii) after exerting investigation effort and receiving the signal is:

$$\mathbb{E}[D_1|e^*, s, f, r] = \min_m \quad \mathbb{E}_{\Psi}\left[\frac{d_1^{\Psi}}{\mu_1(m)} + T(f, r, m) + k\mathbb{1}_{e^*=h}\bigg|e^*, s\right], \tag{4.2}$$

where e^* is the optimal effort client 1 decided to exert. We define the expected value of T to be conditional on the investigation effort exerted by client 1 as well as the signal that she received, such that:

$$T(f, r, m) \equiv \begin{cases} f - r\mathbb{P}_{m=\Psi} & \text{if} \quad m \in \{l, g\} \\ 0 & \text{if} \quad m = \emptyset. \end{cases} \tag{4.3}$$

When deciding on the investigation effort level, client 1 must evaluate expected cost by also taking an expectation over the signal outcome. For simplicity we assume that, ex ante, the signal for a local or global disruption occurs with equal likelihood. Therefore, client 1 evaluates her expected cost in phase (ii) before exerting investigation effort as:

$$\mathbb{E}[D_1|f, r] = \min_e \quad \mathbb{E}_s\left[\mathbb{E}_{\Psi}\left[\frac{d_1^{\Psi}}{\mu_1(m(e|s))} + T(f, r, m(e|s)) + k\mathbb{1}_{e=h}\right]\right], \tag{4.4}$$

where we make explicit that the message to be sent is a function of the investigation effort and conditional on the signal received.

Initiating an allocation of response capacity costs the provider χ for each client the provider deploys resources to, e.g. the cost of sending engineers to the client's site. For tractability we

assume this cost does not depend on the size of allocation and that costs are incurred when response capacity is allocated in phase (ii). Let D_p be the provider's disruption cost net of payment and costs associated initiating response and recall that he is uninformed with regards to the type of disruption. The provider's cost function is:

$$
\mathbb{E}[D_p] = \min_{f,r} \quad \mathbb{E}_s \left[\mathbb{E}_\Psi \left[\frac{d_1^\Psi}{\mu_1(m(e|s))} - T(f, r, m(e|s)) + \chi \mathbb{1}_{m \in \{l,g\}} \right] \right]
$$
$$
+ \mathbb{E}_s \left[\mathbb{E}_\Psi \left[\frac{d_2^\Psi}{\mu_2(m(e|s)) + \alpha \mathbb{1}_{m=l}} + \chi (\mathbb{1}_{m=g} + \mathbb{1}_{m=l} \mathbb{1}_{\Psi=g}) \right] \right], \qquad (4.5)
$$

where the first term consists of the net costs incurred by the provider through client 1 and the second term consists of the net costs incurred through client 2. Note that, provided client 1 calls for a response, the provider incurs the cost of response allocation to client 1 regardless of the signal. The provider incurs allocations costs to client 2 if either client 1 reports a global disruption or if client 1 reports a local disruption and the disruption turns out to be a global disruption. The provider thus optimizes system costs including costs incurred from allocation response capacity. The notation for the model is summarized in order of introduction in Table 4.1.

The cost functions above capture a dual tension between the clients' objectives and the provider's objective. On the one hand, the clients each mininimize costs only accounting for their own disruption length, whereas the provider cares about the combined disruption length across both client. On the other hand, the clients base their decisions on an informative signal or fully verified nature of the disruption, whereas the provider decides on the contract parameters without direct information. We next analyze how these tensions play into inefficient allocation of response capacity.

Table 4.1 Summary of notation

i	Index of client
Ψ	True type of the disruption
s	Signal of the disruption type
q	Signal precision
e	Investigation effort
k	Cost of high investigation effort
m	Message sent by client 1
μ_0	Baseline response capacity of client i
μ_p	Provider's available response capacity
μ_i	Client's effective response capacity after allocation
$\bar{\mu}$	Maximum allocation of response capacity to one client
α	Patching benefit
d_i^Ψ	Disruption cost rate to client i in case of disruption type Ψ
f	Fixed fee
r	Disclosure reward
χ	Cost of response capacity allocation per client

4.4 Analysis

In this section we first solve the problem for the case in which the provider and clients act as one firm and have no misaligned incentives. The solution to this problem serves as a benchmark against which we can compare results in case the provider and clients act on their own and on the basis of private information.

4.4.1 Centralized Response

Let the provider take the role of a centralized decision maker ('the company') who directly allocates response capacity to two different company locations (rather than clients) after a disruption hits one (and possibly two) of the locations. We will use subscript C as a label for the company. Because the company allocates response capacity directly, the contract defined in the previous section does not play a role, nor does communication, although the company does incur the cost of initiating response. Each location has a local response capacity μ_0 like the clients in the decentralized case.

The company decides on the investigation effort and subsequently receives a single signal s_C with precision q through one of the two locations concerning the nature of the disruption. The company receives this signal directly; neither location plays a role in manipulating the signal through communication with the provider. Given the signal, the company has to commit the

response capacity across the two locations and decide on the allocation scheme in phase (ii) before verifying the true nature of the disruption in phase (iii). Allocating response capacity directly, the company faces a choice between two alternatives in phase (ii): (a) allocating all response capacity μ_p[7] and splitting evenly between locations early given the information from the signal ('broad' allocation of response capacity) and; (b) allocating maximum response capacity $\bar{\mu}$ to the location which issued the signal first and postponing allocation of the remaining response capacity to the next phase if necessary ('focused' allocation of response capacity). Rationalizing the expected cost associated with each option, the company can select the optimal decision on the basis of parameter settings. Solving backwards, the company can then decide on the optimal investigation effort strategy, again given parameter settings.

Given that the cost of initiating response does not depend on the size of the response capacity allocated, note that the company has no incentive to allocate less than μ_p in case of broad allocation or allocate less than $\bar{\mu}$ to the first location in case of focused allocation. Note that in case both locations are affected, an equal split of response capacity between the two locations is dominant to any other division of resources,[8] hence in case of broad allocation it is always best to allocate $\mu_p/2$ to each location. To evaluate the first-best response capacity allocation decision by the company in the centralized case, we first need to formalize the expected cost of each of the two options in phase (ii). Let $\mathbb{E}[D_C]$ be the company's expected disruption cost and let s ($-s$) be the type of disruption (not) indicated by the signal. With high investigation effort exerted in phase (ii), the company receives a signal with precision $1/2 < q < 1$. Conditional on high investigation effort, broad allocation has an expected cost of:

$$\mathbb{E}[D_C|e = h, s] = \left(\frac{2(q \sum_i d_i^s + (1-q) \sum_i d_i^{-s})}{2\mu_0 + \mu_p} \right) + k + 2\chi. \tag{4.6}$$

The expected cost of focused allocation depends on the signal, i.e.:

$$\mathbb{E}[D_C|e = h, s = g] = \left(\frac{q d_1^g + (1-q) d_1^l}{\mu_0 + \bar{\mu}} \right) + \left(\frac{q d_2^g}{\mu_0 + \mu_p - \bar{\mu} + \alpha} \right) + k + (1+q)\chi, \tag{4.7}$$

$$\mathbb{E}[D_C|e = h, s = l] = \left(\frac{q d_1^l + (1-q) d_1^g}{\mu_0 + \bar{\mu}} \right) + \left(\frac{(1-q) q d_2^g}{\mu_0 + \mu_p - \bar{\mu} + \alpha} \right) + k + (2-q)\chi. \tag{4.8}$$

The expected disruption costs following low investigation effort are found similarly and left to the Appendix. Lemma 2 characterizes the solution to the optimal response capacity allocation strategy for phase (ii) conditional on the signal received and investigation effort exerted.

---

[7]We retain the subscript p in this section for consistence in notation.

[8]Consider a simple example with $\mu_k = 10$ and disruption cost rate $d = 1$. Splitting the response capacity equal between two locations yields a total disruption cost of $\frac{1}{5} + \frac{1}{5} = \frac{2}{5}$. It is easy to show that any other division yields a higher total cost, e.g. $\frac{1}{6} + \frac{1}{4} = \frac{5}{12} > \frac{2}{5}$.

Lemma 2 (FIRST-BEST CONDITIONAL RESPONSE CAPACITY ALLOCATION). *Define:*

$$\zeta_1 \equiv \frac{1}{1-q} \left(\frac{qd^g + (1-q)d_1^l}{\mu_0 + \bar{\mu}} + \frac{qd^g}{\mu_0 + \mu_p - \bar{\mu} + \alpha} - \frac{2(2qd^g + (1-q)d_1^l)}{2\mu_0 + \mu_p} \right),$$

$$\zeta_2 \equiv \frac{1}{q} \left(\frac{(1-q)d^g + qd_1^l}{\mu_0 + \bar{\mu}} + \frac{(1-q)d^g}{\mu_0 + \mu_p - \bar{\mu} + \alpha} - \frac{2(2(1-q)d^g + qd_1^l)}{2\mu_0 + \mu_p} \right),$$

$$\zeta_3 \equiv \frac{d^g + d_1^l}{\mu_0 + \bar{\mu}} + \frac{d^g}{\mu_0 + \mu_p - \bar{\mu} + \alpha} - \frac{2(2d^g + d_1^l)}{2\mu_0 + \mu_p},$$

$$\hat{\alpha} \equiv \frac{(\mu_p - 2\bar{\mu})^2}{2(\mu_0 + 2\bar{\mu}) - \mu_p}.$$

The optimal response capacity allocation strategy conditional on the investigation effort and the signal received is characterized as follows:
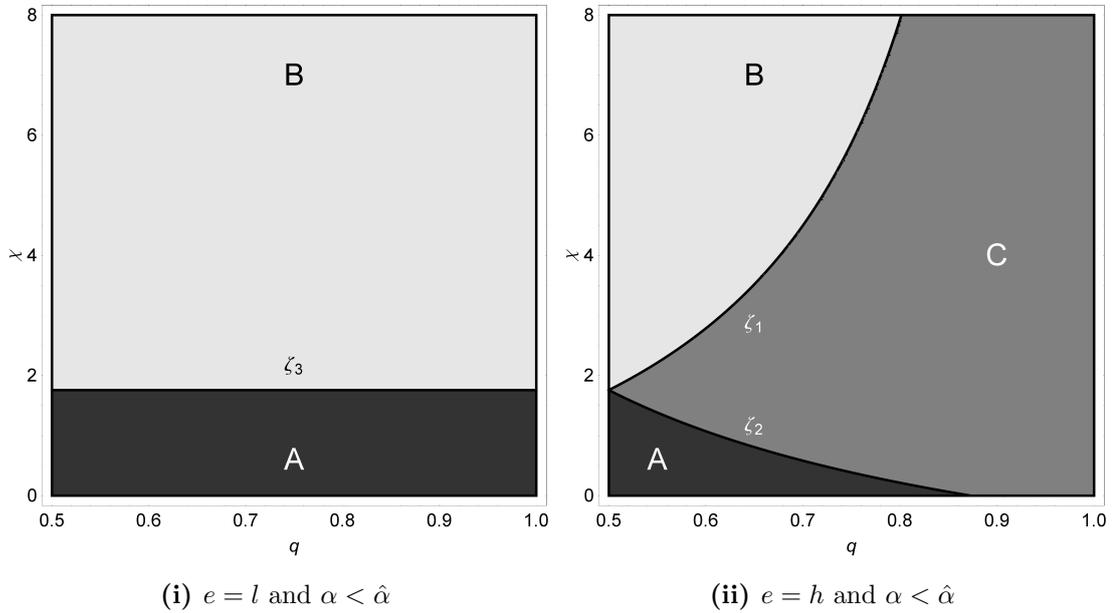
(i) when $e = h$ and $\alpha < \hat{\alpha}$ for $\chi \leq \zeta_2$ broad response capacity allocation is optimal and for $\chi > \zeta_1$ focused response capacity allocation is optimal, in both cases regardless of the signal. For $\zeta_2 < \chi \leq \zeta_1$ broad response capacity allocation is optimal if $s = g$ and focused allocation of response capacity is optimal if $s = l$;

(ii) when $e = h$ and $\alpha \geq \hat{\alpha}$ focused allocation is regardless of the signal;

(iii) when $e = l$ there exists an $\tilde{\alpha}$ such that when $\alpha < \tilde{\alpha}$ broad allocation is optimal if $\chi < \zeta_3$ and focused allocation is optimal otherwise, regardless of the signal. If $\alpha \geq \tilde{\alpha}$, focused allocation is optimal regardless of the signal and allocation cost.

Lemma 2 shows that if the company exerts a high investigation effort there are three different regions that characterize the optimal response capacity allocation. In region A, the cost of allocating response capacity is low enough for the company to choose broad allocation and split the response capacity across both locations regardless of the signal. In region B, the cost of allocating response capacity is too high to warrant deploying response capacity to both locations immediately in phase (ii). Therefore, in region B, it is optimal for the company to choose focused allocation regardless of the signal. Region C is different from the other two regions as for the intermediate costs of allocation within this region, the optimal allocation strategy depends on the signal received. If the signal is $s = g$, broad allocation is optimal; if the signal is $s = l$, it is optimal to focus allocation on the location where the signal originated and delay deploying further resources to the other location.

Figure 4.1 illustrates these results for the case when $d^g > d^l$ and $\alpha < \hat{\alpha}$, for the parameter settings indicated. Three observations can be made from the figure. A first observation is that signal precision has no role in case client 1 exerts low investigation effort, hence we find the simple threshold on χ between regions A and B in the left-hand panel. A second observation is that, after exerting high investigation effort (right-hand panel), using the signal to guide response capacity allocation (region C) is optimal for a wider range of allocation costs the higher the signal precision after high investigation effort. Intuitively, should investigation effort lead to a signal with perfect precision ($q = 1$), following a signal-dependent strategy is the only optimal strategy. A third observation is that while broad allocation regardless of the signal (region A)

is an optimal strategy under certain circumstances, this only holds for low cost of response capacity allocation and low signal precision.

Figure 4.1 First-best conditional response capacity allocation

(i) $e = l$ and $\alpha < \hat{\alpha}$	**(ii)** $e = h$ and $\alpha < \hat{\alpha}$

Notes. Regions denote optimal response capacity allocation strategies. Region A: unconditional broad allocation; region B: unconditional focused allocation; region C: broad allocation when $s = g$ and focused allocation when $s = l$. In this example: $d^l = 2$, $d^g = 10$, $\mu_k = 2$, $\bar{\mu} = 1.85$, $\mu_0 = 1$.

The next step in analyzing the case of centralized response capacity allocation requires examining how the optimal investigation effort depends on q, χ and α. The investigation effort decision is made prior to receiving the signal. Therefore, to compare the expected returns of high and low investigation effort and subsequently optimal response capacity allocation decisions, the company must take expectations over the ex-ante equally likely signals. Proposition 8 characterizes the solution.
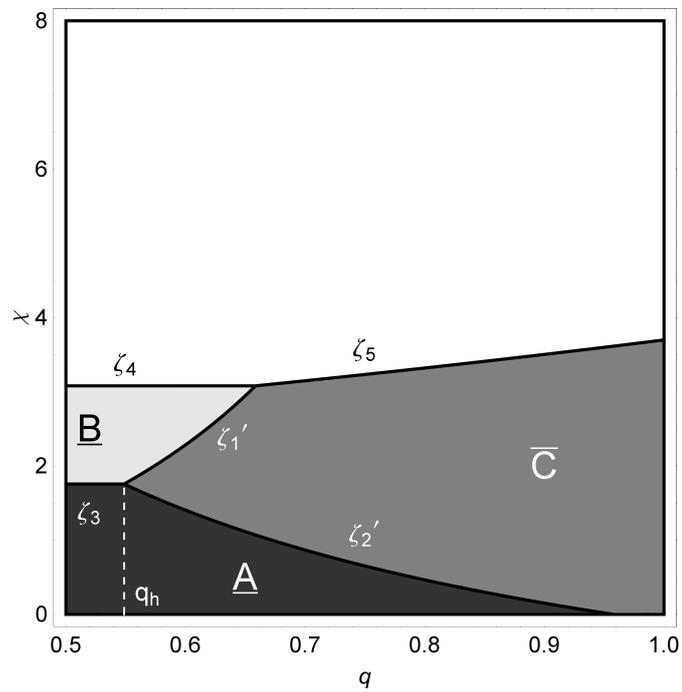
Proposition 8 (FIRST-BEST INVESTIGATION EFFORT AND RESPONSE CAPACITY ALLOCATION). *Let $M = \{A, B, C\}$ be the optimal conditional response capacity allocation regions as identified in Lemma 2 and let $\hat{\alpha}$ be as defined in the same Lemma. Also, let $\mathbb{E}[A|h]$, $\mathbb{E}[B|h]$ and $\mathbb{E}[C|h]$ respectively denote the expected disruption cost for each strategy given high investigation effort and let $\mathbb{E}[A|l]$, $\mathbb{E}[B|l]$ denote the expected disruption costs given low investigation effort. Finally, define:*

$$\zeta_{1\prime} \equiv \zeta_1 - \frac{2k}{1-q},$$

$$\zeta_{2\prime} \equiv \zeta_2 + \frac{2k}{q},$$

$$\zeta_4 \equiv \frac{1}{3}\left(\frac{2d^g + d_1^l}{\mu_0} - \frac{d^g}{\mu_0 + \mu_p - \bar{\mu} + \alpha} - \frac{d^g + d_1^l}{\mu_0 + \bar{\mu}}\right),$$

$$\zeta_5 \equiv \frac{1}{4-q}\left(\frac{2d^g + d_1^l}{\mu_0} - \frac{2(1-q)d_1^l + 4qd^g}{2\mu_0 + \mu_p} - \frac{(1-q)d^g}{\mu_0 + \mu_p - \bar{\mu} + \alpha} - \frac{(1-q)d^g + qd_1^l}{\mu_0 + \bar{\mu}} - 2k\right),$$

$$q_h \equiv \frac{1}{2} + k\left(\frac{(2\mu_0 + \mu_p)(\mu_0 + \bar{\mu})(\mu_0 + \mu_p - \bar{\mu} + \alpha)}{d^g((\mu_0 - 2\bar{\mu})^2 + \alpha(\mu_p - 2(\mu_0 + 2\bar{\mu})))}\right).$$

We find that: i) $\mathbb{E}[A|l] < \mathbb{E}[A|h]$ and $\mathbb{E}[B|l] < \mathbb{E}[B|h]$ for any $k > 0$, otherwise $\mathbb{E}[A|l] = \mathbb{E}[A|h]$ and $\mathbb{E}[B|l] = \mathbb{E}[B|h]$; ii) When $\alpha < \hat{\alpha}$, $q > q_h$, $\zeta_{2\prime} < \chi \leq \zeta_{1\prime}$ and $\chi \leq \zeta_5$ it is optimal to exert high investigation effort and follow a signal-dependent allocation strategy (i.e. broad allocation when $s = g$ and focused allocation when $s = l$). When $\alpha < \hat{\alpha}$, $q > q_h$, but $\chi \leq \zeta_{2\prime}$ ($\chi > \zeta_{1\prime}$) and $\chi \leq \zeta_4$, broad (focused) allocation regardless of the signal is optimal; iii) When $q \leq q_h$ only low investigation effort is optimal, such that subsequently broad allocation (focused allocation) is optimal for $\chi \leq \zeta_3$ ($\chi > \zeta_3$) by Lemma 2, provided $\chi \leq \zeta_4$; and iii) When $\chi \leq \zeta_4$, but $\alpha \geq \hat{\alpha}$ focused allocation is the only optimal response capacity allocation strategy, hence by (i), low investigation effort is optimal; Otherwise no allocation effort and allocation strategy is feasible.

Proposition 8 shows three results. Firstly, when investigation effort is costly, it can never be optimal to exert high investigation effort if it leads to a signal-independent response capacity allocation strategy. This gives a region $\underline{A}$, where low investigation effort is followed by broad allocation of response capacity regardless of the signal and a region $\underline{B}$ where low investigation effort is followed by focused response capacity allocation regardless of the signal. Secondly, when the patching benefit is small (i.e. $\alpha < \hat{\alpha}$), high investigation effort can be first-best provided the cost of investigation effort and response capacity allocation are sufficiently low, but only beyond a threshold signal precision q_h. This gives a region $\overline{C}$, where high investigation effort is followed by broad allocation if $s = g$ and focused allocation if $s = l$. Said differently, in $\overline{C}$ is optimal to invest in a better signal and decide on allocating response capacity after receiving it. Thirdly, when the patching benefit is sufficiently large, we show that high investigation effort cannot be first-best regardless of the marginal cost of each disruption type and focused allocation is the optimal response capacity allocation strategy. Figure 4.2 below illustrates the results of the proposition for $k = 0.1$ and otherwise the same parameter settings as in Figure 4.1.

Figure 4.2 First-best investigation effort and allocation

Notes. Regions denote optimal investigation effort and response capacity allocation strategies. Region $\underline{A}$: low investigation effort and unconditional broad allocation; region $\underline{B}$: low investigation effort and unconditional focused allocation; region $\overline{C}$: high investigation effort and broad allocation when $s = g$ and focused allocation when $s = l$. In this example: $d^l = 2$, $d^g = 10$, $\mu_k = 2$, $\bar{\mu} = 1.85$, $\mu_0 = 1$, $k = 0.1$, $\alpha = 0.02$.

From a managerial perspective, Lemma 2 and Proposition 8 together make it clear that deciding on whether or not to invest in an informative signal and subsequently deciding on the best way to allocate limited response capacity across the affected locations is not trivial, even in a centralized capacity without the possibility for misrepresentation of information. It turns out that the patching benefit plays an important role in driving which strategy is optimal. Provided patching benefit is hight ($\alpha \geq \hat{\alpha}$), it is best to delay deployment of resources to location 2 until after responding to issues raised at location 1.

Naturally, it may be difficult to determine whether the problems faced may be characterized by cross-applicability of solutions when they are eventually found. Alternatively, one can interpret α and q as representing centralized and decentralized learning respectively. Whereas q captures the potential for the nature of the disruption to be detected at each location decentrally; α captures how centrally undertaken response can facilitate understanding the nature of the disruption. Both forms of learning have to be considered in conjunction when considering whether investment in signal quality is necessary and pivotal to response capacity allocation.

4.4.2 Decentralized Response

In a decentralized setting, the provider does not have access to a direct signal to decide on the response capacity allocation. Instead, he has to design a contract with terms f (the pay-per-service fee) and r (the disclosure reward) to set incentives such that the clients exert high investigation effort and communicate the signal they receive truthfully. Without any such incentives, clearly neither client has any interest in reporting a global disruption as they will always receive more help when reporting a local disruption — without incurring any costs. As this strategy is optimal regardless of the signal, it follows that exerting investigation effort is suboptimal from the perspective of the client. Hence when $f, r = 0$ (the provider offered a 'null contract'), client 1 always exerts low investigation effort and communicates $m = l$ regardless of the signal received. However, as demonstrated in the previous section, under the right conditions a signal dependent strategy can be first-best and hence optimal from the perspective of the provider. The result is a difference in expected system disruption costs between a setting with centralized decision making and a setting with decentralized decision making without the right incentives for the clients. The intention of the contract terms offered by the provider to both clients is to control their incentives and minimize this cost difference.

We find the optimal contract terms using backward induction, starting from the perspective of client 1, who, given a set of contract terms, an exerted effort level and received signal, decides on what message $m \in \{l, q\}$ to send to the provider. Recall that the client also has the option of not sending a message, defined as $m = \emptyset$. Let $\mathbb{E}[D_1(m)]$ be client 1's expected disruption costs as a function of the message m she sends. Following high investigation effort, sending message

$m = g$ after receiving $s = g$ yields:

$$\mathbb{E}[D_1(g)|e = h, s = g] = \left(\frac{qd_1^g + (1-q)d_1^l}{\mu_0 + \mu_p/2}\right) + f - qr + k. \tag{4.9}$$

Sending message $m = l$ after receiving $s = g$ yields:

$$\mathbb{E}[D_1(l)|e = h, s = g] = \left(\frac{qd_1^g + (1-q)d_1^l}{\mu_0 + \bar{\mu}}\right) + f - (1-q)r + k. \tag{4.10}$$

Not sending a message (i.e. $m = \emptyset$) after receiving $s = g$ yields:

$$\mathbb{E}[D_1(\emptyset)|e = h, s = g] = \left(\frac{qd_1^g + (1-q)d_1^l}{\mu_0}\right) + k. \tag{4.11}$$

Using the same approach we can find the expected disruption costs associated with each message message m following high investigation effort and receiving a signal $s = l$ as well as following low investigation effort and receiving either possible signal. These functions are left to the Appendix. Comparing the expected cost functions gives the results in Lemma 3, s $(-s)$ is the type of disruption (not) indicated by the signal as before.

Lemma 3 (CONDITIONALLY OPTIMAL COMMUNICATION BY CLIENT 1). *i) Considering the case $e = l$, define:*

$$\vartheta_l \equiv \frac{\bar{\mu}(d_1^g + d_1^l)}{\mu_0(\mu_0 + \bar{\mu})}.$$

In this case the communication decision is independent from the signal received. Between sending $m = g$ or sending $m = l$, the latter is always optimal. Between sending $m = l$ or not sending a message, client 1 prefers not to send a message if $2f - r > \vartheta_l$.

ii) Considering the case $e = h$, define:

$$\vartheta_{h1}(s) \equiv \frac{(2\bar{\mu} - \mu_p)(qd_1^s + (1-q)d_1^{-s})}{(\mu_0 + \bar{\mu})(2\mu_0 + \mu_p)},$$

$$\vartheta_{h2}(s) \equiv \frac{\bar{\mu}(qd_1^s + (1-q)d_1^{-s})}{\mu_0(\mu_0 + \bar{\mu})},$$

$$\vartheta_{h3}(s) \equiv \frac{\mu_p(qd_1^s + (1-q)d_1^{-s})}{\mu_0(2\mu_0 + \mu_p)}.$$

Now in case the client receives a signal $s = g$, we find the following. If $(2q - 1)r > \vartheta(g)_{h1}$ and $f - qr < \vartheta_{h3}(g)$ is optimal for client 1 to communicate truthfully, i.e. $m = g$. If $(2q-1)r < \vartheta_{h1}(g)$ and $f - (1-q)r < \vartheta_{h2}(g)$ it is optimal for client 1 to miscommunicate, i.e. $m = l$. Otherwise it is optimal for the client not to send a message, i.e. $m = \emptyset$. In case the client receives a signal $s = l$ instead, there is no incentive to miscommunicate. Not sending a message can still be optimal, which is the case if $f - qr > \vartheta_{h3}(l)$.

Lemma 3 highlights two motives for miscommunication by client 1, by which we mean motives for the client to send a message that contradicts the signal she received. The first motive follows low investigation effort, after which it is always optimal to report a local disruption even if the signal indicates otherwise. The second motive arises when client has exerted high investigation effort and receives $s = g$, but the disclosure reward is not large enough to offset the increase in expected disruption cost by relinquishing response capacity through reporting a global disruption. Note, however, the incentive to miscommunicate is partially offset by the signal quality: having invested in a good signal adds to the expected value of the disclosure reward, making truthful communication more appealing. Regardless of the investigation effort exerted, client 1 has no incentive to communicate $m = g$ in case $s = l$, though she may prefer not to alert the provider at all. Not sending a message is preferable when the fee to call for response is sufficiently larger than the disclosure reward. If the expected disclosure reward is larger than the fee, even when deliberately miscommunicating, i.e. $(1 - q)r > f$, it is always preferable to send a message.

Let us characterize client 1's strategy as a 3-tuple: (m_g, m_l, e), where m_g is the message sent in case the signal is g, m_l is the message sent in case the signal is l, and e is the investigation effort. Let $\mathcal{S}$ be the client's strategy set under decentralized decision making. From Lemma 3 we know that $|S| = 8$, with:

$$\mathcal{S} = \{(\emptyset, \emptyset, l), (l, l, l), (\emptyset, \emptyset, h), (\emptyset, l, h), (l, \emptyset, h), (l, l, h), (g, \emptyset, h), (g, l, h)\}. \tag{4.12}$$

In $\mathcal{S}$ the strategies $(\emptyset, \emptyset, l)$ and $(\emptyset, \emptyset, h)$ constitute non-partipation, i.e. when client 1 rejects the contract offered (and by symmetry client 2 as well). Said differently, the constraints subject to which the other strategies are preferable to non-participation are client 1's participation constraints. The strategy (g, l, h) is the only strategy that equates to truth-telling by client 1. In the remaining five strategies client 1 either miscommunicates the signal received or does not communicate with the provider given a particular signal. Note that there is no complete miscommunication is the sense that the client communicates $m = l$ when $s = g$ and vice versa. Rather, two of the communication strategies only feature partial miscommunication where $m_g = l$. Next observe that two of the communication strategies only feature 'omission': for one of the two possible signals client 1 decides not to communicate. Such omission can only follow high investigation effort by client 1. The fifth remaining communication strategy $(l, \emptyset, h)$ is special in that it features both miscommunication, as $m_g = l$, and omission, as $m_l = \emptyset$.

In the following we first solve the provider's contracting problem in a simplified setting in which omission is not possible and let $\mathcal{S}' = \mathcal{S} \setminus \{(\emptyset, l, h), (l, \emptyset, h), (g, \emptyset, h)\}$. After that, we reintroduce the possibility of omission and find the necessary contract adjustments to ensure both high investigation effort and truth-telling by client 1.

4.4.2.1 Optimal contract design when omission is not possible

Without omission, the only relevant communication strategies to consider are (l, l, l), (l, l, h) and (l, g, h), subject to meeting client 1's participation constraint. Recall that the strategy (l, l, l) can be induced with a null contract. Also note that, given the results in Lemma 3, no contract terms result in the client's strategy (g, g, h) or (g, g, l) such that the provider will never deploy response capacity to both clients simultaneously regardless of the signal.

To incentivize client 1 to exert high investigation effort, a non-zero disclosure reward ($r > 0$) will be required, without affecting client's communication strategy. Without needing a formal proof, it follows that that $\mathbb{E}[D_p(f', r')|(l, l, l)] < \mathbb{E}[D_p(f'', r'')|(l, l, h)]$, where f' and r' are the optimal contract parameters that induce the (l, l, l) strategy and f'' and r'' are the optimal contract parameters that induce the (l, l, h) strategy. Hence, the provider only needs to evaluate the expected disruption costs Equation 4.5 in Section 4.3.3 between inducing communication strategies (l, l, l) and (g, l, h). Following a similar argument we can disregard contract parameters that induce $(\emptyset, \emptyset, h)$ as opposed to $(\emptyset, \emptyset, l)$. All cost functions are evaluated in the Appendix. It follows that the provider's contracting problem in a context without omission $(\mathcal{P}')$ reduces to:

$$(\mathcal{P}') \quad \min_{f, r \geq 0}, \quad \mathbb{E}[D_p]$$

$$\text{s.t.} \quad \mathbb{E}[D_1(g, l, h)|f, r] \leq \mathbb{E}[D_1(l, l, l)|f, r], \qquad (IC_{E+T})$$

$$\mathbb{E}[D_1(g, l, h)|f, r] \leq \mathbb{E}[D_1(\emptyset, \emptyset, l)|f, r], \qquad (IR_1)$$

$$\mathbb{E}[D_p|(g, l, h)] \leq \mathbb{E}[D_p|(\emptyset, \emptyset, l)]. \qquad (IR_p)$$

In this reduced problem statement, the condition (IC_{E+T}) simultaneously solves the moral hazard problem to do with client 1's investigation effort as well as the adverse selection problem to do with client 1's communication. The conditions (IR_1) and (IR_p) are the individual rationality constraints for client 1 and the provider respectively. Proposition 9 characterizes the solution to $(\mathcal{P}')$.

Proposition 9 (OPTIMAL CONTRACT DESIGN WITHOUT OMISSION). *Define:*

$$\overline{f} \equiv \frac{1}{2}\left(\frac{d_1^l + d^g}{\mu_0} - \frac{2(qd^g + (1-q)d_1^l)}{2\mu_0 + \mu_p} + \frac{qd_1^l + (1-q)d^g}{\mu_0 + \bar{\mu}}\right) + q\underline{r} - k,$$

$$\underline{r} \equiv \frac{(2\bar{\mu} - \mu_p)(qd^g + (1-q)d_1^l)}{(2q-1)(2\mu_0 + \mu_p)(\mu_0 + \bar{\mu})} + \frac{2k}{2q-1},$$

$$\zeta_1'' \equiv \frac{2}{1-q}\left(\frac{qd^g + (1-q)d_1^l}{\mu_0 + \bar{\mu}} + \frac{qd^g}{2(\mu_0 + \mu_p - \bar{\mu} + \alpha)} - \frac{3qd^g + 2(1-q)d_1^l}{2\mu_0 + \mu_p} - k\right),$$

$$\zeta_4' \equiv \frac{1}{3}\left(\frac{d^g}{\mu_0} + \frac{d^g}{\mu_0 + \mu_p - \bar{\mu} + \alpha} - \frac{2(d^g + d_1^l)\bar{\mu}}{\mu_0(\mu_0 + \bar{\mu})}\right),$$

$$\zeta_5' \equiv \frac{2}{4-q}\left(\frac{3d^g + 2d_1^l}{\mu_0} - \frac{4(1-q)d_1^l + 6qd^g}{2\mu_0 + \mu_p} - \frac{(1-q)d^g}{\mu_0 + \mu_p - \bar{\mu} + \alpha} - \frac{2(1-q)d^g + 2qd_1^l}{\mu_0 + \bar{\mu}} - k\right).$$

We find that: i) It is optimal for the provider to set $T(\overline{f}, \underline{r})$ and induce both high investigation effort and thruth-telling by client 1 so long as $\chi \leq \zeta_1''$ and $\chi \leq \zeta_5$, ii) It is optimal for the provider to set $T(\overline{f}', 0)$ and induce client 1 to communicate while allowing for miscommunication so long as $\chi > \zeta_1''$ and $\chi \leq \zeta_4$. Otherwise contracting is not feasible. Finally, there exists an $\alpha^\dagger$ such that for $\alpha \geq \alpha^\dagger$ and $\chi \leq \zeta_4$ only offering $T(\overline{f}', 0)$ can be optimal.

Proposition 9 shows that in a context where response capacity allocation is costly, yet high investigation effort leads to a high precision signal, the provider issues a contract to ensure both high investigation and truthful communication from the client. In case high investigation effort does not lead to a sufficiently precise signal, it is not optimal for the provider to provide the client with incentives to put in this effort. However, provided the cost of allocation is not too high, it is still optimal to induce the client to communicate in case of disruption, albeit with the risk of the client reporting a local disruption if she receives a signal for a global disruption. Although client 1 may thus cause misallocation of response capacity, without communication from the client the provider would not be alerted to the disruption in time to allocate any response capacity. Moreover, ineffective allocation is at least partially offset by the potential for patching. This results in conditions under which offering $T(\overline{f}', 0)$ is optimal from the provider's perspective.

Figure 4.3 illustrates these results. In region $\overline{C}$ (dark gray), the provider offers $T(\overline{f}, \underline{r})$ and ensures high investigation effort and truth-telling by client 1. In region $\underline{B}$ (light gray), the provider offers $T(\overline{f}', 0)$ and induces client 1 to report a local disruption regardless of the signal. For the white region it is infeasible for the provider to offer either contract. In addition, the plot is overlaid with the threshold curves from Figure 4.2. Given the same parameter conditions, we observe three interesting regions where decentralized decision making diverges from centralized,

first-best decision making. First, under decentralized decision making, a much higher signal precision is required in order for high investigation effort to be worthwhile than under centralized decision making. This means in the where the region of $\overline{C}$ from Figure 4.2 overlaps with region $\underline{B}$ in Figure 4.3, the provider underreacts to signals potentially indicating a global disruption. Second, the incentive for client 1 to report a local disruption prevents the provider from inducing client 1 to report such that broad allocation despite the signal is possible. This means where region $\underline{A}$ from Figure 4.2 overlaps with regions $\underline{B}$ and $\overline{C}$ in Figure 4.3 we again find the provider in the decentralized setting is likely not to allocate response capacity globally when it is efficient to do so. Both observations speak to the observed delay in deploying response capacity globally (i.e. executing a worldwide recall) after initial firm awareness of a disruption (FDA 2013). Third, we find disruption response to be feasible for higher χ under contracting than under centralized decision making. The optimal fixed fee that the clients are willing to pay in return for response capacity under either contract is large enough to offset a higher cost of allocation.

Figure 4.3 Investigation effort and allocation under contracting (no omission)
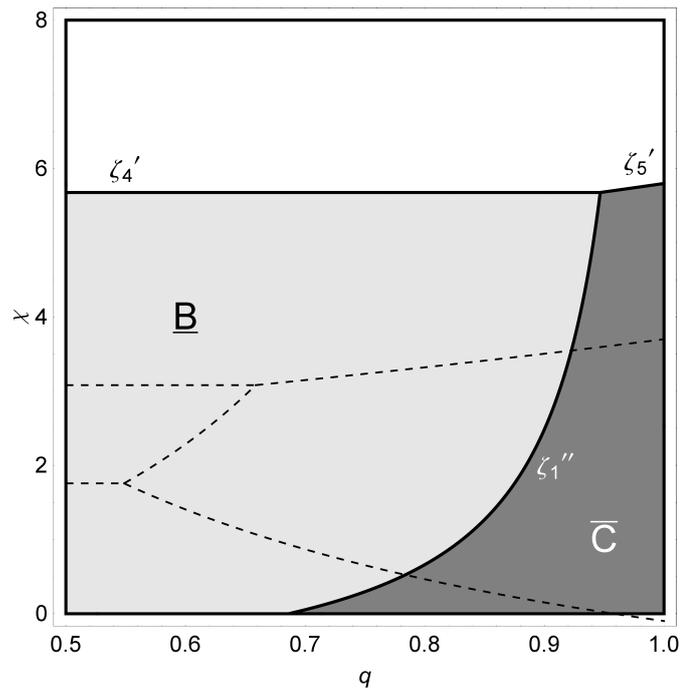
Notes. Regions denote investigation effort and response capacity allocation strategies induced by contracts. Region $\underline{B}$: low investigation effort and unconditional focused allocation, induced by contract $T(\overline{f}', 0)$; region $\overline{C}$: high investigation effort and broad allocation when $s = g$ and focused allocation when $s = l$, induced by $T(\overline{f}, \underline{r})$. In the white region contracting is infeasible. In this example: $d^l = 2$, $d^g = 10$, $\mu_k = 2$, $\bar{\mu} = 1.85$, $\mu_0 = 1$, $k = 0.1$, $\alpha = 0.02$.

4.4.2.2   Optimal contract design when omission is possible

Now we turn to solving for the optimal contract terms in case omission is possible, meaning all eight of client 1's communication strategies in $\mathcal{S}$ are admissible. We need to determine whether the option for client 1 to omit has a bearing on the optimal contract terms to induce both high investigation effort and truthful communication. From the perspective of the provider, the contracting problem is equivalent to $(\mathcal{P})$:

$$(\mathcal{P}) \quad \min_{f,r \geq 0}, \quad \mathbb{E}[D_p]$$

$$\text{s.t.} \quad \mathbb{E}[D_1(g,l,h)|f,r] \leq \mathbb{E}[D_1(l,l,l)|f,r], \qquad (IC_E)$$

$$\mathbb{E}[D_1(g,l,h)|f,r] \leq \mathbb{E}[D_1(\emptyset,l,h)|f,r], \qquad (IC_{T1})$$

$$\mathbb{E}[D_1(g,l,h)|f,r] \leq \mathbb{E}[D_1(l,\emptyset,h)|f,r], \qquad (IC_{T2})$$

$$\mathbb{E}[D_1(g,l,h)|f,r] \leq \mathbb{E}[D_1(g,\emptyset,h)|f,r], \qquad (IC_{T3})$$

$$\mathbb{E}[D_1(g,l,h)|f,r] \leq \mathbb{E}[D_1(\emptyset,\emptyset,l)|f,r], \qquad (IR_1)$$

$$\mathbb{E}[D_p|(g,l,h)] \leq \mathbb{E}[D_p|(\emptyset,\emptyset,l)], \qquad (IR_p)$$

where the first incentive compatibility constraint (IC_E) ensure high investigation effort, the next three incentive compatibility constraints $(IC_{T1:3})$ ensure truth-telling and the individual rationality constraint ensures participation by client 1. Like before the conditions (IR_1) and (IR_p) are the individual rationality constraints for client 1 and the provider respectively.

Rather than solving $(\mathcal{P})$ we can set the contract to $T(\overline{f}, \underline{r})$ as found in Proposition 9 and determine whether under this contract there are conditions under which the constraints $(IC_{T1:3})$ in $(\mathcal{P})$ are violated. Pursuing this approach, we can then determine how the original contract terms should be adjusted to prevent client 1 from not communicating in case she receives a particular signal. Proposition 10 characterizes the necessary adjustment to the contract terms when omission by client 1 is possible.

Proposition 10 (OPTIMAL CONTRACT DESIGN WITH OMISSION). *Consider the contracts $T(\overline{f}, \underline{r})$, $T(\overline{f}', 0)$ and ζ_1'', ζ_4' and ζ_5' as described in Proposition 9 and let:*

$$\delta \equiv \frac{(2q-1)\mu_p\bar{\mu} + q\mu_0\mu_p - (1-q)\mu_0\bar{\mu}}{(2q-1)\mu_p\bar{\mu} + 2q\mu_0\bar{\mu} - (1-q)\mu_0\mu_p},$$

$$x \equiv \frac{2k\mu_0(2\mu_0 + \mu_p)(\mu_0 + \bar{\mu})}{(2q-1)\mu_p\bar{\mu} + 2q\mu_0\bar{\mu} - (1-q)\mu_0\mu_p}.$$

Under both contracts, client 1 has an incentive not to communicate when $s = g$ and $d^l > (\delta + x)d^g$ and when $s = l$ and $d^l < (\delta - x)d^g$. To prevent omission, the provider needs to adjust the fixed

fee downward in either contract by:

$$
\Delta f = \begin{cases} \dfrac{d^l - (\delta + x)d^g}{4}\left(\dfrac{2q(\mu_0 + \mu_p) - \mu_p}{\mu_0(2\mu_0 + \mu_p)} - \dfrac{q}{\mu_0 + \bar{\mu}}\right) & \text{when } d^l > (\delta + x)d^g \\[3ex] \dfrac{(\delta - x)d^g - d^l}{4}\left(\dfrac{2q(\mu_0 + \mu_p) - \mu_p}{\mu_0(2\mu_0 + \mu_p)} - \dfrac{q}{\mu_0 + \bar{\mu}}\right) & \text{when } d^l < (\delta - x)d^g \\[3ex] 0 & \text{otherwise.} \end{cases}
$$

Following a downward adjustment of the fixed fee, the thresholds $\zeta_4\prime$ and $\zeta_5\prime$ shift down respectively by $\Delta f/3$ and $2\Delta f/(4-q)$.

Proposition 10 shows that the possibility of omission by client 1 necessitates an adjustment to the expected contract payment between the provider and the client, conditional on the ratio of the cost of local disruption to the cost of global disruption. Interestingly, this adjustment is non-monotone in this ratio. From Δf, we see that when $d^l < (\delta - x)d^g$ or $d^l > (\delta + x)d^g$, the provider should decrease the fixed fee, the larger the difference between d^l and δd^g. In between the two thresholds, it is still optimal for the client to adhere to truthful communication under the original contract and no adjustment is needed. Whereas under $T(\overline{f}, \underline{r})$, the fixed fee $\overline{f}$ is set to offset the disclosure reward in expectation, the adjusted fixed fee when omission by client 1 is possible effectively reinstates the expected disclosure reward to the client. This means when omission is possible, the client can extract information rent when d^l is small or large enough with respect to d^g. Also note that $\delta < 1$ as $\mu_p > \bar{\mu}$ and $2\bar{\mu} > \mu_p$ by definition. This means there exist instances in which it is worthwhile for the client to not communicate if $s = g$, even when $d^l < d^g$, unless the contract is adjusted. Figure 4.4 illustrates these results. In shaded regions, the provider should lower the fixed fee; between the two solid vertical lines no adjustment is necessary.

4.5 Conclusions

Increasingly large client pools have come to depend on continuous availability and security of equipment provided and maintained by a single specialist provider. Together these firms form connected technology networks at risk of disruption affecting multiple parties simultaneously while the provider's ability to provide response measure may be limited. The scale of the disruption stemming from technology failure is typically unknown at the onset of disruption. Moreover, faulty equipment can in some cases cause life threatening circumstances to consumers or patients downstream. Although a growing stream of work in operations management literature addresses problems surrounding investment in preventative and restorative measures, few works speak to the problem of effective allocation of response capacity once a disruption has hit, but its nature not yet properly understood.

Figure 4.4 Conditions for contract adjustment to counteract omission by client 1
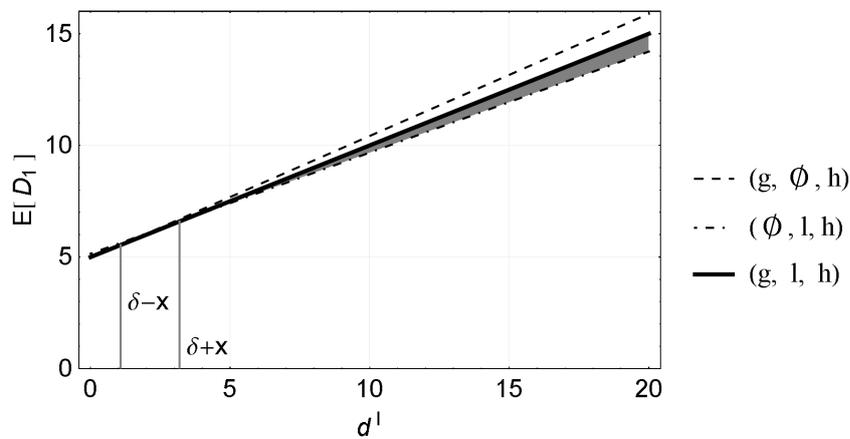
Notes. Contract adjustment is necessary in the shaded region.

In this example: $d^g = 10$, $\mu_k = 2$, $\bar{\mu} = 1.85$, $\mu_0 = 1$, $k = 0.1$,

$\chi = 0.5$, $q = 0.85$, $\alpha = 0.02$.

Motivated by a widespread problem with Baxter International's dialysis equipment affecting hospitals across Europe and the US in 2001, this study examines how detection of latent disruption risks in a network of clients like hospitals using and relying on the same technology for service performance can be improved through contract design. Medical equipment used in hospitals is generally serviced by the manufacturer acting as a service provider through either a full protection plan at a fixed price or through a pay-per-service plan. In the context of bilateral relations particularly the latter is often the most suitable. However, when multiple clients may be at risk of competing over the same resources in case of disruption resulting from a common technology failure, standard contracting practices may no longer suffice. For one, clients may not have the incentive to investigate the nature of a disruption after problems are first recognized. Compounding this problem, adverse incentives can exist for clients to underreport a potentially widespread problem to claim the necessary response capacity. From the perspective of the equipment provider it is crucial to receive and follow the right warning signs in deploying costly response capacity. This is achievable by using contract design to control the clients' incentives to disclose disruptions voluntarily, in line with the findings in Plambeck and Taylor (2015) and Wang et al. (2016).

By means of a contracting game between a single provider and two clients we examined the effect of a disclosure reward to ensure both high investigation effort and truthful communication by the client who first raises alarm in case of a disruption. We modelled how the cost of allocating response capacity as well as the signal precision obtained after investigation factor into optimal decisions, as well as how these decisions are moderated by possible patching benefits. Through this patching benefit we model the potential for focused allocation of response capacity to lead to more expedited resolutions of related disruptions at other locations or clients. The patching

benefit and signal quality can be interpreted as respectively reflecting intelligence gathering by the provider and the clients. We studied the problem in two parts. First we solved the investigation effort and response capacity allocation problem from the perspective of a centralized decision maker. Next we examined decentralized decision making where the provider offers a contract to influence the clients' decisions on investigation effort and communication, after which the provider allocates response capacity to the two clients.

Under centralized decision making we show that for a sufficiently high patching benefit, exerting costly investigation effort is never efficient. In turn, without a precise signal on the nature of the disruption this means adhering to a signal-dependent response capacity allocation strategy is never efficient. For lower patching benefits we show that there are also conditions under which the centralized decision make optimally chooses to ignore signals on the disruption type and adheres to either a focused or broad response capacity allocation strategy regardless of the signal. This holds particularly when the potential signal precision after high investigation effort is low and the cost of allocating response capacity is either sufficiently low (in which case broad allocation regardless of the signal is optimal) or sufficiently high (in which case focused allocation regardless of the signal is optimal). These results imply that even without adverse selection and moral hazard, seemingly naive response capacity allocation that does not take account of disruptions signal can be viable.

Our results for the decentralized setting highlight the benefit of including a disclosure reward next to the service fee in the contract. We show that while clients may be incentivized to identify and report network disruptions, competition for scarce emergency resources and the required investment in understanding their own exposure may still lead clients to deliberately miscommunicate with the provider. A negative byproduct of creating the right incentives for clients to investigate issues seriously is that it opens up avenues for omission given a particular signal, similar to the partial disclosure strategies observed in Kim (2015). Having information gives agents power to be selective in communication, which may hurt system allocation of response capacity. This implies that the provider should take caution in allocating response capacity following alarms raised by clients. We solve for the optimal adjustment in contract design, so that the provider can ensure truth-telling by the client and rely on accurate signals to deploy response measures effectively when a disruption has happened.

Combined, our findings give a game theoretical justification for observed miscommunication of adverse events by hospitals as well delayed response to widespread problems by medical equipment manufacturers. More generally our findings highlight the problems associated with achieving the effective response to disruption in settings where information on the nature of disruptions is both fragmented and affected by interference from misaligned objectives. Although incentives can be adjusted by means of the right contract design, the findings of this study underline that even under theoretically optimal contracts, systematic underreaction to warning signs is difficult to avoid.

Appendix

4.A Proofs

Proof of Lemma 2. We complete the proof for this lemma in three steps. In step 1 we specify the expected disruption costs for the centralized firm for each combination of investigation effort $e \in \{h, l\}$, signal $s \in \{l, g\}$ and response capacity allocation strategy (option (a): broad allocation or option (b): focused allocation). In step 2, we find the indifference functions between strategies and solve for χ. In step 3, we use these indifference functions to find constraints that define which allocation strategy is preferred under which parameter conditions.

Step 1: Conditional Expected Disruption Costs. There are eight different expected cost functions to compare. Let G1 and G2 be the expected cost functions for option (a) and (b) respectively, given $e = h$ and $s = g$. Let G3 and G4 be the expected costs functions for (a) and (b), given $e = h$ and $s = l$; then G5 and G6 given $e = l$ and $s = g$; and G7 and G8 given $e = l$ and $s = l$. We find:

$$\mathbb{E}[G1] = q(2d^g/(\mu_0 + \mu_p/2)) + (1-q)(d_1^l/(\mu_0 + \mu_p/2)) + k + 2\chi, \tag{4.13}$$

$$\mathbb{E}[G2] = q(d^g/(\mu_0 + \bar{\mu})) + (1-q)(d_1^l/(\mu_0 + \bar{\mu})) + q(d^g/(\mu_0 + \mu_p - \bar{\mu} + \alpha)) + k + (1+q)\chi, \tag{4.14}$$

$$\mathbb{E}[G3] = q(d_1^l/(\mu_0 + \mu_p/2)) + (1-q)(2d^g/(\mu_0 + \mu_p/2)) + k + 2\chi, \tag{4.15}$$

$$\mathbb{E}[G4] = q(d_1^l/(\mu_0 + \bar{\mu})) + (1-q)(2d^g/(\mu_0 + \bar{\mu})) + (1-q)(d^g/(\mu_0 + \mu_p - \bar{\mu} + \alpha)) + k + (2-q)\chi, \tag{4.16}$$

$$\mathbb{E}[G5] = 1/2(2d^g/(\mu_0 + \mu_p/2)) + 1/2(d_1^l/(\mu_0 + \mu_p/2)) + 2\chi, \tag{4.17}$$

$$\mathbb{E}[G6] = 1/2(d^g/(\mu_0 + \bar{\mu})) + 1/2(d_1^l/(\mu_0 + \bar{\mu})) + 1/2(d^g/(\mu_0 + \mu_p - \bar{\mu} + \alpha)) + 3\chi/2, \tag{4.18}$$

$$\mathbb{E}[G7] = 1/2(d_1^l/(\mu_0 + \mu_p/2)) + 1/2(2d^g/(\mu_0 + \mu_p/2)) + 2\chi, \tag{4.19}$$

$$\mathbb{E}[G8] = 1/2(d_1^l/(\mu_0 + \bar{\mu})) + 1/2(d^g/(\mu_0 + \bar{\mu})) + 1/2(d^g/(\mu_0 + \mu_p - \bar{\mu} + \alpha)) + 3\chi/2. \tag{4.20}$$

Step 2: Indifference Functions. From step 1 it is easy to observe that $\mathbb{E}[G5] = \mathbb{E}[G7]$ and $\mathbb{E}[G6] = \mathbb{E}[G8]$ as in both cases low investigation effort yields an uninformative signal such that the expected cost conditional on $s = g$ and $s = l$ are exactly the same. Hence we can proceed to find the indifference functions between $\mathbb{E}[G1]$ and $\mathbb{E}[G2]$, $\mathbb{E}[G3]$ and $\mathbb{E}[G4]$ and $\mathbb{E}[G5]$ and $\mathbb{E}[G6]$, but not have to consider $\mathbb{E}[G7]$ and $\mathbb{E}[G8]$ separately. Setting $\mathbb{E}[G1] = \mathbb{E}[G2]$, $\mathbb{E}[G3] = \mathbb{E}[G4]$

and $\mathbb{E}[G5] = \mathbb{E}[G6]$ and solving for χ in each case respectively gives:

$$\zeta_1 \equiv \frac{1}{1-q}\left(\frac{qd^g + (1-q)d_a^l}{\mu_0 + \bar{\mu}} + \frac{qd^g}{\mu_0 + \mu_p - \bar{\mu} + \alpha} - \frac{2(2qd^g + (1-q)d_a^l)}{2\mu_0 + \mu_p}\right), \tag{4.21}$$

$$\zeta_2 \equiv \frac{1}{q}\left(\frac{(1-q)d^g + qd_a^l}{\mu_0 + \bar{\mu}} + \frac{(1-q)d^g}{\mu_0 + \mu_p - \bar{\mu} + \alpha} - \frac{2(2(1-q)d^g + qd_a^l)}{2\mu_0 + \mu_p}\right), \tag{4.22}$$

$$\zeta_3 \equiv \frac{d^g + d_a^l}{\mu_0 + \bar{\mu}} + \frac{d^g}{\mu_0 + \mu_p - \bar{\mu} + \alpha} - \frac{2(2d^g + d_a^l)}{2\mu_0 + \mu_p}. \tag{4.23}$$

Recall the tie-breaking assumption that in case the company is indifferent between options, it is preferable to allocate response capacity as much and as early as possible. From the indifference functions we therefore find that after exerting high investigation effort, if $\chi \leq \zeta_1$ ($\chi > \zeta_1$), broad (focused) allocation is optimal if $s = g$. If $\chi \leq \zeta_2$ ($\chi > \zeta_2$) broad (focused) allocation is optimal if $s = l$. If $\chi \leq \zeta_1$ and $\chi > \zeta_2$ ($\chi > \zeta_1$ and $\chi > \zeta_2$), then broad (focused) allocation is optimal regardless of the signal. After exerting low investigation effort, we find broad (focused) allocation is optimal if $\chi \leq \zeta_3$ ($\chi > \zeta_3$).

Step 3: Optimal Conditional Response Capacity Allocation. From step 2 we can see ζ_3 does not depend on q. To determine the signs of ζ_1 and ζ_2 in q, we find derivatives $\partial\zeta_1/\partial q$ and $\partial\zeta_2/\partial q$. Equating both derivatives to zero and solving for α gives a single solution:

$$\hat{\alpha} \equiv \frac{(\mu_p - 2\bar{\mu})^2}{2(\mu_0 + 2\bar{\mu}) - \mu_p}. \tag{4.24}$$

Evaluating ζ_1 and ζ_2 at $\hat{\alpha}$ gives $\zeta_1 = \zeta_2 = d^l(1/(\mu_0 + \bar{\mu}) - 2/(2\mu_0 + \mu_p)) < 0$ as $2\bar{\mu} > \mu_p$ and by extension, for $\alpha \geq \hat{\alpha}$, $\zeta_1, \zeta_2 < 0$. Next we derive:

$$\frac{\partial\zeta_1}{\partial q \partial\alpha} = -\frac{d^g}{(q-1)^2(\mu_0 + \mu_p - \bar{\mu} + \alpha)^2} < 0, \tag{4.25}$$

$$\frac{\partial\zeta_2}{\partial q \partial\alpha} = \frac{d^g}{q^2(\mu_0 + \mu_p - \bar{\mu} + \alpha)^2} > 0. \tag{4.26}$$

Noting that $\hat{\alpha} > 0$, this means that for $0 \leq \alpha < \hat{\alpha}$, $\partial\zeta_1/\partial q > 0$ and $\partial\zeta_2/\partial q < 0$. Also note that, for $q = 1/2$ we have $\zeta_1 = \zeta_2 = \zeta_3$. With these results, we find that for $q \leq 1/2$ and if $\alpha < \hat{\alpha}$, $\zeta_1 > \zeta_2$. By contrast, if $\alpha \geq \hat{\alpha}$, $\zeta_1 < \zeta_2 < 0$.

If the company has exerted high investigation effort (e=h), it follows that when $0 \leq \alpha < \hat{\alpha}$ for $\chi > \zeta_1$ focused allocation is optimal, regardless of the signal, and for $\chi \leq \zeta_2$ broad allocation is optimal, regardless of the signal. Denote the region in which broad allocation is always optimal as region A and denote the region in which focused allocation is always optimal as region B. For $0 \leq \alpha < \hat{\alpha}$ and $\zeta_2 < \chi < \zeta_1$, it is broad allocation if the signal is $s = g$ and focused allocation is optimal if the signal is $s = l$. Denote the region in which a signal-dependent strategy is optimal region C. When $\alpha \geq \hat{\alpha}$, it holds that $\chi \geq 0 > \zeta_2 > \zeta_1$, hence focused allocation is the only optimal strategy after exerting high investigation effort. If instead the company has exerted low investigation effort (e=l), it follows that when $\chi > \zeta_3$, focused allocation of response capacity is optimal. Solving $\zeta_3 = 0$ for α gives:

$$\tilde{\alpha} = \frac{(2\bar{\mu} - \mu_p)(d^g(\mu_p - 2\bar{\mu}) + d^l(\mu_0 + \mu_p - \bar{\mu}))}{d^l(\mu_p - 2\bar{\mu}) + d^g(\mu_p - 2(\mu_0 + 2\bar{\mu}))}, \tag{4.27}$$

which means that for $\alpha \geq \tilde{\alpha}$ focused response capacity allocation is the only optimal strategy after exerting low investigation effort. $\square$

Proof of Proposition 8. We complete this proof in four steps. In step 1 we characterize the ex-ante expected cost functions given the optimal allocation strategies, which are necessary to determine optimal investigation effort exertion. In step 2 we show that a signal-independent strategy is never preceded by high investigation effort, i.e. when low investigation effort is dominant. In step 3, parts a and b, we characterize the conditions under which high investigation is dominant. In step 4 we evaluate the conditions that ensure feasibility of the investigation effort and response capacity allocation strategies. In all parts of the proof, $\hat{\alpha}$ is the result from Lemma 2 step 3.

Step 1: Ex-ante Expected Disruption Costs Given Optimal Allocation Strategies. Through Lemma 2 we identified three regions $M = \{A, B, C\}$ characterized by distinct response capacity allocation strategies given high investigation effort. Recall that region A is characterized by unconditional broad allocation (a); region B is characterized by unconditional focused allocation (b); region C is characterized by allocation conditional on the signal. Let $\Gamma(M|e, s)$ be the optimal resource allocation strategy for a particular region, given the investigation effort and signal, such that we have $\Gamma(A|h, g) = \Gamma(A|h, l) = a$, $\Gamma(B|h, g) = \Gamma(B|h, l) = b$, but in case of region C:

$$\Gamma(C|h, s) = \Gamma(C|h, g) \cup \Gamma(C|h, l) = \{a, b\}.$$

Also let $\mathbb{E}[A|h]$, $\mathbb{E}[B|h]$ and $\mathbb{E}[C|h]$ respectively denote the expected disruption costs for each strategy across the two company locations given high investigation effort. Recall that by assumption the two signal types are equally likely, i.e. $\mathbb{P}(s = g) = \mathbb{P}(s = l) = 1/2$. We can then compute:

$$\mathbb{E}[A|h] = \left(\mathbb{E}_\Psi[D_C|\Gamma(A|h, g)] + \mathbb{E}_\Psi[D_C|\Gamma(A|h, l)] \right)/2, \tag{4.28}$$

$$\mathbb{E}[B|h] = \left(\mathbb{E}_\Psi[D_C|\Gamma(B|h, g)] + \mathbb{E}_\Psi[D_C|\Gamma(B|h, l)] \right)/2, \tag{4.29}$$

$$\mathbb{E}[C|h] = \left(\mathbb{E}_\Psi[D_C|\Gamma(C|h, g)] + \mathbb{E}_\Psi[D_C|\Gamma(C|h, l)] \right)/2. \tag{4.30}$$

Substituting the expected cost functions from the proof of Lemma 2 and simplifying yields:

$$\mathbb{E}[A|h] = \frac{2d^g + d^l}{2\mu_0 + \mu_p} + 2\chi + k, \tag{4.31}$$

$$\mathbb{E}[B|h] = \frac{1}{2} \left(\frac{d^g + d^l}{\mu_0 + \bar{\mu}} + \frac{d^g}{\mu_0 + \mu_p - \bar{\mu} + \alpha} + 3\chi + 2k \right), \tag{4.32}$$

$$\mathbb{E}[C|h] = \frac{1}{2} \left(\frac{(1-q)d^g + qd^l}{\mu_0 + \bar{\mu}} + \frac{(1-q)d^g}{\mu_0 + \mu_p - \bar{\mu} + \alpha} + \frac{2(2qd^g + (1-q)d^l)}{2\mu_0 + \mu_p} + (4-q)\chi + 2k \right). \tag{4.33}$$

Turning to the case of low investigation effort, we need only compute the expected disruption cost for two regions. Again region A is characterized by unconditional broad allocation and region B is characterized by unconditional focused allocation, i.e. $\Gamma(A|l, g) = \Gamma(A|l, l) = a$,

$\Gamma(B|l, g) = \Gamma(B|l, l) = b$. Let $\mathbb{E}[A|l]$ and $\mathbb{E}[B|l]$ respectively denote the expected disruption cost for each region given low investigation effort. We find:

$$\mathbb{E}[A|l] \equiv \left(\mathbb{E}_\Psi[D_C|\Gamma(A|l, g)] + \mathbb{E}_\Psi[D_C|\Gamma(A|l, l)] \right)/2 = \frac{2d^g + d^l}{2\mu_0 + \mu_p} + 2\chi, \tag{4.34}$$

$$\mathbb{E}[B|l] \equiv \left(\mathbb{E}_\Psi[D_C|\Gamma(B|l, g)] + \mathbb{E}_\Psi[D_C|\Gamma(B|l, l)] \right)/2 = \frac{1}{2} \left(\frac{d^g + d^l}{\mu_0 + \bar{\mu}} + \frac{d^g}{\mu_0 + \mu_p - \bar{\mu} + \alpha} + 3\chi \right). \tag{4.35}$$

Step 2: Dominant Low Investigation Effort. First we compare regions A and B given high investigation effort with the same regions given low investigation effort. Comparing equation 4.31 with equation 4.34 and comparing 4.32 and 4.35 directly shows that $\mathbb{E}[A|l] < \mathbb{E}[A|h]$ and $\mathbb{E}[B|l] < \mathbb{E}[B|h]$ for any $k > 0$, otherwise $\mathbb{E}[A|l] = \mathbb{E}[A|h]$ and $\mathbb{E}[B|l] = \mathbb{E}[B|h]$. This means when investigation effort is costly, it can never be optimal to exert high investigation effort if it leads to a signal-independent response capacity allocation strategy.

Step 3: Dominant High Investigation Effort. We compare region C given high investigation effort with region A and B given low investigation effort. Note that by 2 step 3, region C only exists for $\alpha < \hat{\alpha}$. Solving $\mathbb{E}[C|h] = \mathbb{E}[B|l]$ and $\mathbb{E}[C|h] = \mathbb{E}[A|l]$, respectively gives the threshold functions $\zeta_1\prime$ and $\zeta_2\prime$:

$$\zeta_1\prime \equiv \zeta_1 - \frac{2k}{1 - q}, \tag{4.36}$$

$$\zeta_2\prime \equiv \zeta_2 + \frac{2k}{q}, \tag{4.37}$$

where we take ζ_1 and ζ_2 from the results in Lemma 2 step 2. Next, solving $\zeta_1\prime = \zeta_2\prime$ in q, gives the intersection which we denote by q_h:

$$q_h \equiv \frac{1}{2} + k \left(\frac{(2\mu_0 + \mu_p)(\mu_0 + \bar{\mu})(\mu_0 + \mu_p - \bar{\mu} + \alpha)}{d^g((\mu_0 - 2\bar{\mu})^2 + \alpha(\mu_p - 2(\mu_0 + 2\bar{\mu})))} \right). \tag{4.38}$$

Taking the partial derivative of q_h in k, we are left with the expression in the parentheses on the RHS of 4.38. Note that the numerator of this expression is positive and $2(\mu_0 + 2\bar{\mu}) > \mu_p$ by assumption. Using $\hat{\alpha}$ from Lemma 2 step 3 and also noting that $\hat{\alpha} > 0$ we find that:

$$\lim_{\alpha \to \hat{\alpha}^-} \frac{\partial q_h(\alpha)}{\partial k} = \infty, \tag{4.39}$$

such that for $\alpha < \hat{\alpha}$, it holds that $\partial q_h/\partial k > 0$. It turn, this means $q_h > 1/2$ for $k > 0$ and there exists a $\bar{k}$ such that for $k > \bar{k}$, $q_h > 1$, outside of the feasible range for q. Next, consider that:

$$\frac{\partial \zeta_1\prime}{\partial q} = \frac{\partial \zeta_1}{\partial q} - \frac{2k}{(1 - q)^2} \le \frac{\partial \zeta_1}{\partial q}, \tag{4.40}$$

$$\frac{\partial \zeta_2\prime}{\partial q} = \frac{\partial \zeta_2}{\partial q} - \frac{2k}{q^2} \le \frac{\partial \zeta_1}{\partial q}. \tag{4.41}$$

From Lemma 2 step 3 we know $\partial \zeta_1/\partial q > 0$ and $\partial \zeta_2/\partial q < 0$. Because $\zeta_1\prime < \zeta_2\prime$ at $q = 1/2$ and $\zeta_1\prime = \zeta_2\prime$ at $q = q_h > 1/2$ when $k > 0$, it holds that $\partial \zeta_1\prime/\partial q > \partial \zeta_2\prime/\partial q$. Following this, provided $k > 0$, for $q_h < q$ we find $\zeta_1\prime > \zeta_2\prime$ and for $q < q_h$ we find $\zeta_1\prime < \zeta_2\prime$. Taken together we can

then conclude that for $q > q_h$ and $\zeta_{2\prime} < \chi \leq \zeta_{1\prime}$ it is optimal to exert high investigation effort and follow a signal-dependent allocation strategy (i.e. broad response capacity allocation when $s = g$ and focused response capacity allocation when $s = l$). For $q \leq q_h$ only low investigation effort is optimal and we retrieve ζ_3 from Lemma 2 step 2 as the threshold between broad and focused allocation at optimality.

Step 4: Feasibility of investigation effort and response capacity allocation strategies Having established the optimal investigation effort and response capacity allocation strategies in the preceding steps, note that we need only check for feasibility constraints in χ for region $\mathbb{E}[B|l]$ and region $\mathbb{E}[C|h]$. Next note that the expected disruption cost to the provider of not allocating any response capacity is equal to:

$$\mathbb{E}[D_C] = (d_1^g + d_1^l)/(2\mu_0). \tag{4.42}$$

Equating this expected cost with $\mathbb{E}[B|l]$ and $\mathbb{E}[C|h]$, then solving for χ, we respectively find the threshold functions:

$$\zeta_4 \equiv \frac{1}{3}\left(\frac{2d^g + d_1^l}{\mu_0} - \frac{d^g}{\mu_0 + \mu_p - \bar{\mu} + \alpha} - \frac{d^g + d_1^l}{\mu_0 + \bar{\mu}}\right), \tag{4.43}$$

$$\zeta_5 \equiv \frac{1}{4-q}\left(\frac{2d^g + d_1^l}{\mu_0} - \frac{2(1-q)d_1^l + 4qd^g}{2\mu_0 + \mu_p} - \frac{(1-q)d^g}{\mu_0 + \mu_p - \bar{\mu} + \alpha} - \frac{(1-q)d^g + qd_1^l}{\mu_0 + \bar{\mu}} - 2k\right). \tag{4.44}$$

Consider three cases: 1) $q \leq q_h$ and $\chi > \zeta_4$; 2) $q > q_h$, $\chi > \zeta_{1\prime}$ and $\chi > \zeta_4$; and 3) $q > q_h$, $\chi < \zeta_{1\prime}$ and $\chi > \zeta_5$. In all three cases we find no investigation effort and response capacity allocation strategy is feasible. $\qquad\square$

Proof of Lemma 3. We complete this proof in two steps, first considering the case of client 1 communicating after exerting low investigation effort (step 1) and then considering the case of client communicating after exerting high investigation effort (step 2).

Step 1: Message Thresholds After Exerting Low Investigation Effort. In the case $e = l$, expected costs are independent from the signal, so there are three expected cost functions to compare: one for each type of message and one associated with the decision not to send a message. Let H1, H2 and H3 respectively be the cost to client 1 when communicating $m = g$, communicating $m = l$ and not sending a message. We find:

$$\mathbb{E}[H1] = (d_1^g + d_1^l)/(2\mu_0 + \mu_p) + f - r/2, \tag{4.45}$$

$$\mathbb{E}[H2] = (d_1^g + d_1^l)/(\mu_0 + \bar{\mu}) + f - r/2, \tag{4.46}$$

$$\mathbb{E}[H3] = (d_1^g + d_1^l)/(2\mu_0). \tag{4.47}$$

Comparing $\mathbb{E}[H1]$ and $\mathbb{E}[H2]$, we find $\mathbb{E}[H2] < \mathbb{E}[H1]$ as $\bar{\mu} > \mu_p/2$ by assumption. Therefore we just have to consider the inequality $\mathbb{E}[H3] < \mathbb{E}[H2]$. Solving for f and r on one side of the inequality gives $2f - r > \vartheta_l$, where $\vartheta_l \equiv \bar{\mu}(d_1^g + d_1^l)/(\mu_0(\mu_0 + \bar{\mu}))$.

Step 2: Message Thresholds After Exerting High Investigation Effort. Because in the case $e = h$ the signal does matter there are six expected cost functions to compare. For the case when $e = h$ and $s = g$, let H4, H5 and H6 respectively be the cost to client 1 when communicating $m = g$, communicating $m = l$ and not sending a message. In the same respective order, let H7, H8 and H9 be the expected cost for the case when $e = h$ and $s = l$. We find:

$$\mathbb{E}[H4] = 2(qd_1^g + (1-q)d_1^l)/(2\mu_0 + \mu_p) + f - qr + k, \tag{4.48}$$

$$\mathbb{E}[H5] = (qd_1^g + (1-q)d_1^l)/(\mu_0 + \bar{\mu}) + f - (1-q)r + k, \tag{4.49}$$

$$\mathbb{E}[H6] = (qd_1^g + (1-q)d_1^l)/(\mu_0) + k, \tag{4.50}$$

$$\mathbb{E}[H7] = 2(qd_1^l + (1-q)d_1^g)/(2\mu_0 + \mu_p) + f - (1-q)r + k, \tag{4.51}$$

$$\mathbb{E}[H8] = (qd_1^l + (1-q)d_1^g)/(\mu_0 + \bar{\mu}) + f - qr + k, \tag{4.52}$$

$$\mathbb{E}[H9] = (qd_1^l + (1-q)d_1^g)/(\mu_0) + k. \tag{4.53}$$

Similar to the comparison between $\mathbb{E}[H3]$ and $\mathbb{E}[H2]$ in step 1 of this proof, we compare $\mathbb{E}[H4]$ and $\mathbb{E}[H5]$ to find the inequality $\mathbb{E}[H4] < \mathbb{E}[H5]$ simplifies to $(2q - 1)r > \vartheta_{h1}(g)$, where $\vartheta_{h1}(s) \equiv (2\bar{\mu} - \mu_p)(qd_1^s + (1-q)d_1^{-s})/((\mu_0 + \bar{\mu})(2\mu_0 + \mu_p))$. Next, $\mathbb{E}[H6] < \mathbb{E}[H5]$ simplifies to $f - (1-q)r > \vartheta_{h2}(g)$ where $\vartheta_{h2}(s) \equiv \bar{\mu}(qd_1^s + (1-q)d_1^{-s})/(\mu_0(\mu_0 + \bar{\mu}))$. Finally $\mathbb{E}[H6] < \mathbb{E}[H4]$ simplifies to $f - qr > \vartheta_{h3}(g)$, where $\vartheta_{h3}(s) \equiv \mu_p(qd_1^s + (1-q)d_1^s)/(\mu_0(2\mu_0 + \mu_p))$. Taken together, simultaneously satisfying $f - qr > \vartheta_{h3}(g)$ and $f - (1-q)r > \vartheta_{h2}(g)$ makes not sending a message the preferred choice for client 1. Taking the difference between the two inequalities for each side of the inequality, i.e. $f - (1-q)r - (f - qr) > \vartheta_{h2}(g) - \vartheta_{h3}(g)$, returns $(2q - 1)r > \vartheta_{h1}(g)$. This means if between $m = g$ and $m = l$, $m = g$ is preferred and therefore $(2q - 1)r > \vartheta_{h1}(g)$, then if $f - (1-q)r < \vartheta_{h2}(g)$ it must also hold that $f - qr < \vartheta_{h3}(g)$ such that sending $m = g$ is also preferred over not sending a message. Similarly if $(2q-1)r < \vartheta_{h1}(g)$ and $f - qr < \vartheta_{h3}(g)$, sending $m = l$ is the preferred option. Only three inequalities are therefore needed to characterize the communication strategy in case $s = g$. Inequalities for the case $s = l$ are analogously derived and simplified. When $s = l$, if $(1 - 2q)r < \vartheta(l)_{h1}$ and $f - qr < \vartheta(l)_{h3}$ is optimal for client 1 to communicate truthfully, i.e. $m = l$. Given the assumption that $q > 1/2$ the first condition always holds, which means it is never optimal to communicate $m = g$. In case $f - qr > \vartheta(l)_{h3}$ it is optimal for the client not to send a message. $\qquad \square$

Proof of Proposition 9. We complete this proof in three steps. Step 1 involves finding the disclosure reward necessary to incentivize high investigation effort by client 1. Step 2 involves finding the service fee that guarantees participation by client 1. Step 3 involves evaluating conditions on the cost of investigation effort and response capacity allocation that ensure feasibility of the contract from the provider's perspective.

Step 1: Incentive for high investigation effort. In phase (ii), client 1 decides on what level of investigation effort to exert by evaluating:

$$\mathbb{E}[D_1|f,r] = \min_e \quad \mathbb{E}_s \left[\mathbb{E}_\Psi \left[\frac{d_1^\Psi}{\mu_1(m(e|s))} + T(f,r,m(e|s)) + k\mathbb{1}_{e=h} \right] \right]. \tag{4.54}$$

Given the results in Lemma 3 and restricting the client's strategy set to exclude the option of omission, client 1 will choose to communicate $m = l$ regardless of the signal if there is no incentive to exert high investigation effort. In case there is such an incentive, the incentive should also be such that client 1 will choose to communicate truthfully $m = s$. That is, provided the participation constraint for the client 1 can be met in each instance. Evaluating the expected disruption cost to client 1 in each instance gives:

$$\mathbb{E}[D_1(e = h)] = \frac{1}{2}\left(\frac{qd_1^l + (1-q)d^g}{\mu_0 + \bar{\mu}} + f - qr + k\right) + \frac{1}{2}\left(\frac{qd^g + (1-q)d_1^l}{\mu_0 + \mu_p/2} + f - qr + k\right)$$

$$= \frac{1}{2}\left(\frac{qd_1^l + (1-q)d^g}{\mu_0 + \bar{\mu}} + \frac{qd^g + (1-q)d_1^l}{\mu_0 + \mu_p/2}\right) + f - qr + k, \tag{4.55}$$

$$\mathbb{E}[D_1(e = l)] = \frac{1}{2}\left(\frac{1}{2}\left(\frac{d_1^l + d^g}{\mu_0 + \bar{\mu}}\right) + f - r/2\right) + \frac{1}{2}\left(\frac{1}{2}\left(\frac{d_1^l + d^g}{\mu_0 + \bar{\mu}}\right) + f - r/2\right)$$

$$= \frac{1}{2}\left(\frac{d_1^l + d^g}{\mu_0 + \bar{\mu}} + f - r/2\right). \tag{4.56}$$

Then evaluating $\mathbb{E}[D_1(e = h)] \leq \mathbb{E}[D_1(e = l)]$, solving for r and simplifying the resulting expression gives:

$$r \geq \underline{r} \equiv \frac{(2\bar{\mu} - \mu_p)(qd^g + (1-q)d_1^l)}{(2q-1)(2\mu_0 + \mu_p)(\mu_0 + \bar{\mu})} + \frac{2k}{2q - 1}. \tag{4.57}$$

Step 2: Meeting the participation constraint. Regardless of whether client 1 has the incentive to exert high investigation effort, if the service fee is too high, client 1 has no incentive to call for response in the first place. This is technically the same as client 1 rejecting the provider's contract offer in phase (i). Client 1's expected disruption cost in this situation is:

$$\mathbb{E}[D_1|m = \emptyset] = \frac{1}{2}\left(\frac{d^g + d^1}{\mu_0}\right). \tag{4.58}$$

Evaluating $\mathbb{E}[D_1(e = h)] \leq \mathbb{E}[D_1|m = \emptyset]$ and solving for f gives:

$$f \leq \overline{f} \equiv \frac{d_1^l + d^g}{2\mu_0} - \frac{1}{2}\left(\frac{2(qd^g + (1-q)d_1^l)}{2\mu_0 + \mu_p} + \frac{qd_1^l + (1-q)d^g}{\mu_0 + \bar{\mu}}\right) + qr - k, \tag{4.59}$$

which makes the IR constraint binding for $f = \overline{f}$. The provider thus incentivizes client 1 to exert high investigation effort by offering a contract $T(\overline{f}, \underline{r})$. Although offering a contract $T(0, 0)$ would result in client 1 not exerting investigation effort, there exists a non-zero service fee for which the client still accepts the contract for the benefit of the provider's response capacity. Following this, evaluating $\mathbb{E}[D_1(e = l)] \leq \mathbb{E}[D_1|m = \emptyset]$ gives:

$$f \leq \overline{f}' \equiv \frac{\bar{\mu}(d_1^l + d^g)}{2\mu_0(\mu_0 + \bar{\mu})}, \tag{4.60}$$

such that the contract $T(\overline{f}', 0)$ is optimal to the provider in case it is also optimal to induce low investigation effort and accept miscommunication (i.e. $m = l$) in case $s = g$.

Step 3: Investigation and allocation cost thresholds. What remains to be shown is at which thresholds for the cost of investigation k and cost of response capacity allocation χ it is no

longer feasible or optimal for the provider to induce high investigation effort and truth-telling by client 1. Following step 2 of this proof we can compare the provider's expected disruption cost having offered $T(\overline{f}, \underline{r})$ with the expected disruption cost having offered $T(\overline{f}', 0)$. Evaluating the expected disruption costs to the provider as in Equation 4.5 in Section 4.3.3 conditional on (l, l, l) and (g, l, h) and substituting the relevant contract parameters, we have:

$$\mathbb{E}[D_p|(l,l,l)]] = \frac{1}{2}\left(\frac{d_1^l + d^g}{\mu_0 + \bar{\mu}} + \frac{d^g}{\mu_0 + \mu_p - \bar{\mu} + \alpha} + 3\chi - 2\overline{f}'\right), \tag{4.61}$$

$$\mathbb{E}[D_p|(g,l,h)] = \frac{1}{2}\left(\frac{2((1-q)d_1^l + 2qd^g)}{2\mu_0 + \mu_p} + \frac{(1-q)d^g}{\mu_0 + \mu_p - \bar{\mu} + \alpha} + \frac{(1-q)d^g + qd_1^l}{\mu_0 + \bar{\mu}}\right)$$
$$+ \frac{1}{2}\left((4-q)\chi + 2(q\underline{r} - \overline{f})\right). \tag{4.62}$$

The expected disruption cost to the provider in case client 1 rejects the contract is:

$$\mathbb{E}[D_p|(\emptyset,\emptyset,l)] = \frac{1}{2}\left(\frac{d_1^l + 2d^g}{\mu_0}\right). \tag{4.63}$$

Evaluating $\mathbb{E}[D_p|(g,l,h)] = \mathbb{E}[D_p|(l,l,l)]$, $\mathbb{E}[D_p|(l,l,l)] = \mathbb{E}[D_p|(\emptyset,\emptyset,l)]$, and $\mathbb{E}[D_p|(g,l,h)] = \mathbb{E}[D_p|(\emptyset,\emptyset,l)]$ at the relevant contract terms found in step 1 and step 2 of this proof and solving for χ respectively give:

$$\zeta_1'' \equiv \frac{2}{1-q}\left(\frac{qd^g + (1-q)d_1^l}{\mu_0 + \bar{\mu}} + \frac{qd^g}{2(\mu_0 + \mu_p - \bar{\mu} + \alpha)} - \frac{3qd^g + 2(1-q)d_1^l}{2\mu_0 + \mu_p} - k\right), \tag{4.64}$$

$$\zeta_4' \equiv \frac{1}{3}\left(\frac{d^g}{\mu_0} + \frac{d^g}{\mu_0 + \mu_p - \bar{\mu} + \alpha} - \frac{2(d^g + d_1^l)\bar{\mu}}{\mu_0(\mu_0 + \bar{\mu})}\right), \tag{4.65}$$

$$\zeta_5' \equiv \frac{2}{4-q}\left(\frac{3d^g + 2d_1^l}{\mu_0} - \frac{4(1-q)d_1^l + 6qd^g}{2\mu_0 + \mu_p} - \frac{(1-q)d^g}{\mu_0 + \mu_p - \bar{\mu} + \alpha} - \frac{2(1-q)d^g + 2qd_1^l}{\mu_0 + \bar{\mu}} - k\right). \tag{4.66}$$

By means of these threshold functions we find it is optimal for the provider to set $T(\overline{f}, \underline{r})$ and induce both high investigation effort and truth-telling by client 1 so long as $\chi \leq \zeta_5'$ and $\chi \leq \zeta_1''$. It is optimal for the provider to set $T(\overline{f}', 0)$ and induce client 1 to communicate while allowing for miscommunication so long as $\chi > \zeta_1''$ and $\chi \leq \zeta_4$. Otherwise contracting is not feasible, meaning the provider is better off not contracting for response and letting the system absorb any disruption costs. Finally we solve $\partial \zeta_1''/\partial q = 0$ for α to find:

$$\alpha^\dagger \equiv \frac{d^g(2\bar{\mu} - \mu_p)(\mu_0 + \bar{\mu}) - 2(\mu_0 + \mu_p - \bar{\mu})(d^g(2\bar{\mu} - \mu_p) + k(2\mu_0 + 2\mu_p)(\mu_0 + \bar{\mu}))}{2d^g(2\bar{\mu} - \mu_p) + 2(d^g + k(\mu_p + 2\mu_0)(\mu_0 + \bar{\mu}))}. \tag{4.67}$$

And next substitute $\alpha^\dagger$ and $q = 1/2$ into ζ_1'' to find:

$$\zeta_1''\big|_{\alpha=\alpha^\dagger, q=1/2} = 2d^g\left(\frac{1}{\mu + \bar{\mu}} - \frac{2}{2\mu_0 + \mu_p}\right) - 2k < 0, \tag{4.68}$$

where the inequality holds because $2\bar{\mu} > \mu_p$ and $k \geq 0$ by assumption. Following the same argumentation for the finding in step 3 in the proof for Lemma 2, we therefore find that for $\alpha \geq \alpha^\dagger$ offering $T(\overline{f}, \underline{r})$ to induce a truth-telling by client 1 and follow a signal-dependent

response capacity allocation strategy is not beneficial for any signal quality or allocation cost.
□

Proof of Proposition 10. We complete this proof in two steps. In step 1 we evaluate and compare the expected cost functions for client 1 for each of the strategies in $\mathcal{S}$ that include omission (i.e. $(\emptyset, l, h)$, $(l, \emptyset, h)$ and $(g, \emptyset, h)$) given the optimal contract terms found in Proposition 9, to show under which conditions each strategy is dominant to the other two. In step 2 we compare the omission strategies to the truth-telling strategy given the same contract parameters and determine the necessary adjustment of the contract terms to preserve high investigation effort and truth-telling when omission is allowed.

Step 1: Expected disruption cost to client 1 under omission. Evaluating the expected cost to client 1 in light of the three strategies including omission, we find:

$$\mathbb{E}[D_1(\emptyset, l, h)] = \frac{1}{2}\left(\frac{qd_1^l + (1-q)d^g}{\mu_0 + \bar{\mu}} + \frac{qd^g + (1-q)d^l}{\mu_0} + f - qr + 2k\right), \tag{4.69}$$

$$\mathbb{E}[D_1(l, \emptyset, h)] = \frac{1}{2}\left(\frac{qd_1^l + (1-q)d^g}{\mu_0} + \frac{qd^g + (1-q)d^l}{\mu_0 + \bar{\mu}} + f - (1-q)r + 2k\right), \tag{4.70}$$

$$\mathbb{E}[D_1(g, \emptyset, h)] = \frac{1}{2}\left(\frac{qd_1^l + (1-q)d^g}{\mu_0} + \frac{qd^g + (1-q)d^l}{\mu_0 + \bar{\mu}} + f - qr + 2k\right). \tag{4.71}$$

Next, substituting $f = \bar{f}$ and $r = \underline{r}$ and simplifying gives:

$$\mathbb{E}[D_1(\emptyset, l, h)|\bar{f}, \underline{r}] = \frac{1}{4}\left(2k + \frac{(3-2q)d^l + (1+2q)d^g}{\mu_0} - \frac{2((1-q)d^l + qd^g)}{2\mu_0 + \mu_p} + \frac{(1-q)d^g + qd^l}{\mu_0 + \bar{\mu}}\right), \tag{4.72}$$

$$\mathbb{E}[D_1(l, \emptyset, h)|\bar{f}, \underline{r}] = \frac{1}{4}\left(6k + \frac{(3-2q)d^l + (1+2q)d^g}{\mu_0} + \frac{2((1-q)d^l + qd^g)}{2\mu_0 + \mu_p} - \frac{(1-q)d^g + qd^l}{\mu_0 + \bar{\mu}}\right), \tag{4.73}$$

$$\mathbb{E}[D_1(g, \emptyset, h)|\bar{f}, \underline{r}] = \frac{1}{4}\left(2k + \frac{(3-2q)d^l + (1+2q)d^g}{\mu_0} + \frac{2((1-q)d^l + qd^g)}{2\mu_0 + \mu_p} - \frac{(1-q)d^g + qd^l}{\mu_0 + \bar{\mu}}\right). \tag{4.74}$$

Which clearly shows that $\mathbb{E}[D_1(g, \emptyset, h)|\bar{f}, \underline{r}] < \mathbb{E}[D_1(l, \emptyset, h)|\bar{f}, \underline{r}]$. Despite omission in case of a signal $s = l$ common to both strategies, client 1 still prefers truthful communication in case of a signal $s = g$, which is the intention of the contract terms. Hence we only need to compare strategies $(g, \emptyset, h)$ and $(\emptyset, l, h)$ with the truth-telling strategy (g, l, h).

Step 2: Contract adjustment under omission. From the proof of Proposition 9 step 2, we know for $f = \bar{f}$, client 1's IR constraint is binding when inducing truth-telling is optimal and feasible. Similarly, we know for $f = \bar{f}'$, client 1's IR constraint is binding when instead inducing client 1's communication strategy (l, l, l) is optimal and feasible. Evaluating the expected cost to client 1 following strategy (g, l, h) under $T(\bar{f}, \underline{r})$ therefore gives:

$$\mathbb{E}[D_1(g, l, h)|\bar{f}, \underline{r}] = \mathbb{E}[D_1(l, l, l)|\bar{f}', 0] = \mathbb{E}[D_1(\emptyset, \emptyset, l)] = \frac{1}{2}\left(\frac{d^g + d^1}{\mu_0}\right). \tag{4.75}$$

In the following, we present the arguments for this proof for $\mathbb{E}[D_1(g, l, h)|\overline{f}, \underline{r}]$, although by Equation 4.75 the same results hold for $\mathbb{E}[D_1(l, l, l)|\overline{f}', 0]$. From Equation 4.72, Equation 4.74 and Equation 4.75 we can see all cost functions have positive and linear, but different slopes in d^l and d^g. Evaluating the partial derivates in d^l gives:

$$\frac{\partial}{\partial d^l}\mathbb{E}[D_1(\emptyset, l, h)|\overline{f}, \underline{r}] = \frac{1}{4}\left(\frac{(3 - 2q)\mu_p + (4 - 2q)\mu_0}{\mu_0(2\mu_0 + \mu_p)} + \frac{q}{\mu_0 + \bar{\mu}}\right), \tag{4.76}$$

$$\frac{\partial}{\partial d^l}\mathbb{E}[D_1(g, \emptyset, h)|\overline{f}, \underline{r}] = \frac{1}{4}\left(\frac{(1 + 2q)\mu_p + (4 + 2q)\mu_0}{\mu_0(2\mu_0 + \mu_p)} - \frac{q}{\mu_0 + \bar{\mu}}\right), \tag{4.77}$$

$$\frac{\partial}{\partial d^l}\mathbb{E}[D_1(g, l, h)|\overline{f}, \underline{r}] = \frac{1}{2\mu_0}, \tag{4.78}$$

which can be shown similarly for derivatives in d^g. Subtracting Equation 4.76 from Equation 4.77 we find:

$$\frac{1}{2}\left(\frac{(2q - 1)\mu_p + 2q\mu_0}{\mu_0(2\mu_0 + \mu_p)} - \frac{q}{\mu_0 + \mu_p}\right) > 0, \tag{4.79}$$

which means the slope is higher for $\mathbb{E}[D_1(g, \emptyset, h)]$. Then, subtracting Equation 4.76 from Equation 4.78 and subtracting Equation 4.77 from Equation 4.78 respectively give:

$$\frac{1}{4}\left(\frac{(2q - 1)\mu_p + 2q\mu_0}{\mu_0(2\mu_0 + \mu_p)} - \frac{q}{\mu_0 + \mu_p}\right) > 0 \quad \text{and} \tag{4.80}$$

$$\frac{1}{4}\left(\frac{(1 - 2q)\mu_p - 2q\mu_0}{\mu_0(2\mu_0 + \mu_p)} + \frac{q}{\mu_0 + \mu_p}\right) < 0, \tag{4.81}$$

which means the slope for $\mathbb{E}[D_1(g, l, h)]$ is exactly halfway between the slope for $\mathbb{E}[D_1(g, \emptyset, h)]$ and $\mathbb{E}[D_1(\emptyset, l, h)]$. Moreover, it means there exist two thresholds for the ratio of d^l to d^g below and above which respectively, client 1 has the incentive to not communicate in case she receives a signal $s = l$ or a signal $s = g$. Next we characterize these thresholds. Solving $\mathbb{E}[D_1(\emptyset, l, h)|\overline{f}, \underline{r}] = \mathbb{E}[D_1(g, \emptyset, h)|\overline{f}, \underline{r}]$ for d^l gives a solution $d^l = \delta d^g$, where:

$$\delta \equiv \frac{(2q - 1)\mu_p\bar{\mu} + q\mu_0\mu_p - (1 - q)\mu_0\bar{\mu}}{(2q - 1)\mu_p\bar{\mu} + 2q\mu_0\bar{\mu} - (1 - q)\mu_0\mu_p}. \tag{4.82}$$

For $k = 0$ and $d^l = \delta d^g$, we find $\mathbb{E}[D_1(\emptyset, l, h)|\overline{f}, \underline{r}] = \mathbb{E}[D_1(g, \emptyset, h)|\overline{f}, \underline{r}] = \mathbb{E}[D_1(g, l, h)|\overline{f}, \underline{r}]$, meaning the thresholds coincide. For $k > 0$ and $d^l = \delta d^g$, (g, l, h) remains dominant and we find the two unique thresholds by solving $\mathbb{E}[D_1(\emptyset, l, h)|\overline{f}, \underline{r}] = \mathbb{E}[D_1(g, l, h)|\overline{f}, \underline{r}]$ and $\mathbb{E}[D_1(g, \emptyset, h)|\overline{f}, \underline{r}] = \mathbb{E}[D_1(g, l, h)|\overline{f}, \underline{r}]$ for d^l. We find the two thresholds respectively at $d^l = (\delta - x)d^g$ and $d^l = (\delta + x)d^g$, where we define:

$$x \equiv \frac{2k\mu_0(2\mu_0 + \mu_p)(\mu_0 + \bar{\mu})}{(2q - 1)\mu_p\bar{\mu} + 2q\mu_0\bar{\mu} - (1 - q)\mu_0\mu_p}. \tag{4.83}$$

Following this, for $d^l < (\delta - x)d^g$, client 1 expects a lower disruption cost under the contract $T(\overline{f}, \underline{r})$ when not communicating if she receives a signal $s = l$. Similarly, for $d^l > (\delta + x)d^g$, client 1 prefers not to communicate if she receives a signal $s = g$. In either case, the provider could adjust the fee downward to correct for the difference in expected disruption cost to client 1. From Equation 4.80 and Equation 4.81 it follows that the necessary adjustment of the fixed

fee in the contract is equal to:

$$
\Delta f = \begin{cases}
\dfrac{d^l - (\delta + x)d^g}{4}\left(\dfrac{2q(\mu_0 + \mu_p) - \mu_p}{\mu_0(2\mu_0 + \mu_p)} - \dfrac{q}{\mu_0 + \bar{\mu}}\right) & \text{when } d^l > (\delta + x)d^g \\[4mm]
\dfrac{(\delta - x)d^g - d^l}{4}\left(\dfrac{2q(\mu_0 + \mu_p) - \mu_p}{\mu_0(2\mu_0 + \mu_p)} - \dfrac{q}{\mu_0 + \bar{\mu}}\right) & \text{when } d^l < (\delta - x)d^g \\[4mm]
0 & \text{otherwise.}
\end{cases} \tag{4.84}
$$

By Equation 4.75, the fixed fee in both $T(\overline{f}, \underline{r})$ and $T(\overline{f}', 0)$ should be adjusted by Δf to prevent omission, i.e. $T((\overline{f} - \Delta f), \underline{r})$ and $T((\overline{f} - \Delta f), 0)$, provided $d^l < (\delta - x)d^g$ or $d^l > (\delta + x)d^g$. Following a downward adjustment of the fixed fee, by the arguments in step 3 of Proposition 9 (equations Equation 4.61-Equation 4.66) the thresholds $\zeta_4\prime$ and $\zeta_5\prime$ must necessary shift down respectively by $\Delta f/3$ and $2\Delta f/(4 - q)$. $\qquad\square$

References

Angalakudati M, Balwani S, Calzada J, Chatterjee B, Perakis G, Raad N, Uichanco J (2014) Business analytics for flexible resource allocation under random emergencies. *Management Science* 60(6):1552–1573.

Argon NT, Ziya S (2009) Priority assignment under imperfect information on customer type identities. *Manufacturing & Service Operations Management* 11(4):674–693.

August T, Tunca TI (2006) Network software security and user incentives. *Management Science* 52(11):1703–1720.

Babich V, Tang CS (2012) Managing opportunistic supplier product adulteration: Deferred payments, inspection, and combined mechanisms. *Manufacturing & Service Operations Management* 14(2):301–314.

Baiman S, Fischer PE, Rajan MV (2000) Information, contracting, and quality costs. *Management Science* 46(6):776–789.

Bakshi N, Kim SH, Savva N (2015) Signaling new product reliability with after-sales service contracts. *Management Science* 61(8):1812–1829.

Baxter (2001) Baxter announces agreement with families affected by recent hemodialysis deaths in spain. URL `https://web.archive.org/web/20060313050723/http://www.baxter.com/about_baxter/news_room/news_releases/2001/11-28spsettlement.html`, retrieved 24/11/2016.

BBC News (2001) Croatia links kidney deaths to machines. URL `http://news.bbc.co.uk/1/hi/world/europe/1600667.stm`, retrieved 7/7/2015.

Chan T, De Vericourt F, Besbes O (2014) Contracting in medical equipment maintenance services: an empirical investigation, working Paper, ESMT.

Cohen MA, Agrawal N, Agrawal V (2006) Winning in the aftermarket. *Harvard business review* 84(5):129.

FDA (2002) Baxter Recalls Dialyzers - Class I Recall. URL `http://www.fda.gov/ICECI/EnforcementActions/EnforcementStory/EnforcementStoryArchive/ucm103377.htm`, retrieved 24/11/2016.

FDA (2013) Medical Device Recall Report: FY2003 to FY2012. Technical report, Food and Drug Administration.

Hammonds K (2002) Harry kraemer's moment of truth. URL `http://www.fastcompany.com/45545/harry-kraemers-moment-truth`, retrieved 7/7/2015.

Jain N, Hasija S, Popescu DG (2013) Optimal contracts for outsourcing of repair and restoration services. *Operations Research* 61(6):1295–1311.

Kim SH (2015) Time to come clean? disclosure and inspection policies for green production. *Operations Research* 63(1):1–20.

Kim SH, Cohen MA, Netessine S (2007) Performance contracting in after-sales service supply chains. *Management Science* 53(12):1843–1858.

Kim SH, Cohen MA, Netessine S, Veeraraghavan S (2010) Contracting for infrequent restoration and recovery of mission-critical systems. *Management Science* 56(9):1551–1567.

Kim SH, Tomlin B (2013) Guilt by association: Strategic failure prevention and recovery capacity investments. *Management Science* 59(7):1631–1649.

Lee HH, Pinker EJ, Shumsky RA (2012) Outsourcing a two-level service process. *Management Science* 58(8):1569–1584.

Levitt SD, Snyder CM (1997) Is no news bad news? information transmission and the role of "early warning" in the principal-agent model. *The RAND Journal of Economics* 641–661.

McKinsey Center for Government (2013) Medical Device Recall Report: FY2003 to FY2012. Technical report, McKinsey Center for Government.

Paluch S, Wünderlich NV (2016) Contrasting risk perceptions of technology-based service innovations in inter-organizational settings. *Journal of Business Research* 69(7):2424–2431.

Philips (2014) Philips Remote Services Brochure. Technical Report November, Philips.

Plambeck EL, Taylor TA (2015) Supplier evasion of a buyer's audit: Implications for motivating supplier social and environmental responsibility. *Manufacturing & Service Operations Management* 18(2):184–197.

Roels G, Karmarkar US, Carr S (2010) Contracting for collaborative services. *Management Science* 56(5):849–863.

Schlapp J, Oraiopoulos N, Mak V (2015) Resource allocation decisions under imperfect evaluation and organizational dynamics. *Management Science* 61(9):2139–2159.

Sheffi Y (2005) *The Resilient Enterprise: Overcoming Vulnerability for Competitive Advantage* (The MIT Press), 1st paperback edition, chapter 9.

Shumsky RA, Pinker EJ (2003) Gatekeepers and referrals in services. *Management Science* 49(7):839–856.

Tang CS (2006) Perspectives in supply chain risk management. *International Journal of Production Economics* 103(2):451–488.

Terhune C (2016) FDA Faults Hospitals for Failing to Report Safety Problems with Medical Devices. URL http://www.npr.org/sections/health-shots/2016/10/28/499732331/fda-faults-hospitals-for-failing-to-report-safety-problems-with-medical-devices, retrieved 1/12/2016.

Tomlin B (2006) On the value of mitigation and contingency strategies for managing supply chain disruption risks. *Management Science* 52(5):639–657.

Van Mieghem JA (1998) Investment strategies for flexible resources. *Management Science* 44(8):1071–1078.

Wang S, Sun P, de Véricourt F (2016) Inducing environmental disclosures: A dynamic mechanism design approach. *Operations Research* 64(2):371–389.

Wang Y, Gilland W, Tomlin B (2010) Mitigating supply risk: Dual sourcing or process improvement? *Manufacturing & Service Operations Management* 12(3):489–510.

Wex F, Schryen G, Feuerriegel S, Neumann D (2014) Emergency response in natural disaster management: Allocation and scheduling of rescue units. *European Journal of Operational Research* 235(3):697–708.

Wu OQ, Babich V (2012) Unit-contingent power purchase agreement and asymmetric information about plant outage. *Manufacturing & Service Operations Management* 14(2):245–261.

Yang Z, Aydin G, Babich V, Beil DR (2009) Supply disruptions, asymmetric information, and a backup production option. *Management Science* 55(2):192–209.

Yang Z, Aydin G, Babich V, Beil DR (2012) Using a dual-sourcing option in the presence of asymmetric information about supplier reliability: Competition vs. diversification. *Manufacturing & Service Operations Management* 14(2):202–217.

Young E (2001a) Dozens of dialysis deaths across europe. URL `http://www.newscientist.com/article/dn1435-dozens-of-dialysis-deaths-across-europe.html\#.VZu7tfmrSUk`, retrieved 7/7/2015.

Young E (2001b) Further deaths prompt worldwide recall of dialysis filters. URL `http://www.newscientist.com/article/dn1458-further-deaths-prompt-worldwide-recall-of-dialysis-filters.html\#.VZu7j\_mrSUk`, retrieved 7/7/2015.

Young E (2001c) Manufacturing fluid linked to kidney dialysis deaths. URL `http://www.newscientist.com/article/dn1521-manufacturing-fluid-linked-to-kidney-dialysis-deaths.html\#.VZu7w\_mrSUk`, retrieved 7/7/2015.

Young E (2002) Dialysis filters test positive for poison gas. URL `http://www.newscientist.com/article/dn1753-dialysis-filters-test-positive-for-poison-gas.html\#.VZu7zvmrSUk`, retrieved 7/7/2015.

Chapter 5

Conclusions

This dissertation studies firms' strategic interactions in anticipation of random service disruption following technology failure. In particular it is aimed at understanding how firms invest in detection and response measures, making sure disruption response and recovery are managed as efficiently as possible in with all stakeholders.

A central theme to this dissertation is the need for vendors and clients to collaborate in responding to and recovering from technology failure. When it comes to mission-critical technology (e.g. IT or medical technology) the client operates the system on a daily basis, but the vendor understands the architecture of the system. As such, each party plays a vital role responding to red flags and tracking down the root cause of any problem. Appropriate contract design can help align incentives for disruption risk mitigation. Each of the main chapters highlights different findings to justify the importance of structuring the right incentives to minimize disruption costs and seek to contribute to the literature on contract design for service operations and technology management in operations management.

Chapter 2 in this dissertation shows how in an IT maintenance outsourcing relationship a client should balance penalizing the vendor for downtime with investments in facilitating response to improve system performance, particularly when either party may not be risk neutral. To understand how a client should balance the need to support the vendor while setting the right incentives for the vendor to invest, we develop a model that combines the key characteristics of value co-creation (i.e. complementarity between the firms' investments in response capacity) with maintenance contract practices (i.e. penalty contracts that penalize the vendor for system downtime). We study the difference in the client's expected utility between a case in which investment in response capacity is observable and a case in which it is not. These two cases reflect two extremes in system architecture: simple systems require straightforward measures to respond to problem, whereas responsibilities and readiness to respond are hard to determine in complex systems. We refer to the difference in outcomes between the two cases as the cost of complexity. Firstly, we show that the cost of complexity to the client is decreasing in the risk aversion of vendor but increasing in her own risk aversion. Secondly, we find that a larger difference in risk aversion between client and vendor leads to underinvestment in system uptime in case the client's investment is observable, yet the opposite happens when the client's investment is not observable. The managerial implications of these findings are that when the client is highly risk-averse, she has a lot to gain by making her efforts more observable (e.g., by investing in monitoring mechanisms or processes and systems that make her efforts more

transparent). On the contrary, when the client is working with a highly risk-averse vendor, investing in increasing the observability of her efforts will not have a substantial impact on her profits. Moreover, higher observed uptime may actually be a signal of inefficient investment in high complexity settings. Thus, the relative risk preferences of the two firms play a critical role in the extent to which asymmetric information affects the efficiency of collaborative response.

Contracts are the prime conduit of an outsourcing relationship between two or multiple companies, but leave room for behavioral factors to be in the way efficient contracting decisions. For Chapter 3, we therefore designed an experiment to build on findings in the first chapter and demonstrate that differences in risk preference indeed form a barrier to collaborative response to disruptions, exacerbated by biases in decision making. We built a proprietary software environment in which subjects take the role of a client in a technology outsourcing relationship with a single, computerized vendor. We simulated a contracting game in which subjects decided on both the downtime penalty to the vendor and an investment in response capacity in order to minimize disruption costs. We focused this study on two sets of hypotheses, to do with the effect of differences in risk aversion on subject decisions as well how cognitive feedback on the vendor's risk profile may improve subject decision making. Comparing decisions with the conditionally optimal benchmarks we arrive at two observations that highlight possible heuristic decision biases. Firstly, subjects tend to set and hold on to an inefficiently high investment level even though it is theoretically optimal to adjust decisions under changing differences in risk preferences. Secondly, subjects tend to set and hold on to a penalty that is too high when interacting with more risk averse vendors and too low in case the vendor is equally risk averse, again suggesting subjects stick to inefficient decisions, a well-known finding in the behavioral operations management literature. Surprisingly, cognitive feedback on the vendor's risk aversion appears to have counterproductive effects on subject's performance in the experiment, suggesting cognitive overload can have a reinforcing effect on the heuristic decision biases observed. This study comes with two key managerial implications. The first is that in adjusting to different contracting partners, managers responsible for contract design and implication should be wary of anchoring on previous decisions, particularly when conditions may have shifted. The second is that investing effort in understanding the vendor's risk aversion may make contracting decisions all the more difficult. Particular when it comes to understanding how the vendor's preferences relate to her own, a possible 'cognitive overload' leads to over-reliance on existing heuristics where adjustments are needed to correct for changing conditions.

Chapters 2 and 3 exclusively examined strategic interactions in bilateral relations. However, in reality increasingly large client pools have come to depend on continuous availability and security of equipment provided and maintained by a single specialist vendor. The bilateral relations studied in the first chapter thus only represents a special case among generally larger and more connected technology networks. The scale of the disruption in technology networks is typically unknown at the onset of disruption. Moreover, faulty equipment can in some cases cause life threatening circumstances to consumers or patients downstream. Motivated by a

widespread problem with dialysis equipment from a single provider affecting hospitals across Europe and the US in 2001, the second chapter studies how detection of latent disruption risks in a network of firms reliant on the same technology can be improved. By means of a simple contracting game between a single vendor and two clients we examined the effect of a disclosure reward to ensure both high investigation effort and truthful communication by the client who first raises alarm in case of a disruption. The results show that while clients may be incentivized to identify and report network disruptions, competition for scarce emergency resources and the required investment in understanding their own exposure may incentivize clients to deliberately miscommunicate with the vendor. This implies that the provider should take caution in allocating response capacity following alarms raised by clients. Even when it is feasible to implement the optimal contract terms, differences between optimal allocation strategies from a centralized perspective and optimal allocation strategies from a decentralized perspective highlight that miscommunication and delay of response can turn out to be optimal, echoing real world observations surrounding response to medical equipment failures.

All three chapters that form the core of this dissertation are still works in progress and limitations in each of the papers deserve to be recognized. Particular concerns raised by external reviewers for Chapters 2 and 3 were reflected in the chapters' respective discussion sections. Recognizing limitations to Chapter 2 with regards to the restriction to bilateral contracts and possibly overlooking the role of information exchange at the onset of disruption was what led to the development of Chapter 4. Nonetheless, it bears repeating that the complexity surrounding the outsourcing of mission-critical technology and investments in mitigating future disruptions makes for a difficult context to capture either in a one-shot contracting game with a stylized decision space, let alone an experiment that attempts to reflect these conditions while controlling factors of interest. Successful further development of all three works will rest on finding representative cases that bring to light the tapestry of risk preferences, information gathering and exchange and interdependencies in response to disruptions. Redefining both the models and experimental design across the chapter to match this tapestry may allow for results and insights to follow more naturally, and with direct reference to the motivating cases.

To conclude, effective response to disruptions requires having *sufficient* response capacity ready in advance and *efficient* deployment when these resources are needed. This requires a vendor and its clients to act collaboratively, both in anticipation and in the aftermath of disruption. Contracts govern the relationships between each client and vendor, but designing and implementing the right contract is a difficult task. Depending on the context, we demonstrate penalties on downtime and disclosure rewards can be powerful tools for disruption risk mitigation, as long as both resource collaboration and competition as well differences in objectives and preferences can be taken into consideration.