PERFORMANCE-BASED INCENTIVES IN PUBLIC SECTOR SERVICES

Vanitha Virudachalam

A DISSERTATION

in

Operations, Information and Decisions

For the Graduate Group in Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2020

Supervisor of Dissertation

Sergei Savin, Associate Professor of Operations, Information and Decisions

Graduate Group Chairperson

Nancy Zhang, Ge Li and Ning Zhao Professor, Professor of Statistics

Dissertation Committee

Serguei Netessine, Dhirubhai Ambani Professor of Innovation and Entrepreneurship, Professor of Operations, Information and Decisions, Vice Dean for Global Initiatives

Xuanming Su, Murrel J. Ades Professor, Professor of Operations, Information and Decisions

PERFORMANCE-BASED INCENTIVES IN PUBLIC SECTOR SERVICES

For Raj, who has supported me every step of the way, and for Kailash, who teaches to me to find joy in each day.

ACKNOWLEDGMENT

Above all, I wish to thank my advisor, Sergei Savin, for his thoughtful guidance, his seemingly infinite patience, and for believing in me and in this work. I am grateful to have been mentored by someone whom I admire both as a researcher and a person.

I would also like to thank Maria Rieders for her encouragement throughout this journey. When the realities of being both a mother and a doctoral student seemed intractable, her support and example gave me strength.

I am lucky to have worked with coauthors who provided invaluable feedback and contributions; thank you to Hessam Bavafa, Lerzan Örmeci, and Matthew Steinberg. I am also grateful to Serguei Netessine and Xuanming Su for their many useful suggestions, especially for their help nurturing the germ of an idea for a first-year seminar paper into what became the first chapter of this dissertation.

I am grateful to my sister, Senbagam Virudachalam, who, despite not knowing exactly what I've been doing for these last several years, is still convinced I am the best at it. She has always been my loudest cheerleader, and her presence made Philly feel like home.

Finally, thank you to my parents, Ramasamy and Subbulakshmi Virudachalam, whose love, graciousness, and incomparable work ethic have always been an inspiration. They are my first and most important role models. I can only hope to live up to their example.

ABSTRACT

PERFORMANCE-BASED INCENTIVES IN PUBLIC SECTOR SERVICES

Vanitha Virudachalam

Sergei Savin

This dissertation studies the role of performance-based incentives in public sector service operations. In particular, we consider incentives in the context of K-12 education and healthcare. These systems share three key characteristics: the quality of service provision depends both on the efforts of the service provider and customer, the provider is paid by a third-party, and measuring the quality of service provision is difficult. In three chapters, we study different facets of the incentive problem. In the first chapter, we consider the problem faced by school districts seeking to maximize student performance, as measured by annual state exams. Using a dynamic two-period principal-agent model, we study the interaction between two levers that the school district can use to improve performance: a midyear "interim assessment," which will more accurately gauge whether students are on track to perform well, and performance-based incentives for teachers. We show that investing in an interim assessment is beneficial in only a limited number of scenarios; our results suggest that the growing dependence on third-party assessments may be misplaced. In the second chapter, we turn to healthcare. We study the problem faced by a profit-maximizing, resource-constrained hospital that controls patient inflows by designing a case-mix of its elective procedures and patient outflows via patient discharges. We consider a hospital that makes these decisions in the presence of bundled payments, which implicitly penalize high readmission rates. In our analysis, we focus on assessing the benefits associated with the hospital employing a coordinated decision-making process, in which both portfolio and discharge decisions are made in tandem. We compare a coordinated decision-making structure to two commonly utilized decision-making structures and characterize when the hospital can most benefit from coordination. Finally, in the third chapter, we return to K-12

education. We study the impact of coproduction on the student performance in the presence of merit-based rewards for both teachers and students, using a Cobb-Douglas formulation of the education production function and a Stackelberg model of teachers' and students' effort decisions. We characterize students' and teachers' optimal effort levels for a given reward allocation, and we illustrate the impact on student performance.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

PREFACE

In this dissertation, I study public sector service operations, specifically, K-12 education and healthcare. While clearly distinct, healthcare and education share many key similarities. Both are vast systems staffed with multiple providers that must coordinate their efforts. For both, service provision and its outcome are determined by the joint efforts of the customer (student or patient) and the provider (educator or healthcare worker). In both systems, strategies to improve service provision have been transformed by technology and an increasing ability to collect data. And, simply put, the efficacy of both systems is vital to ensure and sustain the well-being of a population.

In addition, both systems share a similar payment and incentive structure, in which a service provider serves a customer who does not directly provide compensation. Rather, a third-party payer, whether that is the government or an insurance company, must pay the provider – but this is complicated by the difficulty of measuring the quality of service provision. Thus, the payer must determine a compensation scheme that incentivizes the provider to exert the appropriate level of effort.

In the chapters that follow, I focus specifically on the role of performance-based incentives in these two systems. In the first chapter, I study the interaction between end-of-the year performance-based incentives and the provision of more accurate midyear information in K-12 education in the United States. In the second chapter, I turn to healthcare and analyze the optimal patient portfolio and discharge threshold for a resource-constrained hospital under a reimbursement policy that penalizes readmissions. Finally, in the third chapter, I return to education, and, recognizing the coproduced nature of education, I consider the optimal allocation of performance-based incentives between teachers and students.

In Chapter 1, the focus is on the problem faced by school districts seeking to maximize student performance, as measured by annual state exams. Legislation over the last two decades has led to a dramatic increase in the frequency of state standardized testing in schools. More-

over, technological advances that have made possible the collection of tremendous amounts of student performance data. A consequence of this has been the explosive growth of the for-profit testing industry, with school districts spending hundreds of thousands of dollars on "interim assessments," formal, point-in-time tests designed to gauge student progress during the school year. This additional informational input into the educational process is being introduced at a time when a number of school districts are also experimenting with performance-based incentives for teachers. Thus, school districts must consider the relative benefits and interaction of these two strategies.

We examine the relationship between information on student performance and incentives for teachers using a dynamic two-period principal-agent model. In our model, the school district (principal) chooses whether to invest in interim assessments, and, also, how much merit-based compensation to offer teachers, while the teachers (agents) decide on the level of effort to exert in each period. Our results indicate that interim assessments are beneficial in only a limited number of scenarios. For schools that begin the year already behind and not on track to meet state proficiency standards, interim assessments have a limited impact on teachers' efforts if the baseline probability of moving to the proficient state is low. If there is higher baseline probability of moving to the proficient state, an interim assessment may be valuable under low-to-moderate budget levels. For schools that start the year on track to achieve proficiency on the state tests, the school district should invest in an interim assessment if the budget is moderate, the formative assessments are reasonably accurate, and the probability of transitioning to the proficient state from the not-proficient state is significantly lower than the probability of remaining in the proficient state.

Next, in Chapter 2, we turn to the healthcare domain. Specifically, we study the problem faced by a profit-maximizing, resource-constrained hospital that controls patient inflows by designing a case-mix of its elective procedures and patient outflows via patient discharges. We consider a hospital that makes these decisions in the presence of bundled payments, which implicitly penalize high readmission rates. Our model analyzes the impact of pa-

tient flow management decisions on the utilization of two main classes of hospital resources, "front-end" (such as operating rooms), and "backroom" (such as recovery beds). We introduce a new approach for modeling the patient recovery process and use it to characterize the relationship between a patient's length of stay and probability of readmission. On the basis of this modeling approach we develop a two-moment approximation for the utilization of front-end and backroom resources.

Front-end and backroom resources are typically managed by different teams of providers, with varying degrees of coordination between the two. In our analysis, we focus on assessing the benefits associated with the hospital employing a coordinated decision-making process, in which both portfolio and discharge decisions are made in tandem. Specifically, we compare the hospital's profits in the coordinated setting to those under two decentralized approaches: a "front-end" approach, under which both decisions are made based exclusively on the front-end costs, and a "siloed" approach, where discharge decisions are made based on backroom costs, and the case-mix is determined as the optimal match for the discharge policy. We show that hospitals operating under the front-end policy can significantly benefit from coordination when backroom costs are sufficiently high, even if they do not exceed surgical costs. On the other hand, for hospitals operating under the siloed policy, coordination brings significant benefits only when surgical costs are high and significantly dominate the cost structure.

Finally, in Chapter 3, we build upon the work in the first chapter to explicitly investigate the impact of coproduction on performance-based financial incentives in education. In particular, we formulate a model to characterize the impact of joint production on a school's probability of meeting state-specified proficiency standards using a Cobb-Douglas formulation. We include the effort decisions made by both teachers and students in the presence of performance-based incentives offered to both parties by a school district. We assume that these decisions are made in turn, as in a Stackelberg model, where teachers lead and students follow. We characterize students' and teachers' optimal effort levels for a given

reward allocation, and we illustrate the impact on student performance. Although teacher incentives and student incentives have been studied separately in the existing literature, there is limited work understanding the dynamics that occur when both types of incentives are offered simultaneously. Our work bridges this gap.

CHAPTER 1 : Too Much Information: When Does Additional Testing Benefit Schools?

1.1. Introduction

Performance-based contracts have long been identified as a way to incentivize workers when direct oversight is not possible. Yet the effectiveness of such contracts depends on far more than simply the level of incentives offered — in particular, workers must have access to the resources and information necessary to do their jobs, which often require additional monetary expenditures by the company. In this chapter, we study one system where employers combine monetary incentives with investments in additional information for employees: K-12 education in the United States. Specifically, we investigate the relationship between the frequent provision of information on student progress and end-of-the-year merit-based bonuses for teachers.

For decades, student assessments have been an integral source of information on the quality of education provided in the U.S. education system (Linn, 2000), but the form and extent of testing have varied considerably over time and across individual states. With the passage of the No Child Left Behind Act of 2001 (NCLB), a greater emphasis was placed on frequent testing based on a well-defined set of standards. In particular, states were required to ensure student "proficiency" on state tests in reading and math by 2013-14, and individual schools were required to show "adequate yearly progress" (AYP) for the overall student population as well as key subgroups, such as minority students and students in special education (United States General Accounting Office, 2003; Klein, 2015; Dillon and Rotherham, 2007).

State tests are just one example of the many different assessments educators use. A commonly-cited typology of assessments is that given by Perie et al. (2007), who define three categories of assessments used in the U.S. K-12 system: summative, formative, and interim. The annual state tests mandated by NCLB are an example of a *summative* assessment, which are given at the end of an instructional period to check student knowledge

against a broad set of content standards determined by an external entity. On the other hand, *formative* assessments are continuously used by educators to obtain ongoing feedback about how well students understand the material being taught. Finally, *interim* assessments lie between these two types of assessments in terms of both frequency and scope. These are used to evaluate students against a specific set of achievement goals, and the results guide both teaching within the classroom and decision-making more broadly at the school and district levels. Although NCLB only emphasizes summative assessments, it "enshrined the logic of data-driven decision-making in education" (Young and Kim, 2010), leading school districts to increasingly rely on costly third-party interim assessments as a way to evaluate students' and schools' progress towards meeting year-end proficiency goals. These assessments are widely considered to be more reliable indicators of student performance than formative assessments, despite a lack of research that supports this claim (Bulkley et al., 2010).

Additionally, in 2009, shortly after the passage of NCLB, the $4.35 billion Race to the Top Fund (RTTT) was launched with a focus on four core goals: updating educational standards; building data systems to measure student performance and inform educators; "recruit[ing], develop[ing], reward[ing], and retain[ing]" effective teachers and principals; and turning around low-performing schools (U.S. Department of Education, 2009). In part due to RTTT, school districts have been experimenting with pay-for-performance contracts for teachers designed to improve educational outcomes.

Thus, school districts have been both investing in a midyear interim assessment to get an accurate measure of student progress and exploring merit-based rewards. Anecdotal evidence suggests that overall more money is allocated towards ongoing assessments than towards additional teacher compensation. For example, the Teacher Incentive Fund, established by Congress in 2006 to provide grants to support performance-based teacher and principal compensation in high-needs schools, allocated $225 million for such awards in 2016. At the same time, total U.S. spending on classroom assessments in 2017-18 exceeded $1.63 billion

(Raugust et al., 2019), up from $434 million in 2001-02 (Cavanagh, 2015). Moreover, despite the vast sums of money being spent on these two approaches to improve educational outcomes, to the best of our knowledge, there are no studies of their comparative benefits and the interaction between them.

In this chapter, we seek to address this gap in the extant literature. We analyze the problem faced by a school district that wants to maximize the probability of students being "proficient" on the end-of-the-year standardized test by allocating a fixed budget between an "interim" assessment and a merit-based incentive for teachers. In our model, the district plays the role of a principal that can provide an agent (teachers) with financial incentives tied to the achieved educational outcome, and can, at a cost, improve the quality of the information set under which the agent operates. Since the educational process unfolds over a protracted period of time (e.g., a year) and the additional information on the state of student proficiency is provided by the interim assessment in the middle of this process, our model uses a two-period dynamic principal-agent framework. If the district invests in an interim assessment midway through the school year, both the district and the teachers will know the state of student proficiency at that time; otherwise, they will rely only on the less-accurate information from the low-cost formative assessments. In addition to the information provided by an interim assessment, the district may offer teachers a merit-based bonus, which they earn if a sufficient portion of their students show "proficiency" on the year-end standardized test. Teachers respond to both the merit-based incentive and the information they possess about midyear student progress by choosing a dynamic policy that defines their effort levels. In our model, we use a scalar as a simplified representation of the multiple levers a teacher can use to influence educational outcomes, such as spending extra time working with students or creating detailed lesson plans.

We seek to answer two key questions for a school district that is considering using both an interim assessment and merit-based rewards to improve school performance:

1. When should a school district invest in a mid-year interim assessment?

2. What level of merit-based reward should a school offer teachers?

Our results show that, for low-performing schools that are already behind, investing in interim assessments is usually not an effective strategy. Rather, the school district should invest solely in merit-based incentives. The exception to this is when there is a high baseline level of effort that increases the probability that the school will achieve proficiency even in the absence of additional effort induced by the merit-based incentive. For schools that start the year on track to achieve proficiency, there are a limited number of cases where investing in an interim assessment is valuable, which we outline below.

The rest of the chapter is organized as follows. In Section 1.2, we discuss the relevant literature. Our model is presented in Section 1.3, followed by its analysis in Section 1.4. Finally, in Section 1.5, we discuss our results and future avenues of research.

1.2. Literature Review

Our analysis draws on the principal-agent model literature in economics and operations management, as well as the literature on performance pay in K-12 education.

The role of information in principal-agent models has been studied extensively. In particular, there is much work that considers the optimal policy when the agent has, or can independently gain, private information about the production environment. For example, Baron and Myerson (1982) derive the optimal regulatory policy for a monopolistic firm with privately-known costs. Lewis and Sappington (1997) determine the optimal contract to incentivize the agent to acquire and reveal information. Crémer et al. (1998) study when it is optimal for the principal to induce the agent to gather additional information at a cost. On the other hand, multi-period dynamic models account only for a rather small, and a more recent, fraction of the vast principal-agent literature. Fudenberg et al. (1990) introduce stochastic elements to a dynamic principal-agent model and identify conditions under which a long-term contract can be implemented as a sequence of short-term contracts. Plambeck and Zenios (2000) provide an analysis of a previously intractable setting relying on assumptions about the "economic structure" of principal-agent interaction, and

Fuloria and Zenios (2001) build upon this dynamic model in the context of healthcare contracting. Shumsky and Pinker (2003) study the compensation system a firm should offer a gatekeeper who has private knowledge about the complexity of a customer's problem and their ability to treat it. Zhang and Zenios (2008) use a dynamic principal-agent model with hidden information, where the state is known to the agent but not to the principal. Chu and Sappington (2009) characterize the optimal contract when a principal and agent begin with symmetric information but the agent will ultimately acquire superior information.

Our approach follows the spirit of these models: in our model, the agent (teachers) solves a dynamic program to determine the optimal effort allocation policy. The dynamic nature of the agent's response in our model is dictated by the setting we describe: the information brought in by additional costly testing is revealed "in the middle" of a protracted instruction period, potentially altering the agent's decision-making process and, thus, requiring a "closed-loop" modeling approach. The principal's ability to invest in the enhancement of the information set used by the agent is a distinguishing, novel feature of our analysis.

More broadly, our work is also related to the supply chain literature on asymmetric cost information. For example, Corbett and De Groote (2000) study a supplier's optimal quantity discount policy when the buyer's cost is unknown. Ha (2001) analyzes a supplier-buyer relationship with asymmetric cost information under stochastic, price-sensitive demand. Lutze and Özer (2008) characterize a promised-lead time contract, which includes an optimal promised lead time and corresponding payments, that a supplier should offer a retailer who has private information about shortage costs. Additionally, our research relates to work that characterizes the relationship between worker effort and labor costs, e.g. Tan and Netessine (2014).

In our model, the overall effort level of the agent is affected by the monetary incentive offered by the principal and contingent on achieving a pre-specified performance level. In this regard, our work adds to a rich stream of papers focused on contracting and performance pay, also commonly called "merit pay," in K-12 education in the United States. On the

theoretical side, Murnane and Cohen (1986) use the microeconomics contracting framework to argue that merit-pay contracts may be difficult to implement in education settings. They argue that the very nature of the teaching process makes it difficult for supervisors to articulate why some teachers may receive merit pay but others do not, which can lead to dissatisfaction for teachers that do not receive the reward. Similarly, Johnson (1984) points out potential negative effects of teacher-level merit compensation such as harmful competition among teachers and low morale. The lack of precise guidelines that teachers can follow to earn the reward is another complicating factor for the use of merit-based incentives. The concern about rewarding only some teachers within a school can be ameliorated through the use of school-level, rather than teacher-level, incentives (Clotfelter and Ladd, 1996). Still, empirical work suggests that teacher-level incentives remain common, and in the presence of such incentives schools can mitigate the potential negative effects by making merit pay inconspicuous or awarding it to almost everyone (Murnane and Cohen, 1986). In our work, we assume that teachers within a school form a homogeneous group that can earn a school-level reward. Additionally, although teaching remains as much an art as a science, providing teachers with timely information about their progress toward achieving performance targets may alleviate teachers' uncertainty about the path to earning a merit-based reward. The growing availability of interim assessments has made it easier for districts to do just that, and, in our model, we explore the new dynamics brought in by these assessments. In more recent work, Barlevy and Neal (2012) propose a pay-for-percentile incentive scheme that overcomes some of the unintended consequences of existing pay-for-performance schemes, such as coaching and scale manipulation. In our work, we take it as given that schools will use a generic incentive scheme and consider how a district should allocate resources to maximize the probability of achieving a goal. We do not specify the precise type of incentive scheme, only the transition probabilities that determine the likelihood that teachers will earn the reward.

The empirical evidence on the effects of merit pay remains mixed. Eberts et al. (2002) find that teacher-level merit pay improved student retention but negatively impacted student

attendance and course passing rates, with student GPAs remaining unchanged. Figlio and Kenny (2007) use survey data from 390 schools to show that merit pay (defined as "at least one [teacher] ... reported having a merit pay bonus") is correlated with higher test scores. Springer et al. (2011) implemented a short-term, experimental teacher-level performance pay program in Metro Nashville Public Schools (MNPS) where teachers were eligible for up to $15,000 per year in bonuses based on student test-score gains. Although the program did not have a significant, lasting effect on student test scores, it did impact the way some teachers approached their jobs: while 80 percent of teachers believed that the program did not change their teaching practices, teachers in the "treatment group" were more likely to collaborate with other teachers and align their instruction with test preparation. Fryer (2013) studies an experimental school-level incentive program in over 200 high-needs New York City public schools, which was implemented as a randomized school-based trial from the 2007-08 school year through the 2009-10 school year. The author finds no evidence that financial incentives lead to improvements in student performance outcomes or in teacher or student behavior.

The results from longer-term merit-pay experiments are more promising, if modest. Dee and Wyckoff (2015) analyze IMPACT, the teacher evaluation reform introduced during the 2009-10 school year in the District of Columbia Public Schools. IMPACT offers strong financial incentives for highly effective teachers, where effectiveness is determined based on multiple components, such as classroom observation scores and students' performance on standardized tests. The researchers find that for highly effective teachers, the base pay incentives for scoring "highly effective" for another year were associated with a seven-percentile increase in teacher effectiveness. Unlike many earlier studies that focused specifically on short-term, experimental performance pay programs, IMPACT is a multi-pronged, long-term program. Chiang et al. (2017) evaluate the Teacher Incentive Fund (TIF), a program which was established by the U.S. Congress in 2006 and "which provides grants to support performance-based compensation systems for teachers and principals in high-need schools." Specifically, they evaluate ten districts in which the pay-for-performance component of TIF

was randomly assigned. The program was implemented over a four-year period, and, by the second year, it led to a slight increase in student achievement that held steady in the remaining years of the program. In this program, although most educators received a bonus, the actual bonus level was differentiated based on the performance of their students.

These empirical studies focus specifically on the impact of a performance-based incentive and do not consider the influence of midyear assessments on teaching practices and student performance. In our analysis, we focus on identifying the school districts that, in the presence of merit-based teacher compensation, may benefit from additional information brought in by interim assessments as well as the the school districts that are better off using formative assessments.

1.3. Model: Combining Assessments and Merit-Based Pay to Achieve Proficiency

In this section, we present a dynamic principal-agent model that captures the interaction between the school district (principal) and the group of teachers at a school (agent). In our model, the district explores the option of investing in additional information on the state of student performance and providing incentives to maintain or achieve standards of performance, and teachers respond to information and incentives by selecting a dynamic policy defining their effort levels.

1.3.1. Time Horizon, System States and Actions

Consider a discrete-time, two-period model, with time indices $t = 0, 1$ corresponding to the beginning and middle of the school year, respectively, and the index $t = 2$ corresponding to the end of the school year and indicating the time at which the proficiency of the student body at the school is measured via a state-administered standardized assessment. At time $t = 0, 1, 2$, the school proficiency is given by $\beta_t \in \{P, N\}$, which indicates whether that school is "proficient" (P) or "not proficient" (N). We define "proficient" to mean that a sufficient fraction of the school's students are either on track to satisfy state-imposed learning standards at $t = 0, 1$ or satisfy these standards at $t = 2$. (We provide a summary of our notation in Table 3 in the Appendix.)

In each period, teachers decide how much effort to allocate towards activities they believe will improve student performance. We use $e_t \geq 0$, $t = 0, 1$ to denote the teachers' effort level in the first and second semesters of the school year, respectively. Two features of our approach to modeling teachers' effort are important to underscore. First, in our model, we assume that the school's teachers are homogeneous and act as a group, with e_t reflecting teachers' joint effort. This assumption approximates a more complex reality where the effort levels of individual teachers will vary, with e_t reflecting the "average" school-level effort. We believe such a modeling simplification is justified since, in practice, the proficient/not-proficient designation is often applied to the entire school, as are the performance-based incentives. Modeling the "average" effort level thus allows us to focus on the "first-order" effect of the incentives. Second, we approximate the multidimensional nature of efforts that teachers make in reality by a single "aggregate" measure represented by a scalar. Although no single measure can be a perfect representation of teachers' efforts, this scalar can be a proxy, for example, for the extra time that teachers may spend working with students.

We use vector $\mathbf{e} = (e_0, e_1)$ to represent the effort level decisions for the two periods. The evolution of the student proficiency state in each period is influenced by the state in the beginning of the period and by the teachers' effort decision in that period. Figure 1 illustrates the state transition diagram for the discrete-time Markov chain in each period, where $\alpha(e)$ represents the effort-dependent probability of transitioning from "N" in the beginning to "P" in the end of the time period, and $\delta(e)$ represents the respective probability of transitioning from "P" to "N."

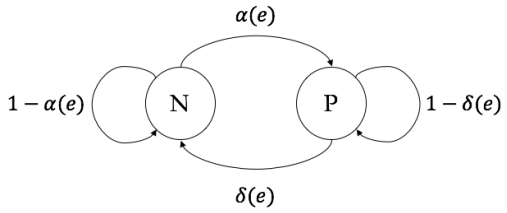

Figure 1: Transition probabilities between proficient ("P") and not proficient ("N") states during each time period as functions of effort level.

In modeling $\alpha(e)$ (the probability of moving to a proficient state) and $1-\delta(e)$ (the probability of remaining in a proficient state), we use the following functional form, which reflects the standard assumptions of monotonicity and non-increasing return-on-effort:

Assumption 1 *a)*

$$\alpha(e) = A(N)e + B(N), \tag{1.1}$$

$$1 - \delta(e) = A(P)e + B(P), \tag{1.2}$$

where $0 \leq e \leq 1$, $0 \leq B(N) < B(P) \leq 1$, $0 \leq A(N) < A(P) \leq 1$, *and* $A(P) + B(P) \leq 1$.

b)

$$\frac{A(N)}{A(P)} = \frac{B(N)}{B(P)} = \mu. \tag{1.3}$$

Under Assumption 1a, probabilities $\alpha(e)$ and $1 - \delta(e)$ are monotonic and concave in e. We normalize to 1 the maximum effort level producing an increase in the probability of being in the proficient state at the end of period. Our model treats e as "additional" effort that can be elicited through merit-based incentives, resulting in enhanced probability of reaching the proficient state. Accordingly, we allow for a "base" effort that can result in a non-zero probability of reaching proficiency even when $e = 0$; this baseline transition probability is captured by $B(N)$ and $B(P)$. In particular, $B(P)$ characterizes the probability that the school remains in the proficient state in the absence of teacher effort, i.e. the "stickiness" of the proficient state.

The maximum probability values $A(N) + B(N)$ and $A(P) + B(P)$ reflect the *co-produced* nature of teaching, where the outcome depends both on the teachers' and students' efforts. Thus, teachers' efforts alone may not guarantee that the proficient state is reached if these probability values are less than 1. The condition $A(N) + B(N) \leq A(P) + B(P)$ implies that it is more difficult to attain proficiency than to maintain it. The literature supports

this assumption: Davison et al. (2004) suggest that groups of students often have difficulty overcoming even small achievement gaps, while Neal and Schanzenbach (2010) argue that the incentives for teachers in many school-accountability systems inevitably lead to students at the lowest end of the achievement distribution getting "left behind." For tractability, we further assume that the ratios of $A(N)$ to $A(P)$ and of $B(N)$ to $B(P)$ are equal to the constant term μ (Assumption 1b). This term captures the extent to which the school can recover when they fall behind, i.e. the "resilience" of the student population.

Furthermore, we assume that the transition dynamic described by (1.1)-(1.2) is stationary and does not depend on the time period. This stationarity assumption is reasonable given that the time periods we consider correspond to several months and that the progress in the previous time period is captured by the baseline transition probability parameters, $B(N)$ and $B(P)$. Finally, for tractability we assume that the marginal impact of effort when the system is in the not-proficient state is less than that when the system is in the proficient state.

1.3.2. Interim and Formative Assessments: Cost and Information Structure

We assume that the teachers' choice of effort levels cannot be directly observed by the school district. Furthermore, the initial state β_0 is known to both the teachers and the district, and the final state β_2 will be made known to both parties after the end-of-year standardized assessment. However, both the teachers and the district may have imperfect knowledge of the intermediate state of the system β_1: students are assessed at the midyear point to determine whether the school is in the proficient or not-proficient state, but this assessment may be inaccurate. The degree of accuracy depends on the type of assessment used. In particular, the district chooses between two options: to administer an *interim* assessment or to rely exclusively on the *formative* assessments already being administered by teachers. The *interim* assessment has a fixed cost F that the district must incur and perfectly reveals β_1, the state of the system at $t = 1$, whereas *formative* assessments do not incur any additional cost for the district but are less accurate than the interim assessment. For either choice, both the teachers and the school district will know the results of the

assessment at $t = 1$.

We use $X_1 \in \{P, N\}$ to denote the proficiency level indicated by the formative assessments. That is, $X_1 = P$ ($X_1 = N$) if the formative assessments indicate that the school is in the proficient (not-proficient) state at $t = 1$. The probability that the formative assessments result X_1 takes a particular value given the true intermediate state β_1 is captured by the parameters $\phi_{P|P}$ and $\phi_{P|N}$, where

$$Pr\left[X_1 = P|\beta_1 = P\right] = \phi_{P|P}, \tag{1.4}$$

$$Pr\left[X_1 = P|\beta_1 = N\right] = \phi_{P|N}. \tag{1.5}$$

These parameters represent the true positive and false positive rates, respectively. We assume that the formative assessments can never perfectly assess the midyear state, i.e. it is never the case that $\phi_{P|P} = 1$ and $\phi_{P|N} = 0$, and that the false positive rate never exceeds the true positive rate:

Assumption 2 $1 > \phi_{P|P} \geq \phi_{P|N} > 0$.

1.3.3. District Decisions and the Timeline of Events

For the analysis of the district's decision problem, we introduce the following notation. At $t = 0$, the district chooses whether to administer an interim assessment in the middle of the year ($t = 1$) and offers teachers a compensation contract that includes a base-pay component that we normalize to zero and a merit (performance-based) pay component:

$$w = \begin{cases} \pi, & \text{if } \beta_2 = P, \\ 0, & \text{if } \beta_2 = N, \end{cases} \tag{1.6}$$

where $\pi > 0$ is the level of merit-based incentive. At the beginning of the school year, teachers know whether the district has chosen to invest in an interim assessment and the terms of the contract, and they receive compensation after the end-of-the-year assessment. The payment of the reward at a single point in time is consistent with current practice in

the field of education. For example, Fryer (2013) describes an incentive scheme in New York City Public Schools in which teachers were given a reward based on annual performance targets. Chiang et al. (2017) study the implementation of the Teacher Incentive Fund (TIF) in ten school districts. They state that in seven out of ten of districts in the study, teachers received their one-time reward during the subsequent school year.

The timeline of events is illustrated in Figure 2. At the beginning of the school year (at $t = 0$), β_0 is known to both the teachers and the district. Based on this information, the district decides whether to implement an interim assessment (z_I) and the merit pay level (π). The teachers respond by determining the policy they will use in selecting their effort levels at the beginning and middle of the school year ($t = 0$ and $t = 1$, respectively). Given the initial state of the system and the effort level teachers select at $t = 0$ (e_0), the system transitions to state β_1. Then, depending on the school district's choice of assessment, the school either conducts an interim assessment or relies on the formative assessment results at the end of the first half of the school year. If the district invests in an interim assessment, both the teachers and the district will know the midyear proficiency state β_1 (Figure 2a). If the district relies on formative assessments, the teachers and district will use the result X_1 to estimate the probability that the proficiency state β_1 is proficient (Figure 2b). Based on the assessment results, teachers select the effort level (e_1) to be applied in the second half of the school year (at $t = 1$). At the end of the school year (at $t = 2$), a standardized assessment is administered and the proficiency state β_2 is revealed to both the teachers and the school district. The teachers are then paid according to the compensation contract (1.6).

We assume that both the district and the teachers are risk-neutral. Below we describe the problems faced by the teachers (agent) and the district (principal).

1.3.4. Teachers' Problem: Dynamic Response to District's Decisions

Given whether the district invests in an interim assessment and the merit-based contract that the district proposes, teachers choose the effort levels that maximize their expected

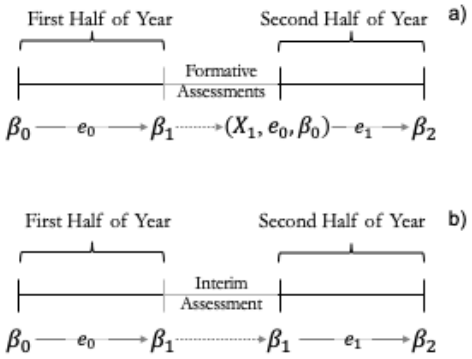

Figure 2: Timeline of events: proficiency states (β_0, β_1, and β_2), the outcome of the formative assessment (X_1), and teachers' actions (e_0 and e_1) when a) the district chooses interim assessment ($z_I = 1$) and when b) the district relies exclusively on formative assessment ($z_I = 0$).

merit-based compensation net of the cost of effort they incur. It follows that, in the absence of a merit-based incentive, teachers will not exert any additional effort, regardless of the district's assessment decision. In modeling the teachers' cost, we use a simple linear functional form representing stationary and constant marginal cost-of-effort.

Assumption 3 *The teachers' cost-of-effort at time* $t = 0, 1$ *is given by* $c(e_t) = \gamma e_t$, *with* $\gamma > 0$.

Note that the teachers' decision problem is represented by a two-period dynamic program, where their effort decision in the second half of the school year depends on the information they receive after the first half of the school year, and the decision in the first half of the year is influenced by the policy they adopt for the second half. In order to provide a formal description of the teachers' problem, we define

$$\mathbf{S}_0 = \beta_0, \tag{1.7}$$

$$\mathbf{S}_1 = \begin{cases} (X_1, e_0, \mathbf{S}_0), & \text{if } z_I = 0, \\ \beta_1, & \text{if } z_I = 1, \end{cases} \tag{1.8}$$

and

$$\mathbf{S}_2 = \beta_2, \tag{1.9}$$

to describe the states of the system at $t = 0, 1$, and 2, respectively. The teachers will use the states at $t = 0$ and $t = 1$ to make their effort decisions at $t = 0$ and $t = 1$, respectively. Note that the state of the system at $t = 1$ has different "content" depending on whether or not an interim assessment is used. In particular, if teachers' information about the proficiency at $t = 1$ is imprecise ($z_I = 0$), they must use both the initial state $\mathbf{S}_0$ as well as their action taken at $t = 0$ (e_0) to calculate the expected net earnings stemming from their action at $t = 1$ (e_1), as we will show below. For each combination of the district's decisions (π, z_I), we can use the notation in (1.7)-(1.9) to express the dynamic program that teachers solve as

$$J_t(\mathbf{S}_t) = \max_{e_t \geq 0} \left[E\left[J_{t+1}\left(h_{t+1}\left(e_t, \mathbf{S}_t \right) \right) \right] - \gamma e_t \right], \quad t = 0, 1, \tag{1.10}$$

where the expectation is taken over the random state of the system $h_{t+1}\left(e_t, \mathbf{S}_t\right)$ at time $t + 1$ and

$$J_2(\mathbf{S}_2) = \begin{cases} \pi, & \text{if } \mathbf{S}_2 = P, \\ 0, & \text{if } \mathbf{S}_2 = N. \end{cases} \tag{1.11}$$

For convenience, we summarize the description of the state $h_{t+1}\left(e_t, \mathbf{S}_t\right)$, for each state-action combination $(e_t, \mathbf{S}_t)$ in Lemma 3 in the Appendix.

To emphasize the connection between the district's decision and the teachers' response, we will use $\left(e_0^*\left(\mathbf{S}_0, \pi, z_I\right), e_1^*\left(\mathbf{S}_1, \pi, z_I\right)\right)$ to denote the optimal effort policy, i.e. the policy that solves the dynamic program (1.10)-(1.11) for a given set of district decisions (π, z_I).

1.3.5. District's Problem: Choosing the Optimal Assessment-Incentive Combination

For each school, the district wants to select the assessment type and matching merit pay compensation to incentivize teachers to choose effort levels that will maximize the school's probability of being in the proficient state when the standardized test is administered. The total amount of investment in the information provided by the interim assessment and

the incentive payments is limited by the school-level budget M. Because each district is likely to manage a number of schools, we assume that it is acceptable for payments to a particular school to exceed the allocated budget, as long as the budget constraint is satisfied in expectation. In order to formulate the district's decision problem, we use, at the slight abuse of notation, $Pr^*[\mathbf{S}_2 = P|\mathbf{S}_0]$ to denote, for fixed π and z_I, the probability that the school is in the proficient state at $t = 2$ under the optimal-response teachers' policy $(e_0^*(\mathbf{S}_0, \pi, z_I), e_1^*(\mathbf{S}_1, \pi, z_I))$, given that the school starts in the state $\mathbf{S}_0$. Then, for a given initial performance state $\mathbf{S}_0$, the district's decision can be expressed as the following optimization problem:

$$\max_{\pi \geq 0, z_I \in \{0,1\}} Pr^*[\mathbf{S}_2 = P|\mathbf{S}_0] \tag{1.12}$$

$$\text{s.t. } \pi Pr^*[\mathbf{S}_2 = P|\mathbf{S}_0] + Fz_I \leq M. \tag{1.13}$$

In summary, (1.10)-(1.11) and (1.12)-(1.13) describe a principal-agent problem where a principal selects the combination of information set and incentives for the agent and the agent's response is represented by a dynamic programming policy.

Below we provide an analysis of this principal-agent problem. First, we describe the optimal merit pay selection and the optimal teachers' response policy under a given assessment decision (Section 1.4.1). We then analyze the problem of the optimal selection of the assessment type (Section 1.4.2).

1.4. Analysis

We begin this section by determining the teachers' optimal effort decision in each semester for any given testing and merit-based incentive scheme. Using these results, we determine the optimal level of merit-based incentive under both assessment decisions and, as a consequence, the districts' optimal assessment decision.

1.4.1. Optimal Teachers' Effort Policy in the Presence of Merit-Based Incentive

The teachers' problem is a two-period dynamic program (1.10)-(1.11). We first characterize, for given π, the teachers' effort decision for the second half of the school year (i.e., at $t = 1$). This effort decision depends on the magnitude of the ratio of the merit-based incentive π to the cost of effort γ. We call this the *scaled incentive* $\hat{\pi}$, where

$$\hat{\pi} = \frac{\pi}{\gamma}. \tag{1.14}$$

Proposition 1 describes the optimal effort level at $t = 1$.

Proposition 1 (Optimal Teachers' Effort in Second Half of School Year) *a) Teachers will exert effort if and only if the scaled incentive is above a certain threshold value. In that case, they will exert maximum effort.*

b) The effort-inducing incentive threshold is decreasing in the level of proficiency observed in the midyear assessment, in the level of student resilience μ, and in the teacher effort levels in the first half of the school year.

The expanded version of the Proposition is presented in the Appendix.

First, as expected, the optimal effort level is an increasing function of the scaled incentive. Second, it is cheaper to incentivize high effort levels when the midyear assessment result is proficient than when the result is not proficient, holding all else constant. This is driven by the greater difficulty of achieving proficiency faced by teachers at a school in the not-proficient state. Thus, in schools that fall behind, teachers will only be incentivized to exert effort if the school district had chosen to offer higher rewards. In practice, this need for extra incentives is further compounded by the fact that more effective teachers tend to be distributed towards more advantaged schools (Clotfelter et al., 2006). Of course, this depends on the degree to which a proficient school differs from a not-proficient school. High levels of student resilience mitigate the difficulty faced by teachers when the school is in the

not-proficient state and thus moderate the optimal level of incentive that the school district should offer.

When teachers do not know the true intermediate state ($z_I = 0$), they rely on both the results of the formative assessments and the first-semester teacher effort level to estimate the probability that the school is in the proficient state and select their second-semester effort level. This reflects the reality that teachers often use multiple inputs throughout the course of the school year to gauge student progress, weighting those inputs based on their experience. Thus, teachers take into account that if they exert high effort levels in the first half of the year, there is a greater likelihood that the school will be in the proficient state by midyear, and therefore a lower incentive is necessary to motivate effort in the second half of the year

We next consider the teachers' effort level decision at $t = 0$.

Proposition 2 (Optimal Teachers' Effort in First Half of School Year) *Suppose that a school's formative assessments are reasonably accurate. Then, the following results hold.*

a) The optimal teachers' effort level in the first half of the school year follows a threshold policy depending on the scaled incentive, where teachers will exert maximum effort if the scaled incentive is above a certain threshold value and zero effort otherwise.

b) If the school starts the year in the not-proficient state, then the effort-inducing threshold is decreasing in the response-to-effort parameter $A(P)$ and stickiness of the proficient state $B(P)$.

c) If the school starts the year in the proficient state, for sufficiently high levels of student resilience, the effort-inducing threshold is increasing in student resilience μ.

The expanded version of the Proposition is presented in the Appendix.

As with teachers' effort levels in the second half of the year, optimal effort in the first half

of the year is positive if and only if the incentive offered by the school district exceeds a threshold. For schools that begin the year in the not-proficient state, this threshold does not depend on the accuracy of the midyear assessment. Due to the high level of incentive necessary to induce positive teacher effort at the beginning of the school year in such schools, the magnitude of this incentive level always exceeds the incentive level necessary to induce positive effort in the second semester, regardless of the midyear assessment accuracy or result. Thus, the quality of midyear information is no longer a factor, and the optimal effort decision is driven entirely by the initial state and teachers' effort levels in the first-semester. While this result may seem extreme, it reflects the magnitude of the challenge faced by teachers in low-performing schools. This is particularly problematic if absolute measures of achievement are used, in which case success can be nearly impossible to attain for students and schools that begin at a disadvantage. In practice, school districts and departments of education have attempted to account for this in various ways, such as achievement measures benchmarked against the previous year's performance or against comparable students and schools.

The costliness of incentivizing positive effort in such schools is mitigated by the efficacy of teachers' effort and the stickiness of the proficient state, since these increase the probability of the school ultimately ending the year in the proficient state. On the other hand, student resilience may have a non-monotonic effect on the scaled incentive threshold. For schools that begin the year in the not-proficient state, the threshold is initially decreasing and then increasing in student resilience levels. The cost of incentivizing positive effort is lowest for moderate levels of student resilience: in this case, the consequence of remaining in the not-proficient state after the first semester is significant but so is the probability of moving to the proficient state.

For schools that begin the year in the proficient state, the effort-inducing threshold depends on the accuracy of the midyear assessment, as well as the response-to-effort parameters and the level of student resilience. In this case, for sufficiently high levels of student resilience,

the cost of incentivizing positive teacher effort is increasing in resilience. This is because, as the implicit penalty from falling into the not-proficient state midyear decreases, teachers become more inclined to wait until the second half of the year to exert high effort levels. Under only formative assessments, the inexact nature of the midyear assessment results leads to non-monotonicity in this relationship. However, under an interim assessment, the effort-inducing threshold is always increasing in student resilience levels.

1.4.2. District's Merit-Based Incentive and Assessment Decisions

Using the characterization of the optimal teachers' response policy in the presence of merit-based incentives and information about midyear student performance, we analyze the district's decision on the optimal incentive level. In particular, we solve the district's optimization problem, (1.12)-(1.13), where we hold the interim assessment decision z_I fixed and maximize the probability that the system will be in the proficient state at the end of the school year under the optimal teachers' response to the merit-based incentive π.

It is straightforward to show that the district's estimate of the probability of achieving proficiency in the final state is a non-decreasing function of the merit-based incentive (see Proposition 14 and Lemma 15 in the Appendix), a property that facilitates the search for the optimal level of the merit-based incentive. Additionally, recall that M represents the school's budget and F represents the cost of the interim assessment. Then, for the analysis below, we use

$$\widehat{M} - z_I \widehat{F} = \frac{M - z_I F}{\gamma} \tag{1.15}$$

to represent the school the district's *scaled available budget*. Proposition 3 describes the optimal choice of the scaled incentive $\frac{\pi^*}{\gamma}$ as a function of the district's scaled available budget for the case where there is zero stickiness of the proficient state ($B(P) = 0$).

Proposition 3 (Optimal Scaled Incentive when $B(P) = 0$) *Suppose that there is no stickiness in the proficient state. Then, it is optimal for the school district to offer a scaled incentive only above certain levels of the scaled budget, and the optimal incentive level is*

monotone increasing in the scaled available budget. Moreover, when the scaled available budget is large, the optimal incentive level is always higher if the school starts the year in the not-proficient state.

The expanded version of the Proposition is presented in the Appendix. Furthermore, Proposition 16 in the Appendix characterizes the optimal choice of the scaled incentive as a function of the district's scaled available budget in the case where the proficient state has non-zero stickiness and the formative assessments are sufficiently accurate.

As expected, the optimal reward is non-decreasing in the scaled budget. Furthermore, the formative assessment accuracy parameters play a role in determining the optimal incentive level only when the school starts the year in the proficient state and student resilience is sufficiently low. This follows from the results in Proposition 2.

Using these results, we now turn to the district's optimal interim assessment decision for a particular school, where the district's optimization problem is given in (1.12)-(1.13). For a given budget M and cost of interim assessment F, the district must decide whether to invest in an interim assessment. Recall that teachers use formative assessments throughout the school year, so the district must determine whether it is advantageous to supplement the information gathered from the formative assessments with an interim assessment. These are widely considered to be more accurate measures of students' proficiency levels, since they are closely modeled after end-of-the-year state exams. We assume that the district will only invest in an interim assessment when the probability that the school is proficient at the end of the year is strictly greater under an interim assessment than when the district relies only on formative assessments.

As one might expect, there are two trivial budget levels for which the interim assessment decision does not change the probability that the school is in the proficient state at the end of the year. First, when the budget is sufficiently small, the school district can never afford to offer a reward that is high enough to induce teachers to exert any additional

effort. Therefore, the probability of achieving proficiency in the final period is smallest in this region. Conversely, when the budget is sufficiently large and the cost of an interim assessment is sufficiently small, the school district can offer a reward that is high enough to incentivize maximum effort levels throughout the school year, regardless of the school's midyear performance. It follows that the school district can maximize the probability of the school achieving proficiency no matter the interim assessment decision.

We characterize the district's optimal decision in the case where each school is allocated a non-trivial budget in the following Propositions. In Proposition 4, we describe the district's optimal decision in the case the school starts the year in the not-proficient state for two cases: when there is zero stickiness of the proficient state ($B(P) = 0$) and when there is non-zero stickiness paired with a reasonably accurate formative assessment. We consider the setting when the school starts the year in the proficient state in the subsequent Proposition.

Proposition 4 (Optimal Assessment Decision when $\mathbf{S}_0 = N$) *Suppose that the school starts the year in the not-proficient state. Then, the following results hold.*

a) If there is no stickiness in the proficient state ($B(P) = 0$), then it is optimal for the school district to forego investing in an interim assessment, regardless of the accuracy of the formative assessments.

b) If there is stickiness in the proficient state ($B(P) > 0$) and the school's formative assessments are reasonably accurate, and if interim assessment costs are sufficiently low, then the optimality of investing in an interim assessment depends on student resilience μ. For sufficiently high levels of student resilience, investing in an interim assessment is optimal for moderate levels of the scaled available budget. For moderate levels of student resilience, investing in an interim assessment is optimal for low-to-moderate scaled available budget levels, and for sufficiently low levels of student resilience, the interim assessment is an optimal investment only for low levels of the scaled available budget.

The expanded version of the Proposition is presented in the Appendix.

Our results suggest that, contrary to the prevailing wisdom, for schools that start the year in the not-proficient state, an interim assessment is not necessarily a worthwhile investment. This outcome is strongest in the case where there is no possibility of moving to the proficient state without additional effort (i.e., when $B(P) = 0$). In this case, it is always more costly to incentivize effort in the first half of the year than the second, and if the merit-based incentive is not high enough to incentivize first-semester effort, then the school will remain "behind" with certainty. Thus, the midyear interim assessment either does not provide any additional information about the state of student proficiency or the information it provides is not relevant to teachers' effort decisions. Thus, the school district should choose to forego investing in an interim assessment.

Next, consider the case where there is a non-zero probability of moving to the proficient state in the absence of additional effort (i.e., $B(P) \geq 0$) and the formative assessment is reasonably accurate. Just as in the previous case, it is more costly to incentivize effort in the first semester than it is in the second semester, and, as one would expect, it is always cheaper to incentivize second-semester effort if the mid-year assessment result is proficient (under either assessment decision) than if it is not proficient. Effort is most expensive to induce under a not-proficient midyear assessment result from an interim assessment. In this case, since teachers recognize that the negative feedback is accurate, their expectation of earning the merit-based incentive is lowest. That is, more-accurate unfavorable information can have a demotivating effect that requires additional compensation to overcome.

Note that when determining the optimal assessment decision, the district must consider both the probability the school achieves proficiency at the end of the year and the total expected cost under a given assessment decision. Based on these factors, the interim assessment is a valuable investment in two cases. The first case occurs when the school has a small budget and the total expected cost of assessments and rewards is lowest (and only affordable) when incentivizing effort after a proficient midyear result from an interim assessment. This holds as long as student resilience levels are sufficiently small. In this case, the difficulty

of transitioning from the not-proficient state is high, which drives the need for a higher reward in order to incentivize effort after a proficient midyear assessment result under the less-accurate formative assessments.

The second case occurs when the district has a slightly larger budget and can afford the expected costs needed to incentivize positive effort after a proficient midyear result under either assessment decision (but not after a not-proficient midyear result) and the probability of achieving proficiency at the end of the year is higher under an interim decision. This occurs for sufficiently large levels of student resilience. (Note, however, that for small values of $B(P)$, the necessary threshold may exceed feasible values of student resilience levels.) In this case, there is a higher probability of moving to the proficient state by the middle of the year from teachers' baseline effort; then, teachers' positive efforts after a proficient midyear assessment result under the interim assessment are more beneficial. Note that for moderate levels of student resilience, both cases hold. Furthermore, once the budget level is high enough that the school district can afford to incentivize effort after a not-proficient midyear, it is never optimal to invest in the interim assessment.

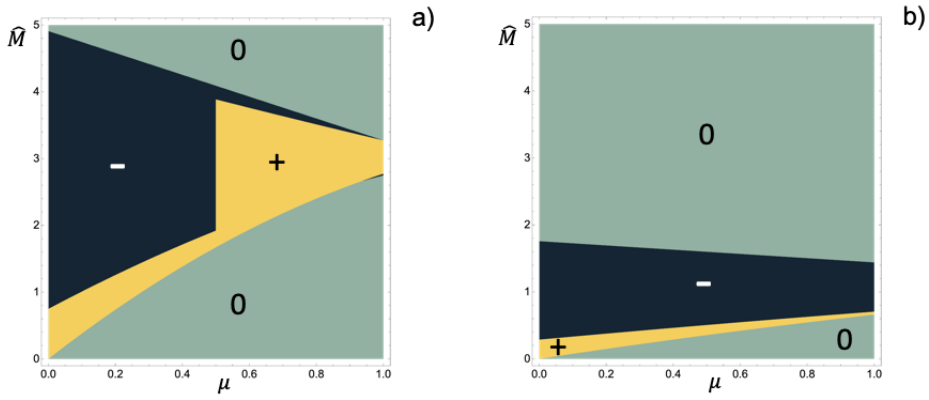

Figure 3: Regions showing where an interim assessment has positive (+), negative (−), or zero value by student resilience μ and scaled budget $\widehat{M}$ when a) $A(P) = 0.22$ and $B(P) = 0.5$ and b) $A(P) = 0.5$ and $B(P) = 0.22$ ($\mathbf{S}_0 = P$, $\phi_{P|N} = 0.1$, $\phi_{P|P} = 0.85$, $F = 0$).

We illustrate this in Figure 3, in which we identify three regions that capture the value of interim assessments. In the darkest shaded region, interim assessments have a negative

value, and relying solely on formative assessments is the unique optimal decision. In the lightest shaded region, interim assessments have a positive value, and investing in an interim assessment is the unique optimal decision. Finally, in the medium shaded region, the district's testing decision does not affect the school's end-of-the-year proficiency level. This region includes the trivial settings of high and low budgets that we describe above. The chosen parameters are based on an estimation of the model parameters using data from the District of Columbia Public Schools, which is described in Section A.2 in the Appendix. Note that the Figure illustrates the best-case scenario for the value of information, since we assume that the cost of an interim assessment is zero.

In Figure 3a, the interim assessment is optimal only for smaller values of student resilience μ. This region becomes larger when μ is sufficiently large, so that the probability of achieving proficiency is higher under the formative assessment. There is also a small region of the highest student resilience levels for which the interim assessment is not optimal for lower budget levels, since the budget cannot support the higher probability of disbursing the merit-based incentive. Comparing Figure 3a to Figure 3b illustrates the sensitivity of the regions of μ to the other parameters. In Figure 3b, the effect of the lower marginal impact of effort in this case is that student resilience levels are never high enough to result in the optimality of the interim assessment for moderate budget levels.

In Proposition 5, we consider the case where school starts the year in the proficient state and the proficient state has zero stickiness ($B(P) = 0$).

Proposition 5 (Optimal Assessment Decision when $\mathbf{S}_0 = P$ and $B(P) = 0$) *Suppose the school starts the year in the proficient state and that there is no stickiness in the proficient state. Then, the following results hold.*

a) If students have sufficiently high levels of resilience, then it is optimal for the school district to forego investing in an interim assessment, regardless of the accuracy of the formative assessments.

b) If there are sufficiently low levels of student resilience, investing in an interim assessment is optimal for moderate levels of the budget and sufficiently low interim assessment costs.

The expanded version of the Proposition is presented in the Appendix.

In the case where the school begins the year on track to achieve proficiency on the state-mandated summative assessments, we again find that interim assessments are beneficial in only a limited number of scenarios. Moreover, even in settings where they do lead to improved proficiency levels, their optimality is fragile: a slight increase in the budget level results in the unique optimal assessment decision being to not invest in an interim assessment.

First, when students have high levels of resilience, there is minimal benefit to teachers from exerting effort in the first half of the school year. Therefore, just as in the case where the school starts the year in the not-proficient state and there no stickiness of the proficient state, the cost of incentivizing first-period effort is always greater than the cost of incentivizing second-period effort. Thus, if the reward is high enough to incentivize first-semester effort, teachers will also exert in the second half of the year, regardless of the state of student proficiency. If, on the other hand, the reward does not incentivize effort in the first half of the year, then teachers know with certainty that the state of student proficiency has deteriorated to the not-proficient state by midyear, regardless of the assessment result.

On the other hand, when there are low levels of student resilience, it is never possible to induce second-period effort, even in limited scenarios, without also inducing first-period effort. Just as in the case when the school begins the year in the not-proficient state, the consideration of both the probability that the school achieves proficiency on the end-of-the-year state assessments and the total expected cost of either assessment decision ultimately determine the optimality of investing in an interim assessment.

We illustrate this in Figure 4 below. As in the previous Figure, we identify three regions that capture the value of interim assessments and illustrate the best-case scenario for the

value of information.

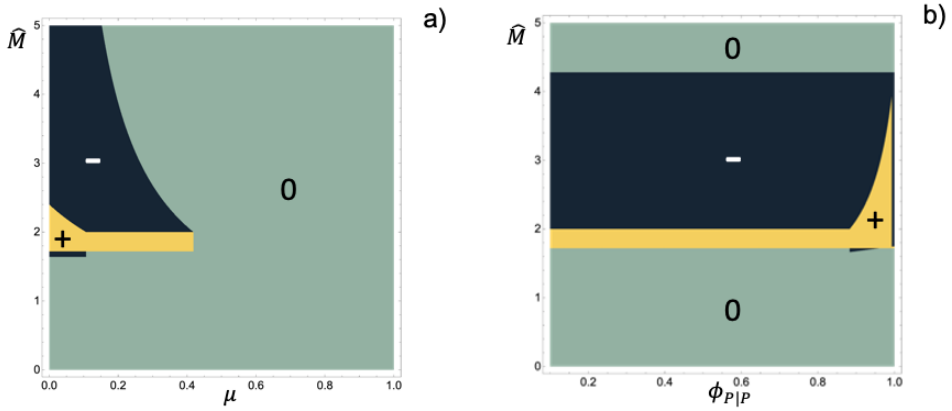

Figure 4: Regions showing where an interim assessment has positive (+), negative (–), or zero value by scaled budget $\widehat{M}$ and a) student resilience μ when $\phi_{P|P} = 0.85$ and b) accuracy parameter $\phi_{P|P}$ when $\mu = 0.18$ ($\mathbf{S}_0 = P$, $A(P) = 0.72$, $\phi_{P|N} = 0.1$, $F = 0$).

Figure 4a shows the value of information as function of student resilience levels and the scaled budget. As described above, for large levels of student resilience, the accuracy of the midyear result does not impact teachers' effort decision; hence, the interim assessment has neither a positive nor negative value in this region. For small levels of student resilience, the region of budget levels for which an interim assessment is valuable is bounded below by a region in which an interim assessment has negative value. In this case, although the optimal reward under an interim assessment is lower than that under only the formative assessments, the higher probability of achieving end-of-the-year proficiency under an interim assessment makes that optimal incentive unaffordable for lower budget levels. Thus, the probability of achieving proficiency at the end of the year is actually lower under the interim assessment. For slightly higher budget levels, the budget can support rewards that incentivize effort after a positive midyear result under either assessment decision, and in this case, the school district should invest in an interim assessment. However, once the budget is sufficiently high, the school district can first afford to incentivize second-semester effort under formative assessments, after either a proficient or not-proficient midyear result. This is because a not-proficient result under the formative assessment preserves the possibility that students are actually in the proficient state. Consequently teachers' expectations that they will receive

the merit-based incentive at the end of the year remain somewhat high, which makes them easier to incentivize.

In Figure 4b, we further illustrate the value of information as a function of the accuracy of the formative assessments result. Consider the scaled budget levels for which the interim assessment is optimal for true-positive rates less than 0.88. In this region, under an interim assessment, the district can choose a reward that incentivizes teachers to always exert effort in the first half of the year and only exert effort in the second half of the year given a positive midyear interim assessment result. Here, paradoxically, relying only on the formative assessments is "too expensive": because the cost of inducing second-semester-effort even after a negative assessment result is less than the cost of incentivizing effort in the first half of the year, offering any non-trivial reward will result in teachers exerting effort throughout the year, and the expected cost for the district will exceed their budget.

When the formative assessment is sufficiently accurate, then the district can incentivize "selective" effort under both the formative and interim assessments. (The effort is selective in the sense that teachers will not exert effort in the second-half of the year after a negative assessment result.) As in Figure 4a,it is more expensive to incentivize second-semester effort under the positive formative assessment result than under a positive interim assessment result, while at the same time, the expected probability of achieving proficiency at the end-of-the-year is higher under the positive interim assessment result. Again, this leads to a small region of low budget levels for which the district relies on the formative assessment only because the expected cost of merit-based incentives under the interim assessment exceeds the budget. For slightly higher budgets, the district can afford the higher probability of achieving proficiency that comes from investing in an interim assessment, and so that becomes the optimal decision. If the budget is sufficiently large, then the district is better off relying solely on the formative assessments, because it can afford to incentivize maximal teacher effort throughout the year, regardless of the midyear result.

Finally, notice that there is a small region where the true positive rate is close to 1 where

it is optimal to forego the interim assessment. In these regions, even if teachers are induced to selectively exert effort, the school has a higher probability of being in the proficient state at the end of the year if it relies only on formative assessments. Here, teachers will almost always exert effort when the true state of midyear student performance is proficient, and they will also exert effort in a subset of cases when the true midyear state is not proficient. Thus, if the district invested in an interim assessment in this case, the additional benefit from always exerting effort when the true state is proficient would be outweighed by the lost potential to recover if the students have fallen to the not-proficient state.

1.5. Discussion

Our research is driven by a desire to understand the strong hold that testing — and, in particular, private testing companies — has taken on public schools in the United States. The assessment industry has experienced exponential growth over the last two decades, with a significant portion of that market arising from the demand for midyear classroom assessments. School districts have invested in such assessments in the hope of improving student learning and, ultimately, performance on end-of-the-year state exams that serve as a key oversight mechanism of schools and school districts. Yet, the efficacy of such a strategy is unclear, particularly when implemented in conjunction with teacher incentive programs. In this chapter, we construct a stylized model to gain insight into the benefits and limitations of these policies.

We find that, in many settings, this growing dependence on such tests is misplaced. Notably, we find that for low-performing schools, i.e. schools that begin the year "behind," perfect accuracy is overrated, and the school district rarely benefits from investing in the interim assessment. In particular, when the probability of students moving to the proficient state is low under baseline levels of effort, the interim assessment is unlikely to be a worthwhile investment. In this case, the challenge of achieving proficiency at the end of the year is great, and offering a commensurately high incentive is the most effective way to incentivize teachers to exert additional effort during the course of the school year. When the probability of transitioning to the proficient state is higher under baseline effort levels, then the interim

assessment is more likely to be a valuable investment, particularly for moderate levels of student resilience.

Similarly, for schools that begin the year on track to achieve proficiency, there are only a limited number of scenarios in which the accurate information from an interim assessment is of value — and, strikingly, there are also a number of scenarios in which this information proves detrimental to a school's performance. To identify which of these scenarios will hold, a school district should consider both the level of student resilience and the level of accuracy of the existing formative assessments used by teachers. For high levels of student resilience, there is little difference between the school being in the not-proficient state and the proficient state. Hence, the accuracy of midyear assessment results does not affect teachers' effort decisions.

For lower levels of student resilience, midyear assessments play a more important role. Specifically, in the absence of clear information, teachers are inclined to believe that the state will remain proficient if they exert effort in the first half of the academic year. This, in turn, makes it easier to motivate teachers using merit-based incentives. In this case, the district is better off foregoing investment in an interim assessment, as long as they can afford the higher probability of disbursing the merit-based incentives. Relying only on the formative assessments enables them to incentivize maximal effort throughout the year for a lower level of merit-based incentive.

On the other hand, when the formative assessments relay the true midyear state of proficiency with reasonable (but imperfect) accuracy, the school district generally benefits from investing in information for moderate budget values. In this case, the relative accuracy of the formative assessment makes teachers less likely to rely on their knowledge of students' high performance at the beginning of the year, which consequently raises the level of merit-based incentive required to motivate them. A counterintuitive exception to this is when the formative assessment almost perfectly predicts when students are proficient; at these high levels of accuracy, investing in information has minimal impact on teachers' behavior when

students perform well on the midyear assessment, but it reduces to zero the probability that teachers' exert effort when students have fallen behind. In such settings, the school district should rely on the formative assessment. Finally, as the budget increases, the school district is first able to afford to incentivize teachers to exert maximal effort throughout the year when they are relying only on formative assessments. In this case, investing in the interim assessment is no longer optimal.

Our stylized model results in a "bang-bang" solution, in which teachers will either exert maximum effort or no effort. We recognize that, in practice, K-12 education is a far more complex system, which our model does not fully capture. First, we characterize effort as a single-dimensional decision made solely by teachers. In practice, teaching and the effort put into it consist of many distinct components, such as lesson planning and professional development. Moreover, students also make an effort decision: they actively determine the type and level of their own effort to exert throughout the year. The information provided by an interim assessment may allow both students and teachers to target their effort more effectively, resulting in a higher probability of students succeeding at no additional "cost" to teachers. We expect that extending our work to include such features will broaden the range of settings for which midyear performance information has positive value. Furthermore, we focus on the level of proficiency attained by a school and do not take into account student "growth." Student growth targets are designed to standardize the level of effort necessary for teachers and district to meet their performance goals, regardless of the initial state of student proficiency. In reality, both measures are important for school accountability. A simple way to adapt our model to this setting is by assuming that every schools starts in the "proficient" state, and letting high-performing and low-performing schools vary in terms of how easily they can transition to the proficient state or recover from falling behind. Finally, although we include a baseline level of effort in our model, we take this as given. In reality, there are a multitude of factors that can increase this baseline effort level, thus diminishing the need for either interim assessments or merit-based incentives.

Still, while these limitations impact the quantitative results in our model, the qualitative insights are robust and, we believe, noteworthy. In practice, school districts often focus on the potential benefits from providing additional information, but not on the potential drawbacks stemming from it. Our analysis identifies settings where extra information is beneficial, but also settings where it may have a demotivating effect. We view our model as the first step in exploring the rich and complex environment of education.

CHAPTER 2 : Surgical Case-Mix and Discharge Decisions: Does Within-Hospital Coordination Matter?

2.1. Introduction

It has long been recognized that the organizational complexity of hospitals requires a high degree of patient care coordination (Georgopoulos and Mann, 1962). To be sure, care coordination can take many forms and impact the provision of healthcare in many ways. Lower mortality rates and other positive outcomes are associated both with coordinated operational decision-making among hospital department heads (Shortell et al., 1976) and physician-nurse collaboration (Mitchell and Shortell, 1997; Nair et al., 2012). Baker et al. (2004) partially attribute a higher incidence of adverse events in teaching hospitals than in community hospitals to miscommunications and poor care coordination. Nagpal et al. (2012) find information transfer and communication failures to be common across the entire surgical pathways, a key source of patient harm. Care coordination can also influence the patient experience indirectly by facilitating the optimal utilization of limited hospital resources, such as operating rooms, staff, and inpatient beds. Indeed, one key benefit of decision support systems is their ability to synthesize information from different parts of the patient care process for the purposes of resource management and admission planning (e.g., Kusters and Groot, 1996; Beliën et al., 2009; Matos and Rodrigues, 2011).

We study one specific type of coordination, namely that between case-mix and patient discharge policies for a profit-maximizing, resource-constrained hospital. Specifically, we consider a hospital with two types of resources: "front-end" resources, such as an operating room, and "backroom" resources, such as recovery beds. The hospital can manage its profitability by restricting the size and composition of its elective patient portfolio and by using discharges to control patients' length of stay. These decisions are often made by different parties: case-mix decisions are the result of policies developed by hospital administrators and surgeons, whereas discharge decisions are often made by attending physicians (Centers for Medicare and Medicaid Services, 2008). For a resource-constrained hospital, there

is a complex interdependence between these two decisions: a given case-mix requires the availability of both types of resources, and discharge decisions must balance the opposing pressures of backroom congestion and the probability of readmission. Moreover, there is ample evidence that patient discharge decisions are already impacted by hospital occupancy levels, both in the emergency department (Forster et al., 2003) and ICU (Kc and Terwiesch, 2012; Long and Mathews, 2017), as are readmission decisions (Fisher et al., 1994). Given the prevalence of this ad hoc approach, it is likely that a hospital would benefit from a process in which the optimal case-mix and discharge policies are determined in tandem. However, in reality such degree of coordinated decision-making among different stakeholders is associated with costs of introducing and maintaining new information flow processes and overcoming the "status quo" organizational norms. The estimation of such organizational costs is a challenging task that often goes beyond the scope of operational investigation. We leave this task outside of the scope of this study, and, instead, focus on the estimation of potential gains that a hospital can realize from coordinating the case-mix and patient discharge policies.

In this chapter, we characterize the benefits of a coordinated decision-making by comparing it to two decision-making approaches likely to arise in realistic hospital settings that employ decentralized patient flow management. The first approach that we label the "front-end" policy is one where both the portfolio and the discharge decisions are dictated exclusively by the hospital's front-end costs. This approach approximates the setting where surgeons' influence make operating room constraints especially salient to the hospital. The second approach is a "siloed" policy, in which for any given portfolio of procedures, discharge decisions are made based on backroom costs, and the elective case-mix is set in response to the patient discharge policy. Our analysis relies on an "open-loop" approach, in which we determine the strategic match between patient flows and hospital capacity. We do not explicitly consider the state of resource utilization in the hospital.

We build on the earlier work by Bavafa et al. (2019) that focuses exclusively on the optimal

case-mix of the elective procedures. Our modeling approach, however, is different in two key aspects. First, we extend the analysis of hospital actions to include the patient discharge policies, and introduce a model of patient recovery that uses a notion of *health state*. In particular, we tie patient health state at discharge to patient readmission probability and length of stay. Second, in order to facilitate the joint analysis of portfolio and discharge decisions, we employ quadratic front-end and backroom cost structures as opposed to a threshold-based cost model in Bavafa et al. (2019). This quadratic cost structure may provide a more realistic representation of the impact of increased resource utilization on hospital expenses. Additionally, in combination with the Central Limit Theorem-based approximation to stochastic resource utilization, such cost structure allows for analytical characterization of the joint portfolio and discharge decisions. In order to connect our modeling approach to hospital realities, we employ two separate data sources to estimate the "base-case" set of modeling parameters. The 2016 Nationwide Readmissions Database enables us to estimate parameters related to patient readmission probability and length of stay, and we leverage data from a medium-sized teaching hospital to estimate the parameters related to surgery durations.

Using our modeling approach, we establish the following results:

1. We derive, for a given portfolio of surgical procedures and patient discharge policy, the closed-form expressions for the first two moments of the front-end and backroom resource utilization (Propositions 6 and 7).

2. We characterize the optimal coordinated policy for a single-specialty hospital, detailing the optimal portfolio of elective procedures and the optimal patient discharge decision (Proposition 8).

3. We provide sufficient conditions for the optimality of "extreme" patient flow policies that combine either "no elective" or "max elective" case-mix with "full recovery" or "minimal recovery" discharge policies for a single-specialty hospital (Proposition 9).

4. We derive the optimal portfolio and discharge decisions under two decentralized policies that are likely to reflect the realities of patient flow management in hospital settings: the "front-end" (Proposition 10) and "siloed" policies (Proposition 11); we then compare their performance to that of the optimal policies.

The remainder of the chapter is organized as follows. Section 2.2 reviews the relevant literature. We present the model and base-case estimates of the parameters in Section 2.3, and we derive the optimal elective case-mix and discharge policies under the coordinated approach in Section 2.4. In Section 2.5, we compare the coordinated approach to the two decentralized decision-making settings likely to be encountered in practice. Finally, Section 2.6 presents a discussion of the results and directions for future research.

2.2. Literature Review

We study the benefits of coordinating the decisions regarding optimal elective case-mix and discharge policies for a profit-maximizing hospital. These decisions have been studied separately in the existing literature, so our work brings together distinct bodies of research.

The case-mix planning problem involves choosing the optimal size and composition of a hospital's portfolio of procedures. This is a key strategic consideration for a hospital that must manage its profitability in the face of resource constraints, and it has become increasingly important in recent decades, as reimbursement schemes have centered on diagnostic related groups (DRGs) (Roth and Van Dierdonck, 1995; Hof et al., 2017). We study this question for a hospital considering the capacity of two resources: the operating room (OR) and recovery beds. Therefore, our work fits into the growing literature on quantity-based revenue management for multiple resources.

Within this area, some authors have explicitly characterized the optimal case-mix, sometimes referred to as the patient mix. Adan and Vissers (2002) develop an integer linear programming model to determine the optimal patient mix for elective procedures based on multiple resource constraints assuming deterministic length of stay (LOS). Ma and Demeulemeester (2013) take a multilevel approach, first determining the optimal case-mix, and

then using these results to build the master surgical schedule. They assume that the mean LOS is fixed, and the number of beds and OR time allocated to each procedure are decision variables. Freeman et al. (2018) develop a multi-phase approach that utilizes mathematical programming to determine a set of solutions for the case-mix and master surgery schedule, assuming stochastic LOS, and then use simulation to assess the quality of each candidate solution.

Our work is closely related to Bavafa et al. (2019) who characterize the optimal number of elective procedures for a single-specialty hospital using a two-moment approximation of OR and recovery bed usage. The authors established the validity of this approach by comparing these approximations to the exact values from a real-life dataset. We extend this work by jointly determining the optimal discharge decision for each procedure in the hospital's portfolio. The consideration of the discharge decision, which in turn determines the distribution of LOS, is the key element that distinguishes our work from the existing literature in this area. Additionally, unlike some of the aforementioned work, we do not explicitly characterize the optimal number of beds or operating rooms, or consider a hard capacity constraint. Rather, we determine the case-mix and discharge policies assuming capacity constraints are nominal and can be exceeded at a cost.

There is also a body of work which focuses on short-term decisions that ultimately determine the case-mix, e.g., the master surgery schedule. Beliën and Demeulemeester (2007) are the first to consider optimal bed usage when building cyclic surgery schedules; they build a master schedule which levels bed usage, using stochastic patients per operating room block and stochastic LOS. Adan et al. (2009) extend their previous work by considering a stochastic, rather than deterministic, patient LOS. They develop a mixed integer linear programming model to generate a master surgical schedule that balances the patient mix to optimize the utilization of multiple resources. Rath et al. (2017) use a two-stage mixed-integer stochastic dynamic programming model to determine the optimal allocation of two parallel resources—operating rooms and anesthesiologists—and the sequencing of surgeries.

Although this work is related to the question we study, our focus is on the long-term case-mix decision. Finally, note that the preceding discussion focuses specifically on quantity-based revenue management for multiple resources in a healthcare setting. Bavafa et al. (2019) provide a broader discussion of the quantity-based revenue management literature, including the single-resource setting. For a recent review of the literature on the case-mix planning problem, including a discussion of how it relates to other hospital planning problems, see Hof et al. (2017).

An additional body of related literature is on discharge decisions in a hospital setting. Within this, there are three broad areas. First, researchers have studied the relationship between patient LOS and the likelihood of readmission. For heart attack and heart failure patients, Carey and Lin (2014) find a negative correlation between LOS and readmission probability. In our analysis, we assume that this negative correlation holds. Taking a cost perspective, Carey (2015) shows that the expected cost savings from avoiding readmission is between 15% to 65% of the cost of an additional day of stay.

Second, there is considerable evidence that physicians respond to congestion when determining discharge levels, which in turn impacts the likelihood that the patient is readmitted. Kc and Terwiesch (2012) show that during busy periods, ICUs ration bed capacity through a more aggressive discharge policy, which leads to an increased likelihood of a "revisit" to the ICU with a longer LOS. Long and Mathews (2017) show that in the ICU, there is a discretionary "boarding" time that increases significantly when ward occupancy levels are high, but that "service" times are unaffected by occupancy levels. This work supports our view that discharge policies should be considered at the strategic level.

Third, due to the complexity of balancing a longer LOS with resource constraints, there is a stream of work on the optimal discharge policy. Chan et al. (2012) devise an ICU discharge policy taking into account readmission risks that results in significant throughput gains without adversely impacting patient mortality rates. This policy focuses on the tactical discharge necessitated by a patient arrival. Shi et al. (2019) create a discharge decision sup-

port tool that dynamically determines how many and which patients should be discharged each day by determining each patient's personalized readmission probability. Although we also study discharge decisions, our focus is on the strategic, long-term policy necessary to most profitably manage resources.

We employ an "open-loop" modeling approach given our focus on the portfolio and discharge decisions at a strategic level. This approach complements the studies that are focused on the tactical management of hospital resources; such dynamic models require "closed-loop" approaches that track the state of the system, e.g., bed occupancy (Ayvaz and Huh 2010, Helm et al. 2011, Shi et al. 2019, Liu et al. 2019). In a similar vein, our analysis leaves out the optimization of intra-day scheduling of procedures. There is an extant literature in operations management that studies such scheduling problems (Patrick et al. 2008, Begen and Queyranne 2011, Helm and Van Oyen 2014, Zacharias and Pinedo 2017, Diamant et al. 2018).

2.3. Managing Elective Procedures and Patient Discharges: A Model

In the presence of any particular reimbursement structure introduced by the payer (such as a government agency or a private insurance company), a hospital can utilize two main levers to manage its profitability. On the patient admissions side, the hospital can control the demand for its resources by limiting the size and the composition of its portfolio of elective procedures. On the patient discharge side, the hospital may use early discharges to modify the distribution of patient LOS and relieve the pressure on its resources. In this section, we present and analyze a model that describes the trade-offs faced by a hospital in using these managerial levers.

Consider a daily portfolio of N types of elective surgical procedures, $\mathbf{a}^e = (a_1^e, \ldots, a_N^e)$, where $a_i^e \in [0, E_i], i = 1, \ldots, N$ is the daily number of type-i procedures that the hospital performs. Here, $\mathbf{E} = (E_1, \ldots, E_N)$ represents the vector of daily demand values for elective procedures, i.e., the maximum numbers of elective procedures that the market in which the hospital operates can support. In our analysis, we assume, for the sake of tractability, that

the daily numbers of elective procedures, $a_i^e, i = 1, \ldots, N$, can take fractional values.

In addition to elective procedures, a hospital may have to perform urgent procedures that represent the unscheduled part of the demand, e.g., coming from patients admitted into the hospital through the emergency department. We represent this daily demand by $\mathbf{a}^u = (a_1^u, \ldots, a_N^u)$, so the hospital's total daily portfolio of procedures is represented by $\mathbf{a} = \mathbf{a}^e + \mathbf{a}^u$. The emergency procedures create additional, exogenous load on hospital resources that translate into additional hospital costs. In our model, we consider the influence of emergency procedures when calculating the hospital's operating costs.

Any hospital manages multiple distinct resources (such as operating rooms, recovery beds, imaging and labs) that can be considered as belonging to one of two broad categories. The first category is "front-end" resources such as operating room; these resources are utilized "infrequently" during the patient's hospital stay, for example, only on the day of surgery. In other words, the utilization of a front-end resource is independent of patient LOS. The second category is backroom resources, e.g., hospital beds; these resources are utilized throughout patient stay at the hospital and their usage is positively correlated with patient LOS. In our model, we focus on two key hospital resources: a single front-end resource, OR capacity, and a single backroom resource, the recovery beds. Our analysis, however, can be readily extended to the setting with arbitrary numbers of front-end and backroom resources.

2.3.1. Modeling Hospital Resource Utilization

Below we present a model that connects the daily portfolio of elective procedures that the hospital chooses and the utilization of its front-end (surgery time) and backroom (recovery beds) resources.

Duration of Surgical Procedures, Health State and Recovery Dynamics

We use $S_{i,j,t}$ to denote the random variable describing the duration of the j-th procedure of type i ($0 \leq j \leq a_i$, $i = 1, \ldots, N$) on day t ($t = 1, 2, \ldots$).

In modeling the patient LOS and discharge process, it is important to realize that, in practice, a patient is discharged from a hospital when, according to the judgment of the hospital's physicians, the patient is likely to have recovered sufficiently to continue follow-up care in a non-hospital environment, e.g., the patient's home. In particular, every discharge decision represents a judgment call that always carries a non-zero probability of less-than-full recovery and subsequent readmission to the hospital. Below we describe our model of patient recovery dynamics that follows the procedure of type i.

More specifically, we consider the following model of patient recovery and discharge. Let $h_i \in \left[h_i^{\min}, h_i^{\max}\right]$ be a scalar quantity, observable to the attending physician, that describes the health state of a patient who underwent a procedure of type $i = 1, \ldots, N$. In practice, the true health state of a patient is a multi-dimensional entity. However, we assume that the attending physician aggregates the numerous factors that determine the true health state of a patient into a single quantity that she uses to determine whether the patient has sufficiently recovered to be discharged. It is this quantity that h_i designates in our model.

Without loss of generality, we normalize the health state scale so that $h_i^{\max} = 1$, for $i = 1, \ldots, N$. Thus, $h_i = 1$ denotes a state of complete recovery without the possibility of readmission. In any real hospital environment, physicians have some flexibility regarding the discharge decisions and must use their judgment to decide on a threshold health state h_i to initiate patient discharge. Thus, discharge decisions involve a trade-off between continuing patient care in the hospital environment and allowing the recovery to be completed outside of the hospital.

In modeling the patient recovery process, we assume that for a patient that undergoes the j-th procedure of type i ($0 \leq j \leq a_i$, $i = 1, \ldots, N$) on day t ($t = 1, 2, \ldots$), the recovery process starts at the state $h_i = 0$ and proceeds to improve the patient's health state at the daily rate $r_{i,j,t}$.

A pair of quantities $(S_{i,j,t}, r_{i,j,t})$ determine the usage of surgery time and recovery bed

resources by the j-th patient of type i on day t. We make the following distributional assumptions regarding these quantities.

Assumption 4 (Surgery Durations and Recovery Rates) *$S_{i,j,t}$ and $r_{i,j,t}$ are distributed according to a joint continuous probability distribution with a finite support, and both the support and the probability density function (PDF) of the distribution depend only on i, but not on j or t:*

$$\text{Prob}\left(S_{i,j,t} \leq x, r_{i,j,t} \leq y\right) = \int_0^x \int_{r_i^{\min}}^y f_i(s,r)drds, x \in [0, S_i^{\max}],\ y \in \left[r_i^{\min}, r_i^{\max}\right],$$
$$i = 1, \ldots, N, j = 1, \ldots, a_i, t = 1, 2, \ldots, \tag{2.1}$$

where $0 < S_i^{\max}$ and $0 < r_i^{\min} \leq r_i^{\max}$, and

$$\int_0^{S_i^{\max}} \int_{r_i^{\min}}^{r_i^{\max}} f_i(s,r)drds = 1, \quad i = 1, \ldots, N. \tag{2.2}$$

Assumption 5 (Inter-Day Independence) *The pairs $(S_{i_1,j_1,t_1}, r_{i_2,j_2,t_1})$ and $(S_{i_1,j_1,t_2}, r_{i_2,j_2,t_2})$ for all procedures performed on different days $t_1 \neq t_2$ are independent random variables.*

Assumption 6 (Inter-Procedure Independence) *The pairs $(S_{i_1,j_1,t}, r_{i_1,j_1,t})$ and $(S_{i_2,j_2,t}, r_{i_2,j_2,t})$ for all procedures of different types ($i_1 \neq i_2$) performed on the same day t are independent random variables.*

For the following analysis, it is convenient to introduce the marginal cumulative distribution

function (CDF) for the procedure duration and the recovery rate:

$$\text{Prob}\left(S_{i,j,t} \leq x\right) = \Phi_i^S\left(x\right) = \int_0^x \int_{r_i^{\min}}^{r_i^{\max}} f_i(s,r)drds,$$

$$i = 1,\ldots,N, j = 1,\ldots,a_i, t = 1,2,\ldots, x \in \left[0, S_i^{\max}\right], \tag{2.3}$$

$$\text{Prob}\left(r_{i,j,t} \leq y\right) = \Phi_i^r\left(y\right) = \int_0^{S_i^{\max}} \int_{r_i^{\min}}^{y} f_i(s,r)drds,$$

$$i = 1,\ldots,N, j = 1,\ldots,a_i, t = 1,2,\ldots, y \in \left[r_i^{\min}, r_i^{\max}\right]. \tag{2.4}$$

We assume that both $\Phi_i^S\left(x\right)$ and $\Phi_i^r\left(y\right)$ are continuously differentiable over their respective domains.

Discharge Threshold and Patient Length of Stay

We assume that hospital physicians select a discharge threshold $h_i^{\text{d}} \in \left[h_i^{\min}, 1\right]$, such that any patient recovering from procedure i is discharged once his health state reaches the level h_i^{d}. Thus, under the discharge policy characterized by the threshold h_i^{d}, the hospital LOS, L_i^{o}, for a patient that undergoes a procedure i is

$$L_i^{\text{o}}\left(h_i^{\text{d}}\right) = \frac{h_i^{\text{d}}}{r_i}, \tag{2.5}$$

a random variable distributed on the interval $\left[L_i^{\min}\left(h_i^{\text{d}}\right), L_i^{\max}\left(h_i^{\text{d}}\right)\right]$ with

$$L_i^{\min}\left(h_i^{\text{d}}\right) = \frac{h_i^{\text{d}}}{r_i^{\max}}, \tag{2.6}$$

$$L_i^{\max}\left(h_i^{\text{d}}\right) = \frac{h_i^{\text{d}}}{r_i^{\min}}, \tag{2.7}$$

according to the CDF

$$\Phi_i^{L,\text{o}}(z|h_i^{\text{d}}) = Pr\left(L_i^{\text{o}}\left(h_i^{\text{d}}\right) \leq z\right) = Pr\left(\frac{h_i^{\text{d}}}{r_i} \leq z\right) = Pr\left(r_i \geq \frac{h_i^{\text{d}}}{z}\right) = 1 - \Phi_i^r\left(\frac{h_i^{\text{d}}}{z}\right). \tag{2.8}$$

Note that in (2.5)-(2.8) we have used a streamlined notation for the recovery rate r_i that does not include indices j and t.

Readmission Dynamics

Hospital readmissions are often used by payers as an indicator for the quality of hospital care. In the US, Medicare keeps track of hospital readmissions that occur within the 30-day period after discharge, considering readmissions related to the original complaint as part of the original "care episode." In our model, we assume that, upon discharge, there exists a likelihood that a patient returns to the hospital for readmission within a certain time period defined by the payer. Under the "bundled" compensation approach, the hospital does not receive any additional payment for care provided upon patient readmission. In other words, from both the payer's and the hospital's perspectives, the readmission care is "rework" caused by a "defect" in the original care and/or patient recovery. While, for any surgical procedure, a certain rate of readmissions cannot be avoided and is treated as acceptable by the payer, a high rate of readmissions may be interpreted as related to hospital discharge policy. In Section 2.3.4 we provide a detailed discussion of the hospital compensation and cost structure. In the current section, we provide a foundation for this discussion and focus on modeling the impact of readmissions on the utilization of hospital resources.

For the hospital discharge policy defined by patient health status at the time of discharge, h_i^{d}, we use $p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)$ to denote the probability that a patient gets readmitted. We formalize our assumptions on the readmission dynamics as follows.

Assumption 7 (Patient Readmission Dynamics) *a) When a patient that underwent a procedure of type i is discharged in the health state h_i^{d}, she will be readmitted with probability $p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)$, where $p_i^{\mathrm{a}}(\cdot)$ is a monotone decreasing function with $p_i^{\mathrm{a}}(1) = 0$.*

b) If a patient is readmitted, she must undergo procedure i and repeat the recovery process.

c) The procedure duration-recovery rate pair of values observed upon readmission are perfectly correlated with the procedure duration-recovery rate pair observed upon original admission.

d) Any patient will be readmitted at most once.

Assumption 7a is intuitive and does not result in a loss of generality. Assumption 7b models the impact of readmissions on the use of front-end and backroom hospital resources by treating the readmission as rework, requiring that the patient care be repeated upon readmission. While our analysis below explicitly relies on this Assumption, it can be extended to handle a more general setting where a readmitted patient may not require the repeat of the original surgical procedure as well as settings where she may require a different surgical procedure. In reality, it is likely that, for a given patient, the readmission surgery duration and the following recovery rate are correlated with the original pair of values for the surgery duration and the recovery rate. Assumption 7c maintains the analytical tractability of the resulting model by treating this correlation as perfect. While this Assumption is admittedly strong, we believe that it allows for substantial analytical headway without altering the qualitative nature of the results. Finally, Assumption 7d avoids modeling complexities caused by potential multiple readmissions that, while possible, are likely to be rare in practice. This Assumption, however, does not impact the generality of our analysis, since, in the settings with substantial presence of multiple readmissions, $p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)$ can be interpreted as the quantity representing the cumulative effect of multiple readmissions.

Our approach to modeling the impact of readmission is designed to allow us to study a trade-off between the "immediate" cost savings resulting from earlier patient discharges and potential "delayed" costs associated with patient readmissions. The advantage of our modeling approach is that these delayed readmission costs are completely characterized by a single function $p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)$. Our analysis, however, can be extended to include more complex modeling of readmission dynamics, such as different distributions of front-end and backroom resource utilizations upon readmission.

Using our Assumptions, we summarize the description of the stochastic utilization by a "type-i" patient of the front-end and the backroom hospital resources, that includes the original admission and a potential readmission, as follows.

Lemma 1 (Distributions of Resource Usage in the Presence of Discharge Management)

Under Assumptions 4-7, when the discharge policy characterized by the threshold h_i^{d} is employed, the total use of operating room time by any "type-i" patient is given by a random variable $F_i\left(h_i^{\mathrm{d}}\right)$ distributed on the interval $[0, 2S_i^{\max}]$ with the CDF

$$\Phi_i^F\left(x|h_i^{\mathrm{d}}\right) = \left(1 - p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)\Phi_i^S\left(x\right) + p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\Phi_i^S\left(\frac{x}{2}\right), \tag{2.9}$$

and the total hospital length of stay $L_i\left(h_i^{\mathrm{d}}\right)$ is distributed on the interval $\left[L_i^{\min}\left(h_i^{\mathrm{d}}\right), 2L_i^{\max}\left(h_i^{\mathrm{d}}\right)\right]$ with the CDF

$$\Phi_i^L(z|h_i^{\mathrm{d}}) = \left(1 - p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)\left(1 - \Phi_i^{\mathrm{r}}\left(\frac{h_i^{\mathrm{d}}}{z}\right)\right) + p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\left(1 - \Phi_i^{\mathrm{r}}\left(\frac{2h_i^{\mathrm{d}}}{z}\right)\right). \tag{2.10}$$

2.3.2. Base-Case Estimation

In this section we use two separate data sources to estimate the base-case parameters for our model. First, we use the 2016 Nationwide Readmissions Database (Healthcare Cost and Utilization Project, 2016) to estimate the parameters related to patient LOS and to characterize the readmission function. Second, we use data on the surgeries performed at a medium-sized teaching hospital between January 1, 2014 and December 31, 2016 to obtain information on surgery durations.

Our modeling approach relies on the hospital being able to reliably estimate the readmission function $p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)$, a novel modeling element we introduced. The 2016 Nationwide Readmissions Database is a comprehensive sample of patient discharges from hospitals across the US, reflecting more than half of all hospitalizations. For each discharge, a wide range of data, including patient and hospital characteristics as well as clinical information, are reported. In particular, for each discharge, the dataset records the Diagnostic Related Group (DRG) that provides a summary of the care received by the patient and that also forms the basis for the compensation the hospital receives. In addition, for each discharged patient, all subsequent readmissions (if any) are documented.

The approach we use for estimating the base-case parameters is as follows. For the purposes

of estimation, we focus on a single DRG = 470 ("Major Joint Replacement or Reattachment of Lower Extremity without Major Complication or Comorbidity") that corresponds to the largest number of discharges across all DRGs and all hospitals. The patient discharges were grouped according to the combination of hospital ownership and teaching status. In particular, the nine hospital groups we use based on the designations in the database are constructed as elements of a 3×3 matrix, reflecting three ownership types ("government, non-federal," "private, not-for-profit," and "private, investor-owned") and three teaching designations ("metropolitan non-teaching," "metropolitan teaching," and "non-metropolitan"). The rationale for using these nine groups is that a hospital's discharge policy is most likely driven by the combination of financial incentives reflected in the ownership type and the practicing style reflected in the hospital's teaching status.

In particular, we assume that each of the nine hospital groups, $k = 1, ..., 9$, share a common level of patient health at discharge, $h^{\mathrm{d},k}$ (for the remainder of this section, in our notation we drop the procedure index i assuming that it corresponds to DRG= 470).

The database contains patient LOS and 30-day readmission events. We assume that, irrespective of the hospital delivering care, all patients share a common distribution of recovery rates, described by a four-parameter beta distribution with the pdf

$$\phi^{\mathrm{r}}\left(x, \alpha, \beta, r_{\min}, r_{\max}\right) = \frac{\left(x - r_{\min}\right)^{\alpha-1}\left(r_{\max} - x\right)^{\beta-1}}{\left(r_{\max} - r_{\min}\right)^{\alpha+\beta-1} B\left(\alpha, \beta\right)}, \tag{2.11}$$

parameterized by α, β, $r_{\min}$, and $r_{\max}$, with $B\left(\alpha, \beta\right)$ representing the beta function. In addition, we consider the following family of readmission functions:

$$p^{\mathrm{a}}\left(h^{\mathrm{d},k}\right) = 1 - \left(h^{\mathrm{d},k}\right)^{\theta}. \tag{2.12}$$

For a given combination of parameters α, β, $r_{\min}$, $r_{\max}$, θ, and discharge thresholds $h^{\mathrm{d},k}, k = 1, ..., 9$, we can compute the expected readmission rate and LOS distribution for each of the nine hospital groups. To estimate these parameters for the base-case, we select them to

Quantity	Value
α	1.639
β	1.893
$r_{\min}$, 1/day	0.008
$r_{\max}$, 1/day	0.229
θ	0.023

Quantity	Value
$h^{\mathrm{d},1}$	0.233
$h^{\mathrm{d},2}$	0.204
$h^{\mathrm{d},3}$	0.212
$h^{\mathrm{d},4}$	0.233
$h^{\mathrm{d},5}$	0.209
$h^{\mathrm{d},6}$	0.235
$h^{\mathrm{d},7}$	0.247
$h^{\mathrm{d},8}$	0.231
$h^{\mathrm{d},9}$	0.245

Table 1: Base-case parameter estimates for DRG = 470.

make sure that the predicted readmission rates and LOS distributions are closely matched by the observed data. This estimation is implemented by minimizing the sum of the squared percentage deviations between the readmission rate, mean LOS, standard deviation of LOS, minimum LOS, and maximum LOS for each of the nine hospital groups.

Table 1 provides a summary of the base-case estimates for the parameters of our model. The resulting probability of patient readmission, p^{a}, as a function of patient health state, $h^{\mathrm{d},k}$, is shown in Figure 5.

Note that our goal here is to provide a simple way to estimate the model parameters based a national dataset. In the case of an individual hospital, a similar estimation can be implemented using detailed data on factors such as bed occupancy and dynamic observable patient health outcomes (Shi et al., 2019).

Finally, as noted earlier, we use data from a medium-sized teaching hospital to obtain information on surgery durations. In particular, we have collected the data for 462 surgeries related to DRG= 470 over the course of three years. The distribution of durations for these surgeries is shown in Figure 6: the mean and the standard deviation of the distribution are 106.98 and 24.81 minutes, respectively.

Below, we will use the base-case parameter set reported in Table 1 and the estimates for the first two moments of the distribution of surgery durations, $\mathrm{E}[S] = 106.98$ and

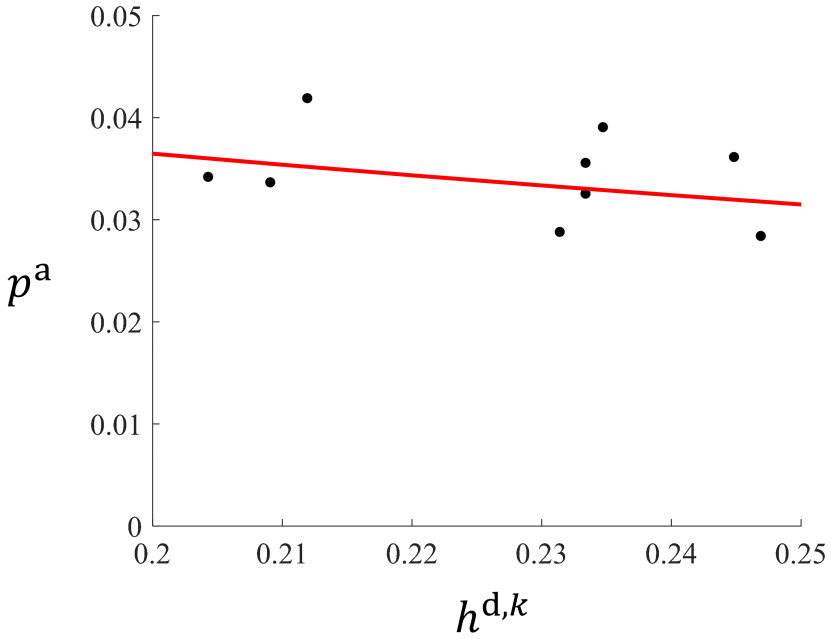

Figure 5: The probability of patient readmission, p^{a}, as a function of the discharge threshold for patient health state, $h^{\mathrm{d},k}$, $k = 1, ..., 9$. Each data point represents one of the nine hospital groups that were formed according to the combination of hospital ownership and teaching status.

$\mathrm{Var}\,[S] = (24.81)^2 = 615.54$ in our numerical examples.

2.3.3. Utilization of Hospital Resources: Central-Limit Approximation

Consider a hospital that uses a daily portfolio of elective procedures $\mathbf{a}^e = (a_1^e, \ldots, a_N^e)$ and the discharge thresholds $\mathbf{h}^{\mathrm{d}} = \left(h_1^{\mathrm{d}}, \ldots, h_N^{\mathrm{d}}\right)$ in the presence of random daily portfolio of urgent procedures $\mathbf{a}^u = (a_1^u, \ldots, a_N^u)$. We assume that the numbers of urgent procedures on any day t follow a stationary distribution that does not depend on t and is described by the PDF $f^u\,(n_1, \ldots, n_N)$, $0 \leq n_i \leq U_i, i = 1, \ldots, N$ with

$$\int_0^{U_N} \ldots \int_0^{U_1} f^u\,(n_1, \ldots, n_N)\, dn_1 \ldots dn_N = 1. \tag{2.13}$$

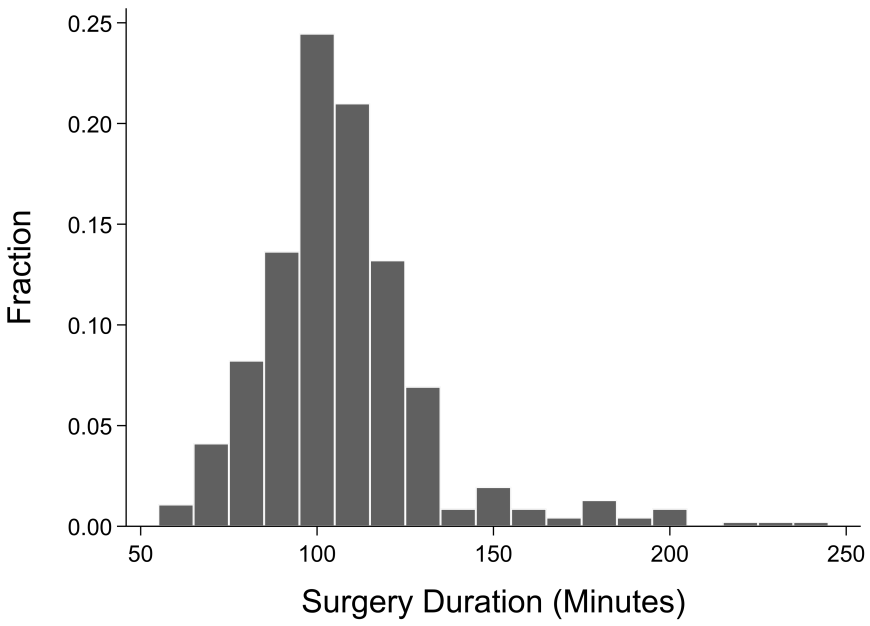

Figure 6: The distribution of surgery durations for DRG= 470 ($n = 462$).

For the analysis below, it is convenient to introduce the following notation for the first and second moments of the distribution of daily numbers of urgent procedures:

$$\mu_i^u = \int_0^{U_N} \dots \int_0^{U_1} (n_i)\, f^u(n_1, \dots, n_N)\, dn_1 \dots dn_N, \quad i = 1, \dots, N, \tag{2.14}$$

$$(\sigma_i^u)^2 = \int_0^{U_N} \dots \int_0^{U_1} (n_i - \mu_i^u)^2\, f^u(n_1, \dots, n_N)\, dn_1 \dots dn_N, \quad i = 1, \dots, N, \tag{2.15}$$

$$\rho_{ij}^u = \frac{\int_0^{U_N} \dots \int_0^{U_1} \left((n_i n_j)\, f^u(n_1, \dots, n_N) - \mu_i^u \mu_j^u \right) dn_1 \dots dn_N}{\sigma_i^u \sigma_j^u}, \quad i, j = 1, \dots, N, i \neq j. \tag{2.16}$$

Then, for $\mathbf{a} = \mathbf{a}^e + \mathbf{a}^u$, let $F_t\left(\mathbf{a}, \mathbf{h}^{\mathrm{d}}\right)$ and $B_t\left(\mathbf{a}, \mathbf{h}^{\mathrm{d}}\right)$ be the total number of hours of OR time and the total number of recovery beds required, respectively, on day t.

For the following analysis, we define the parameters

$$\mu_i^F\left(h_i^{\mathrm{d}}\right) = \int_0^{2S_i^{\max}} \left(1 - \Phi_i^F\left(x | h_i^{\mathrm{d}}\right)\right) dx \tag{2.17}$$

and

$$\left(\sigma_i^F\left(h_i^{\mathrm{d}}\right)\right)^2 = \int_0^{2S_i^{\max}} \left(x - \mu_i^F\left(h_i^{\mathrm{d}}\right)\right)^2 d\Phi_i^F\left(x|h_i^{\mathrm{d}}\right) \tag{2.18}$$

to represent the expected value and the variance of type-i procedure duration. For the recovery beds, we define

$$\mu_i^B\left(h_i^{\mathrm{d}}\right) = L_i^{\min}\left(h_i^{\mathrm{d}}\right) + \int_{L_i^{\min}\left(h_i^{\mathrm{d}}\right)}^{2L_i^{\max}\left(h_i^{\mathrm{d}}\right)} \left(1 - \Phi_i^L(z|h_i^{\mathrm{d}})\right) dz, \tag{2.19}$$

$$\left(\sigma_i^B\left(h_i^{\mathrm{d}}\right)\right)^2 = \int_{L_i^{\min}\left(h_i^{\mathrm{d}}\right)}^{2L_i^{\max}\left(h_i^{\mathrm{d}}\right)} \Phi_i^L(z|h_i^{\mathrm{d}}) \left(1 - \Phi_i^L(z|h_i^{\mathrm{d}})\right) dz. \tag{2.20}$$

Proposition 6 (Limiting Approximations for the Resource Usage) *Let* $\mathbf{a} = \mathbf{a}^e + \mathbf{a}^u \geq \mathbf{0}$ *be the vector of the total numbers of procedures performed each day and* $\mathbf{0} \leq \mathbf{h}^{\mathrm{d}} \leq \mathbf{1}$ *be the vector of discharge thresholds. Define*

$$M^F\left(\mathbf{a}, \mathbf{h}^{\mathrm{d}}\right) = \sum_{i=1}^N a_i \mu_i^F\left(h_i^{\mathrm{d}}\right), \tag{2.21}$$

$$\left(\Sigma^F\left(\mathbf{a}, \mathbf{h}^{\mathrm{d}}\right)\right)^2 = \sum_{i=1}^N a_i \left(\sigma_i^F\left(h_i^{\mathrm{d}}\right)\right)^2, \tag{2.22}$$

$$M^B\left(\mathbf{a}, \mathbf{h}^{\mathrm{d}}\right) = \sum_{i=1}^N a_i \mu_i^B\left(h_i^{\mathrm{d}}\right), \tag{2.23}$$

$$\left(\Sigma^B\left(\mathbf{a}, \mathbf{h}^{\mathrm{d}}\right)\right)^2 = \sum_{i=1}^N a_i \left(\sigma_i^B\left(h_i^{\mathrm{d}}\right)\right)^2. \tag{2.24}$$

Then, for a fixed $\mathbf{h}^{\mathrm{d}}$*, as* $a_i \to \infty, i = 1, \ldots, N$*,*

$$\frac{F_t\left(\mathbf{a}, \mathbf{h}^{\mathrm{d}}\right) - M^F\left(\mathbf{a}, \mathbf{h}^{\mathrm{d}}\right)}{\Sigma^F\left(\mathbf{a}, \mathbf{h}^{\mathrm{d}}\right)} \xrightarrow{d} \mathcal{N}(0,1), \tag{2.25}$$

and

$$\frac{B_t\left(\mathbf{a}, \mathbf{h}^{\mathrm{d}}\right) - M^B\left(\mathbf{a}, \mathbf{h}^{\mathrm{d}}\right)}{\Sigma^B\left(\mathbf{a}, \mathbf{h}^{\mathrm{d}}\right)} \xrightarrow{d} \mathcal{N}(0,1). \tag{2.26}$$

Note that results analogous to those in Proposition 6 can be obtained for settings with an

arbitrary numbers of front-end and backroom resources. As Proposition 6 states, the impact of the discharge threshold choices $h_i^{\mathrm{d}}, i = 1, .., N$, on the utilization of hospital resources under a central-limit approximation is succinctly described by the first and second moment parameters in (2.17)-(2.20). For the following analysis, it is convenient to define

$$\gamma_i^l = \min_{h_i^{\mathrm{d}} \in [h_i^{\min}, 1]} \left(\left| \frac{dp^{\mathrm{a}}}{dh_i^{\mathrm{d}}} \right| \right), \tag{2.27}$$

$$\gamma_i^h = \max_{h_i^{\mathrm{d}} \in [h_i^{\min}, 1]} \left(\left| \frac{dp^{\mathrm{a}}}{dh_i^{\mathrm{d}}} \right| \right). \tag{2.28}$$

The result below contrasts the impact of the discharge threshold on the front-end and the backroom resources.

Proposition 7 (Analytical Characterization of the Moments of Resource Usage) *Consider a procedure of type $i = 1, \ldots, N$.*

a) $\mu_i^F \left(h_i^{\mathrm{d}}\right)$ is given by

$$\mu_i^F \left(h_i^{\mathrm{d}}\right) = \mathrm{E}\left[S_i\right] \left(1 + p_i^{\mathrm{a}} \left(h_i^{\mathrm{d}}\right)\right), \tag{2.29}$$

where $\mathrm{E}\left[S_i\right]$ is the expected value of the duration of a surgical procedure of type i, and is monotone decreasing in h_i^{d}. Also, $\left(\sigma_i^F \left(h_i^{\mathrm{d}}\right)\right)^2$ is given by

$$\left(\sigma_i^F \left(h_i^{\mathrm{d}}\right)\right)^2 = \left(1 + 3p_i^{\mathrm{a}} \left(h_i^{\mathrm{d}}\right)\right) \mathrm{Var}\left[S_i\right] + p_i^{\mathrm{a}} \left(h_i^{\mathrm{d}}\right) \left(1 - p_i^{\mathrm{a}} \left(h_i^{\mathrm{d}}\right)\right) \left(\mathrm{E}\left[S_i\right]\right)^2, \tag{2.30}$$

where $\mathrm{Var}\left[S_i\right]$ is the variance of the type-i surgical procedure duration, and is monotone decreasing in h_i^{d} if and only if

$$p_i^{\mathrm{a}} \left(h_i^{\mathrm{d}}\right) \leq \frac{1}{2} + \frac{3}{2} \left(\frac{\mathrm{Var}\left[S_i\right]}{\left(\mathrm{E}\left[S_i\right]\right)^2} \right). \tag{2.31}$$

b) $\mu_i^B \left(h_i^{\mathrm{d}}\right)$ is given by

$$\mu_i^B \left(h_i^{\mathrm{d}}\right) = h_i^{\mathrm{d}} \left(1 + p_i^{\mathrm{a}} \left(h_i^{\mathrm{d}}\right)\right) \mathrm{E}\left[\frac{1}{r_i}\right], \tag{2.32}$$

where

$$\mathrm{E}\left[\frac{1}{r_i}\right] = \int_{r_i^{\min}}^{r_i^{\max}} \frac{1}{y} d\Phi_i^{\mathrm{r}}(y). \tag{2.33}$$

Moreover, $\left(\sigma_i^B\left(h_i^{\mathrm{d}}\right)\right)^2$ *is given by*

$$\left(\sigma_i^B\left(h_i^{\mathrm{d}}\right)\right)^2 = h_i^{\mathrm{d}}\left(\left(1+p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)G_i^r + p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\left(1-p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)H_i^r\right), \tag{2.34}$$

where

$$G_i^r = \int_{r_i^{\min}}^{r_i^{\max}} \left(\frac{\Phi_i^{\mathrm{r}}(y)\left(1-\Phi_i^{\mathrm{r}}(y)\right)}{y^2}\right) dy, \tag{2.35}$$

$$H_i^r = \int_{\frac{r_i^{\min}}{2}}^{r_i^{\max}} \left(\frac{\Phi_i^{\mathrm{r}}(2y)-\Phi_i^{\mathrm{r}}(y)}{y}\right)^2 dy. \tag{2.36}$$

Finally, both $\mu_i^B\left(h_i^{\mathrm{d}}\right)$ *and* $\left(\sigma_i^B\left(h_i^{\mathrm{d}}\right)\right)^2$ *are monotone increasing in* h_i^{d} *if*

$$\gamma_i^h \leq \frac{G_i^r}{G_i^r + H_i^r}. \tag{2.37}$$

The results of Proposition 7 play a key role in understanding the trade-off faced by the hospital management in setting the patient discharge levels. On the one hand, longer hospital stays always reduce the load that patients place on the "front-end" resources. The mechanism of such reduction is clear: longer stays reduce the probability of readmissions and, consequently, of the rework/repeat of surgical procedures. On the other hand, the higher discharge levels will increase the load on the "backroom" resources unless the delay in patient discharges will significantly reduce the expected impact of readmissions. Figure 7 illustrates the results of Proposition 7 in the case of a single surgical procedure. The Figure underscores, on the one hand, a tension created by the opposing impacts that the discharge threshold may have on the expected utilization of the"front-end" and "backroom" resources, as well as on the behavior of the second moments of the resource utilization.

In a typical hospital setting the "front-end" resources are considered to be "revenue gen-

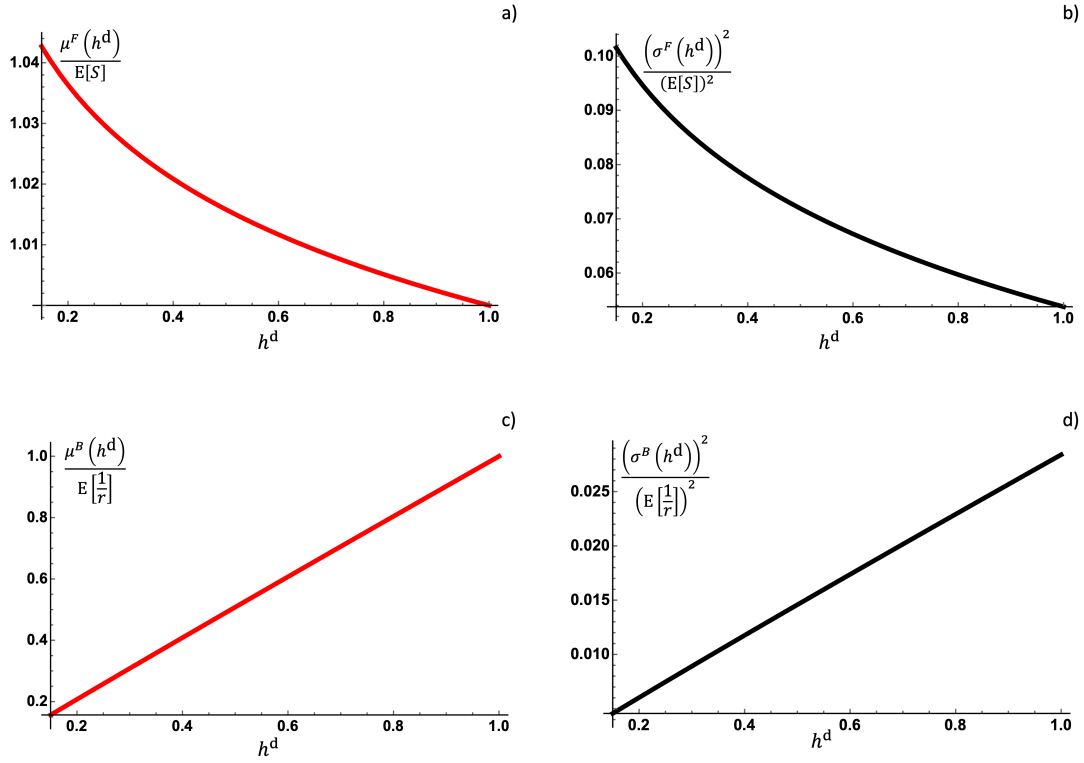

Figure 7: First two moments of the OR and the recovery bed utilization as functions of the discharge threshold h^{d}: a) expected OR utilization, b) variance of the OR utilization, c) expected number of required recovery beds, d) variance of the number of required recovery beds ($h^{\min} = 0.15$, $\mathrm{E}[S] = 106.98$, $\mathrm{Var}[S] = 615.54$, $\mathrm{E}\left[\frac{1}{r}\right] = 12.963$, $\alpha = 1.639$, $\beta = 1.893$, $r^{\min} = 0.008$, $r^{\max} = 0.229$, $\theta = 0.023$).

erating," and their optimal utilization may often be associated with the best policy that a hospital can adopt. Under such policy, a proper accounting for the possibility of patient readmissions will result in keeping patients in the hospital until the readmission probability is 0. This is not unexpected, given the motivation behind the replacement of fee-for-service hospital compensation by the "bundled payments" approach. In particular, while under the fee-for-service approach, the hospital had every incentive to play down the effects of readmissions since it could charge for all services performed upon readmission. As a consequence, it would have been optimal for a hospital to prefer patient discharges at the earliest feasible stage. Under "bundled" payments, a hospital no longer receives any extra compensation for extra work it has to perform if the patient is readmitted, resulting in a hospital

focusing on reducing the readmissions as much as possible. This approach, however, may create a situation in which the "backroom" resources can get overloaded. Resolving this tension between the front-end and backroom resources is a focus of the analysis below.

2.3.4. Hospital Compensation, Operating Costs and Optimization Problem

In our model, we use R_i to denote the "bundled" payment that a hospital receives for an entire episode of care associated with performing a procedure of type i. We assume that the hospital incurs the daily fixed cost of operating its facilities, which we normalize to 0, as well as potential extra costs associated with patient demands for hospital resources exceeding hospital nominal capacity. For a given portfolio of surgical procedures, $\mathbf{a}$, operating under the discharge thresholds $\mathbf{h}^{\mathrm{d}}$, we denote the random daily operating room cost as $C_F\left(\mathbf{a}, \mathbf{h}^{\mathrm{d}}\right)$ and the random recovery bed cost $C_B\left(\mathbf{a}, \mathbf{h}^{\mathrm{d}}\right)$. We treat the costs of using hospital resources as being additive across resources and adopt a widely-used convex functional form to represent the cost of using a particular hospital resource.

Assumption 8 (Resource Cost Functions) *a) For a resource $k = F, B$, let $U_k\left(\mathbf{a}, \mathbf{h}^{\mathrm{d}}\right)$ be the random variable representing the daily usage of that resource under the procedure portfolio $\mathbf{a}$ and discharge levels $\mathbf{h}^{\mathrm{d}}$. Then, the cost associated with the usage of resource k is given by*

$$C_k\left(\mathbf{a}, \mathbf{h}^{\mathrm{d}}\right) = c_k\left(U_k\left(\mathbf{a}, \mathbf{h}^{\mathrm{d}}\right)\right)^2, \tag{2.38}$$

where $c_k \geq 0$ is a resource-specific cost parameter.

b) The hospital's total daily cost is additive across the front-end and backroom resources:

$$C\left(\mathbf{a}, \mathbf{h}^{\mathrm{d}}\right) = C_F\left(\mathbf{a}, \mathbf{h}^{\mathrm{d}}\right) + C_B\left(\mathbf{a}, \mathbf{h}^{\mathrm{d}}\right). \tag{2.39}$$

The hospital's expected daily profit for a given portfolio of surgical procedures and discharge threshold values can be expressed as

$$\Pi\left(\mathbf{a}, \mathbf{h}^{\mathrm{d}}\right) \;=\; \sum_{i=1}^{N} a_i R_i - \mathrm{E}\left[C_F\left(\mathbf{a}, \mathbf{h}^{\mathrm{d}}\right)\right] - \mathrm{E}\left[C_B\left(\mathbf{a}, \mathbf{h}^{\mathrm{d}}\right)\right]. \tag{2.40}$$

Note that under this reimbursement structure, hospitals are implicitly penalized for readmissions. Under a scheme such as the Hospital Readmissions Reduction Program (HRRP), hospitals are explicitly penalized for high readmission rates, via a reduced reimbursement rate R_i.

Consider a setting where a hospital selects a portfolio of elective procedures $\mathbf{a}^e$ in the presence of urgent procedures $\mathbf{a}^u$ it must accommodate, so that $\mathbf{a} = \mathbf{a}^e + \mathbf{a}^u$. For convenience, we express the hospital optimization problem in the form of a Lemma.

Lemma 2 (Hospital Optimization Problem) *Let*

$$\begin{aligned}
\mathcal{A}_i\left(\mathbf{h}^{\mathrm{d}}\right) =& R_i - c_F\left(2\mu_i^F\left(h_i^{\mathrm{d}}\right)\sum_{j=1}^{N}\mu_j^u\mu_j^F\left(h_j^{\mathrm{d}}\right) + \left(\sigma_i^F\left(h_i^{\mathrm{d}}\right)\right)^2\right) \\
& - c_B\left(2\mu_i^B\left(h_i^{\mathrm{d}}\right)\sum_{j=1}^{N}\mu_j^u\mu_j^B\left(h_j^{\mathrm{d}}\right) + \left(\sigma_i^B\left(h_i^{\mathrm{d}}\right)\right)^2\right), \qquad (2.41) \\
\mathcal{B}_i\left(\mathbf{h}^{\mathrm{d}}\right) =& \mu_i^u R_i \\
& - c_F\left(\left((\sigma_i^u)^2 + (\mu_i^u)^2\right)\left(\mu_i^F\left(h_i^{\mathrm{d}}\right)\right)^2 + \mu_i^u\left(\sigma_i^F\left(h_i^{\mathrm{d}}\right)\right)^2\right. \\
& \left. + \sum_{j\neq i}\left(\rho_{ij}^u\sigma_i^u\sigma_j^u + \mu_i^u\mu_j^u\right)\mu_i^F\left(h_i^{\mathrm{d}}\right)\mu_j^F\left(h_j^{\mathrm{d}}\right)\right) \\
& - c_B\left(\left((\sigma_i^u)^2 + (\mu_i^u)^2\right)\left(\mu_i^B\left(h_i^{\mathrm{d}}\right)\right)^2 + \mu_i^u\left(\sigma_i^B\left(h_i^{\mathrm{d}}\right)\right)^2\right. \\
& \left. + \sum_{j\neq i}\left(\rho_{ij}^u\sigma_i^u\sigma_j^u + \mu_i^u\mu_j^u\right)\mu_i^B\left(h_i^{\mathrm{d}}\right)\mu_j^B\left(h_j^{\mathrm{d}}\right)\right). \qquad (2.42)
\end{aligned}$$

The approximate hospital expected daily profit is

$$\Pi_{\mathrm{A}}\left(\mathbf{a}^e,\mathbf{h}^{\mathrm{d}}\right)=\sum_{i=1}^{N}a_i^e\mathcal{A}_i\left(\mathbf{h}^{\mathrm{d}}\right)-c_F\left(\sum_{i=1}^{N}a_i^e\mu_i^F\left(h_i^{\mathrm{d}}\right)\right)^2-c_B\left(\sum_{i=1}^{N}a_i^e\mu_i^B\left(h_i^{\mathrm{d}}\right)\right)^2+\sum_{i=1}^{N}\mathcal{B}_i\left(\mathbf{h}^{\mathrm{d}}\right), \tag{2.43}$$

and the hospital's optimization problem can be expressed as

$$\max_{\mathbf{a}^e,\mathbf{h}^{\mathrm{d}}}\left(\Pi_{\mathrm{A}}\left(\mathbf{a}^e,\mathbf{h}^{\mathrm{d}}\right)\right) \tag{2.44}$$

$$\text{s.t. } 0\leq a_i^e\leq E_i,\ i=1,\ldots,N, \tag{2.45}$$

$$h_i^{\min}\leq h_i^{\mathrm{d}}\leq 1,\ i=1,\ldots,N. \tag{2.46}$$

2.4. Optimal Elective and Discharge Policies: Single-Procedure Setting

In this section we analyze the optimization problem (2.44)-(2.46) for a single-specialty hospital. In the following analysis we drop the surgical procedure index.

Proposition 8 (Optimal Policies for a Single-Specialty Hospital) *For a single-specialty hospital, let*

$$\mathcal{M}\left(h^{\mathrm{d}}\right)=c_F\left(\mu^F\left(h^{\mathrm{d}}\right)\right)^2+c_B\left(\mu^B\left(h^{\mathrm{d}}\right)\right)^2, \tag{2.47}$$

$$\mathcal{V}\left(h^{\mathrm{d}}\right)=c_F\left(\sigma^F\left(h^{\mathrm{d}}\right)\right)^2+c_B\left(\sigma^B\left(h^{\mathrm{d}}\right)\right)^2, \tag{2.48}$$

and

$$f^e\left(h^{\mathrm{d}}\right)=\max\left(0,\min\left(\frac{R-\mathcal{V}\left(h^{\mathrm{d}}\right)}{2\mathcal{M}\left(h^{\mathrm{d}}\right)}-\mu^u,E\right)\right). \tag{2.49}$$

Then, the discharge level and the number of elective procedures that optimize (2.44)-(2.46) are given by

$$\begin{aligned}\hat{h}^{\mathrm{d}}=\underset{h^{\mathrm{d}}\in[h^{\min},1]}{\arg\max}&\left(\left(f^e\left(h^{\mathrm{d}}\right)+\mu^u\right)\left(R-\mathcal{V}\left(h^{\mathrm{d}}\right)\right)\right.\\&\left.-\left(\left(f^e\left(h^{\mathrm{d}}\right)+\mu^u\right)^2+(\sigma^u)^2\right)\mathcal{M}\left(h^{\mathrm{d}}\right)\right)\end{aligned} \tag{2.50}$$

and

$$\hat{a}^e = f^e\left(\hat{h}^{\mathrm{d}}\right). \tag{2.51}$$

Proposition 8 connects the optimal number of elective procedures and the optimal discharge threshold level, and expresses the optimal discharge threshold level in terms of a solution of a one-dimensional optimization problem, for arbitrary readmission function $p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)$. Figure 8 illustrates how the optimal elective and discharge policies change with the expected cost of using "front-end" and "backroom" resources relative to the reimbursement rate.

To determine realistic ranges for these values, we consider the amount hospitals charge for knee-replacement and hip-replacement surgeries, two common procedures that correspond to DRG=470 (Centers for Medicare and Medicaid Services, 2019). In an analysis of the charges for total hip arthoplasty, Bertin (2005) finds that for an inpatient procedure, the total average reimbursement was \$13,950, with an average billed amount for surgery of \$2,874 and an average charge for nursing and room usage of \$4,404. These correspond to 21 and 32 percent of the hospital's total reimbursement, respectively. (Note that the largest charge for such surgeries is typically the implant charge, which is not included in either of these numbers; in this study, the average implant charge was \$12,182, or 87 percent of the total reimbursement.) King et al. (2011) and Richter and Diduch (2017) conduct similar studies on the charges for knee arthroplasty. These studies suggest that the charges corresponding to OR and recovery bed usage relative to the reimbursement rate range from 20 to 45 percent, and that backroom resource usage results in higher charges than that of front-end resources. However, with the exception of King et al. (2011), these studies only consider the amount the hospital bills, not the actual cost of using each resource. Indeed, even for procedures as common as these, actual costs are not well known, and prices and reimbursement rates can depend greatly on the hospital and payer (Blue Cross Blue Shield, 2015; Evans, 2018). We assume the charged amounts are higher than the true costs, but that the allocation of charges for each resource is proportional to the distribution of costs. Further study limitations such as small sample sizes and differing categorization of costs,

in addition to the variety of procedures that correspond to DRG code 470, suggest that the range of expected resource costs to reimbursement rates could be even wider. This is reflected in the Figure.

Note the general trends displayed in the Figure. On the one hand, as the recovery beds become more expensive to manage, the hospital curtails patient LOS and, to a lesser degree, limits the inflow of elective procedures. On the other hand, as the costs associated with OR capacity increase, the hospital, while limiting the patient inflow, allows for longer recovery times. Thus, while both types of costs produce similar "inflow" control, the discharge policies contrast sharply depending on which type of resource is more costly. Figure 9 presents a more detailed picture of the sensitivity of the optimal discharge and case-mix policies to changes in front-end and backroom costs. In particular, as recovery bed costs increase, the hospital should first limit the number of elective procedures. As backroom costs increase, the hospital should also discharge patients prior to a full recovery. If instead the OR becomes more costly to utilize, the hospital's optimal first response is to increase patients' LOS. As costs continue to increase, the hospital should limit the number of elective patients, as well. The Figure also suggests that there are ranges of cost parameters that result in four "limiting" patient-flow management policies: "full recovery" ($\hat{h}^{\mathrm{d}} = 1$) vs. "minimal recovery" ($\hat{h}^{\mathrm{d}} \to h^{\min}$) discharge policies in combination with "max-elective" ($\hat{a}^e = E$) vs. "no-elective" ($\hat{a}^e = 0$) inflow policies.

Below we describe sufficient conditions for the optimality of these "extreme" policies.

Proposition 9 (Sufficient Conditions for the Optimality of "Extreme" Policies) *Consider a single-specialty hospital with* $\gamma^h < 1$.

a) Suppose that

$$R \leq c_F \left(2\mu^u \left(\mathrm{E}\left[S\right] \right)^2 + \mathrm{Var}\left[S\right] \right). \tag{2.52}$$

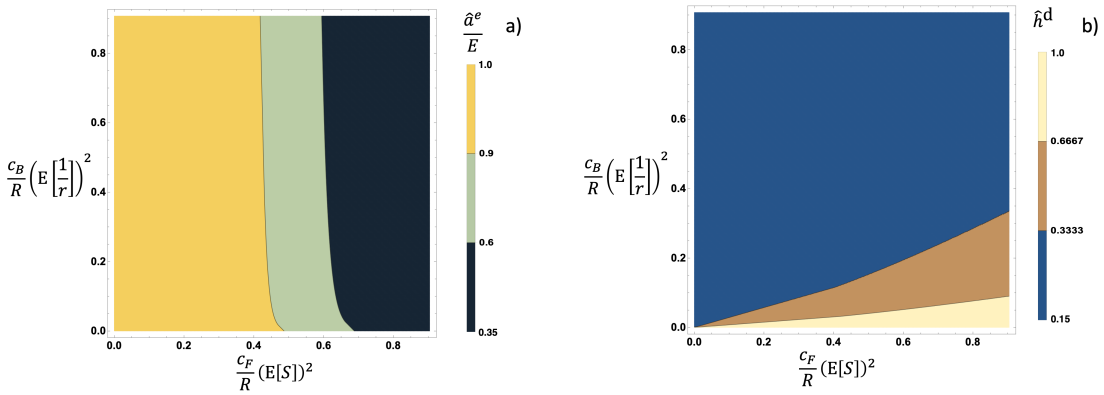

Figure 8: Optimal number of elective procedures (a) and the optimal discharge threshold (b) as functions of cost parameters c_B and c_F ($h^{\min} = 0.15$, $\mathrm{E}[S] = 106.98$, $\mathrm{Var}[S] = 615.54$, $\mathrm{E}\left[\frac{1}{r}\right] = 12.963$, $\alpha = 1.639$, $\beta = 1.893$, $r^{\min} = 0.008$, $r^{\max} = 0.229$, $\theta = 0.023$, $\mu^u = 0.1$, $\sigma^u = 0.2$).

Then, the optimal daily number of elective procedures is $\hat{a}^e = 0$. *If, in addition,*

$$c_B \leq c_F \gamma^l \left(\frac{3 \min\left(\mu^u, (\mu^u)^2 + (\sigma^u)^2\right) \left(\mathrm{Var}\left[S\right] + \left(\mathrm{E}\left[S\right]\right)^2\right)}{\mu^u \left(2G^r + \left(0.25 + \gamma^h\right) H^r\right) + 8\left((\mu^u)^2 + (\sigma^u)^2\right)\left(\mathrm{E}\left[\frac{1}{r}\right]\right)^2} \right), \tag{2.53}$$

where G^r *and* H^r *are defined in (2.35), the optimal patient discharge level is* $\hat{h}^{\mathrm{d}} = 1$. *On the other hand, if*

$$G^r \geq \frac{\gamma^h}{1 - \gamma^h} \left(4 \left(\frac{(\mu^u)^2 + (\sigma^u)^2}{\mu^u} \right) \left(\mathrm{E}\left[\frac{1}{r}\right] \right)^2 + H^r \right) \tag{2.54}$$

and

$$c_B \geq c_F \left(\frac{3\mathrm{Var}\left[S\right] + \left(\mathrm{E}\left[S\right]\right)^2 \left(1 + 4\left(\frac{(\mu^u)^2 + (\sigma^u)^2}{\mu^u}\right)\right)}{\left(\frac{1-\gamma^h}{\gamma^h}\right) G^r - H^r - 4\left(\frac{(\mu^u)^2 + (\sigma^u)^2}{\mu^u}\right)\left(\mathrm{E}\left[\frac{1}{r}\right]\right)^2} \right) \tag{2.55}$$

the optimal patient discharge level $\hat{h}^{\mathrm{d}} = h^{\min}$.

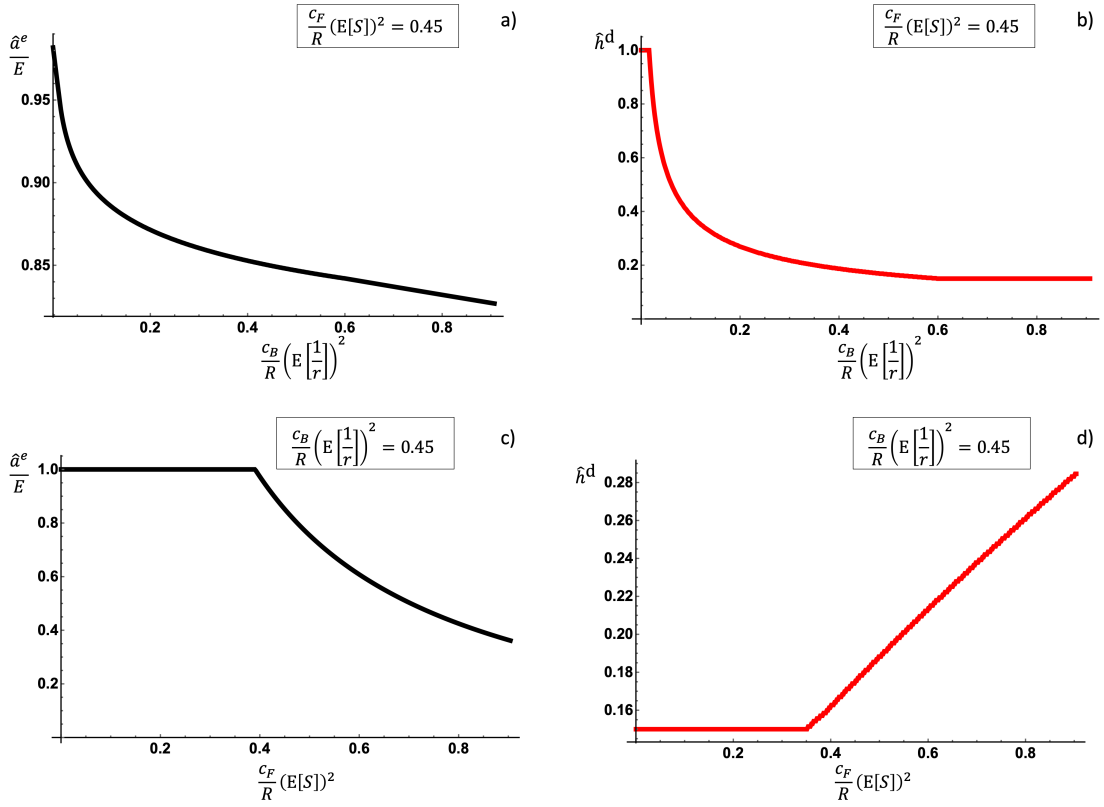

Figure 9: Optimal number of elective procedures and the optimal discharge threshold as functions of cost parameters c_B (a, b) and c_F (c, d) ($h^{\min} = 0.15$, $\mathrm{E}[S] = 106.98$, $\mathrm{Var}[S] = 615.54$, $\mathrm{E}\left[\frac{1}{r}\right] = 12.963$, $\alpha = 1.639$, $\beta = 1.893$, $r^{\min} = 0.008$, $r^{\max} = 0.229$, $\theta = 0.023$, $\mu^u = 0.1$, $\sigma^u = 0.2$).

b) Suppose, for $E+\mu^u \geq 1$, that

$$R \geq c_F \left(8\left(\mathrm{E}\left[S\right]\right)^2 \left(E+\mu^u\right) + \mathrm{Var}\left[S\right] + \left(\frac{3\mathrm{Var}\left[S\right] + \left(\mathrm{E}\left[S\right]\right)^2}{2\mathrm{E}\left[S\right]} \right)^2 \right) + c_B \left(8\left(\mathrm{E}\left[\frac{1}{r}\right]\right)^2 \left(E+\mu^u\right) + G^r + \frac{\left(G^r+H^r\right)^2}{4H^r} \right). \tag{2.56}$$

Then, the optimal daily number of elective procedures is $\hat{a}^e = E$. If, in addition,

$$c_B \leq c_F \gamma^l \left(\frac{3\left(\mathrm{Var}\left[S\right] + \left(\mathrm{E}\left[S\right]\right)^2\right)}{2G^r + \left(0.25+\gamma^h\right)H^r + 8\left(\frac{\left(E+\mu^u\right)^2+\left(\sigma^u\right)^2}{E+\mu^u}\right)\left(\mathrm{E}\left[\frac{1}{r}\right]\right)^2} \right), \tag{2.57}$$

the optimal patient discharge level is $\hat{h}^{\mathrm{d}} = 1$. On the other hand, if

$$G^r \geq \frac{\gamma^h}{1-\gamma^h} \left(4\left(\frac{\left(E+\mu^u\right)^2 + \left(\sigma^u\right)^2}{\left(E+\mu^u\right)} \right) \left(\mathrm{E}\left[\frac{1}{r}\right]\right)^2 + H^r \right) \tag{2.58}$$

and

$$c_B \geq c_F \left(\frac{3\mathrm{Var}\left[S\right] + \left(\mathrm{E}\left[S\right]\right)^2 \left(1 + 4\left(\frac{\left(E+\mu^u\right)^2+\left(\sigma^u\right)^2}{\left(E+\mu^u\right)}\right)\right)}{\left(\frac{1-\gamma^h}{\gamma^h}\right)G^r - H^r - 4\left(\frac{\left(E+\mu^u\right)^2+\left(\sigma^u\right)^2}{\left(E+\mu^u\right)}\right)\left(\mathrm{E}\left[\frac{1}{r}\right]\right)^2} \right) \tag{2.59}$$

the optimal patient discharge level $\hat{h}^{\mathrm{d}} = h^{\min}$.

The results of Proposition 9, illustrated in Figure 10, provide important insights on the revenue-cost trade-offs faced by the hospital. On the one hand, if the hospital compensation rate R is insufficient to cover the procedure costs, the hospital opts for the policy that allows the access to hospital resources only for emergency patients. Whether these emergency patients will be allowed to fully utilize the recovery bed resources depends, however, on the relative values of the backroom and front-end costs, with the discharge policy shifting from "complete recovery" if the front-end costs dominate, to "limited recovery" if the backroom costs become prominent. On the other hand, under generous compensation, the hospital allows as many elective procedures as possible, augmenting this policy by the spectrum of

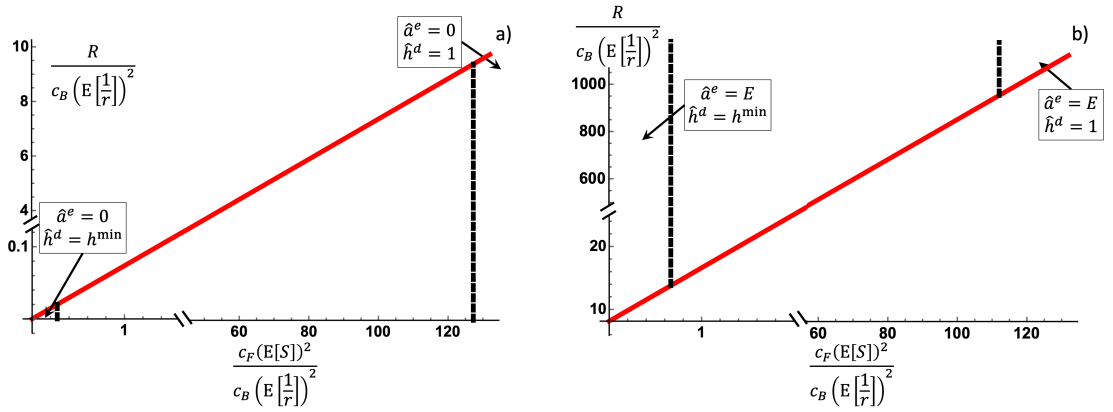

Figure 10: Sufficient conditions for optimality of "extreme" policies: a) $\hat{a}^e = 0$, $\hat{h}^{\mathrm{d}} = 1$ and $\hat{a}^e = 0$, $\hat{h}^{\mathrm{d}} = h^{\min}$, b) $\hat{a}^e = E$, $\hat{h}^{\mathrm{d}} = 1$ and $\hat{a}^e = E$, $\hat{h}^{\mathrm{d}} = h^{\min}$ ($h^{\min} = 0.5$, $\mathrm{E}[S] = 106.98$, $\mathrm{Var}[S] = 615.54$, $\mathrm{E}\left[\frac{1}{r}\right] = 12.963$, $\alpha = 1.639$, $\beta = 1.893$, $r^{\min} = 0.008$, $r^{\max} = 0.229$, $\theta = 0.023$, $\mu^u = 0.01$, $\sigma^u = 0.02$).

discharge management approaches ranging from "complete recovery" to "limited recovery," depending on the relative values of the front-end and the backroom cost parameters.

2.5. Front-End and Siloed Policies

In real hospital settings, the optimal management of patient inflows and discharges may represent a hard-to-achieve theoretical limit of efficiency. In practice, front-end and backroom resources may be managed by different decision makers, and the attainment of hospital-wide optimality relies on the ability of hospital management to achieve a perfect alignment of their objectives and actions. On the one hand, the front-end resources are often considered to be the main engines of hospital revenue generation, and their utilization is managed by surgeons, actors with substantial influence on hospital operations. On the other hand, the backroom resources are often treated as main cost centers, and their management is typically placed in the hands of attending physicians (that may or may not be the surgeons who performed the actual procedure) and head nurses. In addition, the actual patient inflow and discharge decisions are naturally separated in time, a factor that can in practice increase the cost of coordinating these decisions.

In this section we focus on understanding the impact of these coordination costs by considering two alternative policies that are likely to reflect the realities of patient flow management

in hospital settings.

The first policy places the entire decision-making power in the hands of surgeons and assumes that the elective surgery decisions and patient discharge decisions are made exclusively on the basis of front-end costs. Specifically, under this policy, the best values for a^e and h^{d} are chosen by setting $c_B = 0$. We use the term "front-end" (FE) to designate such a policy. The FE policy is designed to approximate the decisions in hospital settings operating under the strong influence of surgeons who ignore the impact of the front-end decisions on the backroom resources. Formally, the elective-discharge decisions under this policy, $\left(\hat{\mathbf{a}}^e_{\mathrm{FE}}, \hat{\mathbf{h}}^{\mathrm{d}}_{\mathrm{FE}}\right)$, are set to optimize (2.44)-(2.46) under $c_B = 0$:

$$\left(\hat{\mathbf{a}}^e_{\mathrm{FE}}, \hat{\mathbf{h}}^{\mathrm{d}}_{\mathrm{FE}}\right) = \underset{\mathbf{a}^e \in [\mathbf{0},\mathbf{E}], \mathbf{h}^{\mathrm{d}} \in [\mathbf{0},\mathbf{1}]}{\arg\max} \left(\Pi_{\mathrm{A}} \left(\mathbf{a}^e, \mathbf{h}^{\mathrm{d}} \right) \middle| c_B = 0 \right). \tag{2.60}$$

The following result provides an analytical description of the hospital's decisions under the FE policy.

Proposition 10 (Front-End Policy: Optimal Portfolio and Discharge Decisions) *Without loss of generality, assume*

$$\frac{R_1 - c_F \mathrm{Var}\left[S_1\right]}{\mathrm{E}\left[S_1\right]} \geq \frac{R_2 - c_F \mathrm{Var}\left[S_2\right]}{\mathrm{E}\left[S_2\right]} \geq \ldots \geq \frac{R_N - c_F \mathrm{Var}\left[S_N\right]}{\mathrm{E}\left[S_N\right]}. \tag{2.61}$$

Under the FE policy, the hospital allows all patients to completely recover,

$$\hat{\mathbf{h}}^{\mathrm{d}}_{\mathrm{FE}} = (1, \ldots, 1), \tag{2.62}$$

and selects the elective portfolio as follows. For

$$\frac{R_1 - c_F \left(\mathrm{Var}\left[S_1\right] + 2\mathrm{E}\left[S_1\right] \left(\sum_{j=1}^{N} \mu_j^u \mathrm{E}\left[S_j\right] \right) \right)}{2c_F \left(\mathrm{E}\left[S_1\right]\right)^2} < E_1, \tag{2.63}$$

the optimal elective portfolio is given by

$$(\hat{a}^e_{\mathrm{FE}})_1 = \frac{\left(R_1 - c_F\left(\mathrm{Var}\left[S_1\right] + 2\mathrm{E}\left[S_1\right]\left(\sum_{j=1}^N \mu_j^u \mathrm{E}\left[S_j\right]\right)\right)\right)^+}{2c_F\left(\mathrm{E}\left[S_1\right]\right)^2}, \tag{2.64}$$

$$(\hat{a}^e_{\mathrm{FE}})_i = 0, \quad i = 2, \ldots, N. \tag{2.65}$$

For

$$\frac{R_N - c_F\left(\mathrm{Var}\left[S_N\right] + 2\mathrm{E}\left[S_N\right]\left(\sum_{j=1}^N \mu_j^u \mathrm{E}\left[S_j\right]\right)\right)}{2c_F \mathrm{E}\left[S_N\right]} > \sum_{j=1}^N E_j \mathrm{E}\left[S_j\right], \tag{2.66}$$

the optimal elective portfolio is given by

$$(\hat{a}^e_{\mathrm{FE}})_i \quad = \quad E_i, i = 1, \ldots, N. \tag{2.67}$$

Finally, for

$$\begin{aligned} E_1 &\leq \frac{R_1 - c_F\left(\mathrm{Var}\left[S_1\right] + 2\mathrm{E}\left[S_1\right]\left(\sum_{j=1}^N \mu_j^u \mathrm{E}\left[S_j\right]\right)\right)}{2c_F\left(\mathrm{E}\left[S_1\right]\right)^2}, \\ \frac{R_N - c_F\left(\mathrm{Var}\left[S_N\right] + 2\mathrm{E}\left[S_N\right]\left(\sum_{j=1}^N \mu_j^u \mathrm{E}\left[S_j\right]\right)\right)}{2c_F \mathrm{E}\left[S_N\right]} &\leq \sum_{j=1}^N E_j \mathrm{E}\left[S_j\right], \end{aligned} \tag{2.68}$$

let

$$i^*_{\mathrm{FE}} = 1 + \max\left(i \in \{1, \ldots, N\} \left| \frac{R_i - c_F\left(\mathrm{Var}\left[S_i\right] + 2\mathrm{E}\left[S_i\right]\left(\sum_{j=1}^N \mu_j^u \mathrm{E}\left[S_j\right]\right)\right)}{2c_F \mathrm{E}\left[S_i\right]} > \sum_{j=1}^i E_j \mathrm{E}\left[S_j\right]\right.\right). \tag{2.69}$$

Then, the optimal elective portfolio is given by

$$(\hat{a}^{e}_{\mathrm{FE}})_i = \begin{cases} E_i, & i = 1, \ldots, i^*_{\mathrm{FE}} - 1, \\ \frac{R_i - c_F \mathrm{Var}[S_i]}{2c_F(\mathrm{E}[S_i])^2} - \frac{\sum_{j=1}^{N}(\mu^u_j \mathrm{E}[S_j]) + \sum_{j=1}^{i-1}(E_j \mathrm{E}[S_j])}{\mathrm{E}[S_i]}, & i = i^*_{\mathrm{FE}}, \\ 0, & i = i^*_{\mathrm{FE}} + 1, \ldots, N. \end{cases} \tag{2.70}$$

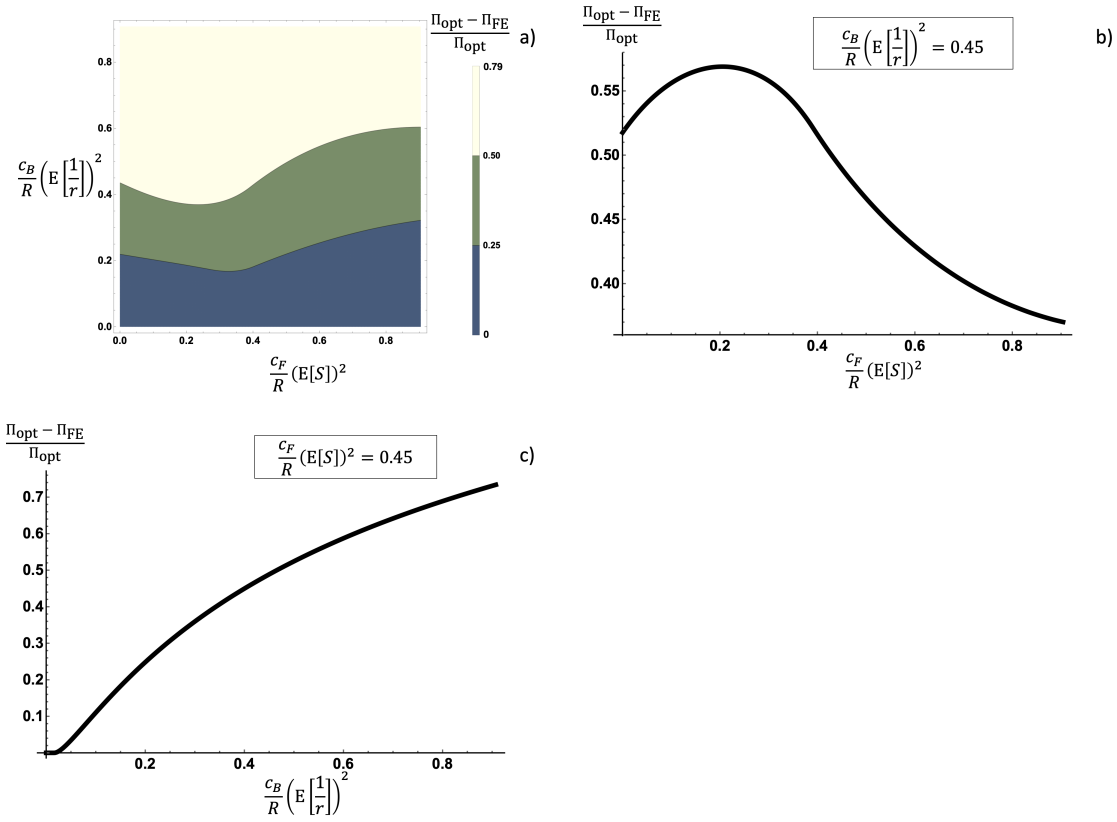

Figure 11: The relative profit gap resulting from the use of the FE policy as a function of cost parameters c_B and c_F ($h^{\min} = 0.15$, $\mathrm{E}[S] = 106.98$, $\mathrm{Var}[S] = 615.54$, $\mathrm{E}\left[\frac{1}{r}\right] = 12.963$, $\alpha = 1.639$, $\beta = 1.893$, $r^{\min} = 0.008$, $r^{\max} = 0.229$, $\theta = 0.023$, $\mu^u = 0.1$, $\sigma^u = 0.2$).

Figure 11 illustrates the potential profit loss resulting from the use of the FE policy as a function of the front-end and backroom cost parameters. As expected, the front-end approach works well in settings where backroom costs are low or, if OR costs are sufficiently high, where surgery-related costs dominate the hospital cost structure. The Figure also highlights the non-monotonicity of the profit gap: for smaller front-end cost values, the

FE policy initially does worse as front-end usage costs increase. In this region, under the optimal policy, patients are discharged as quickly as possible, i.e., $\hat{h}^{\mathrm{d}} = h^{\min}$, resulting in a higher probability that the hospital will incur costs from a readmission. Thus, for low OR costs, the profit under the optimal policy is decreasing in OR costs at a significantly faster rate than under the FE policy.

The second policy, that we call "siloed" (SI), reflects the setting in which the actors managing the backroom resources retain some control over their utilization and are able to select the patient discharge level that minimizes, under any elective policy, the backroom costs. Specifically, we assume that, under the SI policy, the surgeons play the role of the "principal" that determines the portfolio of elective procedures, and the "backroom" managers play the role of an agent that responds to the principal's actions by setting the discharge policies. The SI policy reflects the reality where the elective portfolio decisions are "imposed" on the actors that manage backroom resources within a principal-agent framework, and the front-end profit and backroom costs are managed separately by the principal and the agent, respectively. Formally, the elective-discharge decisions under this policy, $\left(\hat{\mathbf{a}}^e_{\mathrm{SI}}, \hat{\mathbf{h}}^{\mathrm{d}}_{\mathrm{SI}}\right)$, are determined as follows:

$$\hat{\mathbf{a}}^e_{\mathrm{SI}} = \underset{\mathbf{a}^e \in [\mathbf{0}, \mathbf{E}]}{\arg\max} \left(\Pi_{\mathrm{A}} \left(\mathbf{a}^e, \mathbf{h}^{\mathrm{d}}_r \left(\mathbf{a}^e \right) \right) \middle| c_B = 0 \right), \tag{2.71}$$

$$\hat{\mathbf{h}}^{\mathrm{d}}_{\mathrm{SI}} = \mathbf{h}^{\mathrm{d}}_r \left(\hat{\mathbf{a}}^e_{\mathrm{SI}} \right), \tag{2.72}$$

where

$$\begin{aligned} \mathbf{h}^{\mathrm{d}}_r \left(\mathbf{a}^e \right) = \underset{\mathbf{h}^{\mathrm{d}} \in [\mathbf{h}^{\min}, \mathbf{1}]}{\arg\min} & \left(\sum_{i=1}^{N} \left(\left(\sigma^B_i \left(h^{\mathrm{d}}_i \right) \right)^2 + 2 \mu^B_i \left(h^{\mathrm{d}}_i \right) \left(\sum_{k=1}^{N} \mu^u_k \mu^B_k \left(h^{\mathrm{d}}_k \right) \right) \right) a^e_i + \left(\sum_{i=1}^{N} a^e_i \mu^B_i \left(h^{\mathrm{d}}_i \right) \right)^2 \right. \\ & + \sum_{i=1}^{N} \left(\left(\mu^u_i \left(\sigma^B_i \left(h^{\mathrm{d}}_i \right) \right)^2 + (\mu^u_i)^2 + (\sigma^u_i)^2 \right) \left(\mu^B_i \left(h^{\mathrm{d}}_i \right) \right)^2 \right. \\ & \left. \left. + 2 \sum_{j \neq i} \left(\mu^u_i \mu^u_j + \rho^u_{ij} \sigma^u_i \sigma^u_j \right) \mu^B_i \left(h^{\mathrm{d}}_i \right) \mu^B_j \left(h^{\mathrm{d}}_j \right) \right) \right). \end{aligned} \tag{2.73}$$

The following proposition describes the patient discharge and elective portfolio decisions under the SI policy.

Proposition 11 (Siloed Policy: Optimal Portfolio and Discharge Decisions) *For each procedure $i = 1, \ldots, N$, define*

$$\bar{\mathcal{A}}_i^{\mathrm{SI}} = R_i - c_F \left(2\mathrm{E}\left[S_i\right]\left(1 + p_i^{\max}\right) \left(\sum_{j=1}^{N} \mu_j^u \mathrm{E}\left[S_j\right]\left(1 + p_j^{\max}\right) \right) + \left(1 + 3p_i^{\max}\right) \mathrm{Var}\left[S_i\right] + p_i^{\max}\left(1 - p_i^{\max}\right)\left(\mathrm{E}\left[S_i\right]\right)^2 \right), \tag{2.74}$$

and, without loss of generality, assume that

$$\frac{\bar{\mathcal{A}}_1^{\mathrm{SI}}}{\mathrm{E}\left[S_1\right]\left(1 + p_1^{\max}\right)} \geq \cdots \geq \frac{\bar{\mathcal{A}}_N^{\mathrm{SI}}}{\mathrm{E}\left[S_N\right]\left(1 + p_N^{\max}\right)}. \tag{2.75}$$

Under the SI policy, when $\gamma_i^h \leq \frac{G_i^r}{G_i^r + H_i^r}$ for all $i = 1, \ldots, N$, the hospital discharges all patients as soon as possible,

$$\hat{\mathbf{h}}_{\mathrm{SI}}^{\mathrm{d}} = \left(h_1^{\min}, \ldots, h_N^{\min}\right), \tag{2.76}$$

and selects the elective portfolio as follows. For

$$\frac{\bar{\mathcal{A}}_1^{\mathrm{SI}}}{2c_F\left(\mathrm{E}\left[S_1\right]\left(1 + p_1^{\max}\right)\right)^2} < E_1, \tag{2.77}$$

the optimal elective portfolio is given by

$$\left(\hat{a}_{\mathrm{SI}}^e\right)_1 = \frac{\bar{\mathcal{A}}_1^{\mathrm{SI}}}{2c_F\left(\mathrm{E}\left[S_1\right]\left(1 + p_1^{\max}\right)\right)^2}, \tag{2.78}$$

$$\left(\hat{a}_{\mathrm{SI}}^e\right)_i = 0, \quad i = 2, \ldots, N. \tag{2.79}$$

For

$$\frac{\bar{\mathcal{A}}_N^{\mathrm{SI}}}{2c_F \mathrm{E}\left[S_N\right]\left(1+p_N^{\max}\right)} > \sum_{j=1}^{N} E_j \mathrm{E}\left[S_j\right]\left(1+p_j^{\max}\right), \tag{2.80}$$

the optimal elective portfolio is given by

$$(\hat{a}_{\mathrm{SI}}^e)_i \quad = \quad E_i, i = 1, \ldots, N. \tag{2.81}$$

Finally, for

$$E_1 \leq \frac{\bar{\mathcal{A}}_1^{\mathrm{SI}}}{2c_F \left(\mathrm{E}\left[S_1\right]\left(1+p_1^{\max}\right)\right)^2} \quad \textit{and} \quad \frac{\bar{\mathcal{A}}_N^{\mathrm{SI}}}{2c_F \mathrm{E}\left[S_N\right]\left(1+p_N^{\max}\right)} \leq \sum_{j=1}^{N} E_j \mathrm{E}\left[S_j\right]\left(1+p_j^{\max}\right), \tag{2.82}$$

let

$$i_{\mathrm{SI}}^* = 1 + \max\left(i \in \{1, \ldots, N\} \,\middle|\, \frac{\bar{\mathcal{A}}_i^{\mathrm{SI}}}{2c_F \mathrm{E}\left[S_i\right]\left(1+p_i^{\max}\right)} > \sum_{j=1}^{i} E_j \mathrm{E}\left[S_j\right]\left(1+p_j^{\max}\right) \right). \tag{2.83}$$

Then, the optimal elective portfolio is given by

$$(\hat{a}_{\mathrm{SI}}^e)_i = \begin{cases} E_i, & i = 1, \ldots, i_{\mathrm{SI}}^* - 1, \\ \frac{\bar{\mathcal{A}}_i^{\mathrm{SI}}}{2c_F\left(\mathrm{E}[S_i]\left(1+p_i^{\max}\right)\right)^2} - \frac{\sum_{j=1}^{i-1}\left(E_j \mathrm{E}[S_j]\left(1+p_j^{\max}\right)\right)}{\mathrm{E}[S_i]\left(1+p_i^{\max}\right)}, & i = i_{\mathrm{SI}}^*, \\ 0, & i = i_{\mathrm{SI}}^* + 1, \ldots, N. \end{cases} \tag{2.84}$$

Figure 12 shows how the use of the SI policy impacts the hospital's profit under various values of front-end and backroom cost parameters. Note that the SI policy displays performance that is somewhat complementary to that of the FE policy, with near-optimal profits generated in settings where front-end costs are low or the backroom costs are sufficiently high.

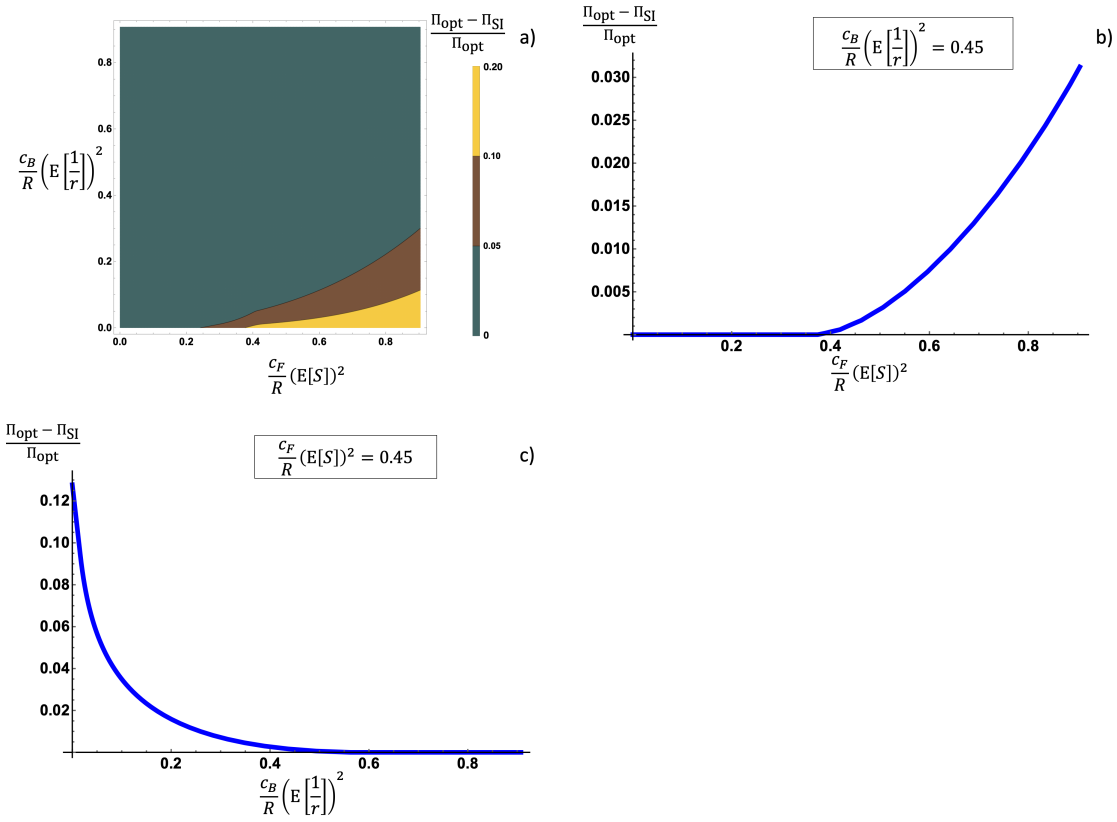

Figure 12: The relative profit gap resulting from the use of the SI policy as a function of cost parameters c_B and c_F ($h^{\min} = 0.15$, $\mathrm{E}[S] = 106.98$, $\mathrm{Var}[S] = 615.54$, $\mathrm{E}\left[\frac{1}{r}\right] = 12.963$, $\alpha = 1.639$, $\beta = 1.893$, $r^{\min} = 0.008$, $r^{\max} = 0.229$, $\theta = 0.023$, $\mu^u = 0.1$, $\sigma^u = 0.2$).

2.6. Discussion

In many hospital environments, the autonomy of decision-makers throughout the patient care process is deeply ingrained in the hospital's culture and is taken as given. Such autonomy is a manifestation of organizational costs associated with maintaining coordinated flow of information and of concerted focus on hospital-wide priorities. Yet, in the absence of coordinated decision-making, hospitals may fail to account for the interconnected nature of patient flow management decisions, potentially resulting in a lower quality of patient care, underutilized resources, and suboptimal financial performance. A combination of sustained pressure on hospitals to maintain their financial viability and proliferation of technology-driven coordination solutions is accentuating the need for careful assessment of costs and benefits of coordination. In this chapter, we present a model that is designed to estimate the potential benefits of coordinated decision-making on two main aspects of patient flow management: the case-mix of elective procedures and patient discharge policies. Our analysis relies on a novel approach to modeling patient recovery and readmission processes that allows for closed-form asymptotic characterization of the first two moments of the utilization of main hospital resources, and, consequently, of hospital expected daily profit, for any combination of elective portfolio and patient discharge decisions. For a single-specialty hospital, we derive patient flow management policies that maximize the hospital's expected profits associated with two main resource groups: "front-end" (e.g., operating rooms) and "backroom" (e.g., recovery beds).

We leverage our model to provide guidance on the settings where the benefits from patient flow coordination may be especially pronounced as well as the settings where those benefits are modest. In particular, we compare the hospital profits under perfect coordination with those achieved under two decentralized policies: the "front-end" policy, where surgeons determine both the patient portfolio and the discharge policy exclusively on the basis of operating room costs, and the "siloed" policy, where patient discharge decisions are driven only by the backroom costs, and are used as the basis for the elective portfolio decisions. Our results establish that, if the existing decision-making process is similar to a front-end

policy, implementing a coordinated decision-making process is beneficial when recovery bed costs are sufficiently high relative to operating room costs, even if recovery bed costs do not dominate the cost structure. On the other hand, hospitals using a siloed policy should move to a coordinated policy only if OR costs are sufficiently high and dominant. Given the high cost of changing established processes in a complex organization, these findings are particularly useful for administrators seeking to assess whether to implement such a change.

Our aim is to construct a parsimonious, strategic-level model of portfolio and discharge decisions that provides closed-form, qualitatively valid prescriptions to hospital managers. In designing such a model, we have to make several assumptions that, while facilitating the analytical tractability of our analysis, may affect the quantitative precision of our recommendations. First, we assume that for a given procedure, the random variables describing the procedure duration and patient recovery process are independent across days. At the same time, in our model, a readmission will result in the patient undergoing the exact same procedure and recovery process as during the original admission. In reality, in both cases these values are likely to be correlated, but are neither identical nor independent.

Second, we assume that any procedure will require an identical pair of resources: a "front-end" OR and a "backroom" recovery bed. Realistically, different procedures will require a combination of specialized and generic resources, which are unlikely to be perfectly categorized into two distinct categories. Third, our analysis focuses on the optimal decisions that are identical day-to-day, whereas a hospital administrator would realistically take a longer, i.e., weekly, view when planning, to account for daily fluctuations in demand, and in physician and resource availability.

Additionally, we take an "open-loop" modeling approach, in which establish the optimal strategic match between patient flows and the hospital's resource capacity. In future work, it will be useful to extend our analysis to consider patient flow management from a dynamic, "closed-loop" perspective.

Finally, we study a hospital that is reimbursed under a "bundled" compensation scheme, which provides an implicit incentive for hospitals to reduce readmissions. However, as the results of Propositions 7 and 9 indicate, this may not be enough to prevent early patient discharges in settings with high costs associated with backroom resources. In practice, the payers for hospital services have been augmenting bundled payments with additional penalties for excessive readmissions. Most notably, with the creation of the Hospital Readmissions Reduction Program (HRRP) in 2010, hospitals with excessive 30-day readmission rates for targeted conditions forfeit a percentage of their Medicare compensation. Extending our analysis to include such performance-based hospital compensation schemes is a promising direction for future research.

CHAPTER 3 : Coproduction in the Classroom: Optimally Allocating Incentives Between Teachers and Students

3.1. Introduction

For as long as coproduction has been studied, education has served as a quintessential example of a service that requires the participation of both the service provider and the consumer. Indeed, the notion of the consumer participating in service production was first discussed in Fuchs (1968), in which the author noted "the importance of the consumer as a cooperating agent in the production process." He subsequently observed that this is common knowledge for educators: "[productivity] in education, as every teacher knows, is determined largely by what the student contributes." Whitaker (1980) expanded on this in the context of public services, stating that "[c]oproduction is essential in services which seek to change the client," in contrast to those policies that can simply be deployed without active participation from citizens. He points out that "[t]he best of lesson plans, instructional materials, and teaching techniques cannot educate the child who will not learn." This has been recognized in the operations management literature as well: Karmarkar and Pitbladdo (1995) highlight education an example of a service where the "customer . . . participat[es] in service production" and which is "complex . . . in terms of output measurement."

Yet, despite the widespread recognition of the crucial role students play in their own education, until recently, this has not carried over to discussions of financial incentives in education. The debate on financial incentives in K-12 education has largely been centered on performance-based rewards for teachers alone, even as teachers' performance is measured based on student performance. More recently, consideration has been given to monetary incentives for students, but the focus has been on student incentives that are offered separately from teacher incentives. There are only a small number of studies that investigate the synergies from concurrently incentivizing teacher and students. This chapter aims to contribute to this nascent literature.

We analyze a model in which a school district allocates a monetary incentive between teachers and students at a school, and teachers and then students respond to this allocation by choosing the level of additional effort to exert. We seek to understand how teachers and students will respond to this incentive.

The rest of this chapter is organized as follows. In Section 3.2, we review the relevant literature. In Section 3.3, we present the model, followed by analysis in Section 3.4. In Section 3.5, we present a numerical study. Finally, we discuss our results and next steps in Section 3.6.

3.2. Literature Review

This work primarily lies at the intersection of two streams of research: the study of co-production in operations management (where it is also known as *joint production*) and the study of financial incentives in K-12 education. We also draw on research on the education production function.

The concept of coproduction was first introduced by Fuchs (1968), in an analysis of the growing service economy in the United States. It was linked to operations management by Chase (1978, 1981), when he used the idea of "consumer contact" to distinguish between different types service systems. Karmarkar and Pitbladdo (1995) expanded on this early work by considering the distinguishing features of a service system, including the degree of joint production, and investigating the implications on service design and competition. They illustrate this using a linear model of joint production. Xue and Harker (2002) introduce the notion of "customer efficiency," recognizing the need to take into account the caliber of a customer's performance when maximizing the quality of service delivery, especially for self-service activities. They propose a customer efficiency management framework. In Xue et al. (2007), the authors draw on the previous work to empirically study the relationship to customer efficiency to measures of firm performance. Xue and Field (2008) study contracts for collaborative services under uncertainty about service needs. They determine optimal pricing and self-service levels under such contracts, assuming efforts between the client

and service provider are substitutes. Roels et al. (2010) determine the preferred contract type for collaborative services given different characteristics of the service environment and assuming complementary efforts. The authors study a single-period Stackelberg game, where either party can offer a contract and, upon acceptance, both parties determine their optimal effort levels. Roels (2014) characterizes how a service provider should structure its coproductive system, based on the degree of standardization of the task, where a high degree of standardization is characterized by "predictably high marginal returns to effort." More recently, coproduction has been studied in the healthcare operations literature to model the importance of patient participation in the healthcare process. Andritsos and Tang (2018) incorporate coproduction when evaluating different hospital reimbursement schemes. Unlike previous work, they consider a three-level Stackelberg problem, in which a payer provides a contract, and the hospital and patient respond by determine their optimal effort levels.

Our work is most closely related to that of Andritsos and Tang (2018), although we study coproduction in the context of public education. We also introduce the element of information asymmetry, which differs from previous work. In particular, we assume that teachers do not know the cost of effort for students. This differs from, e.g. Roels et al. (2010), but is somewhat similar to Andritsos and Tang (2018), who assume the existence of two types of patients with different "effectiveness of the coproductive relation". However, in their model, the hospital can choose a different effort level for each type. In our case, since teachers "treat" an entire class of students concurrently, teachers must choose one effort level that takes into account their beliefs about the distribution of students' cost of effort. Finally, in our work, the customer (student) is directly affected by the allocation decision.

There is a large body of work on the impact of financial incentives in K-12 education in the United States. The majority of this work focuses on performance-based incentives for teachers, also commonly referred to as "merit pay." This literature goes back decades, and the findings tend to be mixed, with some researchers finding positive, if modest, effects

(e.g., Figlio and Kenny (2007); Dee and Wyckoff (2015); Chiang et al. (2017)) and some finding mixed or no significant impact (e.g., Eberts et al. (2002); Springer et al. (2011)) Understanding the efficacy of merit-based incentives is complicated by the varying level of incentives and durations of incentive programs, since many earlier experiments with incentives offered small rewards as part of a short-term program. (See Chapter 1 for a detailed discussion of this literature.)

There is a complementary but smaller body of work on student financial incentives. There are opposing theoretical arguments for the implementation of such programs. On the one hand, Oreopoulos (2007) finds that additional years of compulsory schooling are associated with multiple long-term benefits. The author observes that the fact that high school students drop out in spite of this is "consistent with recent studies in neurology and psychology that suggest adolescents are particularly predisposed to myopic behavior." Financial incentives are put forward as one way to address this blind spot. On the other hand, Gneezy et al. (2011) point out that an important shortcoming of incentives is that they can crowd out intrinsic motivation or change the subject's expectation about the task, thus having no or a negative effect. The authors note that programs are most beneficial when incentives are "for concrete tasks" and "offered to families and not to the children specifically," since such programs offer rewards for tasks that are clearly understood and do not crowd out students' intrinsic motivation to learn.

One way such programs have been implemented in practice are through "conditional cash transfer" (CCT) programs, means-tested programs that "attempt to encourage students to stay in school, rather than simply raising the compulsory school-leaving age" (Dearden et al., 2009). Such programs have been most widely implemented outside of the United States, although they have also been tested in the United States. The most notable example of this is Opportunity NYC, a CCT piloted by New York City from 2007 to 2010, which "tied cash rewards to pre-specified activities and outcomes in children's education, families' preventive health care, and parents' employment." An evaluation of the program found

that it had mixed effects in terms of educational outcomes: it had the most positive impact on high school students "who entered the study more academically prepared than their peers," resulting in higher graduation rates and a higher likelihood of passing the required number of New York State Regents exams for this cohort. However, there was no effect for elementary or middle school students, perhaps because of limited opportunities for rewards for these students (Riccio et al., 2013).

Other researchers have also found mixed results. Fryer (2011) studies the impact of student financial incentives on performance using three different incentive programs in three different cities. He finds no significant impact of such programs in each city on student performance, although in one city, there is a negative effect on the subgroup of students in bilingual classes and a positive effect on the subgroup of students in regular (non-bilingual) classes. The author hypothesizes that this may be due to students' "lack of knowledge of the education production function," i.e., how to improve their performance, and finds qualitative evidence of this. Bettinger (2012) studies a program in Coshocton, Ohio that provided cash incentives to elementary school students for "successful completion of their standardized testing," where students were randomly selected for the program. The author finds that the incentives led to a substantial improvement in students' math scores, but there were no improvements in test scores for other subjects. Levitt et al. (2016) study a program that provided financial incentives to high school freshman based on multiple measures of student achievement. They find that while the overall effects are modest, the program does have "large and significant impact among students on the threshold of meeting the achievement standard."

Finally, there is a nascent literature on the effect of providing monetary incentives to teachers and students simultaneously. Jackson (2010) studies the Advanced Placement Incentive Program (APIP), a unique program targeting underprivileged students in Texas, which "includes cash incentives for both teachers and students for each passing score earned on an advanced placement (AP) exam." The author finds that the program produces positive student achievement outcomes, while avoiding any negative distortions in student or

teacher behavior. For example, students did not "substitut[e] away from other advanced courses toward AP courses." The author concludes that it is likely the combination of both teacher and students incentives as well as "better instruction" that led to these improvements. Behrman et al. (2015) evaluate an experiment in Mexican high schools in which three different incentive schemes were implemented: incentivizing only teachers, incentivizing only students, and incentivizing teachers, students, and administrators. The authors state that this "is the first randomized control trial to incorporate incentive payments to both students and teachers." Note that for the first two incentive schemes, teachers and students were rewarded based on their individual performance, where teachers' performance was measured via the performance of the students they taught. However, in the third incentive scheme, students and teachers were rewarded both for their individual performance and for the performance of their peers. Thus, the third scheme explicitly encouraged active cooperation. The authors find that the joint incentives led to the largest average effects on student achievement. Finally, Todd and Wolpin (2018) develop and estimate a game theoretic model of student and teacher effort decisions, in which the optimal effort levels of students and teachers are a function of student attributes, such as previous knowledge level, and teacher attributes, such as instructional ability. Using data from Mexican high schools, they determine that the main reason for poor performance on end-of-year curriculum-based mathematics examinations was insufficient prior preparation. Our work is closest to that of Todd and Wolpin (2018), although they do not consider merit-based incentives in their model.

Underpinning our work is the idea that education is a system with a series of inputs that lead to some tangible output. This notion was first formally studied in the landmark government report *Equality of Educational Opportunity* (Coleman et al., 1966), more commonly known as the Coleman Report. The Coleman Report analyzed the impact of inputs, such as school and peer characteristics, on student achievement. This led to the conception of an "education production function," which drew on established work on firm production functions in economics. This was first introduced by Hanushek (1968), in which he compared

the educational production function of black sixth-graders to white sixth-graders in order to understand the relative impact of schools on educational achievement. This idea has been widely accepted and studied since then. (See, e.g., Levin, 1974; Hanushek, 1979, 1986; Rivkin et al., 2005, but this is only a partial list.)

Although an educational production function can take many forms, it typically combines inputs related to family background, school and teacher characteristics, peer characteristics, and a student's individual endowments to produce an output, namely, student achievement (Levin, 1974; Hanushek, 2020). Standardized test scores are a widely used, if imperfect, measure of student achievement. As stated by Hanushek (2020), "[t]est scores, or measures of cognitive skills more generally, have been interpreted as proxies for skills that are valued in the labor market and elsewhere and, as such, more immediate measures of human capital differences." Researchers have used a variety of functional forms when empirically estimating the education production function, among them both linear and log-log (Cobb-Douglas) forms. For example, Hanushek (1968) estimated both a linear and a log-log production function and found the log-log model to be superior.

In our work, we assume that the educational production function takes a Cobb-Douglas form, where students' and teachers' effort levels are the inputs and productivity is scaled based on the starting level of student achievement.

3.3. Model

In this section, we present a model that captures the interaction between teachers and students at a given school in the presence of merit-based incentives. We model this as a Stackelberg game with information asymmetry. Specifically, we assume that a school district has exogenously allocated a performance-based reward between the students and teachers to induce additional effort, where the probability of earning the reward is determined by both parties' effort levels. Observing this, teachers determine the total amount of effort they will exert, and subsequently, students determine the total amount of effort they will exert. Teachers have imperfect information about students' cost of exerting effort.

3.3.1. System States and Students' and Teachers' Actions

We consider a single-period game, where $t = 0$ corresponds to the beginning of the period and $t = 1$ corresponds to the end of the period. At time $t = 0, 1$, the school's state of proficiency β_t is either "proficient" (P) or "not proficient" (N). As in Chapter 1, we define proficient to mean that a sufficient fraction of the school's students satisfy or are on track to satisfy state-imposed learning standards. We assume that the state of proficiency is known by both the teachers and students at the start of the period ($t = 0$) and that it will be assessed and made known to both parties at the end of the period ($t = 1$) through a state-administered standardized assessment.

The probability that the school is in the proficient state at the end of the period is determined based on the initial proficiency state β_0 and the total effort exerted by teachers and by students. We assume that within each school, all teachers are homogeneous and act as a group, and similarly, all students are homogeneous and act as a group. Furthermore, teachers and students each exert two types of effort: individual effort and joint effort. For example, for a teacher, individual effort may be additional time spent lesson planning or designing instructional materials, whereas, for a student, individual effort may be additional time spent studying. Joint effort, on the other hand, requires that both parties exert effort, such as with classroom time or additional, after-school tutoring that the teacher may provide to students. We use e_t and e_s to denote teachers' and students' total effort levels, respectively. For both teachers and students, the joint effort level, captured by e_j, is a lower bound on these values that is exogenously determined and known to both players. Because there is a finite amount of effort either player can exert, we further assume that the maximum possible effort is normalized to be 1.

In modeling the transition probability of the schools' proficiency state, we assume this takes the form of a Cobb-Douglas production function, where the teachers' and students' effort levels are inputs, and the productivity of these inputs is scaled by a function of the initial state β_0. The Cobb-Douglas production function has been widely used both in the

operations management literature on coproduction (see, e.g., Leonard and Zivin (2005), Roels et al. (2010), and Andritsos and Tang (2018)) and for education production functions (see, e.g. Hanushek (1968)).

Assumption 9

$$Pr\left[\beta_1 = P|\beta_0\right] = g\left(\beta_0\right)\left(e_s\right)^a\left(e_t\right)^b, \tag{3.1}$$

where $e_j \leq e_s \leq 1$, $e_j \leq e_t \leq 1$, $a + b < 1$, *and* $0 \leq g\left(N\right) \leq g\left(P\right) \leq 1$, *with* $0 \leq e_j$ *and* $0 < a, b$.

Under Assumption 9, a and b capture the relative weights placed on student and teacher effort, respectively, in the transition probability, and there are decreasing returns to scale. The function $g\left(\beta_0\right)$ reflects the maximum probability of transitioning to the proficient state, which is higher when the school starts the period in the proficient state than when it starts in the not-proficient state. This is supported by the literature. Ding and Davison (2005) conduct a longitudinal study of math achievement and find that "disadvantaged students began with a lower initial achievement level, but their rates of gain were not significantly different from those of other students." They observe that "closing gaps once they emerge poses a particularly challenging task." In another article, Davison et al. (2004) note that "data suggest that groups of students seldom make up even small amounts of lost ground." Note that because students and teachers must exert a minimum level of effort, there is always a positive probability of achieving proficiency at the end of the period, as long as $g\left(\beta_0\right)$ is non-zero.

3.3.2. Cost of Effort and Timeline

We assume that the school district cannot directly observe teachers' and students' effort levels. However, in order to incentivize both teachers and students to exert high levels of effort, the district will offer each group a performance-based bonus. Teachers and students will receive the bonus if the school is in the proficient state at the end of the period ($t = 1$) but will otherwise forego it. The school district will allocate a budget M between performance-

based bonuses for teachers and students, which we denote by π_t and π_s, respectively. We take this reward allocation as given.

Teachers and students will observe the reward allocation π_t and π_s and respond sequentially, as in a Stackleberg game where the teacher is the leader and the student is the follower. Both the teacher and student incur a cost from exerting effort.

Assumption 10 *Teachers' and students' cost of effort is given by*

$$C_k(e_k) = c_k e_k, \quad k = s, t, \tag{3.2}$$

where $c_k \in \{c^h, c^l\}$ *and* $c^h > c^l$.

We assume that each player has a convex cost of effort, where this cost can either be high (c^h) or low (c^l). While each player knows their own cost of effort, players do not know each others' cost of effort. However, teachers believe that students' cost of exerting effort has the following distribution:

$$c_s = \begin{cases} c^h \text{ with probability } p_s^h \\ c^l \text{ with probability } 1 - p_s^h. \end{cases} \tag{3.3}$$

Then, upon observing the reward allocation, the teachers (leader) will move first and determine their total effort level e_t, taking into account their belief about the distribution of the students' cost of effort. Next, the students (follower) will observe both the total reward allocation and teachers' effort decision e_t and then determine their total effort level e_s. Students know their own cost of effort and, because they observe the teachers' effort decision, they do not need to form any beliefs about teachers' cost of exerting effort. Finally, the school will transition to a new proficiency state, according to (3.1). This state will be revealed at $t = 1$ when an assessment is administered. If the school is in the proficient state, both teachers and students will receive their merit-based incentives, π_t and π_s, respectively.

3.3.3. Students' and Teachers' Problems

Given the districts' allocation decision and the teachers' effort level decisions, students will choose a total effort level e_s that balances the cost of exerting this additional effort with their expected reward. In particular, students' effort decision satisfies the following:

$$e_s^*(e_t, \pi_s) = \underset{e_j \leq e_s \leq 1}{\arg\max} \{\pi_s Pr[\beta_1 = P|\beta_0] - c_s e_s\}. \tag{3.4}$$

Similarly, given the district's allocation decision and their beliefs about students' cost of exerting effort given in (3.3), teachers will choose a total effort level e_t that balances their expected costs and rewards. That is,

$$e_t^*(\pi_s, \pi_t) = \underset{e_j \leq e_t \leq 1}{\arg\max} \{\pi_t Pr[\beta_1 = P|\beta_0] - c_t e_t\}. \tag{3.5}$$

In summary, (3.4)-(3.5) describe the students' and teachers' maximization problems, where the teachers' beliefs about the cost of effort for students is given in (3.3). In the next section, we present a preliminary analysis of this problem as well as an outline of future directions for this work.

3.4. Analysis

In this section, we present an analysis of the Stackelberg game. We work backwards, first solving for the students' optimal effort decision and then for the teachers' optimal effort decision.

3.4.1. Optimal Effort Levels

The students' optimal effort satisfies (3.4). Recall that, prior to making their effort decision, students observe the school district's incentive allocation, π_s and π_t, and teachers' effort decision e_t, as well as the prespecified joint effort level e_j. Proposition 12 characterizes students' optimal effort level.

Proposition 12 (Students' Optimal Effort Levels) *For any given total effort level ex-*

erted by teachers (e_t) and allocation scheme chosen by the school district (π_s, π_t), the students' best response is to exert total effort $e_s^*(e_t, \pi_s)$, where

$$e_s^*(e_t, \pi_s, \pi_t) = \min\left\{\max\left\{e_j, \left(\frac{\pi_s g(\beta_0) a (e_t)^b}{c_s}\right)^{\frac{1}{1-a}}\right\}, 1\right\}. \tag{3.6}$$

Proposition 12 shows that, as one would expect, student effort levels are increasing in the performance-based reward π_s. However, because students cannot exert an unlimited amount of effort, there is a point at which offering a higher reward becomes ineffective. In reality, this reflects constraints on students' time, as well as natural limitations that arise due to students' initial skills and endowments. Moreover, because the school district and teachers can enforce a minimum level of joint effort, any small reward offered by the district is pointless, since it will not change students' behavior.

Additionally, we see that, as expected, students' optimal effort levels are increasing in teachers' total effort levels e_t. This reflects the complementary nature of teachers' and students' efforts and highlights the importance of an efficient allocation of the reward between teachers and students; namely, direct monetary incentives are not the sole means of incentivizing high levels of student effort.

Observe that, because we assume that teachers are also required to exert a minimum level of effort, it is always possible for the school district to incentivize students to exert the maximum effort level, as long as the budget is sufficiently large. Nevertheless, this may be more easily achieved by concurrently increasing the teachers' reward allocation. To understand these dynamics, we analyze the teachers' optimal effort decision in the subsequent Proposition.

Recall that teachers know their own cost of effort but only know the distribution of students' cost of effort, as given in (3.3). Then, when determining their own optimal effort level, teachers must consider six possible cases of students' effort levels, based on whether students'

effort levels under either cost of effort are the minimum level e_j, the maximum level 1, or an interior solution. In Proposition 13, we describe the teachers' optimal effort level in the case that students' resulting optimal effort level is an interior solution under both high and low costs of effort, c^h and c^l, respectively. For an analysis of each of the six possible cases, see the Appendix.

Proposition 13 (Teachers' Optimal Effort Levels) *Suppose that students' optimal effort level is between e_j and 1 for either possible cost of effort c_s and for any $e_t \in [e_j, 1]$. Then, for any given allocation scheme chosen by the school district (π_s, π_t), the teachers' best response is to exert total effort $e_t^*(\pi_s, \pi_t)$, where*

$$e_t^*(\pi_s, \pi_t) = \min\left\{\max\left\{e_j, \left(g(\beta_0)(\pi_s a)^a \left(\left(\frac{b}{1-a}\right)\left(\frac{\pi_t}{c_t}\right)\left(\frac{p_s^h}{(c^h)^{\frac{a}{1-a}}} + \frac{(1-p_s^h)}{(c^l)^{\frac{a}{1-a}}}\right)\right)^{1-a}\right)^{\frac{1}{1-a-b}}\right\}, 1\right\}. \tag{3.7}$$

In this case, as one would expect, teachers' effort levels are increasing in their performance-based reward π_t. Just as with the students' incentive allocation, the district runs the risk of choosing an inefficient allocation if it is either too high or too low, given the bounds on teachers' effort decision. Additionally, because teachers choose their effort levels first, their decision is also explicitly increasing in the students' performance-based reward π_s: they use this information to refine their expectation of students' effort levels. Observe that the weight given to these incentive levels varies based on the relative weights given to teacher and student effort levels in the production function.

The formulation of the teachers' optimal effort decision in (3.7) belies the complexity of the teachers' optimal effort decision that arises due to the piecewise nature of the students' optimal effort decision. That is, changes in parameter values or reward allocations may require teachers' to choose an optimal effort level that results in students' effort level being

the minimum or maximum value for either cost level – in which case, the closed-form characterization of Proposition 13 is no longer relevant. As shown in the Appendix, closed-form characterizations of the optimal effort level are possible for only a subset of cases. Therefore, we further analyze teachers' optimal effort level through a numerical study.

3.5. Numerical Study

In this section, we numerically analyze the teachers' optimal decision, as well as the probability that the school reaches the proficient state by the end of the year. The parameter values used in this study are summarized in Table (2). The maximum probability of transitioning to the proficient state is higher when the initial state is proficient than when it is not proficient. Furthermore, the minimum level of effort e_j and student and teacher effort weights are fixed, and we allow that the teachers' effort level has a greater impact on the transition probability than the students' effort level. We consider high and low probabilities that students have a high cost of effort (p_s^h). Finally, we consider two scenarios for the cost of effort: one in which the range of effort costs is narrow ($c^h = 7$ and $c^l = 5$) and one in which it is wide ($c^h = 9$ and $c^l = 3$).

Parameters	Values
Maximum probability of transitioning to proficient state $g(\beta_0)$, $\beta_0 = N, P$	$0.6, 0.8$
Minimum effort level e_j	0.1
Student effort weight a	0.35
Teacher effort weight b	0.65
Probability of high cost of effort for students' p_s^h	0.35 or 0.75
High and low cost of effort (c^h, c^l)	$(7, 5)$ or $(9, 3)$

Table 2: Parameter values used in figures.

3.5.1. Teachers' Effort Level Decision

As described in the discussion of Proposition 13, teachers' effort level decisions are complicated by the minimum and maximum bounds on students' effort levels. Figures 13 and 14 illustrate the teachers' optimal effort decision as a function of the teachers' and students' reward allocation, respectively. In each Figure, we compare the four possible scenarios that result when the range of efforts costs is wide (blue) versus narrow (red) and when the

probability that students have a high cost of effort is high (solid line) versus low (dashed line).

In Figure 13, we compare teachers' optimal effort level e_t^* as a function of their reward allocation π_t when their own cost of effort is low (Figure 13a) and high (Figure 13b). The Figure confirms our intuition: teachers' effort levels are increasing in the optimal reward π_t and teachers are cheaper to incentivize when their own cost of effort is low. When teachers have a low cost of effort, there is a clear and intuitive ordering of the costliness of incentivizing them to exert effort: they are easiest to incentivize when the range of costs for student effort is wide and when there is a low probability that students have a high cost of effort. On the other hand, when teachers themselves have a high cost of effort, then they are less sensitive to their beliefs about students' costs and the range of efforts, with the exception of the case where they believe students have a wide range of costs and there is a high probability that students have a high cost of effort.

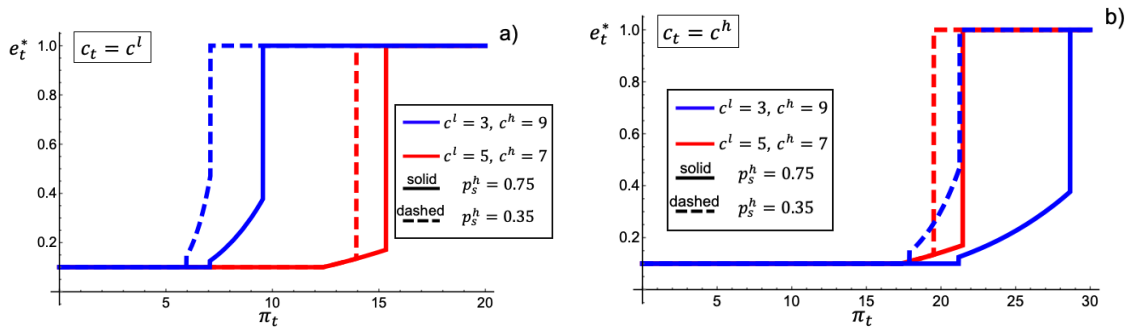

Figure 13: The teachers' optimal effort decision e_t^* as a function of the teachers' reward allocation π_t when teachers have a) a low cost of effort c^l and b) a high cost of effort c^h ($(c^l, c^h) = (3, 9)$ or $(5, 7)$, $p_s^h = 0.35$ or 0.75, $\beta_0 = N$, $g(N) = 0.6$, $p_s^h = 0.35$, $e_j = 0.1$, $a = 0.35$, $b = 0.6$, $\pi_s = 12$).

Figure 14 shows teachers' optimal effort level e_t^* as a function of the students' reward allocation π_s, where Figure 14a shows the case where teachers' own cost of effort is low and Figure 14b shows the case where it is high. In contrast to the previous Figures, here we see that teachers' optimal effort decision does not have a monotonic relationship with

the student reward allocation. Rather, although the optimal effort decision is initially increasing in π_s, once the reward is sufficiently high, teachers' effort levels first decrease before eventually recovering. This behavior is driven by the bounds on students' effort level. In particular, when students' optimal effort level is characterized by an interior solution, i.e. $e_s^* \in (e_j, 1)$, it is increasing in both the reward π_s and teachers' effort levels e_t. However, once their reward allocation is sufficiently high, students will exert the maximum effort level. Anticipating this, as the reward allocation increases, teachers can decrease their effort while maintaining a stable level of student effort. (Specifically, for the regions of π_s where teachers' effort level is decreasing, students will exert maximum effort if their cost of effort is low and a fixed interior effort level if their cost is high.) These Figures highlight the importance of properly allocating rewards: an improper allocation may not only be wasteful – it may be detrimental to the district's ultimate objective.

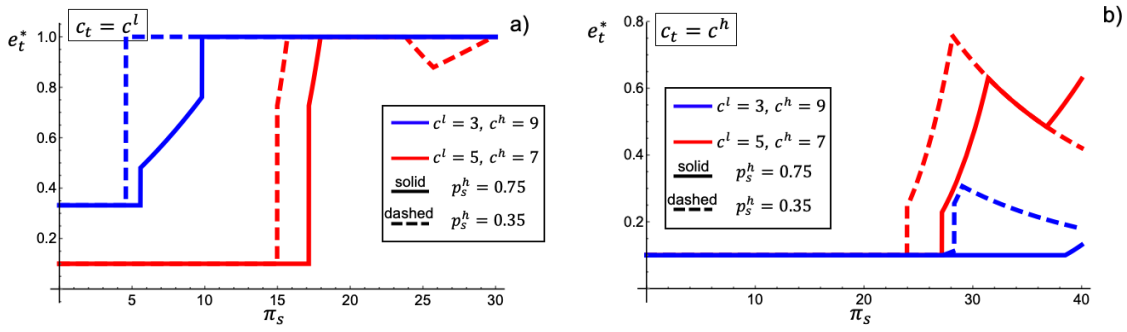

Figure 14: The teachers' optimal effort decision e_t^* as a function of the students' reward allocation π_t when teachers have a) a low cost of effort c^l and b) a high cost of effort c^h ($(c^l, c^h) = (3, 9)$ or $(5, 7)$, $p_s^h = 0.35$ or 0.75, $\beta_0 = N$, $g(N) = 0.6$, $p_s^h = 0.35$, $e_j = 0.1$, $a = 0.35$, $b = 0.6$, $\pi_t = 12$).

3.5.2. Probability of Achieving Proficiency

Finally, we evaluate the probability that the school reaches the proficient state by the end of the year (at $t = 1$) under teachers' and students' optimal effort levels. Using the previous results, we can determine the optimal effort decisions for a given reward allocation and, therefore, the probability of moving to the proficient state. We use $Pr^*[\beta_1 = P|\beta_0]$ to represent the probability that the school is in the proficient state at the end of the period under the teachers' and students' optimal effort decisions, namely $e_s^*(e_t, \pi_s)$ and $e_t^*(\pi_s, \pi_t)$,

for given values of the teachers' and students' actual costs of effort, c_t and c_s.

Figure 15 shows the probability that the school is in the proficient state at the end of the period under the optimal effort decisions as a function of the student and teacher reward allocations. Dashed lines illustrate two possible values of the budget constraint M.

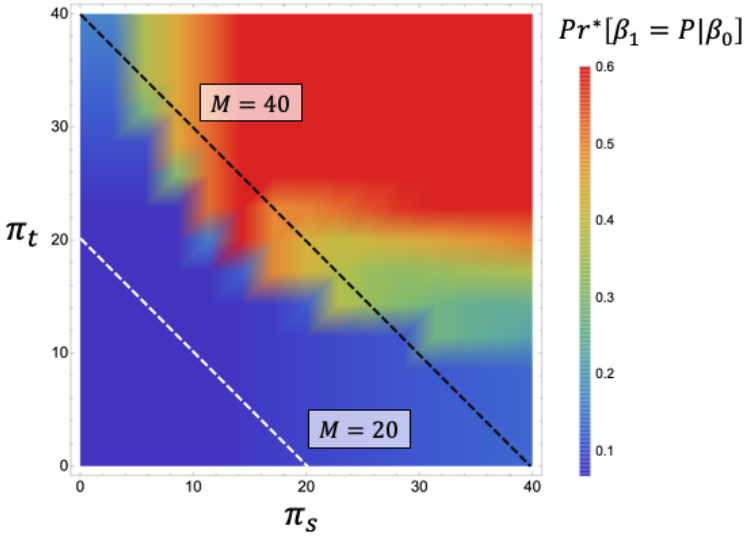

Figure 15: The probability that the school ends the year in the proficient state, $Pr^*[\beta_1 = P|\beta_0]$ as a function of the students' and teachers' reward allocation, π_s and π_t with budget constraint $M = 20$ and $M = 40$ shown ($\beta_0 = N$, $g(N) = 0.6$, $p_s^h = 0.35$, $e_j = 0.1$, $a = 0.35$, $b = 0.6$, $c^l = 3$, $c^h = 9$, $c_t = c^h$, and $c_s = c^l$).

The Figures make clear that for lower budgets ($M = 20$), the allocation may be irrelevant, since it will not be possible to incentivize effort beyond the minimum requirement. We see that when the district has a larger budget ($M = 40$), the reward allocation appears to be more consequential. In this illustration, allocating the entire reward to teachers ($\pi_t = 40, \pi_s = 0$) results in less than a 20 percent probability of ending the year in the proficient state, whereas an optimal allocation ($\pi_t = 26, \pi_s = 14$) can result in a 60 percent probability, a threefold improvement.

3.6. Discussion

In this chapter, we present a Stackelberg model of coproduction in a classroom setting. We assume that the school district offers performance-based incentives to teachers and students

at the beginning of the year, which they will earn if a sufficient proportion of students in the school are assessed to be proficient or higher on the end-of-the-year assessments. Observing that reward, teachers and students, in turn, determine whether to exert effort beyond an exogenously-determined minimum effort level.

Our initial results highlight the differing levels of complexity of the teachers' and students' effort decisions. In particular, students, who make their decision having observed both the district's allocation decision and teachers' effort level decision, will make an effort decision that is always increasing in both their reward allocation and teachers' effort level decision. On the other hand, teachers makes their effort level decision having observed the reward allocation but uncertain about students' cost of exerting effort. Thus, while their effort level will increase in their own reward, it is non-monotone in students' reward allocation. Specifically, it will temporarily decrease in students' reward allocation when it is high enough to induce students to exert a maximum level of effort under low costs.

These results give us the building blocks for further analysis. In particular, in this work, we consider only the students' and teachers' optimal effort decisions and take the performance-based incentives as given. We leave to future work the study of the optimal allocation of incentives by the school district.

APPENDIX

A.1. Notation Table for Chapter 1

Notation	Description
$t = 0, 1, 2$	Time indices corresponding to beginning, middle, and end of the school year, respectively
β_t	State of proficiency at time t, either proficient (P) or not proficient (N)
$\alpha(e)$	Probability of transitioning from N to P as a function of effort e at any period t
$1 - \delta(e)$	Probability of transitioning from P to P as a function of effort e at any period t
$A(P)$	Response-to-effort parameter
$B(P)$	"Stickiness" of the proficient state
μ	Student resilience parameter
z_I	District's binary choice of whether to invest in an interim assessment
X_1	Midyear result from formative assessments, either proficient (P) or not proficient (N)
$\phi_{P\|P}$, $\phi_{P\|N}$	True positive rate and false positive rate of the formative assessments result, respectively
π	Merit-based incentive
γ	Marginal cost of effort at time t
$\mathbf{S}_t$	State of the system at time t
M	School's available budget
F	Cost of interim assessment
$e_t(\mathbf{S}_t, \pi, z_I)$	Effort at time t as a function of state $\mathbf{S}_t$, merit-based incentive π, and assessment decision z_I
$(e_0^*(\mathbf{S}_0, \pi, z_I), e_1^*(\mathbf{S}_1, \pi, z_I))$	Optimal teachers' response policy
$Pr^*[\mathbf{S}_2\|\mathbf{S}_0]$	Probability school will be in proficient state at the end of the year under optimal teachers' response policy for fixed π and z_I
$Pr^*_{z_I}[\mathbf{S}_2\|\mathbf{S}_0]$	Probability school will be in proficient state at the end of the year under optimal teachers' response policy and optimal merit-based incentive for fixed z_I

Table 3: Description of model's notation.

A.2. Parameter Estimation for Chapter 1

In the Figure above, we use "base-case" parameters that we estimated using data from several sources. First, we use results from the DC Comprehensive Assessment System (DC CAS), the end-of-year standardized tests for District of Columbia Public Schools (DCPS)

(District of Columbia Public Schools, 2018a). This is a publicly-available dataset that includes eight years of school performance data (2006-07 through 2013-14) for both reading and math. For each school year, the number and percentage of test takers that fall into each proficiency category (below basic, basic, proficient, and advanced) are given by school and grade level for both the reading and math assessments.

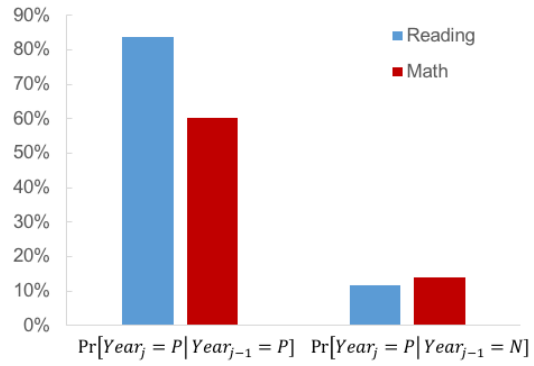

Figure 16: Probability of achieving proficiency (P) at the end of a given year based on the state of proficiency (P or N) in the previous year by subject for District of Columbia Public Schools from 2006-2014.

We base school-level proficiency targets on the 2006-07 annual measurable objectives for DCPS, given in the *Assessment and Accountability Manual* (District of Columbia Office of the State Superintendent of Education, 2011). In particular, for each subject, we average the elementary and secondary targets for the 2006-07 school year and round to the nearest integer. Thus, we define a school as being in the proficient state if at least 45 percent of students are proficient or above in reading and at least 40 percent of students are proficient or above in math. We use these target values for all of the school years in the dataset. Using these values, we calculate that the overall probability that a proficient school remains in the proficient state the following year is 84 percent for the reading assessment and 60 percent for the math assessment, and the probability that a not-proficient school moves to the proficient state is 12 percent for reading and 14 percent for math. We average these and use 72 percent as an estimate of $A(P)+B(P)$ and 13 percent as an estimation of $A(N)+B(N)$.

Therefore, the average level of student resilience μ is 0.18. We recognize the inexact nature of these estimates. First, these transition probabilities are likely to depend on each school, but we rely on the aggregate district-wide data in the absence of a sufficient number of data points for each individual school. Moreover, our model specifically considers the impact of the *additional* effort teachers might exert given appropriate incentives, but we are unable to pinpoint that effect by estimating $A(P)$ and $B(P)$ separately, nor are we able to estimate the effort level that resulted in these transition probabilities. Although this data does include a short period before and after the implementation of IMPACT, accurately accounting for the effect of that program would require teacher- and classroom-level information which we do not have.

Additionally, according to Topol et al. (2012), "school districts are spending an average of \$15–\$20 or more per student on interim assessments and data management systems to house their test data." DCPS enrollment was approximately $45,000$ during this period across 115 schools (District of Columbia Public Schools, 2018b), which suggests a district-wide cost of at least \$675,000–\$900,000 if the district chooses to implement interim assessments, or approximately \$5,870–\$7,826 per school. Although Figures 3 and 4 show the extreme case of a free interim assessment, this number provides a useful point of comparison.

The most difficult value to parameterize is the marginal cost of effort each semester (γ), which represents the maximum cost to teachers for exerting additional effort. Recall that IMPACT is one of the few programs where financial incentives produced measurable results. In that program, teachers were eligible for rewards that ranged from \$5,000 to \$25,000 per teacher, although only a subset of teachers were eligible for rewards at the upper end of that range. Furthermore, the average number of full-time teachers at a DC public school is 35 (District of Columbia Public Schools, 2018c). Based on these values, we assume that the average maximum cost of effort per teacher is \$10,000 over the entire school year. Then, for one school, 2γ is \$350,000.

Finally, we recognize that the accuracy of formative assessments can vary greatly depending

on the teacher and setting.

A.3. Proofs of Analytical Results

Lemma 3 *Define*

$$m(e_0, \mathbf{S}_0) = \begin{cases} 1 - \delta(e_0), & \text{if } \mathbf{S}_0 = P, \\ \alpha(e_0), & \text{if } \mathbf{S}_0 = N. \end{cases} \tag{A.1}$$

Then,

$$Pr\left[h_1\left(e_0, \mathbf{S}_0\right) = \left(P, e_0, \mathbf{S}_0\right)\right] = \phi_{P|P} m(e_0, \mathbf{S}_0) + \phi_{P|N}\left(1 - m(e_0, \mathbf{S}_0)\right), \tag{A.2}$$

and

$$Pr\left[h_2\left(e_1, \mathbf{S}_1\right) = P\right] = \left(1 - \delta\left(e_1\right)\right) Pr\left[\beta_1 = P|\mathbf{S}_1\right] + \alpha\left(e_1\right)\left(1 - Pr\left[\beta_1 = P|\mathbf{S}_1\right]\right), \tag{A.3}$$

where

$$Pr\left[\beta_1 = P|\mathbf{S}_1\right] = \begin{cases} \frac{\phi_{P|P} m(e_0, \mathbf{S}_0)}{\phi_{P|P} m(e_0, \mathbf{S}_0) + \phi_{P|N}(1 - m(e_0, \mathbf{S}_0))}, & \text{if } \mathbf{S}_1 = \left(P, e_0, \mathbf{S}_0\right), \\ \frac{\left(1 - \phi_{P|P}\right) m(e_0, \mathbf{S}_0)}{\left(1 - \phi_{P|P}\right) m(e_0, \mathbf{S}_0) + \left(1 - \phi_{P|N}\right)(1 - m(e_0, \mathbf{S}_0))}, & \text{if } \mathbf{S}_1 = \left(N, e_0, \mathbf{S}_0\right). \end{cases} \tag{A.4}$$

Proof of Lemma 3

When $z_I = 0$, the value of β_1 is not known with certainty, and at $t = 1$, the result of the formative assessment, X_1, is revealed. Focusing on $X_1 = P$, we have to consider two possible combinations for $\mathbf{S}_1$: (P, e_0, P) and (P, e_0, N). The probability of having the first combination is the sum of two probabilities: the one having $\beta_1 = P$ and then generating $X_1 = P$ $\left(\left(1 - \delta\left(e_0\right)\right)\phi_{P|P}\right)$ and the one having $\beta_1 = N$ and then generating $X_1 = P$ $\left(\delta\left(e_0\right)\phi_{P|N}\right)$. Thus, we obtain the first line in (A.2). The derivation of the second line in (A.2) follows the same steps.

The analysis for $t = 2$ involves two steps. First, (A.3) connects the probability that $\beta_1 = P$

to the probability that $\beta_2 = P$ as well, accounting for transitions between $\beta_1 = P$ and $\beta_1 = N$ and $\beta_2 = P$. Second, (A.4) looks at four possible values of the state at $t = 1$, $(X_1, e_0, \mathbf{S}_0)$: $\mathbf{S}_1 = (P, e_0, P)$, $\mathbf{S}_1 = (N, e_0, P)$, $\mathbf{S}_1 = (P, e_0, N)$, and $\mathbf{S}_1 = (N, e_0, N)$. For each of these states, (A.1)-(A.4) express the probability that $\beta_1 = P$. Below we show the derivation of this probability for the case of $\mathbf{S}_1 = (P, e_0, P)$; the derivations for the remaining cases follows the same steps. Using Bayes' rule, we have

$$\begin{aligned} & Pr\left[\beta_1 = P | X_1 = P, e_0, \mathbf{S}_0 = P\right] \\ = \; & \left(Pr\left[X_1 = P | \beta_1 = P, e_0, \mathbf{S}_0 = P\right] Pr\left[\beta_1 = P | e_0, \mathbf{S}_0 = P\right]\right) \\ & \div \left(Pr\left[X_1 = P | \beta_1 = P, e_0, \mathbf{S}_0 = P\right] Pr\left[\beta_1 = P | e_0, \mathbf{S}_0 = P\right]\right. \\ & \left. + Pr\left[X_1 = P | \beta_1 = N, e_0, \mathbf{S}_0 = P\right] Pr\left[\beta_1 = N | e_0, \mathbf{S}_0 = P\right]\right) \\ = \; & \frac{\phi_{P|P}\left(1 - \delta\left(e_0\right)\right)}{\phi_{P|P}\left(1 - \delta\left(e_0\right)\right) + \phi_{P|N}\delta\left(e_0\right)}. \end{aligned} \tag{A.5}$$

When $z_I = 1$, the probabilities are obtained by letting $\phi_{P|P} = 1$ and $\phi_{P|N} = 0$. In this case, β_1 is known exactly, and the response to the effort levels at both $t = 0$ and $t = 1$ is described by (1.1)-(1.2). $\square$

Proof of Proposition 1

The teachers' "profit-to-go" function at $t = 1$ is

$$J_1\left(\mathbf{S}_1\right) = \max_{e_1 \geq 0}\left[\pi Pr\left[\mathbf{S}_2 = P | \mathbf{S}_1\right] - \gamma e_1\right], \tag{A.6}$$

where the state of the system at $t = 1$ is given in (1.8), i.e.,

$$\mathbf{S}_1 = \begin{cases} \left(X_1, e_0, \mathbf{S}_0\right), & \text{if } z_I = 0, \\ \beta_1, & \text{if } z_I = 1. \end{cases} \tag{A.7}$$

The maximization in the expression for the profit-to-go function for $t = 1$ when $z_I = 0$ is carried over a concave function of e_1 for any value of $\mathbf{S}_1$. In particular, using (A.3) and

(A.6), this is represented by

$$J_1(\mathbf{S}_1) = \max_{e_1 \in [0,1]} \left[\pi \left((1 - \delta(e_1)) \, Pr\left[\beta_1 = P | \mathbf{S}_1\right] + \alpha(e_1) \left(1 - Pr\left[\beta_1 = P | \mathbf{S}_1\right]\right) \right) - \gamma e_1 \right], \tag{A.8}$$

with $\alpha(e_1)$ and $\delta(e_1)$ given by (1.1) and (1.2), respectively, and $Pr\left[\beta_1 = P|\mathbf{S}_1\right]$ given by (A.4). Then, the effort level maximizing (A.8) is given by

$$e_1^*(\mathbf{S}_1) = \begin{cases} 0, & \text{if } \frac{\pi}{\gamma} < \mathcal{T}_1(\mathbf{S}_1), \\ 1, & \text{if } \mathcal{T}_1(\mathbf{S}_1) \leq \frac{\pi}{\gamma}, \end{cases} \tag{A.9}$$

and the corresponding profit-to-go function is

$$J_1(\mathbf{S}_1) = \begin{cases} \pi \left((B(P) - B(N)) \, Pr\left[\beta_1 = P|\mathbf{S}_1\right] + B(N) \right), & \text{if } \frac{\pi}{\gamma} < \mathcal{T}_1(\mathbf{S}_1), \\ \pi \left((A(P) - A(N) + B(P) - B(N)) \, Pr\left[\beta_1 = P|\mathbf{S}_1\right] \right. & \\ \left. + A(N) + B(N) \right) - \gamma, & \text{if } \mathcal{T}_1(\mathbf{S}_1) \leq \frac{\pi}{\gamma}, \end{cases} \tag{A.10}$$

where

$$\mathcal{T}_1(\mathbf{S}_1) = \frac{1}{(A(P) - A(N)) \, Pr\left[\beta_1 = P|\mathbf{S}_1\right] + A(N)} \tag{A.11}$$

and

$$Pr\left[\beta_1 = P|\mathbf{S}_1\right] = \begin{cases} \frac{\phi_{P|P} m(e_0, \mathbf{S}_0)}{\phi_{P|P} m(e_0, \mathbf{S}_0) + \phi_{P|N} (1 - m(e_0, \mathbf{S}_0))}, & \text{if } \mathbf{S}_1 = (P, e_0, \mathbf{S}_0), \\ \frac{\left(1 - \phi_{P|P}\right) m(e_0, \mathbf{S}_0)}{\left(1 - \phi_{P|P}\right) m(e_0, \mathbf{S}_0) + \left(1 - \phi_{P|N}\right) (1 - m(e_0, \mathbf{S}_0))}, & \text{if } \mathbf{S}_1 = (N, e_0, \mathbf{S}_0). \end{cases} \tag{A.12}$$

We refer to $\mathcal{T}_1(\mathbf{S}_1)$ as the effort-inducing incentive threshold for the second half of the year. The optimal effort level at $t = 1$ when the district invests in an interim assessment ($z_I = 1$) is the special case when $\phi_{P|P} = 1$ and $\phi_{P|N} = 0$. (Here and below we omit, for simplicity, the designation (π, z_I) when referring to the profit-to-go functions and the optimal effort

levels.)

Under Assumptions 1 and 2 and using (A.4), we have that

$$Pr\left[\beta_1 = P|\mathbf{S}_1 = (P, e_0, \mathbf{S}_0)\right] \geq Pr\left[\beta_1 = P|\mathbf{S}_1 = (N, e_0, \mathbf{S}_0)\right], \tag{A.13}$$

which implies that $\mathcal{T}_1\left(P, e_0, \mathbf{S}_0\right) \leq \mathcal{T}_1\left(N, e_0, \mathbf{S}_0\right)$. Thus, the threshold is decreasing in the observed midyear level of proficiency.

Next, to show that the threshold is decreasing in the level of student resiliency, first rewrite the threshold, combining (A.11) and (A.12):

$$\mathcal{T}_1\left(\mathbf{S}_1\right) = \begin{cases} \frac{\phi_{P|P}(A(\mathbf{S}_0)e_0+B(\mathbf{S}_0))+\phi_{P|N}(1-(A(\mathbf{S}_0)e_0+B(\mathbf{S}_0)))}{A(P)\left(\phi_{P|P}(A(\mathbf{S}_0)e_0+B(\mathbf{S}_0))+\mu\phi_{P|N}(1-(A(\mathbf{S}_0)e_0+B(\mathbf{S}_0)))\right)}, & \text{if } \mathbf{S}_1 = (P, e_0, \mathbf{S}_0), \\ \frac{\left(1-\phi_{P|P}\right)(A(\mathbf{S}_0)e_0+B(\mathbf{S}_0))+\left(1-\phi_{P|N}\right)(1-(A(\mathbf{S}_0)e_0+B(\mathbf{S}_0)))}{A(P)\left(\left(1-\phi_{P|P}\right)(A(\mathbf{S}_0)e_0+B(\mathbf{S}_0))+\mu\left(1-\phi_{P|N}\right)(1-(A(\mathbf{S}_0)e_0+B(\mathbf{S}_0)))\right)}, & \text{if } \mathbf{S}_1 = (N, e_0, \mathbf{S}_0). \end{cases} \tag{A.14}$$

Applying Assumption 1:

$$\mathcal{T}_1\left(X_1, e_0, N\right) = \begin{cases} \frac{\mu(A(P)e_0+B(P))\phi_{P|P}+(1-\mu(A(P)e_0+B(P)))\phi_{P|N}}{A(P)\left(\mu(A(P)e_0+B(P))\phi_{P|P}+\mu(1-\mu(A(P)e_0+B(P)))\phi_{P|N}\right)}, \\ \qquad\qquad \text{if } \mathbf{S}_1 = (P, e_0, N), \\ \frac{\mu(A(P)e_0+B(P))\left(1-\phi_{P|P}\right)+(1-\mu(A(P)e_0+B(P)))\left(1-\phi_{P|N}\right)}{A(P)\left(\mu(A(P)e_0+B(P))\left(1-\phi_{P|P}\right)+\mu(1-\mu(A(P)e_0+B(P)))\left(1-\phi_{P|N}\right)\right)}, \\ \qquad\qquad \text{if } \mathbf{S}_1 = (N, e_0, N). \end{cases} \tag{A.15}$$

and

$$\mathcal{T}_1\left(X_1, e_0, P\right) = \begin{cases} \frac{(A(P)e_0+B(P))\phi_{P|P}+(1-(A(P)e_0+B(P)))\phi_{P|N}}{A(P)\left((A(P)e_0+B(P))\phi_{P|P}+\mu(1-(A(P)e_0+B(P)))\phi_{P|N}\right)}, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{if } \mathbf{S}_1 = (P, e_0, P)\,, \\ \frac{(A(P)e_0+B(P))\left(1-\phi_{P|P}\right)+(1-(A(P)e_0+B(P)))\left(1-\phi_{P|N}\right)}{A(P)\left((A(P)e_0+B(P))\left(1-\phi_{P|P}\right)+\mu(1-(A(P)e_0+B(P)))\left(1-\phi_{P|N}\right)\right)}, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{if } \mathbf{S}_1 = (N, e_0, P)\,. \end{cases} \tag{A.16}$$

It is clear that $\mathcal{T}_1\left(X_1, e_0, P\right)$ is decreasing in μ. Furthermore,

$$\begin{aligned} \frac{\partial}{\partial\mu}\mathcal{T}_1\left(P, e_0, N\right) = &\left(A(P)e_0 + B(P)\right)\left(\phi_{P|P} - \phi_{P|N}\right) \\ &\times \left(\mu A(P)\left(\phi_{P|N} + \left(\phi_{P|P} - \mu\phi_{P|N}\right)\left(A(P)e_0 + B(P)\right)\right)\right)^{-1} \\ &- \left(A(P)\phi_{P|N} + \left(A(P)\phi_{P|P} - 2\mu A(P)\phi_{P|N}\right)\left(A(P)e_0 + B(P)\right)\right) \\ &\times \left(\phi_{P|N} + \mu\left(A(P)e_0 + B(P)\right)\left(\phi_{P|P} - \phi_{P|N}\right)\right) \\ &\times \left(\mu A(P)\left(\phi_{P|N} + \left(\phi_{P|P} - \mu\phi_{P|N}\right)\left(A(P)e_0 + B(P)\right)\right)\right)^{-2}. \end{aligned} \tag{A.17}$$

Then,

$$\begin{aligned} &\frac{\partial}{\partial\mu}\mathcal{T}_1\left(P, e_0, N\right) < 0 \\ \iff &0 < \left(1 - \mu\left(A(P)e_0 + B(P)\right)\right)\phi_{P|N} \\ &+ \left(A(P)e_0 + B(P)\right)\left(\phi_{P|P} - \mu\phi_{P|N} - \mu^2\left(\phi_{P|P} - \phi_{P|N}\right)\left(A(P)e_0 + B(P)\right)\right), \end{aligned} \tag{A.18}$$

which clearly holds. Furthermore,

$$\begin{aligned}
&\frac{\partial}{\partial \mu}\mathcal{T}_1\left(N, e_0, N\right) \\
&= -\left(A(P)e_0 + B(P)\right)\left(\phi_{P|P} - \phi_{P|N}\right) \\
&\times \left(\mu A(P)\left(\left(1-\phi_{P|N}\right) + \left(\left(1-\phi_{P|P}\right) - \left(1-\phi_{P|N}\right)\mu\right)\left(A(P)e_0 + B(P)\right)\right)\right)^{-1} \\
&- \left(\left(1-\phi_{P|N}\right) - \mu\left(A(P)e_0 + B(P)\right)\left(\phi_{P|P} - \phi_{P|N}\right)\right) \\
&\times \left(\mu A(P)\left(\left(1-\phi_{P|N}\right) + \left(\left(1-\phi_{P|P}\right) - \left(1-\phi_{P|N}\right)\mu\right)\left(A(P)e_0 + B(P)\right)\right)\right)^{-2} \\
&\times \left(A(P)\left(1-\phi_{P|N}\right) + \left(A(P)\left(1-\phi_{P|P}\right) - 2\mu\left(1-\phi_{P|N}\right)A(P)\right)\left(A(P)e_0 + B(P)\right)\right).
\end{aligned} \tag{A.19}$$

Then,

$$\begin{aligned}
&\frac{\partial}{\partial \mu}\mathcal{T}_1\left(N, e_0, N\right) < 0 \\
\iff &0 < \left(1 - \mu\left(A(P)e_0 + B(P)\right)\right)^2\left(1-\phi_{P|N}\right) \\
&+ \left(A(P)e_0 + B(P)\right)\left(1-\phi_{P|P}\right)\left(1 - \mu^2\left(A(P)e_0 + B(P)\right)\right),
\end{aligned} \tag{A.20}$$

which clearly holds. Therefore, the midyear effort-inducing incentive threshold is decreasing in student resilience.

Finally, it is straightforward to show that

$$\mathcal{T}_1\left(X_1, 1, \mathbf{S}_0\right) \leq \mathcal{T}_1\left(X_1, 0, \mathbf{S}_0\right) \iff 1 \geq \mu, \tag{A.21}$$

for $X_1 = P, N$, which holds under Assumption 1. □

Proof of Proposition 2

To establish the results of this Proposition, we rely on the assumption given in (A.22). Assume that the false positive and true positive rates of the formative-assessment result and the response-to-effort and baseline transition probability parameters are related by the

following equation:

$$\frac{\left(1-\phi_{P|P}\right)\phi_{P|N}}{\phi_{P|P}-\phi_{P|N}} \leq \frac{\left(1-\left(A(P)+B(P)\right)\right)B(P)}{A(P)}. \tag{A.22}$$

Note that this condition holds when $\phi_{P|P}$ is sufficiently high and $\phi_{P|N}$ is sufficiently low, i.e. for a reasonably accurate formative assessment.

Additionally, we use the following results.

Lemma 4 *Let*

$$\begin{aligned}\mathcal{C}_3\left(\mathbf{S}_0\right) = &\left(1+\phi_{P|N}+\left(\phi_{P|P}-\phi_{P|N}\right)\left(A(\mathbf{S}_0)+B(\mathbf{S}_0)\right)\right)\\ &\div\left(A(\mathbf{S}_0)\left(A(P)\phi_{P|P}+B(P)-\mu\left(A(P)\phi_{P|N}+B(P)\right)\right)\right.\\ &\left.+A(P)\left(\phi_{P|P}-\mu\phi_{P|N}\right)B(\mathbf{S}_0)+\mu A(P)\phi_{P|N}\right).\end{aligned} \tag{A.23}$$

Then, for all values of $\mathbf{S}_0$*,*

$$\max\{\mathcal{T}_1\left(P,1,\mathbf{S}_0\right),\mathcal{C}_3\left(\mathbf{S}_0\right)\} = \mathcal{C}_3\left(\mathbf{S}_0\right), \text{ and} \tag{A.24}$$

$$\min\left\{\mathcal{T}_1\left(P,1,\mathbf{S}_0\right), \frac{1}{(1-\mu)A(\mathbf{S}_0)B(P)}\right\} = \mathcal{T}_1\left(P,1,\mathbf{S}_0\right). \tag{A.25}$$

Proof of Lemma 4

Let

$$\mathcal{E}_4\left(\mathbf{S}_0\right) = \frac{A(P)\phi_{P|P}\left(A(\mathbf{S}_0)+B(\mathbf{S}_0)\right)+\mu A(P)\phi_{P|N}\left(1-\left(A(\mathbf{S}_0)+B(\mathbf{S}_0)\right)\right)}{A(\mathbf{S}_0)\left(\phi_{P|P}\left(A(\mathbf{S}_0)+B(\mathbf{S}_0)\right)+\phi_{P|N}\left(1-\left(A(\mathbf{S}_0)+B(\mathbf{S}_0)\right)\right)\right)}+(1-\mu)A(P). \tag{A.26}$$

Then,

$$
\begin{aligned}
&\mathcal{T}_1\left(P,1,\mathbf{S}_0\right) \leq \mathcal{C}_3\left(\mathbf{S}_0\right) \\
\iff &\frac{\phi_{P|P}\left(A(\mathbf{S}_0)+B(\mathbf{S}_0)\right)+\phi_{P|N}\left(1-\left(A(\mathbf{S}_0)+B(\mathbf{S}_0)\right)\right)}{A(P)\phi_{P|P}\left(A(\mathbf{S}_0)+B(\mathbf{S}_0)\right)+\mu A(P)\phi_{P|N}\left(1-\left(A(\mathbf{S}_0)+B(\mathbf{S}_0)\right)\right)} \\
&\leq \left(1+\phi_{P|N}+\left(\phi_{P|P}-\phi_{P|N}\right)\left(A(\mathbf{S}_0)+B(\mathbf{S}_0)\right)\right) \\
&\div \left(A(\mathbf{S}_0)\left(A(P)\phi_{P|P}+B(P)-\mu\left(A(P)\phi_{P|N}+B(P)\right)\right)\right. \\
&\left.+A(P)B(\mathbf{S}_0)\left(\phi_{P|P}-\mu\phi_{P|N}\right)+\mu A(P)\phi_{P|N}\right) \\
\iff &\left(1-\mu\right)\left(A(P)+B(P)\right) \\
&\leq \frac{A(P)\phi_{P|P}\left(A(\mathbf{S}_0)+B(\mathbf{S}_0)\right)+\mu A(P)\phi_{P|N}\left(1-\left(A(\mathbf{S}_0)+B(\mathbf{S}_0)\right)\right)}{A(\mathbf{S}_0)\left(\phi_{P|P}\left(A(\mathbf{S}_0)+B(\mathbf{S}_0)\right)+\phi_{P|N}\left(1-\left(A(\mathbf{S}_0)+B(\mathbf{S}_0)\right)\right)\right)}+(1-\mu)A(P) \\
\iff &\left(1-\mu\right)\left(A(P)+B(P)\right) \leq \mathcal{E}_4\left(\mathbf{S}_0\right). \qquad \text{(A.27)}
\end{aligned}
$$

Note that $\mathcal{E}_4(N) \geq 1$. Furthermore,

$$
\begin{aligned}
&\mathcal{E}_4(P) \geq \left(1-\mu\right)\left(A(P)+B(P)\right) \\
\iff &\frac{\phi_{P|P}\left(A(P)+B(P)\right)+\mu\phi_{P|N}\left(1-\left(A(P)+B(P)\right)\right)}{\phi_{P|P}\left(A(P)+B(P)\right)+\phi_{P|N}\left(1-\left(A(P)+B(P)\right)\right)}+A(P)(1-\mu) \\
&\geq \left(1-\mu\right)\left(A(P)+B(P)\right) \\
\iff &\mu \geq \frac{B(P)\phi_{P|N}-\left(A(P)+B(P)\right)\left(\left(1-B(P)\right)\phi_{P|P}+B(P)\phi_{P|N}\right)}{B(P)\phi_{P|P}\left(A(P)+B(P)\right)+\phi_{P|N}\left(1+B(P)\right)\left(1-\left(A(P)+B(P)\right)\right)}. \qquad \text{(A.28)}
\end{aligned}
$$

Consider the term on the right-hand side of the inequality in (A.28). Under Assumptions 1 and 2, the denominator is clearly positive, and the numerator is negative, since

$$
\frac{B(P)}{A(P)+B(P)} < \frac{\left(1-B(P)\right)\phi_{P|P}+B(P)\phi_{P|N}}{\phi_{P|N}}. \qquad \text{(A.29)}
$$

Thus, (A.28) holds. (A.24) follows.

Additionally, it is straightforward to show that

$$\frac{1}{(1-\mu)A(\mathbf{S}_0)B(P)} \geq \mathcal{T}_1\left(P,1,\mathbf{S}_0\right) \iff \mathcal{E}_4\left(\mathbf{S}_0\right) \geq (1-\mu)\left(A(P)+B(P)\right), \tag{A.30}$$

which we show to be true above. (A.25) follows. □

Lemma 5 *The second-period effort-inducing incentive threshold is decreasing in first-period effort, i.e.*

$$\mathcal{T}_1\left(P,1,\mathbf{S}_0\right) \leq \mathcal{T}_1\left(P,0,\mathbf{S}_0\right) \text{ and } \mathcal{T}_1\left(N,1,\mathbf{S}_0\right) \leq \mathcal{T}_1\left(N,0,\mathbf{S}_0\right). \tag{A.31}$$

Furthermore, when (A.22) *holds, then the cost-to-incentive ratio threshold is always lower under a proficient formative-assessment result, i.e.*

$$\mathcal{T}_1\left(N,1,\mathbf{S}_0\right) \geq \mathcal{T}_1\left(P,0,\mathbf{S}_0\right). \tag{A.32}$$

Proof of Lemma 5

As noted in Proposition 1,

$$\mathcal{T}_1\left(X_1,1,\mathbf{S}_0\right) \leq \mathcal{T}_1\left(X_1,0,\mathbf{S}_0\right). \tag{A.33}$$

Next, compare the boundaries on the cost-of-effort to reward ratio under different values of

the formative assessment result. Specifically,

$$\begin{aligned}
&\min\left\{\mathcal{T}_1\left(N,0,\mathbf{S}_0\right),\mathcal{T}_1\left(N,1,\mathbf{S}_0\right)\right\} \geq \max\left\{\mathcal{T}_1\left(P,0,\mathbf{S}_0\right),\mathcal{T}_1\left(P,1,\mathbf{S}_0\right)\right\} \\
\iff & \mathcal{T}_1\left(N,1,\mathbf{S}_0\right) \geq \mathcal{T}_1\left(P,0,\mathbf{S}_0\right) \\
\iff & \frac{\left(1-\phi_{P|P}\right)\left(A(\mathbf{S}_0)+B(\mathbf{S}_0)\right)+\left(1-\phi_{P|N}\right)\left(1-\left(A(\mathbf{S}_0)+B(\mathbf{S}_0)\right)\right)}{A(P)\left(\left(1-\phi_{P|P}\right)\left(A(\mathbf{S}_0)+B(\mathbf{S}_0)\right)+\mu\left(1-\phi_{P|N}\right)\left(1-\left(A(\mathbf{S}_0)+B(\mathbf{S}_0)\right)\right)\right)} \\
& \quad \geq \frac{\phi_{P|P}B(\mathbf{S}_0)+\phi_{P|N}\left(1-B(\mathbf{S}_0)\right)}{A(P)\left(\phi_{P|P}B(\mathbf{S}_0)+\mu\phi_{P|N}\left(1-B(\mathbf{S}_0)\right)\right)} \\
\iff & \frac{\left(1-\left(A(\mathbf{S}_0)+B(\mathbf{S}_0)\right)\right)B(\mathbf{S}_0)}{A(\mathbf{S}_0)} \geq \frac{\left(1-\phi_{P|P}\right)\phi_{P|N}}{\phi_{P|P}-\phi_{P|N}}.
\end{aligned} \tag{A.34}$$

Then, if $\mathbf{S}_0 = P$,

$$\mathcal{T}_1\left(N,1,P\right) \geq \mathcal{T}_1\left(P,0,P\right) \iff \frac{\left(1-\left(A(P)+B(P)\right)\right)B(P)}{A(P)} \geq \frac{\left(1-\phi_{P|P}\right)\phi_{P|N}}{\phi_{P|P}-\phi_{P|N}}, \tag{A.35}$$

and if $\mathbf{S}_0 = N$,

$$\mathcal{T}_1\left(N,1,N\right) \geq \mathcal{T}_1\left(P,0,N\right) \iff \frac{\left(1-\mu\left(A(P)+B(P)\right)\right)B(P)}{A(P)} \geq \frac{\left(1-\phi_{P|P}\right)\phi_{P|N}}{\phi_{P|P}-\phi_{P|N}}, \tag{A.36}$$

where $\frac{\left(1-\left(A(P)+B(P)\right)\right)B(P)}{A(P)} \leq \frac{\left(1-\mu\left(A(P)+B(P)\right)\right)B(P)}{A(P)}$. Then, (A.32) follows from (A.22). $\square$

Lemma 6 *Define the following constants:*

$$\mathcal{E}_1(\mathbf{S}_0) = \frac{A(P)\left(1-\phi_{P|P}\right)B(\mathbf{S}_0) + \mu A(P)\left(1-\phi_{P|N}\right)(1-B(\mathbf{S}_0))}{A(\mathbf{S}_0)\left(\left(1-\phi_{P|P}\right)B(\mathbf{S}_0) + \left(1-\phi_{P|N}\right)(1-B(\mathbf{S}_0))\right)}, \tag{A.37}$$

$$\begin{aligned}\mathcal{E}_2(\mathbf{S}_0) = &\left(A(P)\left(A(\mathbf{S}_0)+B(\mathbf{S}_0)\right)\left(1-\phi_{P|P}\right) + \mu A(P)\left(1-\left(A(\mathbf{S}_0)+B(\mathbf{S}_0)\right)\right)\left(1-\phi_{P|N}\right)\right.\\ &\left.+A(\mathbf{S}_0)A(P)(1-\mu)\left(1-\phi_{P|N}\right)\left(1-\phi_{P|P}\right)\right)\\ &\div\left(A(\mathbf{S}_0)\left(\left(1-\phi_{P|P}\right)\left(A(\mathbf{S}_0)+B(\mathbf{S}_0)\right)+\left(1-\phi_{P|N}\right)\left(1-\left(A(\mathbf{S}_0)+B(\mathbf{S}_0)\right)\right)\right)\right),\end{aligned} \tag{A.38}$$

$$\begin{aligned}\mathcal{E}_3(\mathbf{S}_0) = &\frac{\left(A(P)\phi_{P|P}B(\mathbf{S}_0) + \mu A(P)\phi_{P|N}\left(1-B(\mathbf{S}_0)\right)\right)\left(1+\left(\phi_{P|P}-\phi_{P|N}\right)A(\mathbf{S}_0)\right)}{A(\mathbf{S}_0)\left(B(\mathbf{S}_0)\phi_{P|P}+(1-B(\mathbf{S}_0))\phi_{P|N}\right)}\\ &+A(P)\left(1-\phi_{P|P}\right)-\mu A(P)\left(1-\phi_{P|N}\right),\end{aligned} \tag{A.39}$$

$$\begin{aligned}\mathcal{C}_1(\mathbf{S}_0) = &\left(2-\left(\phi_{P|P}-\phi_{P|N}\right)B(\mathbf{S}_0)-\phi_{P|N}\right)\\ &\div\left((1-\mu)A(\mathbf{S}_0)\left(A(P)+B(P)\right)\right.\\ &\left.+A(P)B(\mathbf{S}_0)\left(1-\phi_{P|P}\right)+\mu A(P)\left(1-B(\mathbf{S}_0)\right)\left(1-\phi_{P|N}\right)\right),\end{aligned} \tag{A.40}$$

$$\mathcal{C}_2(\mathbf{S}_0) = \frac{1+\left(\phi_{P|P}-\phi_{P|N}\right)A(\mathbf{S}_0)}{A(\mathbf{S}_0)\left(\left(A(P)\phi_{P|P}+B(P)\right)-\mu\left(A(P)\phi_{P|N}+B(P)\right)\right)}, \tag{A.41}$$

$$\begin{aligned}\mathcal{C}_3(\mathbf{S}_0) = &\left(1+\phi_{P|N}+\left(\phi_{P|P}-\phi_{P|N}\right)\left(A(\mathbf{S}_0)+B(\mathbf{S}_0)\right)\right)\\ &\div\left(A(\mathbf{S}_0)\left(A(P)\phi_{P|P}+B(P)-\mu\left(A(P)\phi_{P|N}+B(P)\right)\right)\right.\\ &\left.+A(P)\left(\phi_{P|P}-\mu\phi_{P|N}\right)B(\mathbf{S}_0)+\mu A(P)\phi_{P|N}\right).\end{aligned} \tag{A.42}$$

Then, the following statements are true:

a)

$$\begin{aligned}&\max\left\{\mathcal{T}_1(N,0,\mathbf{S}_0), \frac{1}{(1-\mu)A(\mathbf{S}_0)\left(A(P)+B(P)\right)}\right\}\\ &=\begin{cases}\frac{1}{(1-\mu)A(\mathbf{S}_0)(A(P)+B(P))}, & \text{if } \mathbf{S}_0 = N, \text{ or}\\ & \text{if } \mathbf{S}_0 = P \text{ and } \mathcal{E}_1(P) \geq (1-\mu)\left(A(P)+B(P)\right),\\ \mathcal{T}_1(N,0,\mathbf{S}_0), & \text{if } \mathbf{S}_0 = P \text{ and } \mathcal{E}_1(P) < (1-\mu)\left(A(P)+B(P)\right).\end{cases}\end{aligned} \tag{A.43}$$

b)

$$\min\{\mathcal{T}_1(N,0,\mathbf{S}_0),\mathcal{C}_1(\mathbf{S}_0)\}$$
$$=\begin{cases}\mathcal{T}_1(N,0,\mathbf{S}_0), & \text{if } \mathbf{S}_0=N, \text{ or}\\ & \text{if } \mathbf{S}_0=P \text{ and } \mathcal{E}_1(P)\geq(1-\mu)(A(P)+B(P)),\\ \mathcal{C}_1(\mathbf{S}_0), & \text{if } \mathbf{S}_0=P \text{ and } \mathcal{E}_1(P)<(1-\mu)(A(P)+B(P)).\end{cases} \tag{A.44}$$

c)

$$\max\{\mathcal{T}_1(N,1,\mathbf{S}_0),\mathcal{C}_1(\mathbf{S}_0)\}$$
$$=\begin{cases}\mathcal{C}_1(\mathbf{S}_0), & \text{if } \mathbf{S}_0=N, \text{ or}\\ & \text{if } \mathbf{S}_0=P \text{ and } \mathcal{E}_2(P)\geq(1-\mu)(A(P)+B(P)),\\ \mathcal{T}_1(N,1,\mathbf{S}_0), & \text{if } \mathbf{S}_0=P \text{ and } \mathcal{E}_2(P)<(1-\mu)(A(P)+B(P)).\end{cases} \tag{A.45}$$

d)

$$\min\{\mathcal{T}_1(N,1,\mathbf{S}_0),\mathcal{C}_2(\mathbf{S}_0)\}$$
$$=\begin{cases}\mathcal{C}_2(\mathbf{S}_0), & \text{if } \mathbf{S}_0=P \text{ and } \mathcal{E}_2(P)<(1-\mu)(A(P)+B(P)),\\ \mathcal{T}_1(N,1,\mathbf{S}_0), & \text{if } \mathbf{S}_0=N, \text{ or}\\ & \text{if } \mathbf{S}_0=P \text{ and } \mathcal{E}_2(P)\geq(1-\mu)(A(P)+B(P)).\end{cases} \tag{A.46}$$

e)

$$\begin{aligned}
&\max\{\mathcal{T}_1\left(P,0,\mathbf{S}_0\right),\mathcal{C}_2\left(\mathbf{S}_0\right)\} \\
&= \begin{cases} \mathcal{T}_1\left(P,0,\mathbf{S}_0\right), & \text{if } \mathbf{S}_0 = P \text{ and } \mathcal{E}_3(P) < (1-\mu)\left(A(P)+B(P)\right), \\ \mathcal{C}_2\left(\mathbf{S}_0\right), & \text{if } \mathbf{S}_0 = N, \text{ or} \\ & \text{if } \mathbf{S}_0 = P \text{ and } \mathcal{E}_3(P) \geq (1-\mu)\left(A(P)+B(P)\right). \end{cases}
\end{aligned} \tag{A.47}$$

f)

$$\begin{aligned}
&\min\{\mathcal{T}_1\left(P,0,\mathbf{S}_0\right),\mathcal{C}_3\left(\mathbf{S}_0\right)\} \\
&= \begin{cases} \mathcal{T}_1\left(P,0,\mathbf{S}_0\right), & \text{if } \mathbf{S}_0 = N, \text{ or} \\ & \text{if } \mathbf{S}_0 = P \text{ and } \mathcal{E}_3(P) \geq (1-\mu)\left(A(P)+B(P)\right), \\ \mathcal{C}_3\left(\mathbf{S}_0\right), & \text{if } \mathbf{S}_0 = P \text{ and } \mathcal{E}_3(P) < (1-\mu)\left(A(P)+B(P)\right). \end{cases}
\end{aligned} \tag{A.48}$$

Proof of Lemma 6

We use (A.11) and (A.12) and plug in the relevant values of e_0 to simplify these boundaries. Then, when $e_0 = 0$,

$$\mathcal{T}_1\left(\mathbf{S}_1\right) = \begin{cases} \frac{B(\mathbf{S}_0)\phi_{P|P}+(1-B(\mathbf{S}_0))\phi_{P|N}}{A(P)\left(B(\mathbf{S}_0)\phi_{P|P}+\mu(1-B(\mathbf{S}_0))\phi_{P|N}\right)}, & \text{if } \mathbf{S}_1 = \left(P,0,\mathbf{S}_0\right), \\ \frac{B(\mathbf{S}_0)\left(1-\phi_{P|P}\right)+(1-B(\mathbf{S}_0))\left(1-\phi_{P|N}\right)}{A(P)\left(B(\mathbf{S}_0)\left(1-\phi_{P|P}\right)+\mu(1-B(\mathbf{S}_0))\left(1-\phi_{P|N}\right)\right)}, & \text{if } \mathbf{S}_1 = \left(N,0,\mathbf{S}_0\right). \end{cases} \tag{A.49}$$

and when $e_0 = 1$,

$$\mathcal{T}_1\left(\mathbf{S}_1\right) = \begin{cases} \frac{(A(\mathbf{S}_0)+B(\mathbf{S}_0))\phi_{P|P}+(1-(A(\mathbf{S}_0)+B(\mathbf{S}_0)))\phi_{P|N}}{A(P)\left((A(\mathbf{S}_0)+B(\mathbf{S}_0))\phi_{P|P}+\mu(1-(A(\mathbf{S}_0)+B(\mathbf{S}_0)))\phi_{P|N}\right)}, & \text{if } \mathbf{S}_1 = \left(P,1,\mathbf{S}_0\right), \\ \frac{(A(\mathbf{S}_0)+B(\mathbf{S}_0))\left(1-\phi_{P|P}\right)+(1-(A(\mathbf{S}_0)+B(\mathbf{S}_0)))\left(1-\phi_{P|N}\right)}{A(P)(A(\mathbf{S}_0)+B(\mathbf{S}_0))\left(1-\phi_{P|P}\right)+\mu A(P)(1-(A(\mathbf{S}_0)+B(\mathbf{S}_0)))\left(1-\phi_{P|N}\right)}, & \text{if } \mathbf{S}_1 = \left(N,1,\mathbf{S}_0\right). \end{cases} \tag{A.50}$$

a) From (A.37),

$$\mathcal{E}_1(\mathbf{S}_0) = \frac{A(P)B(\mathbf{S}_0)\left(1-\phi_{P|P}\right) + \mu A(P)\left(1-B(\mathbf{S}_0)\right)\left(1-\phi_{P|N}\right)}{A(\mathbf{S}_0)\left(B(\mathbf{S}_0)\left(1-\phi_{P|P}\right) + \left(1-B(\mathbf{S}_0)\right)\left(1-\phi_{P|N}\right)\right)}. \tag{A.51}$$

Then,

$$\begin{aligned}
&\max\left\{\mathcal{T}_1(N,0,\mathbf{S}_0), \frac{1}{(1-\mu)A(\mathbf{S}_0)\left(A(P)+B(P)\right)}\right\} \\
&= \begin{cases} (1-\mu)\left(A(P)+B(P)\right)A(\mathbf{S}_0), & \text{if } \mathcal{E}_1(\mathbf{S}_0) \geq (1-\mu)\left(A(P)+B(P)\right), \\ \mathcal{T}_1(N,0,\mathbf{S}_0), & \text{if } \mathcal{E}_1(\mathbf{S}_0) < (1-\mu)\left(A(P)+B(P)\right). \end{cases}
\end{aligned} \tag{A.52}$$

Note that $\mathcal{E}_1(N) \geq 1$. (A.43) follows.

b) The following result, which always holds when $\mathbf{S}_0 = N$, gives (A.44):

$$\begin{aligned}
&\mathcal{C}_1(\mathbf{S}_0) \geq \mathcal{T}_1(N,0,\mathbf{S}_0) \\
\iff &\frac{2-\left(\phi_{P|P}-\phi_{P|N}\right)B(\mathbf{S}_0)-\phi_{P|N}}{(1-\mu)A(\mathbf{S}_0)\left(A(P)+B(P)\right)+A(P)B(\mathbf{S}_0)\left(1-\phi_{P|P}\right)+\mu A(P)\left(1-B(\mathbf{S}_0)\right)\left(1-\phi_{P|N}\right)} \\
&\quad \geq \frac{\left(1-\phi_{P|P}\right)B(\mathbf{S}_0)+\left(1-\phi_{P|N}\right)\left(1-B(\mathbf{S}_0)\right)}{A(P)\left(B(\mathbf{S}_0)\left(1-\phi_{P|P}\right)+\mu\left(1-B(\mathbf{S}_0)\right)\left(1-\phi_{P|N}\right)\right)} \\
\iff &\mathcal{E}_1(\mathbf{S}_0) \geq (1-\mu)\left(A(P)+B(P)\right).
\end{aligned} \tag{A.53}$$

c) From (A.40),

$$\begin{aligned}
\mathcal{C}_1(\mathbf{S}_0) = &\left(2-\left(\phi_{P|P}-\phi_{P|N}\right)B(\mathbf{S}_0)-\phi_{P|N}\right) \\
&\div\left((1-\mu)A(\mathbf{S}_0)\left(A(P)+B(P)\right)+A(P)B(\mathbf{S}_0)\left(1-\phi_{P|P}\right)\right. \\
&\left.+\mu A(P)\left(1-B(\mathbf{S}_0)\right)\left(1-\phi_{P|N}\right)\right),
\end{aligned} \tag{A.54}$$

and from (A.38),

$$\begin{aligned}\mathcal{E}_2(\mathbf{S}_0) = &\left(A(P)\left(A(\mathbf{S}_0)+B(\mathbf{S}_0)\right)\left(1-\phi_{P|P}\right)+\mu A(P)\left(1-\left(A(\mathbf{S}_0)+B(\mathbf{S}_0)\right)\right)\left(1-\phi_{P|N}\right)\right.\\ &+A(\mathbf{S}_0)A(P)(1-\mu)\left(1-\phi_{P|N}\right)\left(1-\phi_{P|P}\right)\Big)\\ &\div\left(A(\mathbf{S}_0)\left(\left(1-\phi_{P|P}\right)\left(A(\mathbf{S}_0)+B(\mathbf{S}_0)\right)+\left(1-\phi_{P|N}\right)\left(1-\left(A(\mathbf{S}_0)+B(\mathbf{S}_0)\right)\right)\right)\right).\end{aligned} \tag{A.55}$$

Then,

$$\begin{aligned}&\mathcal{C}_1(\mathbf{S}_0) \geq \mathcal{T}_1(N,1,\mathbf{S}_0)\\ \iff &\frac{2-\left(\phi_{P|P}-\phi_{P|N}\right)B(\mathbf{S}_0)-\phi_{P|N}}{(1-\mu)A(\mathbf{S}_0)\left(A(P)+B(P)\right)+A(P)B(\mathbf{S}_0)\left(1-\phi_{P|P}\right)+\mu A(P)\left(1-B(\mathbf{S}_0)\right)\left(1-\phi_{P|N}\right)}\\ &\geq \frac{\left(A(\mathbf{S}_0)+B(\mathbf{S}_0)\right)\left(1-\phi_{P|P}\right)+\left(1-\left(A(\mathbf{S}_0)+B(\mathbf{S}_0)\right)\right)\left(1-\phi_{P|N}\right)}{A(P)\left(A(\mathbf{S}_0)+B(\mathbf{S}_0)\right)\left(1-\phi_{P|P}\right)+\mu A(P)\left(1-\left(A(\mathbf{S}_0)+B(\mathbf{S}_0)\right)\right)\left(1-\phi_{P|N}\right)}\\ \iff &\mathcal{E}_2(\mathbf{S}_0) \geq (1-\mu)\left(A(P)+B(P)\right).\end{aligned} \tag{A.56}$$

Note that $\mathcal{E}_2(N) \geq 1$. (A.45) follows.

d) Recall from (A.41)

$$\mathcal{C}_2(\mathbf{S}_0) = \frac{1+\left(\phi_{P|P}-\phi_{P|N}\right)A(\mathbf{S}_0)}{A(\mathbf{S}_0)\left(\left(A(P)\phi_{P|P}+B(P)\right)-\mu\left(A(P)\phi_{P|N}+B(P)\right)\right)}. \tag{A.57}$$

Then,

$$\begin{aligned}&\mathcal{C}_2(\mathbf{S}_0) \geq \mathcal{T}_1(N,1,\mathbf{S}_0)\\ \iff &\frac{1+\left(\phi_{P|P}-\phi_{P|N}\right)A(\mathbf{S}_0)}{A(\mathbf{S}_0)\left(\left(A(P)\phi_{P|P}+B(P)\right)-\mu\left(A(P)\phi_{P|N}+B(P)\right)\right)}\\ &\geq \frac{\left(A(\mathbf{S}_0)+B(\mathbf{S}_0)\right)\left(1-\phi_{P|P}\right)+\left(1-\left(A(\mathbf{S}_0)+B(\mathbf{S}_0)\right)\right)\left(1-\phi_{P|N}\right)}{A(P)\left(A(\mathbf{S}_0)+B(\mathbf{S}_0)\right)\left(1-\phi_{P|P}\right)+\mu A(P)\left(1-\left(A(\mathbf{S}_0)+B(\mathbf{S}_0)\right)\right)\left(1-\phi_{P|N}\right)}\\ \iff &\mathcal{E}_2(\mathbf{S}_0) \geq (1-\mu)\left(A(P)+B(P)\right).\end{aligned} \tag{A.58}$$

Note that $\mathcal{E}_2(N) \geq 1$. (A.46) follows.

e) From (A.39),

$$\mathcal{E}_3(\mathbf{S}_0) = \frac{\left(A(P)\phi_{P|P}B(\mathbf{S}_0) + \mu A(P)\phi_{P|N}\left(1 - B(\mathbf{S}_0)\right)\right)\left(1 + \left(\phi_{P|P} - \phi_{P|N}\right)A(\mathbf{S}_0)\right)}{A(\mathbf{S}_0)\left(B(\mathbf{S}_0)\phi_{P|P} + \left(1 - B(\mathbf{S}_0)\right)\phi_{P|N}\right)}$$
$$+ A(P)\left(1 - \phi_{P|P}\right) - \mu A(P)\left(1 - \phi_{P|N}\right). \tag{A.59}$$

Then,

$$\begin{aligned}
&\mathcal{C}_2(\mathbf{S}_0) \geq \mathcal{T}_1(P, 0, \mathbf{S}_0) \\
\iff &\frac{1 + \left(\phi_{P|P} - \phi_{P|N}\right)A(\mathbf{S}_0)}{A(\mathbf{S}_0)\left(\left(A(P)\phi_{P|P} + B(P)\right) - \mu\left(A(P)\phi_{P|N} + B(P)\right)\right)} \\
&\quad \geq \frac{B(\mathbf{S}_0)\phi_{P|P} + \left(1 - B(\mathbf{S}_0)\right)\phi_{P|N}}{A(P)\left(B(\mathbf{S}_0)\phi_{P|P} + \mu\left(1 - B(\mathbf{S}_0)\right)\phi_{P|N}\right)} \\
\iff &\mathcal{E}_3(\mathbf{S}_0) \geq (1 - \mu)\left(A(P) + B(P)\right).
\end{aligned} \tag{A.60}$$

Note that $\mathcal{E}_3(N) \geq 1$. (A.47) follows.

f) Recall from (A.42) that

$$\begin{aligned}
\mathcal{C}_3(\mathbf{S}_0) = &\left(1 + \phi_{P|N} + \left(\phi_{P|P} - \phi_{P|N}\right)\left(A(\mathbf{S}_0) + B(\mathbf{S}_0)\right)\right) \\
&\div \left(A(\mathbf{S}_0)\left(A(P)\phi_{P|P} + B(P) - \mu\left(A(P)\phi_{P|N} + B(P)\right)\right)\right) \\
&+ A(P)\left(\phi_{P|P} - \mu\phi_{P|N}\right)B(\mathbf{S}_0) + \mu A(P)\phi_{P|N}\Big).
\end{aligned} \tag{A.61}$$

Then,

$$
\begin{aligned}
& \mathcal{C}_3\left(\mathbf{S}_0\right) \geq \mathcal{T}_1\left(P, 0, \mathbf{S}_0\right) \\
\Longleftrightarrow & \left(1+\phi_{P|N}+\left(\phi_{P|P}-\phi_{P|N}\right)\left(A(\mathbf{S}_0)+B(\mathbf{S}_0)\right)\right) \\
& \div\left(A(\mathbf{S}_0)\left(A(P)\phi_{P|P}+B(P)-\mu\left(A(P)\phi_{P|N}+B(P)\right)\right)\right. \\
& \left.+A(P)\left(\phi_{P|P}-\mu\phi_{P|N}\right)B(\mathbf{S}_0)+\mu A(P)\phi_{P|N}\right) \\
& \geq \frac{B(\mathbf{S}_0)\phi_{P|P}+\left(1-B(\mathbf{S}_0)\right)\phi_{P|N}}{A(P)\left(B(\mathbf{S}_0)\phi_{P|P}+\mu\left(1-B(\mathbf{S}_0)\right)\phi_{P|N}\right)} \\
\Longleftrightarrow & \mathcal{E}_3\left(\mathbf{S}_0\right) \geq (1-\mu)\left(A(P)+B(P)\right). \qquad \text{(A.62)}
\end{aligned}
$$

Since $\mathcal{E}_3(N) \geq 1$, (A.48) follows. $\square$

Lemma 7 *When* (A.22) *holds, then*

$$
\mathcal{E}_1(P) \leq \mathcal{E}_2(P) \leq \mathcal{E}_3(P). \qquad \text{(A.63)}
$$

Proof of Lemma 7

From (A.37)–(A.39),

$$
\begin{aligned}
\mathcal{E}_1(P) =& \frac{B(P)\left(1-\phi_{P|P}\right)+\mu\left(1-\phi_{P|N}\right)\left(1-B(P)\right)}{B(P)\left(1-\phi_{P|P}\right)+\left(1-B(P)\right)\left(1-\phi_{P|N}\right)}, \\
\mathcal{E}_2(P) =& \left(\left(A(P)+B(P)\right)\left(1-\phi_{P|P}\right)+\mu\left(1-\left(A(P)+B(P)\right)\right)\left(1-\phi_{P|N}\right)\right. \\
& \left.+A(P)(1-\mu)\left(1-\phi_{P|N}\right)\left(1-\phi_{P|P}\right)\right) \\
& \div\left(\left(1-\phi_{P|P}\right)\left(A(P)+B(P)\right)+\left(1-\phi_{P|N}\right)\left(1-\left(A(P)+B(P)\right)\right)\right), \\
\mathcal{E}_3(P) =& \frac{\left(B(P)\phi_{P|P}+\mu\left(1-B(P)\right)\phi_{P|N}\right)\left(1+\left(\phi_{P|P}-\phi_{P|N}\right)A(P)\right)}{B(P)\phi_{P|P}+\left(1-B(P)\right)\phi_{P|N}} \\
& +A(P)\left(\left(1-\phi_{P|P}\right)-\mu\left(1-\phi_{P|N}\right)\right). \qquad \text{(A.64)}
\end{aligned}
$$

Then,

$$
\begin{aligned}
& \mathcal{E}_1(P) < \mathcal{E}_2(P) \\
\iff & \frac{B(P)\left(1-\phi_{P|P}\right)+\mu\left(1-\phi_{P|N}\right)\left(1-B(P)\right)}{B(P)\left(1-\phi_{P|P}\right)+\left(1-B(P)\right)\left(1-\phi_{P|N}\right)} \\
& \leq \left(\left(A(P)+B(P)\right)\left(1-\phi_{P|P}\right)+\mu\left(1-\left(A(P)+B(P)\right)\right)\left(1-\phi_{P|N}\right)\right. \\
& \left.+A(P)(1-\mu)\left(1-\phi_{P|N}\right)\left(1-\phi_{P|P}\right)\right) \\
& \div\left(\left(1-\phi_{P|P}\right)\left(A(P)+B(P)\right)+\left(1-\phi_{P|N}\right)\left(1-\left(A(P)+B(P)\right)\right)\right) \\
\iff & B(P)\phi_{P|P}+\left(1-B(P)\right)\phi_{P|N} \leq 2, \qquad \text{(A.65)}
\end{aligned}
$$

which holds under Assumptions 1 and 2. Moreover,

$$
\begin{aligned}
& \mathcal{E}_2(P) \leq \mathcal{E}_3(P) \\
\iff & \left(\left(A(P)+B(P)\right)\left(1-\phi_{P|P}\right)+\mu\left(1-\left(A(P)+B(P)\right)\right)\left(1-\phi_{P|N}\right)\right. \\
& \left.+A(P)(1-\mu)\left(1-\phi_{P|N}\right)\left(1-\phi_{P|P}\right)\right) \\
& \div\left(\left(1-\phi_{P|P}\right)\left(A(P)+B(P)\right)+\left(1-\phi_{P|N}\right)\left(1-\left(A(P)+B(P)\right)\right)\right) \\
& \leq \frac{\left(B(P)\phi_{P|P}+\mu\left(1-B(P)\right)\phi_{P|N}\right)\left(1+\left(\phi_{P|P}-\phi_{P|N}\right)A(P)\right)}{B(P)\phi_{P|P}+\left(1-B(P)\right)\phi_{P|N}} \\
& + A(P)\left(\left(1-\phi_{P|P}\right)-\mu\left(1-\phi_{P|N}\right)\right) \\
\iff & \frac{\phi_{P|N}\left(1-\phi_{P|P}\right)}{\phi_{P|P}-\phi_{P|N}} \leq \frac{B(P)\left(1-A(P)-B(P)\right)}{A(P)}, \qquad \text{(A.66)}
\end{aligned}
$$

which holds by assumption, as given in (A.22). □

The profit-to-go function for $t = 0$ under the formative assessment ($z_I = 0$) is

$$
\begin{aligned}
J_0(\mathbf{S}_0) = \max_{e_0 \in [0,1]} & \left[Pr\left[h_1\left(e_0, \mathbf{S}_0\right) = (N, e_0, \mathbf{S}_0)\right] J_1\left(N, e_0, \mathbf{S}_0\right)\right. \\
& \left.+Pr\left[h_1\left(e_0, \mathbf{S}_0\right) = (P, e_0, \mathbf{S}_0)\right] J_1\left(P, e_0, \mathbf{S}_0\right) - \gamma e_0\right], \qquad \text{(A.67)}
\end{aligned}
$$

with $Pr\left[h_1\left(e_0, \mathbf{S}_0\right)\right]$ given in (A.2) and $J_1\left(\mathbf{S}_1\right)$ given in (A.10). Using the results from

Lemma 3 and (A.10), the first term in the expression is

$$
\begin{aligned}
&Pr\left[h_1\left(e_0,\mathbf{S}_0\right)=\left(N,e_0,\mathbf{S}_0\right)\right]J_1\left(N,e_0,\mathbf{S}_0\right)\\
&=\begin{cases}
\pi\left(B(P)\left(1-\phi_{P|P}\right)m(e_0,\mathbf{S}_0)+B(N)\left(1-\phi_{P|N}\right)\left(1-m(e_0,\mathbf{S}_0)\right)\right), & \\
\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{if } \frac{\pi}{\gamma}<\mathcal{T}_1\left(N,e_0,\mathbf{S}_0\right), & \\
\left(\left(A(P)+B(P)\right)\pi-\gamma\right)\left(1-\phi_{P|P}\right)m(e_0,\mathbf{S}_0) & \\
+\left(\mu\left(A(P)+B(P)\right)\pi-\gamma\right)\left(1-\phi_{P|N}\right)\left(1-m(e_0,\mathbf{S}_0)\right), & \\
\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{if } \mathcal{T}_1\left(N,e_0,\mathbf{S}_0\right)\leq\frac{\pi}{\gamma}, &
\end{cases}
\end{aligned}
\tag{A.68}
$$

and the second term is

$$
\begin{aligned}
&Pr\left[h_1\left(e_0,\mathbf{S}_0\right)=\left(P,e_0,\mathbf{S}_0\right)\right]J_1\left(P,e_0,\mathbf{S}_0\right)\\
&=\begin{cases}
\pi\left(B(P)\phi_{P|P}m(e_0,\mathbf{S}_0)+B(N)\phi_{P|N}\left(1-m(e_0,\mathbf{S}_0)\right)\right), & \text{if } \frac{\pi}{\gamma}<\mathcal{T}_1\left(P,e_0,\mathbf{S}_0\right), \\
\left(\left(A(P)+B(P)\right)\pi-\gamma\right)\phi_{P|P}m(e_0,\mathbf{S}_0) & \\
+\left(\mu\left(A(P)+B(P)\right)\pi-\gamma\right)\phi_{P|N}\left(1-m(e_0,\mathbf{S}_0)\right), & \text{if } \mathcal{T}_1\left(P,e_0,\mathbf{S}_0\right)\leq\frac{\pi}{\gamma}.
\end{cases}
\end{aligned}
\tag{A.69}
$$

Both terms are clearly linear functions of e_0, as is the third term, γe_0. Therefore, the optimal value of e_0 must be 0 or 1.

We next characterize the conditions under which each value is optimal. Recall from (A.13) that for any e_0,

$$
\mathcal{T}_1\left(P,e_0,\mathbf{S}_0\right)\leq\mathcal{T}_1\left(N,e_0,\mathbf{S}_0\right), \tag{A.70}
$$

where from (A.11) and (A.12)

$$
\mathcal{T}_1\left(\mathbf{S}_1\right)=\begin{cases}
\frac{\phi_{P|P}m(e_0,\mathbf{S}_0)+\phi_{P|N}(1-m(e_0,\mathbf{S}_0))}{A(P)\left(\phi_{P|P}m(e_0,\mathbf{S}_0)+\mu\phi_{P|N}(1-m(e_0,\mathbf{S}_0))\right)}, & \text{if } \mathbf{S}_1=\left(P,e_0,\mathbf{S}_0\right), \\
\frac{\left(1-\phi_{P|P}\right)m(e_0,\mathbf{S}_0)+\left(1-\phi_{P|N}\right)(1-m(e_0,\mathbf{S}_0))}{A(P)\left(\left(1-\phi_{P|P}\right)m(e_0,\mathbf{S}_0)+\mu\left(1-\phi_{P|N}\right)(1-m(e_0,\mathbf{S}_0))\right)}, & \text{if } \mathbf{S}_1=\left(N,e_0,\mathbf{S}_0\right).
\end{cases}
\tag{A.71}
$$

and from (1.1), (1.2), and (A.1),

$$m(e_0, \mathbf{S}_0) = \begin{cases} B(\mathbf{S}_0), & \text{if } e_0 = 0, \\ A(\mathbf{S}_0) + B(\mathbf{S}_0), & \text{if } e_0 = 1. \end{cases} \tag{A.72}$$

Furthermore, by Lemma 5,

$$\mathcal{T}_1(P, 1, \mathbf{S}_0) \leq \mathcal{T}_1(P, 0, \mathbf{S}_0) \leq \mathcal{T}_1(N, 1, \mathbf{S}_0) \leq \mathcal{T}_1(N, 0, \mathbf{S}_0). \tag{A.73}$$

Then, we must consider the following regions of the scaled incentive $\frac{\pi}{\gamma}$. First, if $\mathcal{T}_1(N, 0, \mathbf{S}_0) \leq \frac{\pi}{\gamma}$, then $e_1^*(\mathbf{S}_1) = 1$ for all $\mathbf{S}_1$, so (A.67) is

$$\begin{aligned} J_0(\mathbf{S}_0) = \max_{e_0 \in [0,1]} [&((A(P) + B(P))\pi - \gamma) m(e_0, \mathbf{S}_0) \\ &+ (\mu(A(P) + B(P))\pi - \gamma)(1 - m(e_0, \mathbf{S}_0)) - \gamma e_0]. \end{aligned} \tag{A.74}$$

Comparing the case where $e_0 = 0$ to $e_0 = 1$ gives the following result:

$$e_0^* = \begin{cases} 0, & \text{if } \mathcal{T}_1(N, 0, \mathbf{S}_0) \leq \frac{\pi}{\gamma} < \frac{1}{(1-\mu)A(\mathbf{S}_0)(A(P)+B(P))}, \\ 1, & \text{if } \max\left\{\mathcal{T}_1(N, 0, \mathbf{S}_0), \frac{1}{(1-\mu)A(\mathbf{S}_0)(A(P)+B(P))}\right\} \leq \frac{\pi}{\gamma}. \end{cases} \tag{A.75}$$

Second, if $\mathcal{T}_1(N, 1, \mathbf{S}_0) \leq \frac{\pi}{\gamma} < \mathcal{T}_1(N, 0, \mathbf{S}_0)$, then

$$e_1^*(\mathbf{S}_1) = \begin{cases} 0, & \text{if } \mathbf{S}_1 = (N, 0, \mathbf{S}_0), \\ 1, & \text{if } \mathbf{S}_1 = (N, 1, \mathbf{S}_0) \text{ or } \mathbf{S}_1 = (P, e_0, \mathbf{S}_0), \end{cases} \tag{A.76}$$

so (A.67) is $J_0(\mathbf{S}_0) = \max\left[\mathcal{J}^0, \mathcal{J}^1\right]$, where

$$\mathcal{J}^0 = \left(B(P)\pi + (A(P)\pi - \gamma)\phi_{P|P}\right) B(\mathbf{S}_0) + \left(\mu B(P)\pi + (\mu A(P)\pi - \gamma)\phi_{P|N}\right)(1 - B(\mathbf{S}_0)) \tag{A.77}$$

corresponds to the case where the function being maximized in the profit-to-go function is evaluated at $e_0 = 0$, and

$$\begin{aligned}\mathcal{J}^1 =& \left(\left(A(P) + B(P)\right)\pi - \gamma\right)\left(A(\mathbf{S}_0) + B(\mathbf{S}_0)\right) \\ &+ \left(\mu\left(A(P) + B(P)\right)\pi - \gamma\right)\left(1 - \left(A(\mathbf{S}_0) + B(\mathbf{S}_0)\right)\right) - \gamma \end{aligned} \tag{A.78}$$

corresponds to the case where $e_0 = 1$. Now,

$$\mathcal{J}^1 \geq \mathcal{J}^0 \iff \frac{\pi}{\gamma} \geq \mathcal{C}_1\left(\mathbf{S}_0\right), \tag{A.79}$$

where, from (A.40),

$$\begin{aligned}&\mathcal{C}_1\left(\mathbf{S}_0\right) \\ =&\frac{2 - \left(\phi_{P|P} - \phi_{P|N}\right)B(\mathbf{S}_0) - \phi_{P|N}}{(1-\mu)A(\mathbf{S}_0)\left(A(P) + B(P)\right) + A(P)B(\mathbf{S}_0)\left(1 - \phi_{P|P}\right) + \mu A(P)\left(1 - B(\mathbf{S}_0)\right)\left(1 - \phi_{P|N}\right)}.\end{aligned} \tag{A.80}$$

Therefore,

$$e_0^* = \begin{cases} 0, & \text{if } \mathcal{T}_1\left(N, 1, \mathbf{S}_0\right) \leq \frac{\pi}{\gamma} < \min\{\mathcal{T}_1\left(N, 0, \mathbf{S}_0\right), \mathcal{C}_1\left(\mathbf{S}_0\right)\}, \\ 1, & \text{if } \max\{\mathcal{T}_1\left(N, 1, \mathbf{S}_0\right), \mathcal{C}_1\left(\mathbf{S}_0\right)\} \leq \frac{\pi}{\gamma} < \mathcal{T}_1\left(N, 0, \mathbf{S}_0\right). \end{cases} \tag{A.81}$$

Third, if $\mathcal{T}_1\left(P, 0, \mathbf{S}_0\right) \leq \frac{\pi}{\gamma} < \mathcal{T}_1\left(N, 1, \mathbf{S}_0\right)$, then

$$e_1^*\left(\mathbf{S}_1\right) = \begin{cases} 0, & \text{if } \mathbf{S}_1 = \left(N, e_0, \mathbf{S}_0\right), \\ 1, & \text{if } \mathbf{S}_1 = \left(P, e_0, \mathbf{S}_0\right), \end{cases} \tag{A.82}$$

so (A.67) is

$$J_0(\mathbf{S}_0) = \max_{e_0 \in [0,1]} \big[\big(B(P)\pi + (A(P)\pi - \gamma)\, \phi_{P|P} \big)\, m(e_0, \mathbf{S}_0) \\ + \big(\mu B(P)\pi + (\mu A(P)\pi - \gamma)\, \phi_{P|N} \big)\, (1 - m(e_0, \mathbf{S}_0)) - \gamma e_0 \big]. \tag{A.83}$$

Therefore,

$$e_0^* = \begin{cases} 0, & \text{if } \mathcal{T}_1(P, 0, \mathbf{S}_0) \leq \frac{\pi}{\gamma} < \min\{\mathcal{T}_1(N, 1, \mathbf{S}_0), \mathcal{C}_2(\mathbf{S}_0)\}, \\ 1, & \text{if } \max\{\mathcal{T}_1(P, 0, \mathbf{S}_0), \mathcal{C}_2(\mathbf{S}_0)\} \leq \frac{\pi}{\gamma} < \mathcal{T}_1(N, 1, \mathbf{S}_0), \end{cases} \tag{A.84}$$

where, from (A.41),

$$\mathcal{C}_2(\mathbf{S}_0) = \frac{1 + \big(\phi_{P|P} - \phi_{P|N}\big) A(\mathbf{S}_0)}{A(\mathbf{S}_0) \big(\big(A(P)\phi_{P|P} + B(P) \big) - \mu \big(A(P)\phi_{P|N} + B(P) \big) \big)}. \tag{A.85}$$

Fourth, if $\mathcal{T}_1(P, 1, \mathbf{S}_0) \leq \frac{\pi}{\gamma} < \mathcal{T}_1(P, 0, \mathbf{S}_0)$, then

$$e_1^*(\mathbf{S}_1) = \begin{cases} 0, & \text{if } \mathbf{S}_1 = (N, e_0, \mathbf{S}_0) \text{ or } \mathbf{S}_1 = (P, 0, \mathbf{S}_0), \\ 1, & \text{if } \mathbf{S}_1 = (P, 1, \mathbf{S}_0), \end{cases} \tag{A.86}$$

so (A.67) is $J_0(\mathbf{S}_0) = \max\big[\mathcal{J}^0, \mathcal{J}^1\big]$, where

$$\mathcal{J}^0 = \pi B(P) \left(B(\mathbf{S}_0) + \mu \left(1 - B(\mathbf{S}_0) \right) \right) \tag{A.87}$$

corresponds to the case where $e_0 = 0$, and

$$\mathcal{J}^1 = \big(B(P)\pi + (A(P)\pi - \gamma)\, \phi_{P|P} \big) \left(A(\mathbf{S}_0) + B(\mathbf{S}_0) \right) \\ + \big(\mu B(P)\pi + (\mu A(P)\pi - \gamma)\, \phi_{P|N} \big) \left(1 - \left(A(\mathbf{S}_0) + B(\mathbf{S}_0) \right) \right) - \gamma \tag{A.88}$$

and corresponds to the case where $e_0 = 1$. Now,

$$\mathcal{J}^1 \geq \mathcal{J}^0 \iff \frac{\pi}{\gamma} \geq \mathcal{C}_3\left(\mathbf{S}_0\right), \tag{A.89}$$

where, from (A.42),

$$\begin{aligned}\mathcal{C}_3\left(\mathbf{S}_0\right) = &\left(1 + \phi_{P|N} + \left(\phi_{P|P} - \phi_{P|N}\right)\left(A(\mathbf{S}_0) + B(\mathbf{S}_0)\right)\right) \\ &\div \left(A(\mathbf{S}_0)\left(A(P)\phi_{P|P} + B(P) - \mu\left(A(P)\phi_{P|N} + B(P)\right)\right)\right. \\ &\left.+A(P)\left(\phi_{P|P} - \mu\phi_{P|N}\right)B(\mathbf{S}_0) + \mu A(P)\phi_{P|N}\right).\end{aligned} \tag{A.90}$$

and from (A.24) in Lemma 4,

$$\max\{\mathcal{T}_1\left(P, 1, \mathbf{S}_0\right), \mathcal{C}_3\left(\mathbf{S}_0\right)\} = \mathcal{C}_3\left(\mathbf{S}_0\right). \tag{A.91}$$

Therefore,

$$e_0^* = \begin{cases} 0, & \text{if } \mathcal{T}_1\left(P, 1, \mathbf{S}_0\right) \leq \frac{\pi}{\gamma} < \min\{\mathcal{T}_1\left(P, 0, \mathbf{S}_0\right), \mathcal{C}_3\left(\mathbf{S}_0\right)\}, \\ 1, & \text{if } \mathcal{C}_3\left(\mathbf{S}_0\right) \leq \frac{\pi}{\gamma} < \mathcal{T}_1\left(P, 0, \mathbf{S}_0\right). \end{cases} \tag{A.92}$$

Finally, if $\frac{\pi}{\gamma} < \mathcal{T}_1\left(P, 1, \mathbf{S}_0\right)$, then $e_1^*\left(\mathbf{S}_1\right) = 0$ for all $\mathbf{S}_1$, so (A.67) is

$$J_0(\mathbf{S}_0) = \max_{e_0 \in [0,1]} \left[\pi B(P)\left(m(e_0, \mathbf{S}_0) + \mu\left(1 - m(e_0, \mathbf{S}_0)\right)\right) - \gamma e_0\right]. \tag{A.93}$$

From (A.25) in Lemma 4,

$$\min\left\{\mathcal{T}_1\left(P, 1, \mathbf{S}_0\right), \frac{1}{(1-\mu)B(P)A(\mathbf{S}_0)}\right\} = \mathcal{T}_1\left(P, 1, \mathbf{S}_0\right). \tag{A.94}$$

Therefore,

$$e_0^* = 0, \quad \text{if } \frac{\pi}{\gamma} < \mathcal{T}_1\left(P, 1, \mathbf{S}_0\right). \tag{A.95}$$

Combining this and applying Lemmas 6 and 7 gives the optimal effort level at $t = 0$:

$$e_0^*(\mathbf{S}_0) = \begin{cases} 0, & \text{if } \frac{\pi}{\gamma} < \mathcal{T}_0(\mathbf{S}_0), \\ 1, & \text{if } \mathcal{T}_0(\mathbf{S}_0) \leq \frac{\pi}{\gamma}, \end{cases} \tag{A.96}$$

where

$$\mathcal{T}_0(P) = \begin{cases} \frac{1}{(1-\mu)A(P)(A(P)+B(P))}, & \text{if } 0 \leq (1-\mu)(A(P)+B(P)) \leq \mathcal{E}_1(P), \\ \mathcal{C}_1(P), & \text{if } \mathcal{E}_1(P) < (1-\mu)(A(P)+B(P)) \leq \mathcal{E}_2(P), \\ \mathcal{C}_2(P), & \text{if } \mathcal{E}_2(P) < (1-\mu)(A(P)+B(P)) \leq \mathcal{E}_3(P), \\ \mathcal{C}_3(P), & \text{if } \mathcal{E}_3(P) < (1-\mu)(A(P)+B(P)), \end{cases} \tag{A.97}$$

and

$$\mathcal{T}_0(N) = \frac{1}{\mu(1-\mu)A(P)(A(P)+B(P))}. \tag{A.98}$$

Under the interim assessment, the threshold when school is in the proficient state at the beginning of the year simplifies to

$$\mathcal{T}_0(P) = \begin{cases} \frac{1}{(1-\mu)A(P)(A(P)+B(P))}, & \text{if } 0 \leq (1-\mu)(A(P)+B(P)) \leq \mu, \\ \frac{1+A(P)}{A(P)(A(P)+(1-\mu)B(P))}, & \text{if } \mu < (1-\mu)(A(P)+B(P)). \end{cases} \tag{A.99}$$

We claim that if the school starts the year in the not-proficient state, the threshold, given in (A.98), is decreasing in the response-to-effort and stickiness of the proficient state. To see this, note that

$$\frac{\partial}{\partial A(P)}\mathcal{T}_0(N) = -\frac{2A(P)+B(P)}{\mu(1-\mu)A(P)(A(P)+B(P))^2} < 0, \tag{A.100}$$

and it is clear that $\mathcal{T}_0(N)$ is decreasing in $B(P)$. On the other hand, note that

$$\frac{\partial}{\partial \mu}\mathcal{T}_0(N) = -\frac{1-2\mu}{\mu^2(1-\mu)^2 A(P)\left(A(P)+B(P)\right)}, \tag{A.101}$$

which is negative for $0 \leq \mu \leq \frac{1}{2}$ and positive for $\frac{1}{2} \leq \mu \leq 1$. Additionally, we claim that $\mathcal{T}_1\left(X_1, e_0, N\right) \leq \mathcal{T}_0(N)$. From Lemma 5, $\mathcal{T}_1\left(P, e_0, N\right) \leq \mathcal{T}_1\left(N, e_0, N\right)$, so it suffices to show that $\mathcal{T}_1\left(N, e_0, N\right) \leq \mathcal{T}_0(N)$. Applying Assumption 1 to (A.11),

$$\mathcal{T}_1\left(N, e_0, N\right) = \frac{\mu\left(A(P)e_0 + B(P)\right)\left(1-\phi_{P|P}\right) + \left(1-\mu\left(A(P)e_0+B(P)\right)\right)\left(1-\phi_{P|N}\right)}{\mu A(P)\left(\left(A(P)e_0+B(P)\right)\left(1-\phi_{P|P}\right) + \left(1-\mu\left(A(P)e_0+B(P)\right)\right)\left(1-\phi_{P|N}\right)\right)}. \tag{A.102}$$

Then,

$$\begin{aligned}
&\mathcal{T}_1\left(N, e_0, N\right) \leq \mathcal{T}_0(N) \\
\iff &-\left(1-\mu\left(A(P)e_0+B(P)\right)\right)\left(1-(1-\mu)\left(A(P)+B(P)\right)\right)\left(1-\phi_{P|N}\right) \\
&\leq \left(A(P)e_0+B(P)\right)\left(1-\mu(1-\mu)\left(A(P)+B(P)\right)\right)\left(1-\phi_{P|P}\right),
\end{aligned} \tag{A.103}$$

which always holds under Assumptions 1 and 2. Thus, the claim is true.

Next, consider the change in the threshold levels with respect to μ when the school begins the year in the proficient case, first note that under Assumptions 1 and 2, $\mathcal{E}_1(P) > 0$ and

$$\begin{aligned}
&\mathcal{E}_1(P) \leq A(P) + B(P) \\
\iff &\mu \leq A(P) + B(P) - \frac{B(P)\left(1-\left(A(P)+B(P)\right)\right)\left(1-\phi_{P|P}\right)}{\left(1-B(P)\right)\left(1-\phi_{P|N}\right)},
\end{aligned} \tag{A.104}$$

where the right-hand side of the above inequality is positive:

$$\begin{aligned}
&A(P) + B(P) - \frac{B(P)\left(1-\left(A(P)+B(P)\right)\right)\left(1-\phi_{P|P}\right)}{\left(1-B(P)\right)\left(1-\phi_{P|N}\right)} > 0 \\
\iff &A(P) > -B(P)\left(1-B(P)\right)\left(\frac{\phi_{P|P}-\phi_{P|N}}{\left(1-B(P)\right)\left(1-\phi_{P|N}\right) + B(P)\left(1-\phi_{P|P}\right)}\right).
\end{aligned} \tag{A.105}$$

Therefore, there exists $\bar{\mu} \in (0,1)$ such that for $\mu \in [\bar{\mu}, 1]$, $0 \leq (1-\mu)(A(P)+B(P)) \leq \mathcal{E}_1(P)$, and therefore, $\mathcal{T}_0(P) = \frac{1}{(1-\mu)A(P)(A(P)+B(P))}$. Now, the threshold is clearly increasing in μ in this region. Furthermore, we can solve for $\bar{\mu}$:

$$\begin{aligned}
&(1-\bar{\mu})(A(P)+B(P)) = \mathcal{E}_1(P) \\
\iff \bar{\mu} &= \frac{(1-B(P))(A(P)+B(P))\left(1-\phi_{P|N}\right) - B(P)\left(1-(A(P)+B(P))\right)\left(1-\phi_{P|P}\right)}{(1-B(P))\left(1+(A(P)+B(P))\right)\left(1-\phi_{P|N}\right) + B(P)(A(P)+B(P))\left(1-\phi_{P|P}\right)}.
\end{aligned} \tag{A.106}$$

Based on the first terms in the numerator and denominator, it is clear that the denominator is at least twice the numerator, so $\bar{\mu} \leq \frac{1}{2}$ for all values of the parameters. In the case of the interim assessment, this result is even stronger: the threshold is always increasing in μ.

□

Lemma 8 *Recall that $e_0^*(\mathbf{S}_0)$ and $e_1^*(\mathbf{S}_1)$ are the optimal teachers' effort policies. Then, for the case where the district chooses to rely only on the formative assessment,*

$$\begin{aligned}
Pr^*[\mathbf{S}_2 = P|\mathbf{S}_0 = P] = &\left(1-\delta\left(e_0^*(P)\right)\right)\left(1-\delta\left(e_1^*\left(P, e_0^*(P), P\right)\right)\right)\phi_{P|P} \\
&+ \delta\left(e_0^*(P)\right)\alpha\left(e_1^*\left(P, e_0^*(P), P\right)\right)\phi_{P|N} \\
&+ \left(1-\delta\left(e_0^*(P)\right)\right)\left(1-\delta\left(e_1^*\left(N, e_0^*(P), P\right)\right)\right)\left(1-\phi_{P|P}\right) \\
&+ \delta\left(e_0^*(P)\right)\alpha\left(e_1^*\left(N, e_0^*(P), P\right)\right)\left(1-\phi_{P|N}\right),
\end{aligned} \tag{A.107}$$

$$\begin{aligned}
Pr^*[\mathbf{S}_2 = P|\mathbf{S}_0 = N] = &\alpha\left(e_0^*(N)\right)\left(1-\delta\left(e_1^*\left(P, e_0^*(N), N\right)\right)\right)\phi_{P|P} \\
&+ \left(1-\alpha\left(e_0^*(N)\right)\right)\alpha\left(e_1^*\left(P, e_0^*(N), N\right)\right)\phi_{P|N} \\
&+ \alpha\left(e_0^*(N)\right)\left(1-\delta\left(e_1^*\left(N, e_0^*(N), N\right)\right)\right)\left(1-\phi_{P|P}\right) \\
&+ \left(1-\alpha\left(e_0^*(N)\right)\right)\alpha\left(e_1^*\left(N, e_0^*(N), N\right)\right)\left(1-\phi_{P|N}\right).
\end{aligned} \tag{A.108}$$

The probability that the final state is proficient if the district relies on the interim assessment is obtained by letting $\phi_{P|P} = 1$ and $\phi_{P|N} = 0$ in (A.107) *and* (A.108).

Proof of Lemma 8

When $\mathbf{S}_0 = P$, we have

$$Pr\left[\beta_1 = P|\mathbf{S}_1\right] = \begin{cases} \frac{\phi_{P|P}\left(1-\delta\left(e_0^*(P)\right)\right)}{\phi_{P|P}\left(1-\delta\left(e_0^*(P)\right)\right)+\phi_{P|N}\left(\delta\left(e_0^*(P)\right)\right)}, & \text{if } \mathbf{S}_1 = (P, e_0^*(P), \mathbf{S}_0)\,, \\ \frac{\left(1-\phi_{P|P}\right)\left(1-\delta\left(e_0^*(P)\right)\right)}{\left(1-\phi_{P|P}\right)\left(1-\delta\left(e_0^*(P)\right)\right)+\left(1-\phi_{P|N}\right)\left(\delta\left(e_0^*(P)\right)\right)}, & \text{if } \mathbf{S}_1 = (N, e_0^*(P), \mathbf{S}_0)\,. \end{cases} \tag{A.109}$$

Then, with (A.2) and (A.3),

$$\begin{aligned} Pr^*[\mathbf{S}_2 = P|\mathbf{S}_0 = P] =& \left(1 - \delta\left(e_1^*\left(X_1, e_0^*(P), P\right)\right)\right) Pr\left[\beta_1 = P|\mathbf{S}_1\right] \\ &+ \alpha\left(e_1^*\left(X_1, e_0^*(P), P\right)\right)\left(1 - Pr\left[\beta_1 = P|\mathbf{S}_1\right]\right) \\ =& Pr\left[\mathbf{S}_1 = (P, e_0^*(P), P)\right] \\ &\times \left(\left(1 - \delta\left(e_1^*\left(P, e_0^*\left(P\right), P\right)\right)\right) Pr\left[\beta_1 = P|\mathbf{S}_1 = (P, e_0^*(P), \mathbf{S}_0)\right]\right. \\ &\left.+\alpha\left(e_1^*\left(P, e_0^*(P), P\right)\right)\left(1 - Pr\left[\beta_1 = P|\mathbf{S}_1 = (P, e_0^*(P), \mathbf{S}_0)\right]\right)\right) \\ &+ Pr\left[\mathbf{S}_1 = (N, e_0^*(P), P)\right] \\ &\times \left(\left(1 - \delta\left(e_1^*\left(N, e_0^*\left(P\right), P\right)\right)\right) Pr\left[\beta_1 = P|\mathbf{S}_1 = (N, e_0^*(P), \mathbf{S}_0)\right]\right. \\ &\left.+\alpha\left(e_1^*\left(N, e_0^*(P), P\right)\right)\left(1 - Pr\left[\beta_1 = P|\mathbf{S}_1 = (N, e_0^*(P), \mathbf{S}_0)\right]\right)\right). \end{aligned} \tag{A.110}$$

Simplifying this gives (A.107). Following similar steps for when $\mathbf{S}_0 = N$ gives (A.108). □

Proposition 14 *Suppose the condition in* (A.22) *holds. Then, the probability that the state at* $t = 2$ *is proficient* $(\mathbf{S}_2 = P)$ *can be described as follows.*

a) For $\mathbf{S}_0 = N$,

$$Pr^*[\mathbf{S}_2 = P|\mathbf{S}_0 = N]$$
$$= \begin{cases} \mu B(P)\left(1+(1-\mu)B(P)\right), & \text{if } 0 \le \frac{\pi}{\gamma} < \mathcal{T}_1(P,0,N), \\ \mu\left(B(P)\left(A(P)\phi_{P|P}+B(P)\right)\right. & \\ \quad \left.+(1-\mu B(P))\left(A(P)\phi_{P|N}+B(P)\right)\right), & \text{if } \mathcal{T}_1(P,0,N) \le \frac{\pi}{\gamma} < \mathcal{T}_1(N,0,N), \\ \mu\left(A(P)+B(P)\right)\left(1+(1-\mu)B(P)\right), & \text{if } \mathcal{T}_1(N,0,N) \le \frac{\pi}{\gamma} < \mathcal{T}_0(N), \\ \mu\left(A(P)+B(P)\right)\left(1+(1-\mu)\left(A(P)+B(P)\right)\right), & \text{if } \mathcal{T}_0(N) \le \frac{\pi}{\gamma}. \end{cases} \tag{A.111}$$

b) For $\mathbf{S}_0 = P$ *and* $0 \le (1-\mu)\left(A(P)+B(P)\right) \le \mathcal{E}_1(P)$,

$$Pr^*[\mathbf{S}_2 = P|\mathbf{S}_0 = P]$$
$$= \begin{cases} B(P)\left(\mu+(1-\mu)B(P)\right), & \text{if } 0 \le \frac{\pi}{\gamma} < \mathcal{T}_1(P,0,P), \\ B(P)\left(A(P)\phi_{P|P}+B(P)\right) & \\ \quad +\mu\left(1-B(P)\right)\left(A(P)\phi_{P|N}+B(P)\right), & \text{if } \mathcal{T}_1(P,0,P) \le \frac{\pi}{\gamma} < \mathcal{T}_1(N,0,P), \\ \left(A(P)+B(P)\right)\left(\mu+(1-\mu)B(P)\right), & \text{if } \mathcal{T}_1(N,0,P) \le \frac{\pi}{\gamma} < \mathcal{T}_0(P), \\ \left(A(P)+B(P)\right)\left(\mu+(1-\mu)\left(A(P)+B(P)\right)\right), & \text{if } \mathcal{T}_0(P) \le \frac{\pi}{\gamma}. \end{cases} \tag{A.112}$$

c) For $\mathbf{S}_0 = P$ *and* $\mathcal{E}_1(P) < (1-\mu)\left(A(P)+B(P)\right) \leq \mathcal{E}_2(P)$,

$$Pr^*[\mathbf{S}_2 = P|\mathbf{S}_0 = P]$$
$$= \begin{cases} B(P)\left(\mu + (1-\mu)B(P)\right), & \text{if } 0 \leq \frac{\pi}{\gamma} < \mathcal{T}_1\left(P,0,P\right), \\ B(P)\left(A(P)\phi_{P|P} + B(P)\right) & \\ +\mu\left(1-B(P)\right)\left(A(P)\phi_{P|N} + B(P)\right), & \text{if } \mathcal{T}_1\left(P,0,P\right) \leq \frac{\pi}{\gamma} < \mathcal{T}_0\left(P\right), \\ \left(A(P)+B(P)\right)\left(\mu + (1-\mu)\left(A(P)+B(P)\right)\right), & \text{if } \mathcal{T}_0\left(P\right) \leq \frac{\pi}{\gamma}. \end{cases} \tag{A.113}$$

d) For $\mathbf{S}_0 = P$ *and* $\mathcal{E}_2(P) < (1-\mu)\left(A(P)+B(P)\right) \leq \mathcal{E}_3(P)$,

$$Pr^*[\mathbf{S}_2 = P|\mathbf{S}_0 = P]$$
$$= \begin{cases} B(P)\left(\mu + (1-\mu)B(P)\right), & \text{if } 0 \leq \frac{\pi}{\gamma} < \mathcal{T}_1\left(P,0,P\right), \\ B(P)\left(A(P)\phi_{P|P} + B(P)\right) & \\ +\mu\left(1-B(P)\right)\left(A(P)\phi_{P|N} + B(P)\right), & \text{if } \mathcal{T}_1\left(P,0,P\right) \leq \frac{\pi}{\gamma} < \mathcal{T}_0\left(P\right), \\ \left(A(P)+B(P)\right)\left(A(P)\phi_{P|P} + B(P)\right) & \\ +\mu\left(1-\left(A(P)+B(P)\right)\right)\left(A(P)\phi_{P|N} + B(P)\right), & \text{if } \mathcal{T}_0\left(P\right) \leq \frac{\pi}{\gamma} < \mathcal{T}_1\left(N,1,P\right), \\ \left(A(P)+B(P)\right)\left(\mu + (1-\mu)\left(A(P)+B(P)\right)\right), & \text{if } \mathcal{T}_1\left(N,1,P\right) \leq \frac{\pi}{\gamma}. \end{cases} \tag{A.114}$$

e) For $\mathbf{S}_0 = P$ *and* $\mathcal{E}_3(P) < (1-\mu)\left(A(P)+B(P)\right) \leq 1$,

$$
\begin{aligned}
&Pr^*[\mathbf{S}_2 = P | \mathbf{S}_0 = P] \\
&= \begin{cases}
B(P)\left(\mu + (1-\mu)B(P)\right), & \text{if } 0 \leq \frac{\pi}{\gamma} < \mathcal{T}_0\left(P\right), \\
\left(A(P)+B(P)\right)\left(A(P)\phi_{P|P} + B(P)\right) & \\
+\mu\left(1-\left(A(P)+B(P)\right)\right)\left(A(P)\phi_{P|N} + B(P)\right), & \text{if } \mathcal{T}_0\left(P\right) \leq \frac{\pi}{\gamma} < \mathcal{T}_1\left(N,1,P\right), \\
\left(A(P)+B(P)\right)\left(\mu + (1-\mu)\left(A(P)+B(P)\right)\right), & \text{if } \mathcal{T}_1\left(N,1,P\right) \leq \frac{\pi}{\gamma}.
\end{cases}
\end{aligned}
\tag{A.115}
$$

Proof of Proposition 14

First, consider the probability that the final state is proficient when the district relies on the formative assessment.

Suppose that $\mathbf{S}_0 = N$. Then, from (A.98) in Proposition 2, the cost-to-incentive ratio threshold in period 1 is

$$
\mathcal{T}_0(N) = \frac{1}{\mu(1-\mu)A(P)\left(A(P)+B(P)\right)}, \tag{A.116}
$$

and, from (A.15) , the period 2 threshold when $\mathbf{S}_0 = N$ is

$$
\mathcal{T}_1\left(\mathbf{S}_1\right) = \begin{cases}
\frac{\mu(A(P)e_0+B(P))\phi_{P|P}+(1-\mu(A(P)e_0+B(P)))\phi_{P|N}}{\mu A(P)\left((A(P)e_0+B(P))\phi_{P|P}+(1-\mu(A(P)e_0+B(P)))\phi_{P|N}\right)}, & \text{if } \mathbf{S}_1 = (P, e_0, N), \\
\frac{\mu(A(P)e_0+B(P))\left(1-\phi_{P|P}\right)+(1-\mu(A(P)e_0+B(P)))\left(1-\phi_{P|N}\right)}{\mu A(P)\left((A(P)e_0+B(P))\left(1-\phi_{P|P}\right)+(1-\mu(A(P)e_0+B(P)))\left(1-\phi_{P|N}\right)\right)}, & \text{if } \mathbf{S}_1 = (N, e_0, N).
\end{cases}
\tag{A.117}
$$

From Lemma 5 and from (A.43) in Lemma 6,

$$
\mathcal{T}_1\left(P,1,N\right) \leq \mathcal{T}_1\left(P,0,N\right) \leq \mathcal{T}_1\left(N,1,N\right) \leq \mathcal{T}_1\left(N,0,N\right) \leq \mathcal{T}_0(N). \tag{A.118}
$$

For each possible region of $\frac{\pi}{\gamma}$, it is straightforward to determine the optimal effort in each

period and, plugging that into (A.108), the probability that final state is proficient. Then, (A.111) follows.

Next, suppose that $\mathbf{S}_0 = P$. From (A.97) in Proposition 2, the period 1 cost-to-incentive ratio threshold is

$$\mathcal{T}_0(P) = \begin{cases} \frac{1}{(1-\mu)A(P)(A(P)+B(P))}, & \text{if } 0 \leq (1-\mu)\left(A(P)+B(P)\right) \leq \mathcal{E}_1(P), \\ \mathcal{C}_1(P), & \text{if } \mathcal{E}_1(P) < (1-\mu)\left(A(P)+B(P)\right) \leq \mathcal{E}_2(P), \\ \mathcal{C}_2(P), & \text{if } \mathcal{E}_2(P) < (1-\mu)\left(A(P)+B(P)\right) \leq \mathcal{E}_3(P), \\ \mathcal{C}_3(P), & \text{if } \mathcal{E}_3(P) < (1-\mu)\left(A(P)+B(P)\right) \leq 1, \end{cases} \tag{A.119}$$

where, as given above, in (A.37)–(A.42),

$$\mathcal{E}_1(P) = \frac{B(P)\left(1-\phi_{P|P}\right) + \mu\left(1-B(P)\right)\left(1-\phi_{P|N}\right)}{B(P)\left(1-\phi_{P|P}\right) + \left(1-B(P)\right)\left(1-\phi_{P|N}\right)}, \tag{A.120}$$

$$\begin{aligned} \mathcal{E}_2(P) = &\left(\left(A(P)+B(P)\right)\left(1-\phi_{P|P}\right) + \mu\left(1-\left(A(P)+B(P)\right)\right)\left(1-\phi_{P|N}\right)\right. \\ &\left. + A(P)\left(1-\mu\right)\left(1-\phi_{P|N}\right)\left(1-\phi_{P|P}\right)\right) \\ &\div \left(\left(1-\phi_{P|N}\right) - \left(\phi_{P|P}-\phi_{P|N}\right)\left(A(P)+B(P)\right)\right), \end{aligned} \tag{A.121}$$

$$\begin{aligned} \mathcal{E}_3(P) = &\frac{\left(B(P)\phi_{P|P} + \mu\left(1-B(P)\right)\phi_{P|N}\right)\left(1+\left(\phi_{P|P}-\phi_{P|N}\right)A(P)\right)}{\left(B(P)\phi_{P|P} + \left(1-B(P)\right)\phi_{P|N}\right)} \\ &+ A(P)\left(\left(1-\phi_{P|P}\right) - \mu\left(1-\phi_{P|N}\right)\right), \end{aligned} \tag{A.122}$$

$$\mathcal{C}_1(P) = \left(2 - \left(\phi_{P|P}-\phi_{P|N}\right)B(P) - \phi_{P|N}\right) \tag{A.123}$$

$$\begin{aligned} &\div \left((1-\mu)A(P)\left(A(P)+B(P)\right) + A(P)B(P)\left(1-\phi_{P|P}\right)\right. \\ &\left. +\mu A(P)\left(1-B(P)\right)\left(1-\phi_{P|N}\right)\right), \end{aligned} \tag{A.124}$$

$$\mathcal{C}_2(P) = \frac{1+\left(\phi_{P|P}-\phi_{P|N}\right)A(P)}{A(P)\left(\left(A(P)\phi_{P|P}+B(P)\right) - \mu\left(A(P)\phi_{P|N}+B(P)\right)\right)}, \tag{A.125}$$

$$\begin{aligned} \mathcal{C}_3(P) = &\left(1+\phi_{P|N}+\left(\phi_{P|P}-\phi_{P|N}\right)\left(A(P)+B(P)\right)\right) \\ &\div \left(A(P)\left(A(P)\phi_{P|P}+B(P) - \mu\left(A(P)\phi_{P|N}+B(P)\right)\right)\right. \\ &\left. +A(P)\left(\phi_{P|P}-\mu\phi_{P|N}\right)B(P) + \mu A(P)\phi_{P|N}\right). \end{aligned} \tag{A.126}$$

From (A.11) and (A.12) in Proposition 1, the period 2 threshold is

$$\mathcal{T}_1(\mathbf{S}_1) = \begin{cases} \frac{(A(P)e_0+B(P))\phi_{P|P}+(1-(A(P)e_0+B(P)))\phi_{P|N}}{A(P)\left((A(P)e_0+B(P))\phi_{P|P}+\mu(1-(A(P)e_0+B(P)))\phi_{P|N}\right)}, & \text{if } \mathbf{S}_1 = (P, e_0, P), \\ \frac{(A(P)e_0+B(P))\left(1-\phi_{P|P}\right)+(1-(A(P)e_0+B(P)))\left(1-\phi_{P|N}\right)}{A(P)\left((A(P)e_0+B(P))\left(1-\phi_{P|P}\right)+\mu(1-(A(P)e_0+B(P)))\left(1-\phi_{P|N}\right)\right)}, & \text{if } \mathbf{S}_1 = (N, e_0, P). \end{cases} \tag{A.127}$$

When $0 \leq (1-\mu)(A(P)+B(P)) \leq \mathcal{E}_1(P)$, from Lemma 5 and from (A.43) in Lemma 6,

$$\mathcal{T}_1(P,1,P) \leq \mathcal{T}_1(P,0,P) \leq \mathcal{T}_1(N,1,P) \leq \mathcal{T}_1(N,0,P) \leq \mathcal{T}_0(P). \tag{A.128}$$

Then, applying (A.108), (A.112) follows.

When $\mathcal{E}_1(P) < (1-\mu)(A(P)+B(P)) \leq \mathcal{E}_2(P)$, from Lemma 5 and from (A.44) and (A.45) in Lemma 6,

$$\mathcal{T}_1(P,1,P) \leq \mathcal{T}_1(P,0,P) \leq \mathcal{T}_1(N,1,P) \leq \mathcal{T}_0(P) \leq \mathcal{T}_1(N,0,P). \tag{A.129}$$

Then, applying (A.108), (A.113) follows.

If $\mathcal{E}_2(P) < (1-\mu)(A(P)+B(P)) \leq \mathcal{E}_3(P)$, from Lemma 5 and from (A.46) and (A.47) in Lemma 6,

$$\mathcal{T}_1(P,1,P) \leq \mathcal{T}_1(P,0,P) \leq \mathcal{T}_0(P) \leq \mathcal{T}_1(N,1,P) \leq \mathcal{T}_1(N,0,P). \tag{A.130}$$

Then, applying (A.108), (A.114) follows.

Finally, if $\mathcal{E}_3(P) < (1-\mu)(A(P)+B(P)) \leq 1$, from Lemma 5, from (A.48) in Lemma 6, and from (A.24) in Lemma 4,

$$\mathcal{T}_1(P,1,P) \leq \mathcal{T}_0(P) \leq \mathcal{T}_1(P,0,P) \leq \mathcal{T}_1(N,1,P) \leq \mathcal{T}_1(N,0,P). \tag{A.131}$$

Then, applying (A.108), (A.115) follows.

Note that in each case, it is straightforward to show that the the probability that final state is proficient is non-decreasing in the scaled incentive threshold. □

Proposition 15 *Suppose that* $B(P) = 0$. *Then, the probability that the state at* $t = 2$ *is proficient* $(\mathbf{S}_2 = P)$ *can be described as follows.*

a) For $\mathbf{S}_0 = N$,

$$\tilde{Pr}^*[\mathbf{S}_2 = P|\mathbf{S}_0 = N] = \begin{cases} 0, & \text{if } 0 \leq \frac{\pi}{\gamma} < \frac{1}{\mu A(P)}, \\ \mu A(P), & \text{if } \frac{1}{\mu A(P)} \leq \frac{\pi}{\gamma} < \tilde{\mathcal{T}}_0(N), \\ \mu A(P)\left(1 + (1-\mu)A(P)\right), & \text{if } \tilde{\mathcal{T}}_0(N) \leq \frac{\pi}{\gamma}. \end{cases} \tag{A.132}$$

b) For $\mathbf{S}_0 = P$ *and* $0 \leq (1-\mu)A(P) \leq \mu$,

$$\tilde{Pr}^*[\mathbf{S}_2 = P|\mathbf{S}_0 = P] = \begin{cases} 0, & \text{if } 0 \leq \frac{\pi}{\gamma} < \frac{1}{\mu A(P)}, \\ \mu A(P), & \text{if } \frac{1}{\mu A(P)} \leq \frac{\pi}{\gamma} < \tilde{\mathcal{T}}_0(P), \\ A(P)\left(\mu + (1-\mu)A(P)\right), & \text{if } \tilde{\mathcal{T}}_0(P) \leq \frac{\pi}{\gamma}. \end{cases} \tag{A.133}$$

c) For $\mathbf{S}_0 = P$ *and* $\mu < (1-\mu)A(P) \leq \frac{A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)}{(1-A(P))\left(\phi_{P|P}-\phi_{P|N}\right)}$,

$$\tilde{Pr}^*[\mathbf{S}_2 = P|\mathbf{S}_0 = P] = \begin{cases} 0, & \text{if } 0 \leq \frac{\pi}{\gamma} < \tilde{\mathcal{T}}_0(P), \\ A(P)\left(A(P) + \mu\left(1 - A(P)\right)\right), & \text{if } \tilde{\mathcal{T}}_0(P) \leq \frac{\pi}{\gamma}. \end{cases} \tag{A.134}$$

d) For $\mathbf{S}_0 = P$ *and* $\frac{A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)}{(1-A(P))\left(\phi_{P|P}-\phi_{P|N}\right)} < (1-\mu)A(P) \leq 1$,

$$\begin{aligned}
&\tilde{Pr}^*[\mathbf{S}_2 = P|\mathbf{S}_0 = P] \\
&= \begin{cases} 0, & \text{if } 0 \leq \frac{\pi}{\gamma} < \tilde{\mathcal{T}}_0(P), \\ A(P)\left(A(P)\phi_{P|P} + \mu\left(1-A(P)\right)\phi_{P|N}\right), & \text{if } \tilde{\mathcal{T}}_0(P) \leq \frac{\pi}{\gamma} < \tilde{\mathcal{T}}_1(N,1,P), \\ A(P)\left(A(P) + \mu(1-A(P))\right), & \text{if } \tilde{\mathcal{T}}_1(N,1,P) \leq \frac{\pi}{\gamma}. \end{cases}
\end{aligned} \tag{A.135}$$

Under the interim assessment, the case for $\mathbf{S}_0 = P$ *simplifies as follows.*

a) For $\mathbf{S}_0 = P$ *and* $0 \leq (1-\mu)A(P) \leq \mu$,

$$\tilde{Pr}^*[\mathbf{S}_2 = P|\mathbf{S}_0 = P] = \begin{cases} 0, & \text{if } 0 \leq \frac{\pi}{\gamma} < \frac{1}{\mu A(P)}, \\ \mu A(P), & \text{if } \frac{1}{\mu A(P)} \leq \frac{\pi}{\gamma} < \tilde{\mathcal{T}}_0(P), \\ A(P)\left(\mu + (1-\mu)A(P)\right), & \text{if } \tilde{\mathcal{T}}_0(P) \leq \frac{\pi}{\gamma}. \end{cases} \tag{A.136}$$

b) For $\mathbf{S}_0 = P$ *and* $\mu < (1-\mu)A(P) \leq 1$,

$$\tilde{Pr}^*[\mathbf{S}_2 = P|\mathbf{S}_0 = P] = \begin{cases} 0, & \text{if } 0 \leq \frac{\pi}{\gamma} < \tilde{\mathcal{T}}_0(P), \\ (A(P))^2, & \text{if } \tilde{\mathcal{T}}_0(P) \leq \frac{\pi}{\gamma} < \tilde{\mathcal{T}}_1(N,1,P), \\ A(P)\left(A(P) + \mu\left(1-A(P)\right)\right), & \text{if } \tilde{\mathcal{T}}_1(N,1,P) \leq \frac{\pi}{\gamma}. \end{cases} \tag{A.137}$$

Proof of Proposition 15

To establish the results of this Proposition, we use the following result.

Lemma 9 *Suppose that* $B(P) = 0$. *(In this case, we do not assume that* (A.22) *holds.)*

Then the optimal effort level in the first half of the year is as follows:

$$\tilde{e}_0^*(\mathbf{S}_0) = \begin{cases} 0, & \text{if } \frac{\pi}{\gamma} < \tilde{\mathcal{T}}_0(\mathbf{S}_0), \\ 1, & \text{if } \tilde{\mathcal{T}}_0(\mathbf{S}_0) \leq \frac{\pi}{\gamma}, \end{cases} \tag{A.138}$$

where

$$\tilde{\mathcal{T}}_0(P) = \begin{cases} \frac{1}{(1-\mu)(A(P))^2}, & \text{if } 0 \leq (1-\mu)A(P) \leq \mu, \\ \frac{2}{A(P)(A(P)+\mu(1-A(P)))}, & \text{if } \mu < (1-\mu)A(P) \\ & \leq \frac{A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)}{(1-A(P))\left(\phi_{P|P}-\phi_{P|N}\right)}, \\ \frac{1+A(P)\phi_{P|P}+(1-A(P))\phi_{P|N}}{A(P)\left(A(P)\phi_{P|P}+\mu(1-A(P))\phi_{P|N}\right)}, & \text{if } \frac{A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)}{(1-A(P))\left(\phi_{P|P}-\phi_{P|N}\right)} \\ & < (1-\mu)A(P), \end{cases} \tag{A.139}$$

and

$$\tilde{\mathcal{T}}_0(N) = \frac{1}{\mu(1-\mu)(A(P))^2}. \tag{A.140}$$

Proof of Lemma 9

As in the proof of Proposition 2, the profit-to-go function for $t = 0$ is

$$\begin{aligned} J_0(\mathbf{S}_0) = \max_{e_0 \in [0,1]} \big[& Pr\left[h_1\left(e_0, \mathbf{S}_0\right) = \left(N, e_0, \mathbf{S}_0\right)\right] J_1\left(N, e_0, \mathbf{S}_0\right) \\ & + Pr\left[h_1\left(e_0, \mathbf{S}_0\right) = \left(P, e_0, \mathbf{S}_0\right)\right] J_1\left(P, e_0, \mathbf{S}_0\right) - \gamma e_0 \big], \end{aligned} \tag{A.141}$$

with $Pr\left[h_1\left(e_0, \mathbf{S}_0\right)\right]$ given in (A.2) and $J_1\left(\mathbf{S}_1\right)$ given in (A.10), and e_0 must equal 0 or 1.

For clarity, we use a tilde to denote function values for the case where $B(P) = 0$. Then,

from (A.68) and (A.69),

$$\begin{aligned}
&\tilde{Pr}\left[h_1\left(e_0,\mathbf{S}_0\right)=\left(N,e_0,\mathbf{S}_0\right)\right]\tilde{J}_1\left(N,e_0,\mathbf{S}_0\right)\\
&=\begin{cases}
0, & \text{if } \frac{\pi}{\gamma}<\tilde{\mathcal{T}}_1\left(N,e_0,\mathbf{S}_0\right),\\
\left(A(P)\pi-\gamma\right)\left(1-\phi_{P|P}\right)\tilde{m}(e_0,\mathbf{S}_0) & \\
+\left(\mu A(P)\pi-\gamma\right)\left(1-\phi_{P|N}\right)\left(1-\tilde{m}(e_0,\mathbf{S}_0)\right), & \text{if } \tilde{\mathcal{T}}_1\left(N,e_0,\mathbf{S}_0\right)\leq\frac{\pi}{\gamma},
\end{cases}
\end{aligned} \tag{A.142}$$

and the second term is

$$\begin{aligned}
&\tilde{Pr}\left[h_1\left(e_0,\mathbf{S}_0\right)=\left(P,e_0,\mathbf{S}_0\right)\right]\tilde{J}_1\left(P,e_0,\mathbf{S}_0\right)\\
&=\begin{cases}
0, & \text{if } \frac{\pi}{\gamma}<\tilde{\mathcal{T}}_1\left(P,e_0,\mathbf{S}_0\right),\\
\left(A(P)\pi-\gamma\right)\phi_{P|P}\tilde{m}(e_0,\mathbf{S}_0) & \\
+\left(\mu A(P)\pi-\gamma\right)\phi_{P|N}\left(1-\tilde{m}(e_0,\mathbf{S}_0)\right), & \text{if } \tilde{\mathcal{T}}_1\left(P,e_0,\mathbf{S}_0\right)\leq\frac{\pi}{\gamma}.
\end{cases}
\end{aligned} \tag{A.143}$$

Furthermore, from (A.11) and (A.12)

$$\begin{aligned}
\tilde{\mathcal{T}}_1\left(\mathbf{S}_1\right)&=\begin{cases}
\frac{\phi_{P|P}\tilde{m}(e_0,\mathbf{S}_0)+\phi_{P|N}(1-\tilde{m}(e_0,\mathbf{S}_0))}{A(P)\left(\phi_{P|P}\tilde{m}(e_0,\mathbf{S}_0)+\mu\phi_{P|N}(1-\tilde{m}(e_0,\mathbf{S}_0))\right)}, & \text{if } \mathbf{S}_1=\left(P,e_0,\mathbf{S}_0\right),\\
\frac{\left(1-\phi_{P|P}\right)\tilde{m}(e_0,\mathbf{S}_0)+\left(1-\phi_{P|N}\right)(1-\tilde{m}(e_0,\mathbf{S}_0))}{A(P)\left(\left(1-\phi_{P|P}\right)\tilde{m}(e_0,\mathbf{S}_0)+\mu\left(1-\phi_{P|N}\right)(1-\tilde{m}(e_0,\mathbf{S}_0))\right)}, & \text{if } \mathbf{S}_1=\left(N,e_0,\mathbf{S}_0\right).
\end{cases}\\
&=\begin{cases}
\frac{A(\mathbf{S}_0)\phi_{P|P}+(1-A(\mathbf{S}_0))\phi_{P|N}}{A(P)\left(A(\mathbf{S}_0)\phi_{P|P}+\mu(1-A(\mathbf{S}_0))\phi_{P|N}\right)}, & \text{if } \mathbf{S}_1=\left(P,1,\mathbf{S}_0\right),\\
\frac{A(\mathbf{S}_0)\left(1-\phi_{P|P}\right)+(1-A(\mathbf{S}_0))\left(1-\phi_{P|N}\right)}{A(P)\left(A(\mathbf{S}_0)\left(1-\phi_{P|P}\right)+\mu(1-A(\mathbf{S}_0))\left(1-\phi_{P|N}\right)\right)}, & \text{if } \mathbf{S}_1=\left(N,1,\mathbf{S}_0\right),\\
\frac{1}{\mu A(P)}, & \text{if } \mathbf{S}_1=\left(X_1,0,\mathbf{S}_0\right),
\end{cases}
\end{aligned} \tag{A.144}$$

where from (1.1), (1.2), and (A.1), and by assumption,

$$\tilde{m}(e_0,\mathbf{S}_0)=\begin{cases}
0, & \text{if } e_0=0,\\
A(\mathbf{S}_0), & \text{if } e_0=1.
\end{cases} \tag{A.145}$$

Moreover, from (A.13), for a given e_0,

$$\tilde{\mathcal{T}}_1\left(P, e_0, \mathbf{S}_0\right) \leq \tilde{\mathcal{T}}_1\left(N, e_0, \mathbf{S}_0\right), \tag{A.146}$$

and under Assumption 2, $\tilde{\mathcal{T}}_1\left(P, 0, \mathbf{S}_0\right) \geq \tilde{\mathcal{T}}_1\left(N, 1, \mathbf{S}_0\right)$. Therefore,

$$\tilde{\mathcal{T}}_1\left(P, 1, \mathbf{S}_0\right) \leq \tilde{\mathcal{T}}_1\left(N, 1, \mathbf{S}_0\right) \leq \tilde{\mathcal{T}}_1\left(P, 0, \mathbf{S}_0\right) = \tilde{\mathcal{T}}_1\left(N, 0, \mathbf{S}_0\right). \tag{A.147}$$

Note that the ordering of the middle two terms and the equality of the last two terms differs from the ordering given in (A.73) in the Proposition. Therefore, we rely on our earlier results for select cases.

First, if $\frac{1}{\mu A(P)} \leq \frac{\pi}{\gamma}$, then from (A.75),

$$e_0^* = \begin{cases} 0, & \text{if } \frac{1}{\mu A(P)} \leq \frac{\pi}{\gamma} < \frac{1}{(1-\mu)A(\mathbf{S}_0)A(P)}, \\ 1, & \text{if } \max\left\{\frac{1}{\mu A(P)}, \frac{1}{(1-\mu)A(\mathbf{S}_0)A(P)}\right\} \leq \frac{\pi}{\gamma}. \end{cases} \tag{A.148}$$

It is straightforward to show that

$$\max\left\{\frac{1}{\mu A(P)}, \frac{1}{(1-\mu)A(\mathbf{S}_0)A(P)}\right\} = \begin{cases} \frac{1}{(1-\mu)A(\mathbf{S}_0)A(P)}, & \text{if } \mathbf{S}_0 = N, \text{ or} \\ & \text{if } \mathbf{S}_0 = P \text{ and } \mu \geq (1-\mu)A(P), \\ \frac{1}{\mu A(P)}, & \text{if } \mathbf{S}_0 = P \text{ and } \mu < (1-\mu)A(P). \end{cases} \tag{A.149}$$

Then, when $\mathbf{S}_0 = N$, or when $\mathbf{S}_0 = P$ and $0 \leq (1-\mu)A(P) \leq \mu$,

$$e_0^* = \begin{cases} 0, & \text{if } \frac{1}{\mu A(P)} \leq \frac{\pi}{\gamma} < \frac{1}{(1-\mu)A(\mathbf{S}_0)A(P)}, \\ 1, & \text{if } \frac{1}{(1-\mu)A(\mathbf{S}_0)A(P)} \leq \frac{\pi}{\gamma}, \end{cases} \tag{A.150}$$

and when $\mathbf{S}_0 = P$ and $\mu < (1-\mu)A(P)$,

$$e_0^* = 1, \quad \text{for } \frac{1}{\mu A(P)} \leq \frac{\pi}{\gamma}. \tag{A.151}$$

Second, if $\tilde{\mathcal{T}}_1(N, 1, \mathbf{S}_0) \leq \frac{\pi}{\gamma} < \frac{1}{\mu A(P)}$, then

$$e_1^*(\mathbf{S}_1) = \begin{cases} 0, & \text{if } \mathbf{S}_1 = (X_1, 0, \mathbf{S}_0), \\ 1, & \text{if } \mathbf{S}_1 = (X_1, 1, \mathbf{S}_0), \end{cases} \tag{A.152}$$

and the reward function is

$$J_0(\mathbf{S}_0) = \max\left[\mathcal{J}^0, \mathcal{J}^1\right], \tag{A.153}$$

where $\mathcal{J}^0 = 0$ corresponds to the case where $e_0 = 0$, and

$$\mathcal{J}^1 = A(P)\left(A(\mathbf{S}_0) + \mu\left(1 - A(\mathbf{S}_0)\right)\right)\pi - 2\gamma \tag{A.154}$$

and corresponds to the case where $e_0 = 1$. Now,

$$\mathcal{J}^1 \geq \mathcal{J}^0 \iff \frac{\pi}{\gamma} \geq \frac{2}{A(P)\left(A(\mathbf{S}_0) + \mu\left(1 - A(\mathbf{S}_0)\right)\right)}. \tag{A.155}$$

Therefore,

$$e_0^* = \begin{cases} 0, & \text{if } \tilde{\mathcal{T}}_1(N, 1, \mathbf{S}_0) \leq \frac{\pi}{\gamma} < \min\left\{\frac{1}{\mu A(P)}, \frac{2}{A(P)(A(\mathbf{S}_0)+\mu(1-A(\mathbf{S}_0)))}\right\}, \\ 1, & \text{if } \max\left\{\tilde{\mathcal{T}}_1(N, 1, \mathbf{S}_0), \frac{2}{A(P)(A(\mathbf{S}_0)+\mu(1-A(\mathbf{S}_0)))}\right\} \leq \frac{\pi}{\gamma} < \frac{1}{\mu A(P)}. \end{cases} \tag{A.156}$$

Note that

$$\frac{1}{\mu A(P)} > \frac{2}{A(P)\left(A(\mathbf{S}_0) + \mu\left(1 - A(\mathbf{S}_0)\right)\right)} \iff (1-\mu)\, A(\mathbf{S}_0) > \mu, \tag{A.157}$$

which never holds when $\mathbf{S}_0 = N$. Furthermore,

$$\begin{aligned}
&\tilde{\mathcal{T}}_1\left(N, 1, \mathbf{S}_0\right) > \frac{2}{A(P)\left(A(\mathbf{S}_0) + \mu\left(1 - A(\mathbf{S}_0)\right)\right)} \\
\iff &\frac{\left(1-\phi_{P|P}\right)A(\mathbf{S}_0) + \left(1-\phi_{P|N}\right)\left(1 - A(\mathbf{S}_0)\right)}{A(P)\left(\left(1-\phi_{P|P}\right)A(\mathbf{S}_0) + \mu\left(1-\phi_{P|N}\right)\left(1 - A(\mathbf{S}_0)\right)\right)} \\
&> \frac{2}{A(P)\left(A(\mathbf{S}_0) + \mu\left(1 - A(\mathbf{S}_0)\right)\right)} \\
\iff &(1-\mu)A(\mathbf{S}_0) > \frac{A(\mathbf{S}_0)\left(1-\phi_{P|P}\right) + \mu\left(1 - A(\mathbf{S}_0)\right)\left(1-\phi_{P|N}\right)}{\left(1 - A(\mathbf{S}_0)\right)\left(\phi_{P|P} - \phi_{P|N}\right)}.
\end{aligned} \tag{A.158}$$

When $\mathbf{S}_0 = N$, (A.158) becomes

$$0 > A(P)\left(1-\phi_{P|P}\right) + \left(1 - \mu A(P)\right)\left(\left(1-\phi_{P|N}\right) - (1-\mu)A(P)\left(\phi_{P|P} - \phi_{P|N}\right)\right), \tag{A.159}$$

which never holds. When $\mathbf{S}_0 = P$, (A.158) becomes

$$(1-\mu)A(P) > \frac{A(P)\left(1-\phi_{P|P}\right) + \mu\left(1 - A(P)\right)\left(1-\phi_{P|N}\right)}{\left(1 - A(P)\right)\left(\phi_{P|P} - \phi_{P|N}\right)}, \tag{A.160}$$

Note that under the Assumptions,

$$\frac{A(P)\left(1-\phi_{P|P}\right) + \mu\left(1 - A(P)\right)\left(1-\phi_{P|N}\right)}{\left(1 - A(P)\right)\left(\phi_{P|P} - \phi_{P|N}\right)} \geq \mu. \tag{A.161}$$

Therefore, when $\mathbf{S}_0 = N$ or when $\mathbf{S}_0 = P$ and $0 \leq (1-\mu)A(P) \leq \mu$,

$$e_0^* = 0, \quad \text{for } \tilde{\mathcal{T}}_1\left(N, 1, \mathbf{S}_0\right) \leq \frac{\pi}{\gamma} < \frac{1}{\mu A(P)}, \tag{A.162}$$

when $\mathbf{S}_0 = P$ and $\mu < (1-\mu)A(P) \leq \frac{A(P)\left(1-\phi_{P|P}\right)+\mu\left(1-A(P)\right)\left(1-\phi_{P|N}\right)}{\left(1-A(P)\right)\left(\phi_{P|P}-\phi_{P|N}\right)}$, then

$$e_0^* = \begin{cases} 0, & \text{if } \tilde{\mathcal{T}}_1\left(N, 1, \mathbf{S}_0\right) \leq \frac{\pi}{\gamma} < \frac{2}{A(P)\left(A(\mathbf{S}_0)+\mu\left(1-A(\mathbf{S}_0)\right)\right)}, \\ 1, & \text{if } \frac{2}{A(P)\left(A(\mathbf{S}_0)+\mu\left(1-A(\mathbf{S}_0)\right)\right)} \leq \frac{\pi}{\gamma} < \frac{1}{\mu A(P)}. \end{cases} \tag{A.163}$$

and when $\mathbf{S}_0 = P$ and $\frac{A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)}{(1-A(P))\left(\phi_{P|P}-\phi_{P|N}\right)} < (1-\mu)A(P)$, then

$$e_0^* = 1, \quad \text{for } \tilde{\mathcal{T}}_1\left(N, 1, \mathbf{S}_0\right) \leq \frac{\pi}{\gamma} < \frac{1}{\mu A(P)}. \tag{A.164}$$

Third, if $\tilde{\mathcal{T}}_1\left(P, 1, \mathbf{S}_0\right) \leq \frac{\pi}{\gamma} < \tilde{\mathcal{T}}_1\left(N, 1, \mathbf{S}_0\right)$, then

$$e_1^*\left(\mathbf{S}_1\right) = \begin{cases} 0, & \text{if } \mathbf{S}_1 = \left(P, 0, \mathbf{S}_0\right) \text{ or } \mathbf{S}_1 = \left(N, e_0, \mathbf{S}_0\right), \\ 1, & \text{if } \mathbf{S}_1 = \left(P, 1, \mathbf{S}_0\right), \end{cases} \tag{A.165}$$

and the reward function is

$$J_0(\mathbf{S}_0) = \max\left[\mathcal{J}^0, \mathcal{J}^1\right], \tag{A.166}$$

where $\mathcal{J}^0 = 0$, and corresponds to the case where $e_0 = 0$, and

$$\mathcal{J}^1 = A(\mathbf{S}_0)\left(A(P)\pi - \gamma\right)\phi_{P|P} + \left(1 - A(\mathbf{S}_0)\right)\left(\mu A(P)\pi - \gamma\right)\phi_{P|N} - \gamma, \tag{A.167}$$

and corresponds to the case where $e_0 = 1$. Now,

$$\mathcal{J}^1 \geq \mathcal{J}^0 \iff \frac{\pi}{\gamma} \geq \frac{1 + A(\mathbf{S}_0)\phi_{P|P} + \left(1 - A(\mathbf{S}_0)\right)\phi_{P|N}}{A(P)\left(A(\mathbf{S}_0)\phi_{P|P} + \mu\left(1 - A(\mathbf{S}_0)\right)\phi_{P|N}\right)}. \tag{A.168}$$

Clearly,

$$\frac{1 + A(\mathbf{S}_0)\phi_{P|P} + \left(1 - A(\mathbf{S}_0)\right)\phi_{P|N}}{A(P)\left(A(\mathbf{S}_0)\phi_{P|P} + \mu\left(1 - A(\mathbf{S}_0)\right)\phi_{P|N}\right)} \geq \tilde{\mathcal{T}}_1\left(P, 1, \mathbf{S}_0\right) \tag{A.169}$$

Therefore,

$$e_0^* = \begin{cases} 0, & \text{if } \tilde{\mathcal{T}}_1\left(P,1,\mathbf{S}_0\right) \leq \frac{\pi}{\gamma} < \min\left\{\tilde{\mathcal{T}}_1\left(N,1,\mathbf{S}_0\right), \frac{1+A(\mathbf{S}_0)\phi_{P|P}+(1-A(\mathbf{S}_0))\phi_{P|N}}{A(P)\left(A(\mathbf{S}_0)\phi_{P|P}+\mu(1-A(\mathbf{S}_0))\phi_{P|N}\right)}\right\}, \\ 1, & \text{if } \frac{1+A(\mathbf{S}_0)\phi_{P|P}+(1-A(\mathbf{S}_0))\phi_{P|N}}{A(P)\left(A(\mathbf{S}_0)\phi_{P|P}+\mu(1-A(\mathbf{S}_0))\phi_{P|N}\right)} \leq \frac{\pi}{\gamma} < \tilde{\mathcal{T}}_1\left(N,1,\mathbf{S}_0\right). \end{cases} \tag{A.170}$$

Furthermore,

$$\begin{aligned} &\tilde{\mathcal{T}}_1\left(N,1,\mathbf{S}_0\right) > \frac{1+A(\mathbf{S}_0)\phi_{P|P}+\left(1-A(\mathbf{S}_0)\right)\phi_{P|N}}{A(P)\left(A(\mathbf{S}_0)\phi_{P|P}+\mu\left(1-A(\mathbf{S}_0)\right)\phi_{P|N}\right)} \\ \iff &\frac{\left(1-\phi_{P|P}\right)A(\mathbf{S}_0)+\left(1-\phi_{P|N}\right)\left(1-A(\mathbf{S}_0)\right)}{A(P)\left(\left(1-\phi_{P|P}\right)A(\mathbf{S}_0)+\mu\left(1-\phi_{P|N}\right)\left(1-A(\mathbf{S}_0)\right)\right)} \\ &> \frac{1+A(\mathbf{S}_0)\phi_{P|P}+\left(1-A(\mathbf{S}_0)\right)\phi_{P|N}}{A(P)\left(A(\mathbf{S}_0)\phi_{P|P}+\mu\left(1-A(\mathbf{S}_0)\right)\phi_{P|N}\right)} \\ \iff &(1-\mu)A(\mathbf{S}_0) > \frac{A(\mathbf{S}_0)\left(1-\phi_{P|P}\right)+\mu\left(1-A(\mathbf{S}_0)\right)\left(1-\phi_{P|N}\right)}{\left(1-A(\mathbf{S}_0)\right)\left(\phi_{P|P}-\phi_{P|N}\right)}. \end{aligned} \tag{A.171}$$

The last line of the inequality is identical to (A.158). Therefore, when $\mathbf{S}_0 = N$ or when $\mathbf{S}_0 = P$ and $(1-\mu)A(P) \leq \frac{A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)}{(1-A(P))\left(\phi_{P|P}-\phi_{P|N}\right)}$,

$$e_0^* = 0, \quad \text{for } \tilde{\mathcal{T}}_1\left(P,1,\mathbf{S}_0\right) \leq \frac{\pi}{\gamma} < \tilde{\mathcal{T}}_1\left(N,1,\mathbf{S}_0\right), \tag{A.172}$$

and when $\mathbf{S}_0 = P$ and $\frac{A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)}{(1-A(P))\left(\phi_{P|P}-\phi_{P|N}\right)} < (1-\mu)A(P)$,

$$e_0^* = \begin{cases} 0, & \text{if } \tilde{\mathcal{T}}_1\left(P,1,\mathbf{S}_0\right) \leq \frac{\pi}{\gamma} < \frac{1+A(\mathbf{S}_0)\phi_{P|P}+(1-A(\mathbf{S}_0))\phi_{P|N}}{A(P)\left(A(\mathbf{S}_0)\phi_{P|P}+\mu(1-A(\mathbf{S}_0))\phi_{P|N}\right)}, \\ 1, & \text{if } \frac{1+A(\mathbf{S}_0)\phi_{P|P}+(1-A(\mathbf{S}_0))\phi_{P|N}}{A(P)\left(A(\mathbf{S}_0)\phi_{P|P}+\mu(1-A(\mathbf{S}_0))\phi_{P|N}\right)} \leq \frac{\pi}{\gamma} < \tilde{\mathcal{T}}_1\left(N,1,\mathbf{S}_0\right). \end{cases} \tag{A.173}$$

Finally, if $\frac{\pi}{\gamma} < \tilde{\mathcal{T}}_1\left(P,1,\mathbf{S}_0\right)$, then from (A.95), $e_0^* = 0$.

Combining this gives the optimal effort level at $t = 0$:

$$e_0^*(\mathbf{S}_0) = \begin{cases} 0, & \text{if } \frac{\pi}{\gamma} < \tilde{\mathcal{T}}_0(\mathbf{S}_0), \\ 1, & \text{if } \tilde{\mathcal{T}}_0(\mathbf{S}_0) \leq \frac{\pi}{\gamma}, \end{cases} \tag{A.174}$$

where

$$\tilde{\mathcal{T}}_0(P) = \begin{cases} \frac{1}{(1-\mu)(A(P))^2}, & \text{if } 0 \leq (1-\mu)A(P) \leq \mu, \\ \frac{2}{A(P)(A(P)+\mu(1-A(P)))}, & \text{if } \mu < (1-\mu)A(P) \\ & \leq \frac{A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)}{(1-A(P))\left(\phi_{P|P}-\phi_{P|N}\right)}, \\ \frac{1+A(P)\phi_{P|P}+(1-A(P))\phi_{P|N}}{A(P)\left(A(P)\phi_{P|P}+\mu(1-A(P))\phi_{P|N}\right)}, & \text{if } \frac{A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)}{(1-A(P))\left(\phi_{P|P}-\phi_{P|N}\right)} \\ & < (1-\mu)A(P), \end{cases} \tag{A.175}$$

and

$$\tilde{\mathcal{T}}_0(N) = \frac{1}{\mu(1-\mu)(A(P))^2}. \tag{A.176}$$

□

First, consider the probability that the final state is proficient when the district relies on the formative assessment.

Suppose that $\mathbf{S}_0 = N$. Then, from (A.140), the incentive-to-cost threshold in period 1 is

$$\tilde{\mathcal{T}}_0(N) = \frac{1}{\mu(1-\mu)(A(P))^2}. \tag{A.177}$$

and, from (A.144), the threshold in the second-half of the year is

$$\tilde{\mathcal{T}}_1\left(X_1, e_0, N\right) = \begin{cases} \frac{\mu A(P)\phi_{P|P}+(1-\mu A(P))\phi_{P|N}}{A(P)\left(\mu A(P)\phi_{P|P}+\mu(1-\mu A(P))\phi_{P|N}\right)}, & \text{if } \mathbf{S}_1 = (P,1,N)\,, \\ \frac{\mu A(P)\left(1-\phi_{P|P}\right)+(1-\mu A(P))\left(1-\phi_{P|N}\right)}{A(P)\left(\mu A(P)\left(1-\phi_{P|P}\right)+\mu(1-\mu A(P))\left(1-\phi_{P|N}\right)\right)}, & \text{if } \mathbf{S}_1 = (N,1,N)\,, \\ \frac{1}{\mu A(P)}, & \text{if } \mathbf{S}_1 = (X_1,0,N)\,, \end{cases} \tag{A.178}$$

From (A.147) and applying (A.43) in Lemma 6,

$$\tilde{\mathcal{T}}_1\left(P,1,N\right) \leq \tilde{\mathcal{T}}_1\left(N,1,N\right) \leq \tilde{\mathcal{T}}_1\left(P,0,N\right) = \tilde{\mathcal{T}}_1\left(N,0,N\right) \leq \tilde{\mathcal{T}}_0(N). \tag{A.179}$$

For each possible region of $\frac{\pi}{\gamma}$, it is straightforward to determine the optimal effort in each period and, plugging that into (A.108), the probability that final state is proficient. Then, (A.132) follows.

Next, suppose that $\mathbf{S}_0 = P$. From (A.139), the period 1 cost-to-incentive ratio threshold is

$$\tilde{\mathcal{T}}_0(P) = \begin{cases} \frac{1}{(1-\mu)(A(P))^2}, & \text{if } 0 \leq (1-\mu)A(P) \leq \mu, \\ \frac{2}{A(P)(A(P)+\mu(1-A(P)))}, & \text{if } \mu < (1-\mu)A(P) \\ & \quad \leq \frac{A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)}{(1-A(P))\left(\phi_{P|P}-\phi_{P|N}\right)}, \\ \frac{1+A(P)\phi_{P|P}+(1-A(P))\phi_{P|N}}{A(P)\left(A(P)\phi_{P|P}+\mu(1-A(P))\phi_{P|N}\right)}, & \text{if } \frac{A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)}{(1-A(P))\left(\phi_{P|P}-\phi_{P|N}\right)} \\ & \quad < (1-\mu)A(P), \end{cases} \tag{A.180}$$

and from (A.144), the threshold in the second-half of the year is

$$\tilde{\mathcal{T}}_1\left(X_1, e_0, P\right) = \begin{cases} \frac{A(P)\phi_{P|P}+(1-A(P))\phi_{P|N}}{A(P)\left(A(P)\phi_{P|P}+\mu(1-A(P))\phi_{P|N}\right)}, & \text{if } \mathbf{S}_1 = (P,1,P)\,, \\ \frac{A(P)\left(1-\phi_{P|P}\right)+(1-A(P))\left(1-\phi_{P|N}\right)}{A(P)\left(A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)\right)}, & \text{if } \mathbf{S}_1 = (N,1,P)\,, \\ \frac{1}{\mu A(P)}, & \text{if } \mathbf{S}_1 = (X_1,0,P)\,, \end{cases} \tag{A.181}$$

When $0 \leq (1-\mu)A(P) \leq \mu$, from (A.147) and applying (A.43) in Lemma 6,

$$\tilde{\mathcal{T}}_1(P,1,P) \leq \tilde{\mathcal{T}}_1(N,1,P) \leq \tilde{\mathcal{T}}_1(P,0,P) = \tilde{\mathcal{T}}_1(N,0,P) \leq \tilde{\mathcal{T}}_0(P). \tag{A.182}$$

Then, applying (A.107), (A.133) follows.

When $\mu < (1-\mu)A(P) \leq \frac{A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)}{(1-A(P))\left(\phi_{P|P}-\phi_{P|N}\right)}$, from (A.147) and applying (A.157) and (A.158),

$$\tilde{\mathcal{T}}_1(P,1,P) \leq \tilde{\mathcal{T}}_1(N,1,P) \leq \tilde{\mathcal{T}}_0(P) \leq \tilde{\mathcal{T}}_1(P,0,P) = \tilde{\mathcal{T}}_1(N,0,P). \tag{A.183}$$

Then, applying (A.107), (A.134) follows.

When $\frac{A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)}{(1-A(P))\left(\phi_{P|P}-\phi_{P|N}\right)} < (1-\mu)A(P)$, from (A.147) and applying (A.169) and (A.171),

$$\tilde{\mathcal{T}}_1(P,1,P) \leq \tilde{\mathcal{T}}_0(P) \leq \tilde{\mathcal{T}}_1(N,1,P) \leq \tilde{\mathcal{T}}_1(P,0,P) = \tilde{\mathcal{T}}_1(N,0,P). \tag{A.184}$$

Then, applying (A.107), (A.135) follows.

Note that in each case, it is straightforward to show that the the probability that final state is proficient is non-decreasing in the scaled incentive threshold. $\square$

Proposition 16 (Optimal Merit-Based Incentive) *Define the following constants:*

$$\mathcal{H}_1 = \frac{\mu\left(B(P)\left(A(P)\phi_{P|P} + B(P)\right) + (1-\mu B(P))\left(A(P)\phi_{P|N} + B(P)\right)\right)}{\mathcal{T}_1\left(P,0,N\right)}, \tag{A.185}$$

$$\mathcal{H}_2 = \frac{\mu\left(A(P)+B(P)\right)\left(1+(1-\mu)B(P)\right)}{\mathcal{T}_1\left(N,0,N\right)}, \tag{A.186}$$

$$\mathcal{H}_3 = \frac{B(P)\left(A(P)\phi_{P|P} + B(P)\right) + \mu\left(1-B(P)\right)\left(A(P)\phi_{P|N} + B(P)\right)}{\mathcal{T}_1\left(P,0,P\right)}, \tag{A.187}$$

$$\mathcal{H}_4 = \frac{\left(A(P)+B(P)\right)\left(\mu+(1-\mu)B(P)\right)}{\mathcal{T}_1\left(N,0,P\right)}, \tag{A.188}$$

$$\mathcal{H}_5 = \frac{\left(A(P)+B(P)\right)\left(\mu+(1-\mu)\left(A(P)+B(P)\right)\right)}{\mathcal{C}_1(P)}, \tag{A.189}$$

$$\mathcal{H}_6 = \frac{\left(A(P)+B(P)\right)\left(A(P)\phi_{P|P} + B(P)\right) + \mu\left(1-\left(A(P)+B(P)\right)\right)\left(A(P)\phi_{P|N} + B(P)\right)}{\mathcal{C}_2(P)}, \tag{A.190}$$

$$\mathcal{H}_7 = \frac{\left(A(P)+B(P)\right)\left(\mu+(1-\mu)\left(A(P)+B(P)\right)\right)}{\mathcal{T}_1\left(N,1,P\right)}, \tag{A.191}$$

$$\mathcal{H}_8 = \frac{\left(A(P)+B(P)\right)\left(A(P)\phi_{P|P} + B(P)\right) + \mu\left(1-\left(A(P)+B(P)\right)\right)\left(A(P)\phi_{P|N} + B(P)\right)}{\mathcal{C}_3(P)}. \tag{A.192}$$

Suppose the condition in (A.22) *holds. Then, the optimal scaled merit-based incentive can be characterized as follows.*

a) For $\mathbf{S}_0 = N$,

$$\frac{\pi}{\gamma}^* = \begin{cases} 0, & \text{if } 0 \leq \frac{M-Fz_I}{\gamma} < \mathcal{H}_1, \\ \mathcal{T}_1\left(P,0,N\right), & \text{if } \mathcal{H}_1 \leq \frac{M-Fz_I}{\gamma} < \mathcal{H}_2, \\ \mathcal{T}_1\left(N,0,N\right), & \text{if } \mathcal{H}_2 \leq \frac{M-Fz_I}{\gamma} < \frac{1+(1-\mu)B(P)}{(1-\mu)A(P)} + 1, \\ \mathcal{T}_0\left(N\right), & \text{if } \frac{(1+(1-\mu)B(P))}{(1-\mu)A(P)} + 1 \leq \frac{M-Fz_I}{\gamma}. \end{cases} \tag{A.193}$$

b) For $\mathbf{S}_0 = P$ *and* $0 \leq (1-\mu)\left(A(P)+B(P)\right) \leq \mathcal{E}_1(P)$,

$$\pi^* = \begin{cases} 0, & \text{if } 0 \leq \frac{M-Fz_I}{\gamma} < \mathcal{H}_3, \\ \mathcal{T}_1\left(P,0,P\right), & \text{if } \mathcal{H}_3 \leq \frac{M-Fz_I}{\gamma} < \mathcal{H}_4, \\ \mathcal{T}_1\left(N,0,P\right), & \text{if } \mathcal{H}_4 \leq \frac{M-Fz_I}{\gamma} < \frac{\mu(1-B(P))+B(P)}{(1-\mu)A(P)}, \\ \mathcal{T}_0\left(P\right), & \text{if } \frac{\mu(1-B(P))+B(P)}{(1-\mu)A(P)} \leq \frac{M-Fz_I}{\gamma}. \end{cases} \tag{A.194}$$

c) For $\mathbf{S}_0 = P$ *and* $\mathcal{E}_1(P) < (1-\mu)\left(A(P)+B(P)\right) \leq \mathcal{E}_2(P)$,

$$\pi^* = \begin{cases} 0, & \text{if } 0 < \frac{M-Fz_I}{\gamma} < \mathcal{H}_3, \\ \mathcal{T}_1\left(P,0,P\right), & \text{if } \mathcal{H}_3 \leq \frac{M-Fz_I}{\gamma} < \mathcal{H}_5, \\ \mathcal{T}_0\left(P\right), & \text{if } \mathcal{H}_5 \leq \frac{M-Fz_I}{\gamma}. \end{cases} \tag{A.195}$$

d) For $\mathbf{S}_0 = P$ *and* $\mathcal{E}_2(P) < (1-\mu)\left(A(P)+B(P)\right) \leq \mathcal{E}_3(P)$,

$$\pi^* = \begin{cases} 0, & \text{if } 0 < \frac{M-Fz_I}{\gamma} < \mathcal{H}_3, \\ \mathcal{T}_1\left(P,0,P\right), & \text{if } \mathcal{H}_3 \leq \frac{M-Fz_I}{\gamma} < \mathcal{H}_6, \\ \mathcal{T}_0\left(P\right), & \text{if } \mathcal{H}_6 \leq \frac{M-Fz_I}{\gamma} < \mathcal{H}_7, \\ \mathcal{T}_1\left(N,1,P\right), & \text{if } \mathcal{H}_7 \leq \frac{M-Fz_I}{\gamma}. \end{cases} \tag{A.196}$$

e) For $\mathbf{S}_0 = P$ *and* $\mathcal{E}_3(P) < (1-\mu)\left(A(P)+B(P)\right) \leq 1$,

$$\pi^* = \begin{cases} 0, & \text{if } 0 < \frac{M-Fz_I}{\gamma} < \mathcal{H}_8, \\ \mathcal{T}_0\left(P\right), & \text{if } \mathcal{H}_8 \leq \frac{M-Fz_I}{\gamma} < \mathcal{H}_7, \\ \mathcal{T}_1\left(N,1,P\right), & \text{if } \mathcal{H}_7 \leq \frac{M-Fz_I}{\gamma}. \end{cases} \tag{A.197}$$

Proof of Proposition 16

From (1.12)-(1.13), the school district's maximization problem is

$$\max_{\pi \geq 0} Pr^*[\mathbf{S}_2 = P|\mathbf{S}_0] \tag{A.198}$$

$$\text{s.t. } \pi Pr^*[\mathbf{S}_2 = P|\mathbf{S}_0] + Fz_I \leq M, \tag{A.199}$$

with $Pr^*[\mathbf{S}_2 = P|\mathbf{S}_0]$ given in (A.111)-(A.115).

In the proof of Proposition 14, we show that $Pr^*[\mathbf{S}_2 = P|\mathbf{S}_0]$ is a non-decreasing step function of the scaled incentive threshold. Therefore, the expression on the left-hand side of the district's constraint (A.199) is an increasing function of π, and we must consider the value of the objective function (A.198) at the endpoints of each interval of $\frac{\pi}{\gamma}$ that corresponds to a "step." We assume that if $Pr^*[\mathbf{S}_2 = P|\mathbf{S}_0]$ is constant over a region of π and any value of π in that region is optimal, the district will choose the smallest value of π in that region.

To determine the optimal merit-based incentive π^* and the corresponding probability that the final state is proficient, we must consider the several cases stated in Proposition 14 that determine the characterization of $Pr^*[\mathbf{S}_2 = P|\mathbf{S}_0]$.

We begin with the case where $\mathbf{S}_0 = N$ and $Pr^*[\mathbf{S}_2 = P|\mathbf{S}_0]$ is given by (A.111).

First, $\frac{\pi}{\gamma}^* = 0$ if $0 \leq M - Fz_I$ and

$$\begin{aligned} & M - Fz_I < \mathcal{T}_1(P, 0, N)\gamma\mu \\ & \qquad \times \big(B(P)\big(A(P)\phi_{P|P} + B(P)\big) + (1 - \mu B(P))\big(A(P)\phi_{P|N} + B(P)\big)\big) \\ \iff & \frac{M - Fz_I}{\gamma} < \mathcal{H}_1, \end{aligned} \tag{A.200}$$

where

$$\mathcal{H}_1 = \left(1 + \frac{B(P)\left(1 + (1-\mu)B(P)\right)}{A(P)\left(B(P)\phi_{P|P} + \phi_{P|N}\left(1 - \mu B(P)\right)\right)}\right)\left(\left(\phi_{P|P}\mu B(P) + \phi_{P|N}\left(1 - \mu B(P)\right)\right)\right). \tag{A.201}$$

Second, $\frac{\pi}{\gamma}^* = \mathcal{T}_1(P, 0, N)$ if $\mathcal{H}_1 \leq \frac{M - Fz_I}{\gamma}$ and

$$\begin{aligned} & M - Fz_I < \mathcal{T}_1(N, 0, N)\,\gamma\left(\mu\left(A(P) + B(P)\right)\left(1 + (1-\mu)B(P)\right)\right) \\ \iff & \frac{M - Fz_I}{\gamma} < \mathcal{H}_2, \end{aligned} \tag{A.202}$$

where

$$\mathcal{H}_2 = \frac{\left(A(P) + B(P)\right)\left(1 + (1-\mu)B(P)\right)\left(\left(1 - \phi_{P|P}\right)\mu B(P) + \left(1 - \phi_{P|N}\right)\left(1 - \mu B(P)\right)\right)}{A(P)\left(B(P)\left(1 - \phi_{P|P}\right) + \left(1 - \phi_{P|N}\right)\left(1 - \mu B(P)\right)\right)}. \tag{A.203}$$

Third, $\frac{\pi}{\gamma}^* = \mathcal{T}_1(N, 0, N)$ if $\mathcal{H}_2 \leq \frac{M - Fz_I}{\gamma}$ and

$$\begin{aligned} & M - Fz_I < \mathcal{T}_0(N)\,\gamma\mu\left(A(P) + B(P)\right)\left(1 + (1-\mu)\left(A(P) + B(P)\right)\right) \\ \iff & \frac{M - Fz_I}{\gamma} < \frac{\left(1 + (1-\mu)B(P)\right)}{(1-\mu)A(P)} + 1. \end{aligned} \tag{A.204}$$

Finally, $\frac{\pi}{\gamma}^* = \mathcal{T}_0(N)$ if

$$\frac{\left(1 + (1-\mu)B(P)\right)}{(1-\mu)A(P)} + 1 \leq \frac{M - Fz_I}{\gamma}. \tag{A.205}$$

(A.193) combines these results. We follow similar steps for the cases where $\mathbf{S}_0 = P$, using the functional forms for $Pr^*\left[\mathbf{S}_2 = P|\mathbf{S}_0\right]$ given by (A.112)-(A.115).

Combining these results gives the statement in Proposition 16. From the proof of Proposition 14, it follows that the optimal reward is increasing in the budget. $\square$

Proof of Proposition 3

From (1.12)-(1.13), the school district's maximization problem is

$$\max_{\pi \geq 0} Pr^*[\mathbf{S}_2 = P|\mathbf{S}_0] \tag{A.206}$$

$$\text{s.t. } \pi Pr^*[\mathbf{S}_2 = P|\mathbf{S}_0] \leq M - Fz_I, \tag{A.207}$$

with $Pr^*[\mathbf{S}_2 = P|\mathbf{S}_0]$ given in (A.132)-(A.135).

In the proof of Proposition 15, we show that $\tilde{Pr}^*[\mathbf{S}_2 = P|\mathbf{S}_0]$ is a non-decreasing step function of the scaled incentive threshold, where the tilde denotes that this is the optimal probability in the case where there is no stickiness in the proficient state, i.e. $B(P) = 0$. Therefore, the expression on the left-hand side of the district's constraint (A.207) is an increasing function of the scaled reward, and we must consider the value of the objective function (A.206) at the endpoints of each interval of the scaled reward that corresponds to a "step." We assume that if $Pr^*[\mathbf{S}_2 = P|\mathbf{S}_0]$ is constant over a region of π and any value of π in that region is optimal, the district will choose the smallest value of π in that region.

To determine the optimal scaled incentive and the corresponding probability that the final state is proficient, we must consider the several cases stated in Proposition 15 that determine the characterization of $Pr^*[\mathbf{S}_2 = P|\mathbf{S}_0]$.

When $\mathbf{S}_0 = N$, from (A.140), the incentive-to-cost threshold in the first half of the year is

$$\tilde{\mathcal{T}}_0(N) = \frac{1}{\mu(1-\mu)(A(P))^2}, \tag{A.208}$$

and from (A.178), the threshold in the second-half of the year when $\mathbf{S}_0 = N$ is

$$\tilde{\mathcal{T}}_1(X_1, e_0, N) = \begin{cases} \frac{\mu A(P)\phi_{P|P} + (1-\mu A(P))\phi_{P|N}}{\mu A(P)\left(A(P)\phi_{P|P} + (1-\mu A(P))\phi_{P|N}\right)}, & \text{if } \mathbf{S}_1 = (P, 1, N), \\ \frac{\mu A(P)\left(1-\phi_{P|P}\right) + (1-\mu A(P))\left(1-\phi_{P|N}\right)}{\mu A(P)\left(A(P)\left(1-\phi_{P|P}\right) + (1-\mu A(P))\left(1-\phi_{P|N}\right)\right)}, & \text{if } \mathbf{S}_1 = (N, 1, N), \\ \frac{1}{\mu A(P)}, & \text{if } \mathbf{S}_1 = (X_1, 0, N). \end{cases} \tag{A.209}$$

We begin with the case where $\mathbf{S}_0 = N$ and $Pr^*[\mathbf{S}_2 = P|\mathbf{S}_0]$ is given by (A.132). First, $\frac{\tilde{\pi}^*}{\gamma} = 0$ if

$$0 \leq \frac{M - Fz_I}{\gamma} < \mu A(P)\tilde{\mathcal{T}}_1(N, 0, N) \iff 0 \leq \frac{M - Fz_I}{\gamma} < 1. \tag{A.210}$$

Second, $\frac{\tilde{\pi}^*}{\gamma} = \tilde{\mathcal{T}}_1(N, 0, N)$ if

$$\begin{aligned} &\mu A(P)\tilde{\mathcal{T}}_1(N, 0, N) \leq \frac{M}{\gamma} < \mu A(P)\left(1 + (1-\mu)A(P)\right)\tilde{\mathcal{T}}_0(N) \\ \iff &1 \leq \frac{M - Fz_I}{\gamma} < 1 + \frac{1}{(1-\mu)A(P)}. \end{aligned} \tag{A.211}$$

Third, $\frac{\tilde{\pi}^*}{\gamma} = \tilde{\mathcal{T}}_0(N)$ if

$$\mu A(P)\left(1 + (1-\mu)A(P)\right)\tilde{\mathcal{T}}_0(N) \leq \frac{M - Fz_I}{\gamma} \iff 1 + \frac{1}{(1-\mu)A(P)} \leq \frac{M - Fz_I}{\gamma}. \tag{A.212}$$

Combining this, when $\mathbf{S}_0 = N$,

$$\frac{\tilde{\pi}^*}{\gamma} = \begin{cases} 0, & \text{if } 0 \leq \frac{M-Fz_I}{\gamma} < 1, \\ \frac{1}{\mu A(P)}, & \text{if } 1 \leq \frac{M-Fz_I}{\gamma} < 1 + \frac{1}{(1-\mu)A(P)}, \\ \frac{1}{\mu(1-\mu)(A(P))^2}, & \text{if } 1 + \frac{1}{(1-\mu)A(P)} \leq \frac{M-Fz_I}{\gamma}. \end{cases} \tag{A.213}$$

When $\mathbf{S}_0 = P$, from (A.139), the incentive-to-cost threshold in the first half of the year is

$$\tilde{\mathcal{T}}_0(P) = \begin{cases} \frac{1}{(1-\mu)(A(P))^2}, & \text{if } 0 \le (1-\mu)A(P) \le \mu, \\ \frac{2}{A(P)(A(P)+\mu(1-A(P)))}, & \text{if } \mu < (1-\mu)A(P) \\ & \le \frac{A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)}{(1-A(P))\left(\phi_{P|P}-\phi_{P|N}\right)}, \\ \frac{1+A(P)\phi_{P|P}+(1-A(P))\phi_{P|N}}{A(P)\left(A(P)\phi_{P|P}+\mu(1-A(P))\phi_{P|N}\right)}, & \text{if } \frac{A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)}{(1-A(P))\left(\phi_{P|P}-\phi_{P|N}\right)} \\ & < (1-\mu)A(P), \end{cases} \tag{A.214}$$

and from (A.181), the threshold in the second-half of the year is

$$\tilde{\mathcal{T}}_1\left(X_1, e_0, P\right) = \begin{cases} \frac{A(P)\phi_{P|P}+(1-A(P))\phi_{P|N}}{A(P)\left(A(P)\phi_{P|P}+\mu(1-A(P))\phi_{P|N}\right)}, & \text{if } \mathbf{S}_1 = (P,1,P)\,, \\ \frac{A(P)\left(1-\phi_{P|P}\right)+(1-A(P))\left(1-\phi_{P|N}\right)}{A(P)\left(A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)\right)}, & \text{if } \mathbf{S}_1 = (N,1,P)\,, \\ \frac{1}{\mu A(P)}, & \text{if } \mathbf{S}_1 = (X_1,0,P)\,. \end{cases} \tag{A.215}$$

When $\mathbf{S}_0 = P$ and $0 \le (1-\mu)A(P) \le \mu$, from (A.133),

$$\tilde{Pr}^*[\mathbf{S}_2 = P | \mathbf{S}_0 = P] = \begin{cases} 0, & \text{if } 0 \le \frac{\pi}{\gamma} < \frac{1}{\mu A(P)}, \\ \mu A(P), & \text{if } \frac{1}{\mu A(P)} \le \frac{\pi}{\gamma} < \frac{1}{(1-\mu)(A(P))^2}, \\ A(P)\left(\mu + (1-\mu)A(P)\right), & \text{if } \frac{1}{(1-\mu)(A(P))^2} \le \frac{\pi}{\gamma}. \end{cases} \tag{A.216}$$

Then,

$$\frac{\tilde{\pi}^*}{\gamma} = \begin{cases} 0, & \text{if } 0 \le \frac{M-Fz_I}{\gamma} < 1, \\ \frac{1}{\mu A(P)}, & \text{if } 1 \le \frac{M-Fz_I}{\gamma} < 1 + \frac{\mu}{(1-\mu)A(P)}, \\ \frac{1}{(1-\mu)(A(P))^2}, & \text{if } 1 + \frac{\mu}{(1-\mu)A(P)} \le \frac{M-Fz_I}{\gamma}. \end{cases} \tag{A.217}$$

When $\mathbf{S}_0 = P$ and $\mu < (1-\mu)A(P) \leq \frac{A(P)\big(1-\phi_{P|P}\big)+\mu(1-A(P))\big(1-\phi_{P|N}\big)}{(1-A(P))\big(\phi_{P|P}-\phi_{P|N}\big)}$, from (A.134),

$$\tilde{Pr}^*[\mathbf{S}_2 = P|\mathbf{S}_0 = P] = \begin{cases} 0, & \text{if } 0 \leq \frac{\pi}{\gamma} < \frac{2}{A(P)(A(P)+\mu(1-A(P)))}, \\ A(P)\left(A(P)+\mu\left(1-A(P)\right)\right), & \text{if } \frac{2}{A(P)(A(P)+\mu(1-A(P)))} \leq \frac{\pi}{\gamma}. \end{cases} \tag{A.218}$$

Then,

$$\frac{\tilde{\pi}^*}{\gamma} = \begin{cases} 0, & \text{if } 0 \leq \frac{M-Fz_I}{\gamma} < 2, \\ \frac{2}{A(P)(A(P)+\mu(1-A(P)))}, & \text{if } 2 \leq \frac{M-Fz_I}{\gamma}. \end{cases} \tag{A.219}$$

Finally, when $\mathbf{S}_0 = P$ and $\frac{A(P)\big(1-\phi_{P|P}\big)+\mu(1-A(P))\big(1-\phi_{P|N}\big)}{(1-A(P))\big(\phi_{P|P}-\phi_{P|N}\big)} < (1-\mu)A(P) \leq 1$, from (A.135),

$$\begin{aligned} &\tilde{Pr}^*[\mathbf{S}_2 = P|\mathbf{S}_0 = P] \\ &= \begin{cases} 0, & \text{if } 0 \leq \frac{\pi}{\gamma} < \frac{1+A(P)\phi_{P|P}+(1-A(P))\phi_{P|N}}{A(P)\big(A(P)\phi_{P|P}+\mu(1-A(P))\phi_{P|N}\big)}, \\ A(P)\left(A(P)\phi_{P|P}+\mu\left(1-A(P)\right)\phi_{P|N}\right), & \\ \qquad \text{if } \frac{1+A(P)\phi_{P|P}+(1-A(P))\phi_{P|N}}{A(P)\big(A(P)\phi_{P|P}+\mu(1-A(P))\phi_{P|N}\big)} \leq \frac{\pi}{\gamma} < \frac{A(P)\big(1-\phi_{P|P}\big)+(1-A(P))\big(1-\phi_{P|N}\big)}{A(P)\big(A(P)\big(1-\phi_{P|P}\big)+\mu(1-A(P))\big(1-\phi_{P|N}\big)\big)}, & \\ A(P)\left(A(P)+\mu(1-A(P))\right), \qquad \text{if } \frac{A(P)\big(1-\phi_{P|P}\big)+(1-A(P))\big(1-\phi_{P|N}\big)}{A(P)\big(A(P)\big(1-\phi_{P|P}\big)+\mu(1-A(P))\big(1-\phi_{P|N}\big)\big)} \leq \frac{\pi}{\gamma}. & \end{cases} \end{aligned} \tag{A.220}$$

Then,

$$\frac{\tilde{\pi}^*}{\gamma} = \begin{cases} 0, & \text{if } 0 \leq \frac{M-Fz_I}{\gamma} < 1 + A(P)\phi_{P|P} + (1-A(P))\,\phi_{P|N}, \\ \frac{1+A(P)\phi_{P|P}+(1-A(P))\phi_{P|N}}{A(P)\left(A(P)\phi_{P|P}+\mu(1-A(P))\phi_{P|N}\right)}, & \\ \qquad \text{if } 1 + A(P)\phi_{P|P} + (1-A(P))\,\phi_{P|N} \leq \frac{M-Fz_I}{\gamma} & \\ \qquad\quad < \frac{(\mu+(1-\mu)A(P))\left(A(P)\left(1-\phi_{P|P}\right)+(1-A(P))\left(1-\phi_{P|N}\right)\right)}{A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)}, & \\ \frac{A(P)\left(1-\phi_{P|P}\right)+(1-A(P))\left(1-\phi_{P|N}\right)}{A(P)\left(A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)\right)}, & \\ \qquad \text{if } \frac{(\mu+(1-\mu)A(P))\left(A(P)\left(1-\phi_{P|P}\right)+(1-A(P))\left(1-\phi_{P|N}\right)\right)}{A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)} \leq \frac{M-Fz_I}{\gamma}. & \end{cases} \tag{A.221}$$

Then, the optimal scaled merit-based incentive can be characterized as follows.

a) For $\mathbf{S}_0 = N$,

$$\frac{\tilde{\pi}^*}{\gamma} = \begin{cases} 0, & \text{if } 0 \leq \frac{M-Fz_I}{\gamma} < 1, \\ \frac{1}{\mu A(P)}, & \text{if } 1 \leq \frac{M-Fz_I}{\gamma} < 1 + \frac{1}{(1-\mu)A(P)}, \\ \frac{1}{\mu(1-\mu)(A(P))^2}, & \text{if } 1 + \frac{1}{(1-\mu)A(P)} \leq \frac{M-Fz_I}{\gamma}. \end{cases} \tag{A.222}$$

b) For $\mathbf{S}_0 = P$ and $0 \leq (1-\mu)A(P) \leq \mu$,

$$\frac{\tilde{\pi}^*}{\gamma} = \begin{cases} 0, & \text{if } 0 \leq \frac{M-Fz_I}{\gamma} < 1, \\ \frac{1}{\mu A(P)}, & \text{if } 1 \leq \frac{M-Fz_I}{\gamma} < 1 + \frac{\mu}{(1-\mu)A(P)}, \\ \frac{1}{(1-\mu)(A(P))^2}, & \text{if } 1 + \frac{\mu}{(1-\mu)A(P)} \leq \frac{M-Fz_I}{\gamma}. \end{cases} \tag{A.223}$$

c) For $\mathbf{S}_0 = P$ and $\mu < (1-\mu)A(P) \leq \frac{A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)}{(1-A(P))\left(\phi_{P|P}-\phi_{P|N}\right)}$,

$$\frac{\tilde{\pi}^*}{\gamma} = \begin{cases} 0, & \text{if } 0 \leq \frac{M-Fz_I}{\gamma} < 2, \\ \frac{2}{A(P)(A(P)+\mu(1-A(P)))}, & \text{if } 2 \leq \frac{M-Fz_I}{\gamma}. \end{cases} \tag{A.224}$$

d) For $\mathbf{S}_0 = P$ and $\frac{A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)}{(1-A(P))\left(\phi_{P|P}-\phi_{P|N}\right)} < (1-\mu)A(P) \leq 1$,

$$\frac{\tilde{\pi}^*}{\gamma} = \begin{cases} 0, & \text{if } 0 \leq \frac{M-Fz_I}{\gamma} < 1 + A(P)\phi_{P|P} + (1-A(P))\,\phi_{P|N}, \\ \frac{1+A(P)\phi_{P|P}+(1-A(P))\phi_{P|N}}{A(P)\left(A(P)\phi_{P|P}+\mu(1-A(P))\phi_{P|N}\right)}, & \\ & \text{if } 1 + A(P)\phi_{P|P} + (1-A(P))\,\phi_{P|N} \leq \frac{M-Fz_I}{\gamma} \\ & \quad < \frac{(\mu+(1-\mu)A(P))\left(A(P)\left(1-\phi_{P|P}\right)+(1-A(P))\left(1-\phi_{P|N}\right)\right)}{A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)}, \\ \frac{A(P)\left(1-\phi_{P|P}\right)+(1-A(P))\left(1-\phi_{P|N}\right)}{A(P)\left(A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)\right)}, & \\ & \text{if } \frac{(\mu+(1-\mu)A(P))\left(A(P)\left(1-\phi_{P|P}\right)+(1-A(P))\left(1-\phi_{P|N}\right)\right)}{A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)} \leq \frac{M-Fz_I}{\gamma}. \end{cases} \tag{A.225}$$

In the case of an interim assessment, the optimal incentive for schools that begin the year in the proficient state becomes

a) For $\mathbf{S}_0 = P$ and $0 \leq (1-\mu)A(P) \leq \mu$,

$$\frac{\tilde{\pi}^*}{\gamma} = \begin{cases} 0, & \text{if } 0 \leq \frac{M-F}{\gamma} < 1, \\ \frac{1}{\mu A(P)}, & \text{if } 1 \leq \frac{M-F}{\gamma} < 1 + \frac{\mu}{(1-\mu)A(P)}, \\ \frac{1}{(1-\mu)(A(P))^2}, & \text{if } 1 + \frac{\mu}{(1-\mu)A(P)} \leq \frac{M-F}{\gamma}. \end{cases} \tag{A.226}$$

b) For $\mathbf{S}_0 = P$ and $\mu < (1-\mu)A(P) \leq 1$,

$$\frac{\tilde{\pi}^*}{\gamma} = \begin{cases} 0, & \text{if } 0 \leq \frac{M-F}{\gamma} < 1 + A(P), \\ \frac{1+A(P)}{A(P)^2}, & \text{if } 1 + A(P) \leq \frac{M-F}{\gamma} < 1 + \left(\frac{1-\mu}{\mu}\right) A(P), \\ \frac{1}{\mu A(P)}, & \text{if } 1 + \left(\frac{1-\mu}{\mu}\right) A(P) \leq \frac{M-F}{\gamma}. \end{cases} \tag{A.227}$$

It directly follows from earlier results that the optimal scaled incentive is increasing in the scaled budget.

We next compare the optimal reward under high levels of M for different values of that starting state of proficiency. We claim that the maximum optimal reward when the school starts in the not-proficient state is always higher than the maximum optimal reward when the school starts in the proficient state. For large budget values and when $\mathbf{S}_0 = N$,

$$\frac{\tilde{\pi}^*}{\gamma} = \frac{1}{\mu(1-\mu)(A(P))^2}. \tag{A.228}$$

For large M when $\mathbf{S}_0 = P$,

$$\frac{\tilde{\pi}^*}{\gamma} = \begin{cases} \frac{1}{(1-\mu)(A(P))^2}, & \text{if } 0 \leq (1-\mu)A(P) \leq \mu, \\ \frac{2}{A(P)(A(P)+\mu(1-A(P)))}, & \text{if } \mu < (1-\mu)A(P) \\ & \quad \leq \frac{A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)}{(1-A(P))\left(\phi_{P|P}-\phi_{P|N}\right)}, \\ \frac{A(P)\left(1-\phi_{P|P}\right)+(1-A(P))\left(1-\phi_{P|N}\right)}{A(P)\left(A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)\right)}, & \text{if } \frac{A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)}{(1-A(P))\left(\phi_{P|P}-\phi_{P|N}\right)} \\ & \quad < (1-\mu)A(P) \leq 1. \end{cases} \tag{A.229}$$

Clearly,

$$\frac{1}{\mu(1-\mu)(A(P))^2} \geq \frac{1}{(1-\mu)(A(P))^2} \iff 1 \geq \mu. \tag{A.230}$$

When $\mu < (1-\mu)A(P) \leq \frac{A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)}{(1-A(P))\left(\phi_{P|P}-\phi_{P|N}\right)}$

$$\frac{1}{\mu(1-\mu)(A(P))^2} \geq \frac{2}{A(P)\left(A(P)+\mu\left(1-A(P)\right)\right)} \iff \mu \geq (1-\mu)A(P)\left(2\mu-1\right). \tag{A.231}$$

This clearly holds for $\mu \leq \frac{1}{2}$, and for $\frac{1}{2} < \mu \leq 1$, this becomes

$$\frac{\mu}{2\mu-1} \geq (1-\mu)A(P), \tag{A.232}$$

which holds since $\frac{\mu}{2\mu-1} \geq \mu$. Finally,

$$\begin{aligned} &\frac{1}{\mu(1-\mu)(A(P))^2} \geq \frac{A(P)\left(1-\phi_{P|P}\right)+\left(1-A(P)\right)\left(1-\phi_{P|N}\right)}{A(P)\left(A(P)\left(1-\phi_{P|P}\right)+\mu\left(1-A(P)\right)\left(1-\phi_{P|N}\right)\right)} \\ \iff &A(P)\left(1-\mu(1-\mu)A(P)\right)\left(1-\phi_{P|P}\right)+\mu\left(1-A(P)\right)\left(1-(1-\mu)A(P)\right)\left(1-\phi_{P|N}\right) \\ &\geq 0, \end{aligned} \tag{A.233}$$

which clearly holds. □

Lemma 10 *When $B(P)=0$, the probability that the school is in the proficient state at the end of the year ($\mathbf{S}_2 = P$) when the school district offers teachers the optimal reward, given in Proposition 3, can be described as follows.*

a) For $\mathbf{S}_0 = N$,

$$\tilde{Pr}^*_{z_I}[\mathbf{S}_2 = P|\mathbf{S}_0 = N] = \begin{cases} 0, & \text{if } 0 \leq \frac{M-Fz_I}{\gamma} < 1, \\ \mu A(P), & \text{if } 1 \leq \frac{M-Fz_I}{\gamma} < 1+\frac{1}{(1-\mu)A(P)}, \\ \mu A(P)\left(1+(1-\mu)A(P)\right), & \text{if } 1+\frac{1}{(1-\mu)A(P)} \leq \frac{M-Fz_I}{\gamma}. \end{cases} \tag{A.234}$$

b) For $\mathbf{S}_0 = P$ *and* $0 \leq (1-\mu)A(P) \leq \mu$,

$$\tilde{Pr}^*_{z_I}[\mathbf{S}_2 = P|\mathbf{S}_0 = P] = \begin{cases} 0, & \text{if } 0 \leq \frac{M-Fz_I}{\gamma} < 1, \\ \mu A(P), & \text{if } 1 \leq \frac{M-Fz_I}{\gamma} < 1 + \frac{\mu}{(1-\mu)A(P)}, \\ A(P)\left(\mu + (1-\mu)A(P)\right), & \text{if } 1 + \frac{\mu}{(1-\mu)A(P)} \leq \frac{M-Fz_I}{\gamma}. \end{cases} \tag{A.235}$$

c) For $\mathbf{S}_0 = P$ *and* $\mu < (1-\mu)A(P) \leq \frac{A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)}{(1-A(P))\left(\phi_{P|P}-\phi_{P|N}\right)}$,

$$\tilde{Pr}^*_{z_I}[\mathbf{S}_2 = P|\mathbf{S}_0 = P] = \begin{cases} 0, & \text{if } 0 \leq \frac{M-Fz_I}{\gamma} < 2, \\ A(P)\left(A(P) + \mu\left(1 - A(P)\right)\right), & \text{if } 2 \leq \frac{M-Fz_I}{\gamma}. \end{cases} \tag{A.236}$$

d) For $\mathbf{S}_0 = P$ *and* $\frac{A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)}{(1-A(P))\left(\phi_{P|P}-\phi_{P|N}\right)} < (1-\mu)A(P) \leq 1$,

$$\begin{aligned} &\tilde{Pr}^*_{z_I}[\mathbf{S}_2 = P|\mathbf{S}_0 = P] \\ &= \begin{cases} 0, & \text{if } 0 \leq \frac{M-Fz_I}{\gamma} < 1 + A(P)\phi_{P|P} + (1-A(P))\,\phi_{P|N}, \\ A(P)\left(A(P)\phi_{P|P} + \mu\left(1-A(P)\right)\phi_{P|N}\right), & \\ \quad \text{if } 1 + A(P)\phi_{P|P} + (1-A(P))\,\phi_{P|N} \leq \frac{M-Fz_I}{\gamma} & \\ \quad < \frac{(\mu+(1-\mu)A(P))\left(A(P)\left(1-\phi_{P|P}\right)+(1-A(P))\left(1-\phi_{P|N}\right)\right)}{A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)}, & \\ A(P)\left(A(P) + \mu(1-A(P))\right), & \\ \quad \text{if } \frac{(\mu+(1-\mu)A(P))\left(A(P)\left(1-\phi_{P|P}\right)+(1-A(P))\left(1-\phi_{P|N}\right)\right)}{A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)} \leq \frac{M-Fz_I}{\gamma}. & \end{cases} \end{aligned} \tag{A.237}$$

Proof of Lemma 10

This follows directly from the results in Propositions 15 and 3. □

Proof of Proposition 4

Suppose that $\mathbf{S}_0 = N$. We consider the school district's optimal decision in two scenarios: first, when $B(P) = 0$, and second, when $B(P) \geq 0$ and the formative assessments result is reasonably accurate, i.e., (A.22) holds. When $B(P) \geq 0$, let Z^* denote the set of optimal decisions, z_I^*, for the district, where $Z^* \in \{\{0\}, \{1\}, \{0,1\}\}$. In keeping with earlier notation, when $B(P) = 0$ the set of optimal decisions is represented by $\tilde{Z}^*$.

First, consider the case where $B(P) = 0$. Recall from (A.234) that

$$\tilde{Pr}^*_{z_I}[\mathbf{S}_2 = P|\mathbf{S}_0 = N] = \begin{cases} 0, & \text{if } 0 \leq \frac{M - Fz_I}{\gamma} < 1, \\ \mu A(P), & \text{if } 1 \leq \frac{M - Fz_I}{\gamma} < 1 + \frac{1}{(1-\mu)A(P)}, \\ \mu A(P)\left(1 + (1-\mu)A(P)\right), & \text{if } 1 + \frac{1}{(1-\mu)A(P)} \leq \frac{M - Fz_I}{\gamma}. \end{cases} \tag{A.238}$$

Considering each feasible range of $\frac{M}{\gamma}$ gives the following result.

- If $0 \leq \frac{M}{\gamma} < 1$, then for all $\frac{F}{\gamma} \leq \frac{M}{\gamma}$, $\tilde{Z}^* = \{0,1\}$.
- If $1 \leq \frac{M}{\gamma} < 1 + \frac{1}{(1-\mu)A(P)}$, then

$$\tilde{Z}^* = \begin{cases} \{0,1\}, & \text{if } \frac{F}{\gamma} \leq \frac{M}{\gamma} - 1, \\ \{0\}, & \text{if } \frac{M}{\gamma} - 1 < \frac{F}{\gamma}. \end{cases} \tag{A.239}$$

- If $1 + \frac{1}{(1-\mu)A(P)} \leq \frac{M}{\gamma}$, then

$$\tilde{Z}^* = \begin{cases} \{0,1\}, & \text{if } \frac{F}{\gamma} \leq \frac{M}{\gamma} - 1 - \frac{1}{1-\mu)A(P)}, \\ \{0\}, & \text{if } \frac{M}{\gamma} - 1 - \frac{1}{1-\mu)A(P)} < \frac{F}{\gamma}. \end{cases} \tag{A.240}$$

Next, consider the case where $B(P) > 0$ and the condition in (A.22) holds. Recall from

(A.111) in Proposition 14 that

$$
\begin{aligned}
&Pr^*[\mathbf{S}_2 = P|\mathbf{S}_0 = N] \\
&= \begin{cases}
\mu B(P)\left(1+(1-\mu)B(P)\right), & \text{if } 0 \leq \frac{\pi}{\gamma} < \mathcal{T}_1\left(P,0,N\right), \\
\mu\left(B(P)\left(A(P)\phi_{P|P}+B(P)\right)\right. & \\
\quad \left.+(1-\mu B(P))\left(A(P)\phi_{P|N}+B(P)\right)\right), & \text{if } \mathcal{T}_1\left(P,0,N\right) \leq \frac{\pi}{\gamma} < \mathcal{T}_1\left(N,0,N\right), \\
\mu\left(A(P)+B(P)\right)\left(1+(1-\mu)B(P)\right), & \text{if } \mathcal{T}_1\left(N,0,N\right) \leq \frac{\pi}{\gamma} < \mathcal{T}_0\left(N\right), \\
\mu\left(A(P)+B(P)\right)\left(1+(1-\mu)\left(A(P)+B(P)\right)\right), & \text{if } \mathcal{T}_0\left(N\right) \leq \frac{\pi}{\gamma},
\end{cases}
\end{aligned}
\tag{A.241}
$$

and from (A.193) in Proposition 16 that

$$
\frac{\pi}{\gamma}^* = \begin{cases}
0, & \text{if } 0 \leq \frac{M-Fz_I}{\gamma} < \mathcal{H}_1, \\
\mathcal{T}_1\left(P,0,N\right), & \text{if } \mathcal{H}_1 \leq \frac{M-Fz_I}{\gamma} < \mathcal{H}_2, \\
\mathcal{T}_1\left(N,0,N\right), & \text{if } \mathcal{H}_2 \leq \frac{M-Fz_I}{\gamma} < \frac{1+(1-\mu)B(P)}{(1-\mu)A(P)}+1, \\
\mathcal{T}_0\left(N\right), & \text{if } \frac{(1+(1-\mu)B(P))}{(1-\mu)A(P)}+1 \leq \frac{M-Fz_I}{\gamma},
\end{cases}
\tag{A.242}
$$

where from (A.185) and (A.186)

$$
\begin{aligned}
\mathcal{H}_1 &= \frac{\mu\left(B(P)\left(A(P)\phi_{P|P}+B(P)\right)+(1-\mu B(P))\left(A(P)\phi_{P|N}+B(P)\right)\right)}{\mathcal{T}_1\left(P,0,N\right)}, \\
\mathcal{H}_2 &= \frac{\mu\left(A(P)+B(P)\right)\left(1+(1-\mu)B(P)\right)}{\mathcal{T}_1\left(N,0,N\right)}.
\end{aligned}
\tag{A.243}
$$

Then, the probability that the school ends the year in the proficient state under the optimal

incentive is given by

$$
\begin{aligned}
&Pr^*_{z_I}[\mathbf{S}_2 = P|\mathbf{S}_0 = N] \\
&= \begin{cases}
\mu B(P)\left(1+(1-\mu)B(P)\right), & \text{if } 0 \le \frac{M-Fz_I}{\gamma} < \mathcal{H}_1, \\
\mu\left(B(P)\left(A(P)\phi_{P|P}+B(P)\right)\right. & \\
\quad \left.+(1-\mu B(P))\left(A(P)\phi_{P|N}+B(P)\right)\right), & \text{if } \mathcal{H}_1 \le \frac{M-Fz_I}{\gamma} < \mathcal{H}_2, \\
\mu\left(A(P)+B(P)\right)\left(1+(1-\mu)B(P)\right), & \text{if } \mathcal{H}_2 \le \frac{M-Fz_I}{\gamma} < \frac{1+(1-\mu)B(P)}{(1-\mu)A(P)}+1, \\
\mu\left(A(P)+B(P)\right)\left(1+(1-\mu)\left(A(P)+B(P)\right)\right), & \text{if } \frac{(1+(1-\mu)B(P))}{(1-\mu)A(P)}+1 \le \frac{M-Fz_I}{\gamma}.
\end{cases}
\end{aligned}
\tag{A.244}
$$

Under the interim assessment, this becomes

$$
\begin{aligned}
&Pr^*_{z_I}[\mathbf{S}_2 = P|\mathbf{S}_0 = N] \\
&= \begin{cases}
\mu B(P)\left(1+(1-\mu)B(P)\right), & \text{if } 0 \le \frac{M-Fz_I}{\gamma} < \mathcal{H}^I_1, \\
\mu B(P)\left(\left(A(P)+B(P)\right)+(1-\mu B(P))\right), & \text{if } \mathcal{H}^I_1 \le \frac{M-Fz_I}{\gamma} < \mathcal{H}^I_2, \\
\mu\left(A(P)+B(P)\right)\left(1+(1-\mu)B(P)\right), & \text{if } \mathcal{H}^I_2 \le \frac{M-Fz_I}{\gamma} < \frac{1+(1-\mu)B(P)}{(1-\mu)A(P)}+1, \\
\mu\left(A(P)+B(P)\right)\left(1+(1-\mu)\left(A(P)+B(P)\right)\right), & \text{if } \frac{(1+(1-\mu)B(P))}{(1-\mu)A(P)}+1 \le \frac{M-Fz_I}{\gamma},
\end{cases}
\end{aligned}
\tag{A.245}
$$

where

$$
\begin{aligned}
\mathcal{H}^I_1 &= \mu B(P)\left(\frac{A(P)+1+(1-\mu)B(P)}{A(P)}\right), \\
\mathcal{H}^I_2 &= \frac{\left(A(P)+B(P)\right)\left(1+(1-\mu)B(P)\right)}{A(P)}.
\end{aligned}
\tag{A.246}
$$

Comparing the second cases in (A.244) and (A.245), for $\mu > 0$,

$$\begin{aligned}
&\mu\left(B(P)\left(A(P)\phi_{P|P} + B(P)\right) + (1-\mu B(P))\left(A(P)\phi_{P|N} + B(P)\right)\right) \\
&\leq \mu B(P)\left((A(P) + B(P)) + (1-\mu B(P))\right) \\
\iff &\mu_0 \leq \mu, \text{ where } \mu_0 = \frac{1}{B(P)} - \frac{1-\phi_{P|P}}{\phi_{P|N}}. \qquad \text{(A.247)}
\end{aligned}$$

Note that this holds for $\phi_{P|N}$ sufficiently small and that this is less likely to hold for $B(P)$ small.

Furthermore, note that $\mathcal{H}_2 \leq \mathcal{H}_2^I$ follows from Assumption 1, but $\mathcal{H}_1$ and $\mathcal{H}_1^I$ do not have a clear ordering. In particular,

$$\begin{aligned}
\mathcal{H}_1 \leq \mathcal{H}_1^I \iff &(1-\mu B(P))\,\phi_{P|N} \times \\
&\left(\frac{(1-\mu)\,(1+(1-\mu)B(P))}{A(P)} + \frac{(1-\mu B(P))\,\phi_{P|N}}{B(P)} + \left(\phi_{P|P} - \mu\left(1-\phi_{P|P}\right)\right)\right) \\
&\leq \mu B(P)\phi_{P|P}\left(1-\phi_{P|P}\right). \qquad \text{(A.248)}
\end{aligned}$$

The left-hand side of inequality (A.248) is clearly decreasing in μ and the right-hand side is increasing in μ. Therefore, if (A.248) holds, then there exists $\mu_1 \in [0,1]$ such that (A.248) holds for all $\mu \geq \mu_1$, holding all other parameters constant. In particular, if (A.248) holds for some feasible range of μ values, then it must hold for $\mu = 1$. Plugging this into (A.248), we have

$$\mathcal{H}_1 \leq \mathcal{H}_1^I \Rightarrow \frac{\phi_{P|N}}{\phi_{P|N} + 1 - \phi_{P|P}} \leq B(P). \qquad \text{(A.249)}$$

Finally, we claim that if (A.248) holds, then (A.247) must hold as well. In particular, if there exists $\mu_1 \in [0,1]$ such that (A.248) holds for all $\mu \geq \mu_1$ but $\mathcal{H}_1 \geq \mathcal{H}_1^I$ at μ_0, then

$\mu_1 \geq \mu_0$, hence the claim is true. To see this is indeed the case, note that at μ_0,

$$\begin{aligned}
&\mathcal{H}_1 \geq \mathcal{H}_1^I \\
\iff &\left(1-\mu_0 B(P)\right)\phi_{P|N} \\
&\quad \times \left(\frac{(1-\mu_0)\left(1+(1-\mu_0)B(P)\right)}{A(P)} + \frac{\left(1-\mu_0 B(P)\right)\phi_{P|N}}{B(P)} + \left(\phi_{P|P} - \mu_0\left(1-\phi_{P|P}\right)\right)\right) \\
&\quad \geq \mu_0 B(P)\phi_{P|P}\left(1-\phi_{P|P}\right) \\
\iff &B(P) \geq \frac{\phi_{P|N}}{\phi_{P|N}+1-\phi_{P|P}},
\end{aligned} \tag{A.250}$$

which we know to be true from (A.249). Therefore, $\mathcal{H}_1 \leq \mathcal{H}_1^I \Rightarrow \mu_0 \leq \mu$.

Then, for smaller budgets, i.e. when $\frac{M}{\gamma} < \mathcal{H}_2$, we must consider the following cases. First, suppose $\mathcal{H}_1 \leq \mathcal{H}_1^I$:

- If $0 \leq \frac{M}{\gamma} < \mathcal{H}_1$, then for all $\frac{F}{\gamma} \leq \frac{M}{\gamma}$, $Z^* = \{0,1\}$.
- If $\mathcal{H}_1 \leq \frac{M}{\gamma} < \mathcal{H}_1^I$, then for all $\frac{F}{\gamma} \leq \frac{M}{\gamma}$, $Z^* = \{0\}$.
- If $\mathcal{H}_1^I \leq \frac{M}{\gamma} < \mathcal{H}_2$, then

$$Z^* = \begin{cases} \{1\}, & \text{if } \frac{F}{\gamma} \leq \frac{M}{\gamma} - \mathcal{H}_1^I, \\ \{0\}, & \text{if } \frac{M}{\gamma} - \mathcal{H}_1^I < \frac{F}{\gamma}. \end{cases} \tag{A.251}$$

Second, suppose $\mathcal{H}_1^I \leq \mathcal{H}_1$:

- If $0 \leq \frac{M}{\gamma} < \mathcal{H}_1^I$, then for all $\frac{F}{\gamma} \leq \frac{M}{\gamma}$, $Z^* = \{0,1\}$.
- If $\mathcal{H}_1^I \leq \frac{M}{\gamma} < \mathcal{H}_1$, then

$$Z^* = \begin{cases} \{1\}, & \text{if } \frac{F}{\gamma} \leq \frac{M}{\gamma} - \mathcal{H}_1^I, \\ \{0\}, & \text{if } \frac{M}{\gamma} - \mathcal{H}_1^I < \frac{F}{\gamma}, \end{cases} \tag{A.252}$$

- If $\mathcal{H}_1 \leq \frac{M}{\gamma} < \mathcal{H}_2$,

  - if (A.247) holds, then

$$Z^* = \begin{cases} \{1\}, & \text{if } \frac{F}{\gamma} \leq \frac{M}{\gamma} - \mathcal{H}_1, \\ \{0\}, & \text{if } \frac{M}{\gamma} - \mathcal{H}_1 < \frac{F}{\gamma}, \end{cases} \tag{A.253}$$

  - if (A.247) does not hold, then for all $\frac{F}{\gamma} \leq \frac{M}{\gamma}$, $Z^* = \{0\}$.

Finally, for budgets exceeding $\mathcal{H}_2$,

- If $\mathcal{H}_2 \leq \frac{M}{\gamma} < \mathcal{H}_2^I$, then for all $\frac{F}{\gamma} \leq \frac{M}{\gamma}$, $Z^* = \{0\}$.

- If $\mathcal{H}_2^I \leq \frac{M}{\gamma} < \frac{1+(1-\mu)B(P)}{(1-\mu)A(P)} + 1$, then

$$Z^* = \begin{cases} \{0,1\}, & \text{if } \frac{F}{\gamma} \leq \frac{M}{\gamma} - \mathcal{H}_2^I, \\ \{0\}, & \text{if } \frac{M}{\gamma} - \mathcal{H}_2^I < \frac{F}{\gamma}. \end{cases} \tag{A.254}$$

- If $\frac{1+(1-\mu)B(P)}{(1-\mu)A(P)} + 1 \leq \frac{M}{\gamma}$, then

$$Z^* = \begin{cases} \{0,1\}, & \text{if } \frac{F}{\gamma} \leq \frac{M}{\gamma} - \frac{1+(1-\mu)B(P)}{(1-\mu)A(P)} - 1, \\ \{0\}, & \text{if } \frac{M}{\gamma} - \frac{1+(1-\mu)B(P)}{(1-\mu)A(P)} - 1 < \frac{F}{\gamma}. \end{cases} \tag{A.255}$$

We assume that the district will only invest in the interim assessment if the probability of achieving proficiency in that case is strictly greater than the probably under only the formative assessment results. Furthermore, define $M_L \geq 0$ as the upper bound on the region of budget levels for which the probability of reaching the proficient state at the end of the year is smallest, regardless of the assessment decision, and define $M_U \geq M_L$ as the lower bound on the region of budget levels for which the probability of reaching the proficient

state at the end of the year is highest under either assessment decision for a sufficiently low cost of interim assessment F. Put another way, M_L (M_U) is the upper (lower) bound on the region of trivially-small (large) budget levels.

Then, combining the above results gives the following.

a) If $B(P) = 0$, then $\frac{M_L}{\gamma} = 1$ and $\frac{M_U}{\gamma} = 1 + \frac{1}{(1-\mu)A(P)}$. For all values of the budget, the school district will not invest in the interim assessment.

b) If $B(P) > 0$ and the condition in (A.22) holds, then $\frac{M_L}{\gamma} = \min\left\{\mathcal{H}_1, \mathcal{H}_1^I\right\}$ and $\frac{M_U}{\gamma} = \frac{1+(1-\mu)B(P)}{(1-\mu)A(P)} + 1$. Furthermore, the interim assessment is optimal in two possible scenarios: first, if

$$\mathcal{H}_1^I \leq \frac{M}{\gamma} < \mathcal{H}_1 \text{ and } \frac{F}{\gamma} \leq \frac{M}{\gamma} - \mathcal{H}_1^I, \tag{A.256}$$

and second, if (A.247) holds and

$$\max\left\{\mathcal{H}_1, \mathcal{H}_1^I\right\} \leq \frac{M}{\gamma} < \mathcal{H}_2 \text{ and } \frac{F}{\gamma} \leq \frac{M}{\gamma} - \max\left\{\mathcal{H}_1, \mathcal{H}_1^I\right\}. \quad \square \tag{A.257}$$

Proof of Proposition 5

Suppose that $\mathbf{S}_0 = P$ and $B(P) = 0$. As in the previous Proposition, let $\tilde{Z}^*$ denote the set of optimal decisions, z_I^*, for the district, where $\tilde{Z}^* \in \{\{0\}, \{1\}, \{0,1\}\}$.

For $\mathbf{S}_0 = P$ and $0 \leq (1-\mu)A(P) \leq \mu$, recall from (A.235) that

$$\tilde{Pr}_{z_I}^*[\mathbf{S}_2 = P | \mathbf{S}_0 = P] = \begin{cases} 0, & \text{if } 0 \leq \frac{M - F z_I}{\gamma} < 1, \\ \mu A(P), & \text{if } 1 \leq \frac{M - F z_I}{\gamma} < 1 + \frac{\mu}{(1-\mu)A(P)}, \\ A(P)\left(\mu + (1-\mu)A(P)\right), & \text{if } 1 + \frac{\mu}{(1-\mu)A(P)} \leq \frac{M - F z_I}{\gamma}. \end{cases} \tag{A.258}$$

Then, we have the following result.

- If $0 \leq \frac{M}{\gamma} < 1$, then for all $\frac{F}{\gamma} \leq \frac{M}{\gamma}$, $\tilde{Z}^* = \{0, 1\}$.

- If $1 \leq \frac{M}{\gamma} < 1 + \frac{\mu}{(1-\mu)A(P)}$, then

$$\tilde{Z}^* = \begin{cases} \{0, 1\}, & \text{if } \frac{F}{\gamma} \leq \frac{M}{\gamma} - 1, \\ \{0\}, & \text{if } \frac{M}{\gamma} - 1 < \frac{F}{\gamma}. \end{cases} \tag{A.259}$$

- If $1 + \frac{\mu}{(1-\mu)A(P)} \leq \frac{M}{\gamma}$, then

$$\tilde{Z}^* = \begin{cases} \{0, 1\}, & \text{if } \frac{F}{\gamma} \leq \frac{M}{\gamma} - \frac{\mu}{(1-\mu)A(P)}, \\ \{0\}, & \text{if } \frac{M}{\gamma} - \frac{\mu}{(1-\mu)A(P)} < \frac{F}{\gamma}. \end{cases} \tag{A.260}$$

For $\mathbf{S}_0 = P$ and $\mu \leq (1-\mu)A(P)$, recall from (A.236) that, under only formative assessments, if $\mu < (1-\mu)A(P) \leq \frac{A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)}{(1-A(P))\left(\phi_{P|P}-\phi_{P|N}\right)}$,

$$\tilde{Pr}^*_{z_I}[\mathbf{S}_2 = P | \mathbf{S}_0 = P] = \begin{cases} 0, & \text{if } 0 \leq \frac{M}{\gamma} < 2, \\ A(P)\left(A(P) + \mu\left(1 - A(P)\right)\right), & \text{if } 2 \leq \frac{M}{\gamma}, \end{cases} \tag{A.261}$$

and from (A.237), if $\frac{A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)}{(1-A(P))\left(\phi_{P|P}-\phi_{P|N}\right)} < (1-\mu)A(P) \leq 1$,

$$\tilde{Pr}^*_{z_I}[\mathbf{S}_2 = P|\mathbf{S}_0 = P]$$
$$= \begin{cases} 0, & \text{if } 0 \leq \frac{M}{\gamma} < 1 + A(P)\phi_{P|P} + (1-A(P))\,\phi_{P|N}, \\ A(P)\left(A(P)\phi_{P|P} + \mu\,(1-A(P))\,\phi_{P|N}\right), & \\ & \text{if } 1 + A(P)\phi_{P|P} + (1-A(P))\,\phi_{P|N} \leq \frac{M}{\gamma} \\ & \quad < \frac{(\mu+(1-\mu)A(P))\left(A(P)\left(1-\phi_{P|P}\right)+(1-A(P))\left(1-\phi_{P|N}\right)\right)}{A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)}, \\ A(P)\left(A(P)+\mu(1-A(P))\right), & \\ & \text{if } \frac{(\mu+(1-\mu)A(P))\left(A(P)\left(1-\phi_{P|P}\right)+(1-A(P))\left(1-\phi_{P|N}\right)\right)}{A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)} \leq \frac{M}{\gamma}. \end{cases} \tag{A.262}$$

Under the interim assessment, for all $\mu < (1-\mu)A(P) \leq 1$, these two cases simplify to

$$\tilde{Pr}^*_{z_I}[\mathbf{S}_2 = P|\mathbf{S}_0 = P]$$
$$= \begin{cases} 0, & \text{if } 0 \leq \frac{M-F}{\gamma} < 1 + A(P), \\ (A(P))^2, & \text{if } 1 + A(P) \leq \frac{M-F}{\gamma} < 1 + \left(\frac{1-\mu}{\mu}\right)A(P), \\ A(P)\left(A(P)+\mu\left(1-A(P)\right)\right), & \text{if } 1 + \left(\frac{1-\mu}{\mu}\right)A(P) \leq \frac{M-F}{\gamma}. \end{cases} \tag{A.263}$$

When determining the optimal assessment decision, we must consider two cases. First, consider the case where the accuracy of the formative assessment is such that $\mu < (1-\mu)A(P) \leq \frac{A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)}{(1-A(P))\left(\phi_{P|P}-\phi_{P|N}\right)}$. Comparing two of the bounds on $\frac{M}{\gamma}$, note that

$$2 < 1 + \left(\frac{1-\mu}{\mu}\right)A(P) \iff \mu < (1-\mu)\,A(P). \tag{A.264}$$

Then, we have the following result.

- If $0 \leq \frac{M}{\gamma} < 1 + A(P)$, then for all $\frac{F}{\gamma} \leq \frac{M}{\gamma}$, $\tilde{Z}^* = \{0, 1\}$.

- If $1 + A(P) \leq \frac{M}{\gamma} < 2$, then

$$\tilde{Z}^* = \begin{cases} \{1\}, & \text{if } \frac{F}{\gamma} \leq \frac{M}{\gamma} - 1 - A(P), \\ \{0,1\}, & \text{if } \frac{M}{\gamma} - 1 - A(P) < \frac{F}{\gamma}. \end{cases} \tag{A.265}$$

- If $2 \leq \frac{M}{\gamma} < 1 + \left(\frac{1-\mu}{\mu}\right) A(P)$, then for all $\frac{F}{\gamma} \leq \frac{M}{\gamma}$, $\tilde{Z}^* = \{0\}$.

- If $1 + \left(\frac{1-\mu}{\mu}\right) A(P) \leq \frac{M}{\gamma}$, then

$$\tilde{Z}^* = \begin{cases} \{0,1\}, & \text{if } \frac{F}{\gamma} \leq \frac{M}{\gamma} - 1 - \left(\frac{1-\mu}{\mu}\right) A(P), \\ \{0\}, & \text{if } \frac{M}{\gamma} - 1 - A(P) < \frac{F}{\gamma}. \end{cases} \tag{A.266}$$

Next, consider the case where $\frac{A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)}{(1-A(P))\left(\phi_{P|P}-\phi_{P|N}\right)} < (1-\mu)A(P) \leq 1$. First, rewrite the bound in this case in terms of μ:

$$\begin{aligned} &\frac{A(P)\left(1-\phi_{P|P}\right)+\mu\left(1-A(P)\right)\left(1-\phi_{P|N}\right)}{\left(1-A(P)\right)\left(\phi_{P|P}-\phi_{P|N}\right)} < (1-\mu)A(P) \\ \iff &\mu < \frac{A(P)\left(\left(1-A(P)\right)\left(\phi_{P|P}-\phi_{P|N}\right)-\left(1-\phi_{P|P}\right)\right)}{\left(1-A(P)\right)\left(\left(1-\phi_{P|N}\right)+A(P)\left(\phi_{P|P}-\phi_{P|N}\right)\right)} \end{aligned} \tag{A.267}$$

Then, compare the smaller bound on $\frac{M-F}{\gamma}$ in the interim case to the larger bound on $\frac{M}{\gamma}$ in the formative case:

$$\begin{aligned} 1 + A(P) &\leq \frac{(\mu+(1-\mu)A(P))\left(A(P)\left(1-\phi_{P|P}\right)+\left(1-A(P)\right)\left(1-\phi_{P|N}\right)\right)}{A(P)\left(1-\phi_{P|P}\right)+\mu\left(1-A(P)\right)\left(1-\phi_{P|N}\right)} \\ \iff \mu &\leq \frac{\left(1-A(P)\right)\left(1-\phi_{P|N}\right)-\left(1-\phi_{P|P}\right)}{\left(1-A(P)\right)\left(\left(\phi_{P|P}-\phi_{P|N}\right)+\left(1-\phi_{P|N}\right)\right)}. \end{aligned} \tag{A.268}$$

The bound on μ in this case, (A.267), is stronger than the bound in (A.268); therefore,

(A.268) always holds in this case:

$$\begin{aligned}
&\frac{A(P)\left((1-A(P))\left(\phi_{P|P}-\phi_{P|N}\right)-\left(1-\phi_{P|P}\right)\right)}{(1-A(P))\left(\left(1-\phi_{P|N}\right)+A(P)\left(\phi_{P|P}-\phi_{P|N}\right)\right)} \\
&\leq \frac{(1-A(P))\left(1-\phi_{P|N}\right)-\left(1-\phi_{P|P}\right)}{(1-A(P))\left(\left(\phi_{P|P}-\phi_{P|N}\right)+\left(1-\phi_{P|N}\right)\right)} \\
\iff & A(P)\left(\phi_{P|P}-\phi_{P|N}\right) \leq 1-\phi_{P|N}. \qquad \text{(A.269)}
\end{aligned}$$

Furthermore, the larger bound on $\frac{M}{\gamma}$ in the formative case is smaller than the larger bound on $\frac{M-F}{\gamma}$ in the interim case, since

$$\begin{aligned}
&\frac{(\mu+(1-\mu)A(P))\left(A(P)\left(1-\phi_{P|P}\right)+(1-A(P))\left(1-\phi_{P|N}\right)\right)}{A(P)\left(1-\phi_{P|P}\right)+\mu\left(1-A(P)\right)\left(1-\phi_{P|N}\right)} \leq 1+\left(\frac{1-\mu}{\mu}\right)A(P) \\
\iff & (1-\mu)A(P)\left(1-\phi_{P|P}\right)\left(A(P)\left(1-\frac{1}{\mu}\right)-1\right) \leq 0, \qquad \text{(A.270)}
\end{aligned}$$

which clearly holds. Then, there are two possible orders of the bounds on $\frac{M-Fz_I}{\gamma}$. First, suppose that

$$\begin{aligned}
&1+A(P) \\
&\leq 1+A(P)\phi_{P|P}+\left(1-A(P)\right)\phi_{P|N} \\
&\leq \frac{(\mu+(1-\mu)A(P))\left(A(P)\left(1-\phi_{P|P}\right)+(1-A(P))\left(1-\phi_{P|N}\right)\right)}{A(P)\left(1-\phi_{P|P}\right)+\mu\left(1-A(P)\right)\left(1-\phi_{P|N}\right)} \\
&\leq 1+\left(\frac{1-\mu}{\mu}\right)A(P). \qquad \text{(A.271)}
\end{aligned}$$

Then,

$$1+A(P) \leq 1+A(P)\phi_{P|P}+\left(1-A(P)\right)\phi_{P|N} \iff A(P)\left(1-\phi_{P|P}\right) \leq \left(1-A(P)\right)\phi_{P|N}, \qquad \text{(A.272)}$$

and therefore,

$$\begin{aligned} &(A(P))^2 \leq A(P)\left(A(P)\phi_{P|P} + \mu\left(1 - A(P)\right)\phi_{P|N}\right) \\ \iff &A(P)(1-\phi_{P|P}) \leq \mu\left(1 - A(P)\right)\phi_{P|N} \end{aligned} \tag{A.273}$$

holds, since $\mu \leq 1$. Then, we have the following result.

- If $0 \leq \frac{M}{\gamma} < 1 + A(P)$, then for all $\frac{F}{\gamma} \leq \frac{M}{\gamma}$, $\tilde{Z}^* = \{0, 1\}$.
- If $1 + A(P) \leq \frac{M}{\gamma} < 1 + A(P)\phi_{P|P} + \left(1 - A(P)\right)\phi_{P|N}$, then

$$\tilde{Z}^* = \begin{cases} \{1\}, & \text{if } \frac{F}{\gamma} \leq \frac{M}{\gamma} - 1 - A(P), \\ \{0, 1\}, & \text{if } \frac{M}{\gamma} - 1 - A(P) < \frac{F}{\gamma}. \end{cases} \tag{A.274}$$

- If $1+A(P)\phi_{P|P}+\left(1 - A(P)\right)\phi_{P|N} \leq \frac{M}{\gamma} < \frac{(\mu+(1-\mu)A(P))\left(A(P)\left(1-\phi_{P|P}\right)+(1-A(P))\left(1-\phi_{P|N}\right)\right)}{A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)}$, then for all $\frac{F}{\gamma} \leq \frac{M}{\gamma}$, $\tilde{Z}^* = \{0\}$.
- If $\frac{(\mu+(1-\mu)A(P))\left(A(P)\left(1-\phi_{P|P}\right)+(1-A(P))\left(1-\phi_{P|N}\right)\right)}{A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)} \leq \frac{M}{\gamma} < 1 + \left(\frac{1-\mu}{\mu}\right)A(P)$, then for all $\frac{F}{\gamma} \leq \frac{M}{\gamma}$, $\tilde{Z}^* = \{0\}$.
- If $1 + \left(\frac{1-\mu}{\mu}\right)A(P) \leq \frac{M}{\gamma}$, then

$$\tilde{Z}^* = \begin{cases} \{0, 1\}, & \text{if } \frac{F}{\gamma} \leq \frac{M}{\gamma} - 1 - \left(\frac{1-\mu}{\mu}\right)A(P), \\ \{0\}, & \text{if } \frac{M}{\gamma} - 1 - \left(\frac{1-\mu}{\mu}\right)A(P) < \frac{F}{\gamma}. \end{cases} \tag{A.275}$$

Second, suppose that

$$\begin{aligned}
&1 + A(P)\phi_{P|P} + (1 - A(P))\,\phi_{P|N} \\
&\leq 1 + A(P) \\
&\leq \frac{(\mu + (1-\mu)A(P))\left(A(P)\left(1-\phi_{P|P}\right) + (1-A(P))\left(1-\phi_{P|N}\right)\right)}{A(P)\left(1-\phi_{P|P}\right) + \mu\,(1-A(P))\left(1-\phi_{P|N}\right)} \\
&\leq 1 + \left(\frac{1-\mu}{\mu}\right) A(P)
\end{aligned} \tag{A.276}$$

Then,

$$A(P)\left(A(P)\phi_{P|P} + \mu\,(1-A(P))\,\phi_{P|N}\right) \leq (A(P))^2\,, \tag{A.277}$$

and we have the following result.

- If $0 \leq \frac{M}{\gamma} < 1 + A(P)\phi_{P|P} + (1-A(P))\,\phi_{P|N}$, then for all $\frac{F}{\gamma} \leq \frac{M}{\gamma}$, $\tilde{Z}^* = \{0,1\}$.
- If $1 + A(P)\phi_{P|P} + (1-A(P))\,\phi_{P|N} \leq \frac{M}{\gamma} < 1 + A(P)$, then for all $\frac{F}{\gamma} \leq \frac{M}{\gamma}$, $\tilde{Z}^* = \{0\}$.
- If $1 + A(P) \leq \frac{M}{\gamma} < \frac{(\mu+(1-\mu)A(P))\left(A(P)\left(1-\phi_{P|P}\right)+(1-A(P))\left(1-\phi_{P|N}\right)\right)}{A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)}$, then

$$\tilde{Z}^* = \begin{cases} \{1\}\,, & \text{if } \frac{F}{\gamma} \leq \frac{M}{\gamma} - 1 - A(P), \\ \{0\}\,, & \text{if } \frac{M}{\gamma} - 1 - A(P) < \frac{F}{\gamma}. \end{cases} \tag{A.278}$$

- If $\frac{(\mu+(1-\mu)A(P))\left(A(P)\left(1-\phi_{P|P}\right)+(1-A(P))\left(1-\phi_{P|N}\right)\right)}{A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)} \leq \frac{M}{\gamma} < 1 + \left(\frac{1-\mu}{\mu}\right) A(P)$, then for all $\frac{F}{\gamma} \leq \frac{M}{\gamma}$, $\tilde{Z}^* = \{0\}$.
- If $1 + \left(\frac{1-\mu}{\mu}\right) A(P) \leq \frac{M}{\gamma}$, then

$$\tilde{Z}^* = \begin{cases} \{0,1\}\,, & \text{if } \frac{F}{\gamma} \leq \frac{M}{\gamma} - 1 - \left(\frac{1-\mu}{\mu}\right) A(P), \\ \{0\}\,, & \text{if } \frac{M}{\gamma} - 1 - \left(\frac{1-\mu}{\mu}\right) A(P) < \frac{F}{\gamma}. \end{cases} \tag{A.279}$$

From Proposition 4, recall that we assume that the district will only invest in the interim assessment if the probability of achieving proficiency in that case is strictly greater than the probably under only the formative assessment results, and that M_L (M_U) is the upper (lower) bound for trivially small (large) budget levels.

Then, combining these results gives the school district's optimal assessment decision.

a) If $\mathbf{S}_0 = P$ and $0 \leq (1-\mu)A(P) \leq \mu$, then $\frac{M_L}{\gamma} = 1$ and $\frac{M_U}{\gamma} = 1 + \frac{\mu}{(1-\mu)A(P)}$ with $\frac{F'}{\gamma} = \frac{M}{\gamma} - \frac{\mu}{(1-\mu)A(P)}$. For all values of the budget, the school district will not invest in the interim assessment.

b) If $\mathbf{S}_0 = P$ and $\mu < (1-\mu)A(P) \leq \frac{A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)}{(1-A(P))\left(\phi_{P|P}-\phi_{P|N}\right)}$, then $\frac{M_L}{\gamma} = 1 + A(P)$ and $\frac{M_U}{\gamma} = 1 + \left(\frac{1-\mu}{\mu}\right)A(P)$ with $\frac{F'}{\gamma} = \frac{M}{\gamma} - 1 - \left(\frac{1-\mu}{\mu}\right)A(P)$. Furthermore,

- If $1 + A(P) \leq \frac{M}{\gamma} < 2$, then

$$\tilde{Z}^* = \begin{cases} \{1\}, & \text{if } \frac{F}{\gamma} \leq \frac{M}{\gamma} - 1 - A(P), \\ \{0,1\}, & \text{if } \frac{M}{\gamma} - 1 - A(P) < \frac{F}{\gamma}. \end{cases} \tag{A.280}$$

- If $2 \leq \frac{M}{\gamma} < 1 + \left(\frac{1-\mu}{\mu}\right)A(P)$, then for all $\frac{F}{\gamma} \leq \frac{M}{\gamma}$, $\tilde{Z}^* = \{0\}$.

c) If $\frac{A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)}{(1-A(P))\left(\phi_{P|P}-\phi_{P|N}\right)} < (1-\mu)A(P) \leq 1$ and $1+A(P)\phi_{P|P}+(1-A(P))\,\phi_{P|N} \leq 1 + A(P)$, then $\frac{M_L}{\gamma} = 1 + A(P)$ and $\frac{M_U}{\gamma} = 1 + \left(\frac{1-\mu}{\mu}\right)A(P)$ with $\frac{F'}{\gamma} = \frac{M}{\gamma} - 1 - \left(\frac{1-\mu}{\mu}\right)A(P)$. Furthermore,

- If $1 + A(P) \leq \frac{M}{\gamma} < 1 + A(P)\phi_{P|P} + (1-A(P))\,\phi_{P|N}$, then

$$\tilde{Z}^* = \begin{cases} \{1\}, & \text{if } \frac{F}{\gamma} \leq \frac{M}{\gamma} - 1 - A(P), \\ \{0,1\}, & \text{if } \frac{M}{\gamma} - 1 - A(P) < \frac{F}{\gamma}. \end{cases} \tag{A.281}$$

- If $1+A(P)\phi_{P|P}+(1-A(P))\,\phi_{P|N} \leq \frac{M}{\gamma} < 1+\left(\frac{1-\mu}{\mu}\right)A(P)$, then for all $\frac{F}{\gamma} \leq \frac{M}{\gamma}$, $\tilde{Z}^* = \{0\}$.

d) If $\frac{A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)}{(1-A(P))\left(\phi_{P|P}-\phi_{P|N}\right)} < (1-\mu)A(P) \leq 1$ and $1+A(P) \leq 1+A(P)\phi_{P|P}+(1-A(P))\,\phi_{P|N}$, then $\frac{M_L}{\gamma} = 1+A(P)\phi_{P|P}+(1-A(P))\,\phi_{P|N}$ and $\frac{M_U}{\gamma} = 1+\left(\frac{1-\mu}{\mu}\right)A(P)$ with $\frac{F'}{\gamma} = \frac{M}{\gamma}-1-\left(\frac{1-\mu}{\mu}\right)A(P)$. Furthermore,

- If $1+A(P)\phi_{P|P}+(1-A(P))\,\phi_{P|N} \leq \frac{M}{\gamma} < 1+A(P)$, then for all $\frac{F}{\gamma} \leq \frac{M}{\gamma}$, $\tilde{Z}^* = \{0\}$.

- If $1+A(P) \leq \frac{M}{\gamma} < \frac{(\mu+(1-\mu)A(P))\left(A(P)\left(1-\phi_{P|P}\right)+(1-A(P))\left(1-\phi_{P|N}\right)\right)}{A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)}$, then

$$\tilde{Z}^* = \begin{cases} \{1\}, & \text{if } \frac{F}{\gamma} \leq \frac{M}{\gamma}-1-A(P), \\ \{0\}, & \text{if } \frac{M}{\gamma}-1-A(P) < \frac{F}{\gamma}. \end{cases} \tag{A.282}$$

- If $\frac{(\mu+(1-\mu)A(P))\left(A(P)\left(1-\phi_{P|P}\right)+(1-A(P))\left(1-\phi_{P|N}\right)\right)}{A(P)\left(1-\phi_{P|P}\right)+\mu(1-A(P))\left(1-\phi_{P|N}\right)} \leq \frac{M}{\gamma} < 1+\left(\frac{1-\mu}{\mu}\right)A(P)$, then for all $\frac{F}{\gamma} \leq \frac{M}{\gamma}$, $\tilde{Z}^* = \{0\}$. $\square$

Proof of Lemma 1

Assuming a perfect correlation between the use of resources upon the original admission and upon readmission, and assuming a single potential readmission, we obtain that the total use of operating room time by any "type-i" patient, F_i is equal to S_i, the duration of the original surgery (with probability $1-p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)$), or to $2S_i$ (with probability $p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)$). Thus, F_i is distributed on the interval $[0, 2S_i^{\max}]$, and its CDF is given by

$$\begin{aligned} \Phi_i^F\left(x|h_i^{\mathrm{d}}\right) = \mathrm{P}\left[F_i \leq x\right] &= \left(1-p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)\mathrm{P}\left[S_i \leq x\right] + p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\mathrm{P}\left[2S_i \leq x\right] \\ &= \left(1-p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)\mathrm{P}\left[S_i \leq x\right] + p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\mathrm{P}\left[S_i \leq \frac{x}{2}\right] \\ &= \left(1-p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)\Phi_i^S\left(x\right) + p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\Phi_i^S\left(\frac{x}{2}\right). \end{aligned} \tag{A.283}$$

From the point of view of the total use of recovery beds, under the discharge policy characterized by the threshold h_i^{d}, each type-i patient with the recovery rate r_i has a total hospital

length of stay equal to the sum of the length of stay from the original admission and from a potential readmission, i.e., $L_i\left(h_i^{\mathrm{d}}\right) = L_i^{\mathrm{o}}\left(h_i^{\mathrm{d}}\right) + L_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)$, equal to

$$L_i\left(h_i^{\mathrm{d}}\right) = \begin{cases} \frac{h_i^{\mathrm{d}}}{r_i}, & \text{with probability } 1 - p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right), \\ \frac{2h_i^{\mathrm{d}}}{r_i}, & \text{with probability } p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right). \end{cases} \tag{A.284}$$

Note that $L_i\left(h_i^{\mathrm{d}}\right)$ is distributed on the interval $\left[L_i^{\min}\left(h_i^{\mathrm{d}}\right), 2L_i^{\max}\left(h_i^{\mathrm{d}}\right)\right]$, where from (2.6) and (2.7)

$$L_i^{\min}\left(h_i^{\mathrm{d}}\right) = \frac{h_i^{\mathrm{d}}}{r_i^{\max}}, \tag{A.285}$$

$$L_i^{\max}\left(h_i^{\mathrm{d}}\right) = \frac{h_i^{\mathrm{d}}}{r_i^{\min}}, \tag{A.286}$$

with the CDF

$$\begin{aligned} \Phi_i^L(z|h_i^{\mathrm{d}}) &= \mathrm{P}\left[L_i\left(h_i^{\mathrm{d}}\right) \le z\right] = \left(1 - p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)\mathrm{P}\left[\frac{h_i^{\mathrm{d}}}{r_i} \le z\right] + p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\mathrm{P}\left[\frac{2h_i^{\mathrm{d}}}{r_i} \le z\right] \\ &= \left(1 - p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)\mathrm{P}\left[r_i \ge \frac{h_i^{\mathrm{d}}}{z}\right] + p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\mathrm{P}\left[r_i \ge \frac{2h_i^{\mathrm{d}}}{z}\right] \\ &= \left(1 - p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)\left(1 - \Phi_i^{\mathrm{r}}\left(\frac{h_i^{\mathrm{d}}}{z}\right)\right) + p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\left(1 - \Phi_i^{\mathrm{r}}\left(\frac{2h_i^{\mathrm{d}}}{z}\right)\right). \end{aligned} \tag{A.287}$$

□

Proof of Proposition 6

For simplicity, in the proof we will omit the dependence of the involved quantities on the discharge thresholds h_i^{d}.

We first prove the normal approximation for the usage of the operating room time. Let k_i index each patient undergoing procedure i, where $k_i = 1, \ldots, a_i$, and let $F_t^{k_i}$ be a non-negative random variable describing the total duration of the procedure for patient k_i who is first admitted at time t, where for procedure i, $F_t^{k_i}$ are i.i.d. random variables with mean μ_i^F and variance $\left(\sigma_i^F\right)^2$. Then, the total front-end resource utilization associated with

procedure i is $F_t^i(a_i) = \sum_{k_i=1}^{a_i} F_t^{k_i}$, where $E\left(F_t^i(a_i)\right) = a_i\mu_i^F$ and $V\left(F_t^i(a_i)\right) = a_i\left(\sigma_i^F\right)^2$.

The Lyapunov Central Limit Theorem (Billingsley, 1995, p. 362) states that for a sequence of m independent random variables X_n, $n = 1, \ldots, m$, each with finite expectation μ_n and variance σ_n^2, the expression

$$\frac{\sum_{n=1}^{m}(X_n - \mu_n)}{\sqrt{\sum_{n=1}^{m}\sigma_n^2}} \xrightarrow{d} \mathcal{N}(0,1) \tag{A.288}$$

when $m \to \infty$ provided there exists $\delta > 0$ such that

$$\lim_{m\to\infty} \frac{\sum_{n=1}^{m} E\left[|X_n - \mu_n|^{2+\delta}\right]}{\left(\sum_{n=1}^{m}\sigma_n^2\right)^{\frac{2+\delta}{2}}} = 0. \tag{A.289}$$

Applying this result to the sequence $F_t^{k_i}$, $k_i = 1, \ldots, a_i$, we get that

$$\frac{F_t^i(a_i) - a_i\mu_i^F}{\sqrt{a_i}\sigma_i^F} \xrightarrow{d} \mathcal{N}(0,1) \tag{A.290}$$

provided there exists $\delta > 0$ such that

$$\lim_{a_i\to\infty} \frac{\sum_{k_i=1}^{a_i} E\left[\left|F_t^{k_i} - \mu_i^F\right|^{2+\delta}\right]}{\left(\sum_{k_i=1}^{a_i}\left(\sigma_i^F\right)^2\right)^{\frac{2+\delta}{2}}} = 0. \tag{A.291}$$

Since $F_t^{k_i}$ are i.i.d random variables, the expression in the limit in (A.291) becomes

$$\frac{a_i E\left[\left|F_t^{k_i} - \mu_i^F\right|^{2+\delta}\right]}{\left(a_i\left(\sigma_i^F\right)^2\right)^{\frac{2+\delta}{2}}} = \left(\frac{E\left[\left|F_t^{k_i} - \mu_i^F\right|^{2+\delta}\right]}{\left(\sigma_i^F\right)^{2+\delta}}\right) a_i^{-\frac{\delta}{2}}. \tag{A.292}$$

Then, (A.291) always holds since, for all $\delta > 0$,

$$\lim_{a_i \to \infty} a_i^{-\frac{\delta}{2}} = 0. \tag{A.293}$$

We can use this result to represent the limiting distribution of $F_t^i(a_i)$ as a normal random variable with mean $a_i \mu_i^F$ and variance $a_i \left(\sigma_i^F\right)^2$. Then, given the independence of resource-utilization for different procedure types, the total front-end resource utilization across all procedure types is represented by the normal random variable

$$F_t(\mathbf{a}) = \sum_{i=1}^{N} F_t^i(a_i) \sim \mathcal{N}\left(M^F(\mathbf{a}), \left(\Sigma^F(\mathbf{a})\right)^2\right), \tag{A.294}$$

where

$$M^F(\mathbf{a}) = \sum_{i=1}^{N} a_i \mu_i^F, \tag{A.295}$$

$$\left(\Sigma^F(\mathbf{a})\right)^2 = \sum_{i=1}^{N} a_i \left(\sigma_i^F\right)^2. \tag{A.296}$$

We next prove the normal approximation for the usage of recovery beds for a single procedure. Recall that the minimum and maximum values for patient length of stay (LOS) after a procedure are $L^{\min}$ and $2L^{\max}$ days, respectively. (This includes recovery time after a potential readmission.) Each day can be divided into $\eta \in \mathbb{N}$ equally sized periods of width $\frac{1}{\eta}$. Let L be a non-negative random variable describing patient LOS. Also, let $p^t = P\left(L \geq \frac{t}{\eta}\right)$ be the probability that a patient stays at least $t \in \{1, 2, \ldots, 2\eta L^{\max}\}$ periods at the hospital, with $p^{\eta L^{\min}} = 1$, and $p^t \geq p^{t+1}$. Define X^{dt} to be the number of patients who underwent the procedure on period d and who also spent at least t periods at the hospital.

The hospital performs a procedures each day, so there are $\frac{a}{\eta}$ procedures performed in each period. Therefore, X^{dt} is a binomial random variable with parameters $(\frac{a}{\eta}, p^t)$. Let Y^d be

the number of patients who stay at the hospital in period d. Then,

$$\begin{aligned} Y^d &= X^{d,1} + X^{d-1,2} + X^{d-2,3} + \ldots + X^{d-(2\eta L^{\max}-1),2\eta L^{\max}} \\ &= \left(\frac{a}{\eta}\right)\eta L^{\min} + X^{d-\eta L^{\min},\eta L^{\min}+1} + X^{d-\left(\eta L^{\min}+1\right),\eta L^{\min}+2} + \ldots + X^{d-(2\eta L^{\max}-1),2\eta L^{\max}} \\ &= aL^{\min} + \sum_{t=\eta L^{\min}+1}^{2\eta L^{\max}} X^{d-t+1,t}, \end{aligned} \tag{A.297}$$

where $X^{d-t+1,t}$ are all independent random variables since they refer to patients who underwent the procedure in different periods. Moreover, $X^{d-t+1,t}$ is binomial with $\left(\frac{a}{\eta}, p^t\right)$ for each $t \in \{\eta L^{\min}+1, \eta L^{\min}+2, \ldots, 2\eta L^{\max}\}$. Thus, X^{dt} and Y^d have the same distribution for any period d. Then, we can let Y denote the number of occupied beds in any period, so that

$$Y = aL^{\min} + \sum_{t=\eta L^{\min}+1}^{2\eta L^{\max}} X^t, \tag{A.298}$$

where X^t is binomial with parameters $\left(\frac{a}{\eta}, p^t\right)$.

As (A.298) involves the convolution of η binomial random variables, we can compute the expected value and the variance of Y, the number of occupied beds in a period. Since $X^t \sim B\left(\frac{a}{\eta}, p^t\right)$, we have $E(X^t) = \left(\frac{a}{\eta}\right)p^t$. Then,

$$\mu_Y = E(Y) = aL^{\min} + \sum_{t=\eta L^{\min}+1}^{\eta 2L^{\max}} \left(\frac{a}{\eta}\right)p^t = a\left(L^{\min} + \sum_{t=\eta L^{\min}+1}^{\eta 2L^{\max}} p^t\left(\frac{1}{\eta}\right)\right). \tag{A.299}$$

For the variance, $V(X^t) = \left(\frac{a}{\eta}\right)p^t(1-p^t)$, and

$$\sigma_Y^2 = V(Y) = \sum_{t=\eta L^{\min}+1}^{2\eta L^{\max}} \left(\frac{a}{\eta}\right)p^t(1-p^t) = a\left(\sum_{t=\eta L^{\min}+1}^{2\eta L^{\max}} p^t\left(1-p^t\right)\left(\frac{1}{\eta}\right)\right). \tag{A.300}$$

As $\eta \to \infty$, we can replace p^t with $1 - \Phi^L$ and use the following notation for μ_Y and σ_Y:

$$\mu_Y = a\mu^B, \quad \sigma_Y^2 = a\left(\sigma^B\right)^2, \tag{A.301}$$

where

$$\mu^B = L^{\min} + \int_{L^{\min}}^{2L^{\max}} \left(1 - \Phi^L(x)\right) dx, \tag{A.302}$$

$$\left(\sigma^B\right)^2 = \int_{L^{\min}}^{2L^{\max}} \Phi^L(x)\left(1 - \Phi^L(x)\right) dx. \tag{A.303}$$

As in the case of operating room usage, we apply the Lyapunov CLT to get $Y \sim \mathcal{N}\left(\mu_Y, \sigma_Y^2\right)$ as $a \to \infty$. In the multiple resource case, Y becomes the sum of independent normal distributions for each procedure, so the limiting distribution is $Y \sim \mathcal{N}\left(M^B(\mathbf{a}), \left(\Sigma^B(\mathbf{a})\right)^2\right)$, where

$$M^B(\mathbf{a}) = \sum_{i=1}^{N} a_i \mu_i^B, \tag{A.304}$$

$$\left(\Sigma^B(\mathbf{a})\right)^2 = \sum_{i=1}^{N} a_i \left(\sigma_i^B\right)^2. \tag{A.305}$$

□

Proof of Proposition 7

a) Using (2.9) and $\bar{S}_i = E[S_i]$, we can express (2.17) as

$$\begin{aligned} E[F_i] = \mu_i^F\left(h_i^{\mathrm{d}}\right) &= \int_0^{2S_i^{\max}} \left(1 - \left(1 - p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)\Phi_i^S(x) - p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\Phi_i^S\left(\frac{x}{2}\right)\right) dx \\ &= \left(1 - p^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)\bar{S}_i + p^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\left(2\bar{S}_i\right) \\ &= \bar{S}_i\left(1 + p^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right). \end{aligned} \tag{A.306}$$

Then, since $\frac{dp_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)}{dh_i^{\mathrm{d}}} < 0$, we have $\frac{d\mu_i^S\left(h_i^{\mathrm{d}}\right)}{dh_i^{\mathrm{d}}} < 0$.

Using this result, we can express (2.18) as

$$\begin{aligned}
\mathrm{Var}\left[F_i\right] = \left(\sigma_i^F\left(h_i^{\mathrm{d}}\right)\right)^2 = &\int_0^{2S_i^{\max}} \left(x-\mu_i^F\left(h_i^{\mathrm{d}}\right)\right)^2 d\Phi_i^F\left(x|h_i^{\mathrm{d}}\right) \\
=&p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\int_0^{2S_i^{\max}} \left(x-\mu_i^F\left(h_i^{\mathrm{d}}\right)\right)^2 d\Phi_i^S\left(\frac{x}{2}\right) \\
&+\left(1-p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)\int_0^{2S_i^{\max}} \left(x-\mu_i^F\left(h_i^{\mathrm{d}}\right)\right)^2 d\Phi_i^S\left(x\right) \\
=&p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\int_0^{S_i^{\max}} \left(2x-\left(1+p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)\bar{S}_i\right)^2 d\Phi_i^S\left(x\right) \\
&+\left(1-p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)\int_0^{S_i^{\max}} \left(x-\left(1+p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)\bar{S}_i\right)^2 d\Phi_i^S\left(x\right) \\
=&p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\int_0^{S_i^{\max}} \left(4x^2-4x\left(1+p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)\bar{S}_i+\left(1+p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)^2\bar{S}_i^2\right) d\Phi_i^S\left(x\right) \\
&+\left(1-p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)\int_0^{S_i^{\max}} \left(x^2-2x\left(1+p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)\bar{S}_i+\left(1+p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)^2\bar{S}_i^2\right) d\Phi_i^S\left(x\right) \\
=&\int_0^{S_i^{\max}} \left(\left(1+3p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)x^2-2\left(1+p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)^2\bar{S}_i x+\left(1+p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)^2\bar{S}_i^2\right) d\Phi_i^S\left(x\right) \\
=&\left(1+3p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)\left(\mathrm{Var}\left[S_i\right]+\left(\bar{S}_i\right)^2\right)-2\left(1+p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)^2\bar{S}_i^2+\left(1+p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)^2\bar{S}_i^2 \\
=&\left(1+3p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)\mathrm{Var}\left[S_i\right]+p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\left(1-p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)\bar{S}_i^2, \qquad \text{(A.307)}
\end{aligned}$$

where $\mathrm{Var}\left[S_i\right]$ is the variance of the type-i surgical procedure duration. Then, since $\frac{dp_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)}{dh_i^{\mathrm{d}}} < 0$ we have $\frac{d\left(\sigma_i^F\left(h_i^{\mathrm{d}}\right)\right)^2}{dh_i^{\mathrm{d}}} \leq 0$ if and only if

$$3\mathrm{Var}\left[S_i\right]+\left(1-2p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)\bar{S}_i^2 \geq 0, \qquad \text{(A.308)}$$

or

$$p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right) \leq \frac{1}{2}+\frac{3}{2}\left(\frac{\mathrm{Var}\left[S_i\right]}{\bar{S}_i^2}\right). \qquad \text{(A.309)}$$

b) Using (2.10) and integration by parts, we have

$$
\begin{aligned}
&E\left[L_i\left(h_i^{\mathrm{d}}\right)\right] \\
=&\mu_i^B\left(h_i^{\mathrm{d}}\right)=\int_0^{2L_i^{\max}\left(h_i^{\mathrm{d}}\right)}\left(1-\Phi_i^L(z|h_i^{\mathrm{d}})\right)dz \\
=&\frac{h_i^{\mathrm{d}}}{r_i^{\max}}+\int_{L_i^{\min}\left(h_i^{\mathrm{d}}\right)}^{2L_i^{\max}\left(h_i^{\mathrm{d}}\right)}\left(1-\Phi_i^L(z|h_i^{\mathrm{d}})\right)dz \\
=&\frac{h_i^{\mathrm{d}}}{r_i^{\max}}+ \\
&\int_{L_i^{\min}\left(h_i^{\mathrm{d}}\right)}^{2L_i^{\max}\left(h_i^{\mathrm{d}}\right)}\left(1-\left(\left(1-p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)\left(1-\Phi_i^{\mathrm{r}}\left(\frac{h_i^{\mathrm{d}}}{z}\right)\right)+p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\left(1-\Phi_i^{\mathrm{r}}\left(\frac{2h_i^{\mathrm{d}}}{z}\right)\right)\right)\right)dz \\
=&\frac{h_i^{\mathrm{d}}}{r_i^{\max}}+\left(1-p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)\int_{L_i^{\min}\left(h_i^{\mathrm{d}}\right)}^{2L_i^{\max}\left(h_i^{\mathrm{d}}\right)}\Phi_i^{\mathrm{r}}\left(\frac{h_i^{\mathrm{d}}}{z}\right)dz+p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\int_{L_i^{\min}\left(h_i^{\mathrm{d}}\right)}^{2L_i^{\max}\left(h_i^{\mathrm{d}}\right)}\Phi_i^{\mathrm{r}}\left(\frac{2h_i^{\mathrm{d}}}{z}\right)dz \\
=&h_i^{\mathrm{d}}\left[\frac{1}{r_i^{\max}}-\left(1-p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)\int_{\frac{r_i^{\min}}{2}}^{r_i^{\max}}\Phi_i^{\mathrm{r}}(y)\,d\left(\frac{1}{y}\right)-p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\int_{\frac{r_i^{\min}}{2}}^{r_i^{\max}}\Phi_i^{\mathrm{r}}(2y)\,d\left(\frac{1}{y}\right)\right] \\
=&h_i^{\mathrm{d}}\left[\frac{1}{r_i^{\max}}-\right. \\
&\left.\left(1-p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)\left(\frac{1}{r_i^{\max}}-\int_{\frac{r_i^{\min}}{2}}^{r_i^{\max}}\frac{1}{y}d\Phi_i^{\mathrm{r}}(y)\right)-p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\left(\frac{1}{r_i^{\max}}-\int_{\frac{r_i^{\min}}{2}}^{r_i^{\max}}\frac{1}{y}d\Phi_i^{\mathrm{r}}(2y)\right)\right] \\
=&h_i^{\mathrm{d}}\left[\left(1-p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)\int_{\frac{r_i^{\min}}{2}}^{r_i^{\max}}\frac{1}{y}d\Phi_i^{\mathrm{r}}(y)+p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\int_{\frac{r_i^{\min}}{2}}^{r_i^{\max}}\frac{1}{y}d\Phi_i^{\mathrm{r}}(2y)\right] \\
=&h_i^{\mathrm{d}}\left(1+p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)E\left[\frac{1}{r_i}\right], \qquad \text{(A.310)}
\end{aligned}
$$

where we have denoted

$$
E\left[\frac{1}{r_i}\right]=\int_{r_i^{\min}}^{r_i^{\max}}\frac{1}{y}d\Phi_i^{\mathrm{r}}(y). \qquad \text{(A.311)}
$$

and also used continuous differentiability of $\Phi_i^{\mathrm{r}}(y)$ that, in turn, implies that the distribution of the recovery rates does not have a "finite mass" at $y=r_i^{\min}$.

Thus, $\mu_i^B\left(h_i^{\mathrm{d}}\right)$ is monotone increasing in h_i^{d} on $\left[h_i^{\min}, 1\right]$ if and only if

$$\begin{aligned}\frac{d\mu_i^B\left(h_i^{\mathrm{d}}\right)}{dh_i^{\mathrm{d}}} &= E\left[\frac{1}{r_i}\right]\left[\left(1+p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)+h_i^{\mathrm{d}}\left(\frac{dp_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)}{dh_i^{\mathrm{d}}}\right)\right] \geq 0 \\ &\iff \frac{1+p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)}{h_i^{\mathrm{d}}} \geq \left|\frac{dp_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)}{dh_i^{\mathrm{d}}}\right|, \end{aligned} \tag{A.312}$$

since

$$\frac{dp_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)}{dh_i^{\mathrm{d}}} < 0, \tag{A.313}$$

and (A.312) is ensured by $\gamma_i^h \leq 1$.

Next, consider

$$\begin{aligned}\left(\sigma_i^B\left(h_i^{\mathrm{d}}\right)\right)^2 &= \int_{L_i^{\min}\left(h_i^{\mathrm{d}}\right)}^{2L_i^{\max}\left(h_i^{\mathrm{d}}\right)} \Phi_i^L(z|h_i^{\mathrm{d}})\left(1-\Phi_i^L(z|h_i^{\mathrm{d}})\right)dz \\ &= \int_{\frac{h_i^{\mathrm{d}}}{r_i^{\max}}}^{\frac{2h_i^{\mathrm{d}}}{r_i^{\min}}} \left(\left(1-p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)\left(1-\Phi_i^{\mathrm{r}}\left(\frac{h_i^{\mathrm{d}}}{z}\right)\right)+p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\left(1-\Phi_i^{\mathrm{r}}\left(\frac{2h_i^{\mathrm{d}}}{z}\right)\right)\right) \\ &\quad \times\left(\left(1-p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)\Phi_i^{\mathrm{r}}\left(\frac{h_i^{\mathrm{d}}}{z}\right)+p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\Phi_i^{\mathrm{r}}\left(\frac{2h_i^{\mathrm{d}}}{z}\right)\right)dz. \end{aligned} \tag{A.314}$$

Let $y = \frac{h_i^{\mathrm{d}}}{z}$, so

$$dz = d\left(\frac{h_i^{\mathrm{d}}}{y}\right) = -\frac{h_i^{\mathrm{d}}}{y^2}dy. \tag{A.315}$$

Then,

$$\begin{aligned}
&\left(\sigma_i^B\left(h_i^{\mathrm{d}}\right)\right)^2 \\
=&h_i^{\mathrm{d}}\int_{\frac{r_i^{\min}}{2}}^{r_i^{\max}}\frac{1}{y^2}\left(\left(1-p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)\left(1-\Phi_i^{\mathrm{r}}\left(y\right)\right)+p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\left(1-\Phi_i^{\mathrm{r}}\left(2y\right)\right)\right) \\
&\times\left(\left(1-p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)\Phi_i^{\mathrm{r}}\left(y\right)+p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\Phi_i^{\mathrm{r}}\left(2y\right)\right)dy \\
=&h_i^{\mathrm{d}}\left(\left(1-p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)^2K_1+p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\left(1-p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)\left(K_2+K_3\right)+\left(p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)^2K_4\right),
\end{aligned} \tag{A.316}$$

where

$$K_1=\int_{\frac{r_i^{\min}}{2}}^{r_i^{\max}}\frac{\Phi_i^{\mathrm{r}}\left(y\right)\left(1-\Phi_i^{\mathrm{r}}\left(y\right)\right)}{y^2}dy, \tag{A.317}$$

$$K_2=\int_{\frac{r_i^{\min}}{2}}^{r_i^{\max}}\frac{\Phi_i^{\mathrm{r}}\left(2y\right)\left(1-\Phi_i^{\mathrm{r}}\left(y\right)\right)}{y^2}dy, \tag{A.318}$$

$$K_3=\int_{\frac{r_i^{\min}}{2}}^{r_i^{\max}}\frac{\Phi_i^{\mathrm{r}}\left(y\right)\left(1-\Phi_i^{\mathrm{r}}\left(2y\right)\right)}{y^2}dy, \tag{A.319}$$

$$K_4=\int_{\frac{r_i^{\min}}{2}}^{r_i^{\max}}\frac{\Phi_i^{\mathrm{r}}\left(2y\right)\left(1-\Phi_i^{\mathrm{r}}\left(2y\right)\right)}{y^2}dy. \tag{A.320}$$

Using integration by parts, we can rewrite each term as a function of r_i. In particular,

$$\begin{aligned}
K_1&=\int_{\frac{r_i^{\min}}{2}}^{r_i^{\max}}\frac{\Phi_i^{\mathrm{r}}\left(y\right)\left(1-\Phi_i^{\mathrm{r}}\left(y\right)\right)}{y^2}dy \\
&=E\left[\frac{1-2\Phi_i^{\mathrm{r}}\left(r_i\right)}{r_i}\right],
\end{aligned} \tag{A.321}$$

and

$$\begin{aligned}
K_4&=\int_{\frac{r_i^{\min}}{2}}^{r_i^{\max}}\frac{\Phi_i^{\mathrm{r}}\left(2y\right)\left(1-\Phi_i^{\mathrm{r}}\left(2y\right)\right)}{y^2}dy \\
&=2E\left[\frac{1-2\Phi_i^{\mathrm{r}}\left(r_i\right)}{r_i}\right]=2K_1.
\end{aligned} \tag{A.322}$$

Plugging this back into (A.316) yields

$$\begin{aligned}
&\left(\sigma_i^B\left(h_i^{\mathrm{d}}\right)\right)^2 \\
=&h_i^{\mathrm{d}}\left(\left(\left(1-p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)^2+2\left(p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)^2\right)K_1+p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\left(1-p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)(K_2+K_3)\right) \\
=&h_i^{\mathrm{d}}\left(\left(1+\left(p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)^2\right)K_1+p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\left(1-p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)(K_2+K_3-2K_1)\right) \\
=&h_i^{\mathrm{d}}\left(\left(1+p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)K_1+p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\left(1-p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)(K_2+K_3-3K_1)\right). \qquad \text{(A.323)}
\end{aligned}$$

Furthermore, using (A.317)-(A.320),

$$\begin{aligned}
K_2+K_3-3K_1 =&K_2+K_3-K_1-K_4 \\
=&\int_{\frac{r_i^{\min}}{2}}^{r_i^{\max}}\frac{\Phi_i^{\mathrm{r}}(2y)\left(1-\Phi_i^{\mathrm{r}}(y)\right)}{y^2}dy+\int_{\frac{r_i^{\min}}{2}}^{r_i^{\max}}\frac{\Phi_i^{\mathrm{r}}(y)\left(1-\Phi_i^{\mathrm{r}}(2y)\right)}{y^2}dy \\
&-\int_{\frac{r_i^{\min}}{2}}^{r_i^{\max}}\frac{\Phi_i^{\mathrm{r}}(y)\left(1-\Phi_i^{\mathrm{r}}(y)\right)}{y^2}dy-\int_{\frac{r_i^{\min}}{2}}^{r_i^{\max}}\frac{\Phi_i^{\mathrm{r}}(2y)\left(1-\Phi_i^{\mathrm{r}}(2y)\right)}{y^2}dy \\
=&\int_{\frac{r_i^{\min}}{2}}^{r_i^{\max}}\left(\frac{\left(\Phi_i^{\mathrm{r}}(2y)-\Phi_i^{\mathrm{r}}(y)\right)}{y}\right)^2dy\geq 0. \qquad \text{(A.324)}
\end{aligned}$$

Then,

$$\left(\sigma_i^B\left(h_i^{\mathrm{d}}\right)\right)^2=h_i^{\mathrm{d}}\left(1+p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)G_i^r+h_i^{\mathrm{d}}p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\left(1-p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)H_i^r, \qquad \text{(A.325)}$$

where

$$G_i^r=\int_{\frac{r_i^{\min}}{2}}^{r_i^{\max}}\frac{\Phi_i^{\mathrm{r}}(y)\left(1-\Phi_i^{\mathrm{r}}(y)\right)}{y^2}dy\geq 0, \qquad \text{(A.326)}$$

$$H_i^r=\int_{\frac{r_i^{\min}}{2}}^{r_i^{\max}}\left(\frac{\left(\Phi_i^{\mathrm{r}}(2y)-\Phi_i^{\mathrm{r}}(y)\right)}{y}\right)^2dy\geq 0. \qquad \text{(A.327)}$$

Finally,

$$\frac{d\left(\sigma_i^B\left(h_i^{\mathrm{d}}\right)\right)^2}{dh_i^{\mathrm{d}}} = \left(1 + p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right) + h_i^{\mathrm{d}}\left(\frac{dp_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)}{dh_i^{\mathrm{d}}}\right)\right) G_i^r + \left(p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\left(1 - p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right) + h_i^{\mathrm{d}}\frac{dp_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)}{dh_i^{\mathrm{d}}}\left(1 - 2p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)\right) H_i^r, \quad \text{(A.328)}$$

and $\left(\sigma_i^B\left(h_i^{\mathrm{d}}\right)\right)^2$ is monotone increasing if and only if

$$\begin{aligned} &\left(1 + p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right) + h_i^{\mathrm{d}}\left(\frac{dp_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)}{dh_i^{\mathrm{d}}}\right)\right) G_i^r \\ &+ \left(p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\left(1 - p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right) + h_i^{\mathrm{d}}\left(\frac{dp_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)}{dh_i^{\mathrm{d}}}\right)\left(1 - 2p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right)\right) H_i^r \geq 0 \\ \iff &\left(1 + p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right) G_i^r + p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\left(1 - p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right) H_i^r \\ &\geq h_i^{\mathrm{d}}\left(G_i^r + \left(1 - 2p_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)\right) H_i^r\right)\left|\frac{dp_i^{\mathrm{a}}\left(h_i^{\mathrm{d}}\right)}{dh_i^{\mathrm{d}}}\right|. \end{aligned} \quad \text{(A.329)}$$

Notice that the left-hand side of the above inequality is greater than G_i^r and the right-hand side of the inequality is less than $\left(G_i^r + H_i^r\right)\gamma_i^h$. Therefore, the above inequality is ensured by

$$\frac{G_i^r}{G_i^r + H_i^r} \geq \gamma_i^h. \quad \text{(A.330)}$$

□

Proof of Lemma 2

From Proposition 6, the random usage of a hospital resource k can be expressed as $U_k = \mu_k + \sigma_k\mathcal{N}(0,1)$. Then, under Assumption 8, the expected hospital cost incurred from using the resource can be expressed as $c_k E[U_k^2] = c_k(\mu_k^2 + \sigma_k^2)$ since $E[\mathcal{N}(0,1)^2] = 1$, and $E[\mathcal{N}(0,1)] = 0$. Then, for a resource $k = F, B$, the expected hospital cost incurred under

the procedure portfolio $\mathbf{a}$ and discharge thresholds $\mathbf{h}^{\mathrm{d}}$ is given by

$$C_k\left(\mathbf{a},\mathbf{h}^{\mathrm{d}}\right)=c_k\left(\left(M^k\left(\mathbf{a},\mathbf{h}^{\mathrm{d}}\right)\right)^2+\left(\Sigma^k\left(\mathbf{a},\mathbf{h}^{\mathrm{d}}\right)\right)^2\right). \tag{A.331}$$

The approximate hospital expected daily profit for a given portfolio of elective procedures and discharge threshold values can be expressed as

$$\begin{aligned}\Pi_{\mathrm{A}}\left(\mathbf{a}^e,\mathbf{h}^{\mathrm{d}}\right)=&\sum_{i=1}^{N}\left(a_i^e+\mu_i^u\right)R_i\\&-c_F\left(E\left[\left(M^F\left(\mathbf{a}^e+\mathbf{a}^u,\mathbf{h}^{\mathrm{d}}\right)\right)^2\right]+E\left[\left(\Sigma^F\left(\mathbf{a}^e+\mathbf{a}^u,\mathbf{h}^{\mathrm{d}}\right)\right)^2\right]\right)\\&-c_B\left(E\left[\left(M^B\left(\mathbf{a}^e+\mathbf{a}^u,\mathbf{h}^{\mathrm{d}}\right)\right)^2\right]+E\left[\left(\Sigma^B\left(\mathbf{a}^e+\mathbf{a}^u,\mathbf{h}^{\mathrm{d}}\right)\right)^2\right]\right),\end{aligned} \tag{A.332}$$

where the expectations are taken with respect to the distribution of the daily numbers of urgent procedures.

Note that

$$\begin{aligned}E\left[\left(M^F\left(\mathbf{a}^e+\mathbf{a}^u,\mathbf{h}^{\mathrm{d}}\right)\right)^2\right]=&E\left[\left(\sum_{i=1}^{N}\left(a_i^e+a_i^u\right)\mu_i^F\left(h_i^{\mathrm{d}}\right)\right)^2\right]\\=&\left(\sum_{i=1}^{N}a_i^e\mu_i^F\left(h_i^{\mathrm{d}}\right)\right)^2+2\left(\sum_{i=1}^{N}a_i^e\mu_i^F\left(h_i^{\mathrm{d}}\right)\right)\left(\sum_{j=1}^{N}\mu_j^u\mu_j^F\left(h_j^{\mathrm{d}}\right)\right)\\&+\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\rho_{ij}^u\sigma_i^u\sigma_j^u+\mu_i^u\mu_j^u\right)\mu_i^F\left(h_i^{\mathrm{d}}\right)\mu_j^F\left(h_j^{\mathrm{d}}\right)\\=&\left(\sum_{i=1}^{N}a_i^e\mu_i^F\left(h_i^{\mathrm{d}}\right)\right)^2+2\left(\sum_{i=1}^{N}a_i^e\mu_i^F\left(h_i^{\mathrm{d}}\right)\right)\left(\sum_{j=1}^{N}\mu_j^u\mu_j^F\left(h_j^{\mathrm{d}}\right)\right)\\&+\sum_{i=1}^{N}\left((\sigma_i^u)^2+(\mu_i^u)^2\right)\left(\mu_i^F\left(h_i^{\mathrm{d}}\right)\right)^2\\&+\sum_{i=1}^{N}\sum_{j\neq i}\left(\rho_{ij}^u\sigma_i^u\sigma_j^u+\mu_i^u\mu_j^u\right)\mu_i^F\left(h_i^{\mathrm{d}}\right)\mu_j^F\left(h_j^{\mathrm{d}}\right)\end{aligned} \tag{A.333}$$

and

$$E\left[\left(\Sigma^F\left(\mathbf{a}^e+\mathbf{a}^u,\mathbf{h}^{\mathrm{d}}\right)\right)^2\right]=\sum_{i=1}^{N}a_i^e\left(\sigma_i^F\left(h_i^{\mathrm{d}}\right)\right)^2+\sum_{i=1}^{N}\mu_i^u\left(\sigma_i^F\left(h_i^{\mathrm{d}}\right)\right)^2. \tag{A.334}$$

Using the same method,

$$\begin{aligned}E\left[\left(M^B\left(\mathbf{a}^e+\mathbf{a}^u,\mathbf{h}^{\mathrm{d}}\right)\right)^2\right]=&\left(\sum_{i=1}^{N}a_i^e\mu_i^B\left(h_i^{\mathrm{d}}\right)\right)^2+2\left(\sum_{i=1}^{N}a_i^e\mu_i^B\left(h_i^{\mathrm{d}}\right)\right)\left(\sum_{j=1}^{N}\mu_j^u\mu_j^B\left(h_j^{\mathrm{d}}\right)\right)\\&+\sum_{i=1}^{N}\left((\sigma_i^u)^2+(\mu_i^u)^2\right)\left(\mu_i^B\left(h_i^{\mathrm{d}}\right)\right)^2\\&+\sum_{i=1}^{N}\sum_{j\neq i}\left(\rho_{ij}^u\sigma_i^u\sigma_j^u+\mu_i^u\mu_j^u\right)\mu_i^B\left(h_i^{\mathrm{d}}\right)\mu_j^B\left(h_j^{\mathrm{d}}\right),\end{aligned} \tag{A.335}$$

and

$$E\left[\left(\Sigma^B\right)^2\left(\mathbf{a}^e+\mathbf{a}^u,\mathbf{h}^{\mathrm{d}}\right)\right]=\sum_{i=1}^{N}a_i^e\left(\sigma_i^B\left(h_i^{\mathrm{d}}\right)\right)^2+\sum_{i=1}^{N}\mu_i^u\left(\sigma_i^B\left(h_i^{\mathrm{d}}\right)\right)^2. \tag{A.336}$$

Substituting these expressions into (A.332) gives the following:

$$
\begin{aligned}
&\Pi_{\mathrm{A}}\left(\mathbf{a}^{e},\mathbf{h}^{\mathrm{d}}\right)\\
=&\sum_{i=1}^{N}\left(a_{i}^{e}+\mu_{i}^{u}\right)R_{i}\\
&-c_{F}\left(\left(\sum_{i=1}^{N}a_{i}^{e}\mu_{i}^{F}\left(h_{i}^{\mathrm{d}}\right)\right)^{2}+2\left(\sum_{i=1}^{N}a_{i}^{e}\mu_{i}^{F}\left(h_{i}^{\mathrm{d}}\right)\right)\left(\sum_{j=1}^{N}\mu_{j}^{u}\mu_{j}^{F}\left(h_{j}^{\mathrm{d}}\right)\right)\right.\\
&+\sum_{i=1}^{N}\left(\left(\sigma_{i}^{u}\right)^{2}+\left(\mu_{i}^{u}\right)^{2}\right)\left(\mu_{i}^{F}\left(h_{i}^{\mathrm{d}}\right)\right)^{2}+\sum_{i=1}^{N}\sum_{j\neq i}\left(\rho_{ij}^{u}\sigma_{i}^{u}\sigma_{j}^{u}+\mu_{i}^{u}\mu_{j}^{u}\right)\mu_{i}^{F}\left(h_{i}^{\mathrm{d}}\right)\mu_{j}^{F}\left(h_{j}^{\mathrm{d}}\right)\\
&\left.+\sum_{i=1}^{N}a_{i}^{e}\left(\sigma_{i}^{F}\left(h_{i}^{\mathrm{d}}\right)\right)^{2}+\sum_{i=1}^{N}\mu_{i}^{u}\left(\sigma_{i}^{F}\left(h_{i}^{\mathrm{d}}\right)\right)^{2}\right)\\
&-c_{B}\left(\left(\sum_{i=1}^{N}a_{i}^{e}\mu_{i}^{B}\left(h_{i}^{\mathrm{d}}\right)\right)^{2}+2\left(\sum_{i=1}^{N}a_{i}^{e}\mu_{i}^{B}\left(h_{i}^{\mathrm{d}}\right)\right)\left(\sum_{j=1}^{N}\mu_{j}^{u}\mu_{j}^{B}\left(h_{j}^{\mathrm{d}}\right)\right)\right.\\
&+\sum_{i=1}^{N}\left(\left(\sigma_{i}^{u}\right)^{2}+\left(\mu_{i}^{u}\right)^{2}\right)\left(\mu_{i}^{B}\left(h_{i}^{\mathrm{d}}\right)\right)^{2}+\sum_{i=1}^{N}\sum_{j\neq i}\left(\rho_{ij}^{u}\sigma_{i}^{u}\sigma_{j}^{u}+\mu_{i}^{u}\mu_{j}^{u}\right)\mu_{i}^{B}\left(h_{i}^{\mathrm{d}}\right)\mu_{j}^{B}\left(h_{j}^{\mathrm{d}}\right)\\
&\left.+\sum_{i=1}^{N}a_{i}^{e}\left(\sigma_{i}^{B}\left(h_{i}^{\mathrm{d}}\right)\right)^{2}+\sum_{i=1}^{N}\mu_{i}^{u}\left(\sigma_{i}^{B}\left(h_{i}^{\mathrm{d}}\right)\right)^{2}\right)\\
=&\sum_{i=1}^{N}a_{i}^{e}\mathcal{A}_{i}\left(\mathbf{h}^{\mathrm{d}}\right)-c_{F}\left(\sum_{i=1}^{N}a_{i}^{e}\mu_{i}^{F}\left(h_{i}^{\mathrm{d}}\right)\right)^{2}-c_{B}\left(\sum_{i=1}^{N}a_{i}^{e}\mu_{i}^{B}\left(h_{i}^{\mathrm{d}}\right)\right)^{2}+\sum_{i=1}^{N}\mathcal{B}_{i}\left(\mathbf{h}^{\mathrm{d}}\right),
\end{aligned}
\tag{A.337}
$$

where

$$\begin{aligned}\mathcal{A}_i\left(\mathbf{h}^{\mathrm{d}}\right) =& R_i - c_F\left(2\mu_i^F\left(h_i^{\mathrm{d}}\right)\sum_{j=1}^{N}\mu_j^u\mu_j^F\left(h_j^{\mathrm{d}}\right)+\left(\sigma_i^F\left(h_i^{\mathrm{d}}\right)\right)^2\right)\\ &- c_B\left(2\mu_i^B\left(h_i^{\mathrm{d}}\right)\sum_{j=1}^{N}\mu_j^u\mu_j^B\left(h_j^{\mathrm{d}}\right)+\left(\sigma_i^B\left(h_i^{\mathrm{d}}\right)\right)^2\right), \qquad \text{(A.338)}\\ \mathcal{B}_i\left(\mathbf{h}^{\mathrm{d}}\right) =& \mu_i^u R_i\\ &- c_F\left(\left((\sigma_i^u)^2+(\mu_i^u)^2\right)\left(\mu_i^F\left(h_i^{\mathrm{d}}\right)\right)^2+\mu_i^u\left(\sigma_i^F\left(h_i^{\mathrm{d}}\right)\right)^2\right.\\ &\left.+\sum_{j\neq i}\left(\rho_{ij}^u\sigma_i^u\sigma_j^u+\mu_i^u\mu_j^u\right)\mu_i^F\left(h_i^{\mathrm{d}}\right)\mu_j^F\left(h_j^{\mathrm{d}}\right)\right)\\ &- c_B\left(\left((\sigma_i^u)^2+(\mu_i^u)^2\right)\left(\mu_i^B\left(h_i^{\mathrm{d}}\right)\right)^2+\mu_i^u\left(\sigma_i^B\left(h_i^{\mathrm{d}}\right)\right)^2\right.\\ &\left.+\sum_{j\neq i}\left(\rho_{ij}^u\sigma_i^u\sigma_j^u+\mu_i^u\mu_j^u\right)\mu_i^B\left(h_i^{\mathrm{d}}\right)\mu_j^B\left(h_j^{\mathrm{d}}\right)\right). \qquad \text{(A.339)}\end{aligned}$$

□

Proof of Proposition 8

For $n = 1$, (2.44)-(2.46) becomes

$$\max_{a^e,h^{\mathrm{d}}}\left(a^e\mathcal{A}\left(h^{\mathrm{d}}\right)-(a^e)^2\left(c_F\left(\mu^F\left(h^{\mathrm{d}}\right)\right)^2+c_B\left(\mu^B\left(h^{\mathrm{d}}\right)\right)^2\right)+\mathcal{B}\left(h^{\mathrm{d}}\right)\right) \qquad \text{(A.340)}$$

$$\text{s.t. } 0\leq a^e\leq E, \qquad \text{(A.341)}$$

$$h^{\min}\leq h^{\mathrm{d}}\leq 1, \qquad \text{(A.342)}$$

where we have dropped the index of the procedure, and

$$\begin{aligned}\mathcal{A}\left(h^{\mathrm{d}}\right) =& R - c_F\left(2\mu^u\left(\mu^F\left(h^{\mathrm{d}}\right)\right)^2 + \left(\sigma^F\left(h^{\mathrm{d}}\right)\right)^2\right) \\ & - c_B\left(2\mu^u\left(\mu^B\left(h^{\mathrm{d}}\right)\right)^2 + \left(\sigma^B\left(h^{\mathrm{d}}\right)\right)^2\right), \qquad \text{(A.343)} \\ \mathcal{B}\left(h^{\mathrm{d}}\right) =& \mu^u R - c_F\left(\left((\sigma^u)^2 + (\mu^u)^2\right)\left(\mu^F\left(h^{\mathrm{d}}\right)\right)^2 + \mu^u\left(\sigma^F\left(h^{\mathrm{d}}\right)\right)^2\right) \\ & - c_B\left(\left((\sigma^u)^2 + (\mu^u)^2\right)\left(\mu^B\left(h^{\mathrm{d}}\right)\right)^2 + \mu^u\left(\sigma^B\left(h^{\mathrm{d}}\right)\right)^2\right). \qquad \text{(A.344)}\end{aligned}$$

Then, the objective function (A.340) is

$$\begin{aligned}& a^e\mathcal{A}\left(h^{\mathrm{d}}\right) - (a^e)^2\left(c_F\left(\mu^F\left(h^{\mathrm{d}}\right)\right)^2 + c_B\left(\mu^B\left(h^{\mathrm{d}}\right)\right)^2\right) + \mathcal{B}\left(h^{\mathrm{d}}\right) \\ & = (a^e + \mu^u)\left(R - \mathcal{V}\left(h^{\mathrm{d}}\right)\right) - \left((a^e + \mu^u)^2 + (\sigma^u)^2\right)\mathcal{M}\left(h^{\mathrm{d}}\right), \qquad \text{(A.345)}\end{aligned}$$

where

$$\mathcal{M}\left(h^{\mathrm{d}}\right) = c_F\left(\mu^F\left(h^{\mathrm{d}}\right)\right)^2 + c_B\left(\mu^B\left(h^{\mathrm{d}}\right)\right)^2 \qquad \text{(A.346)}$$

$$\mathcal{V}\left(h^{\mathrm{d}}\right) = c_F\left(\sigma^F\left(h^{\mathrm{d}}\right)\right)^2 + c_B\left(\sigma^B\left(h^{\mathrm{d}}\right)\right)^2 \qquad \text{(A.347)}$$

(A.345) is a concave, quadratic function of a^e, and the stationary point is at

$$a^e = \frac{R - \mathcal{V}\left(h^{\mathrm{d}}\right)}{2\mathcal{M}\left(h^{\mathrm{d}}\right)} - \mu^u. \qquad \text{(A.348)}$$

Therefore, for a given discharge threshold h^{d}, the optimal value of a^e is its stationary point, if it is between 0 and E; otherwise, the optimal value is 0, if a^e is negative, or E, if it is positive. The optimal value is given by $f^e\left(h^{\mathrm{d}}\right)$ in (2.49).

Finally, substituting $f^e\left(h^{\mathrm{d}}\right)$ into (A.345) gives

$$\left(f^e\left(h^{\mathrm{d}}\right) + \mu^u\right)\left(R - \mathcal{V}\left(h^{\mathrm{d}}\right)\right) - \left(\left(f^e\left(h^{\mathrm{d}}\right) + \mu^u\right)^2 + (\sigma^u)^2\right)\mathcal{M}\left(h^{\mathrm{d}}\right). \qquad \text{(A.349)}$$

The optimal discharge threshold is given by (2.50). □

Proof of Proposition 9

In order to establish the results of the Proposition, we first prove the following lemma.

Lemma 11 *In the setting with a single procedure, let* $\mathrm{E}[S]$ *and* $\mathrm{Var}[S]$ *be the expectation and the variance of the surgical procedure durations, and* $\mathrm{E}\left[\frac{1}{r}\right]$ *be the expectation of the inverse of the recovery rate. Then,*

$$\mathrm{E}[S] \leq \mu^F\left(h^{\mathrm{d}}\right) \leq 2\mathrm{E}[S], \tag{A.350}$$

$$\mathrm{Var}[S] \leq \left(\sigma^F\left(h^{\mathrm{d}}\right)\right)^2 \leq \mathrm{Var}[S] + \left(\frac{3\mathrm{Var}[S] + (\mathrm{E}[S])^2}{2\mathrm{E}[S]}\right)^2, \tag{A.351}$$

$$h^{\mathrm{d}}\left(\mathrm{E}\left[\frac{1}{r}\right]\right) \leq \mu^B\left(h^{\mathrm{d}}\right) \leq 2h^{\mathrm{d}}\left(\mathrm{E}\left[\frac{1}{r}\right]\right), \tag{A.352}$$

$$h^{\mathrm{d}}G^r \leq \left(\sigma^B\left(h^{\mathrm{d}}\right)\right)^2 \leq h^{\mathrm{d}}\left(G^r + \frac{(G^r + H^r)^2}{4H^r}\right), \tag{A.353}$$

where G^r *and* H^r *are defined in (2.35).*

Proof of Lemma 11. From (2.29),

$$\mu^F\left(h^{\mathrm{d}}\right) = \mathrm{E}[S]\left(1 + p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\right), \tag{A.354}$$

and (A.350) follows since $p^{\mathrm{a}}\left(h^{\mathrm{d}}\right) \in [0,1]$.

Next, from (2.30),

$$\begin{aligned}\left(\sigma^F\left(h^{\mathrm{d}}\right)\right)^2 &= \left(1 + 3p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\right)\mathrm{Var}[S] + p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\left(1 - p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\right)(\mathrm{E}[S])^2 \\ &= \mathrm{Var}[S] + \left(3\mathrm{Var}[S] + (\mathrm{E}[S])^2\right)p^{\mathrm{a}}\left(h^{\mathrm{d}}\right) - (\mathrm{E}[S])^2\left(p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\right)^2, \end{aligned} \tag{A.355}$$

which is a concave function of $p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)$. Then, (A.355) is minimized at either $p^{\mathrm{a}}\left(h^{\mathrm{d}}\right) = 0$ or $p^{\mathrm{a}}\left(h^{\mathrm{d}}\right) = 1$. Comparing the two values, the minimum is $\mathrm{Var}[S]$. The stationary point of

the function is at

$$p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)=\frac{3\mathrm{Var}\left[S\right]+\left(\mathrm{E}\left[S\right]\right)^{2}}{2\left(\mathrm{E}\left[S\right]\right)^{2}}, \tag{A.356}$$

so the maximum of (A.355) is

$$\mathrm{Var}\left[S\right]+\left(\frac{3\mathrm{Var}\left[S\right]+\left(\mathrm{E}\left[S\right]\right)^{2}}{2\mathrm{E}\left[S\right]}\right)^{2}. \tag{A.357}$$

(A.351) follows. Furthermore, from (2.32),

$$\mu^{B}\left(h^{\mathrm{d}}\right)=h^{\mathrm{d}}\left(1+p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\right)\mathrm{E}\left[\frac{1}{r}\right]. \tag{A.358}$$

(A.352) follows, since $p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\in[0,1]$.

Finally, from (2.34),

$$\begin{aligned}\frac{\left(\sigma^{B}\left(h^{\mathrm{d}}\right)\right)^{2}}{h^{\mathrm{d}}}&=\left(1+p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\right)G^{r}+p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\left(1-p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\right)H^{r}\\&=G^{r}+p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\left(G^{r}+H^{r}\right)-\left(p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\right)^{2}H^{r},\end{aligned} \tag{A.359}$$

where

$$G^{r}=\int_{\frac{r^{\min}}{2}}^{r^{\max}}\frac{\Phi^{\mathrm{r}}\left(y\right)\left(1-\Phi^{\mathrm{r}}\left(y\right)\right)}{y^{2}}dy, \tag{A.360}$$

$$H^{r}=\int_{\frac{r^{\min}}{2}}^{r^{\max}}\left(\frac{\left(\Phi^{\mathrm{r}}\left(2y\right)-\Phi^{\mathrm{r}}\left(y\right)\right)}{y}\right)^{2}dy, \tag{A.361}$$

and both terms are clearly positive.

(A.359) is a quadratic, concave function of $p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)$. The stationary point is at

$$p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)=\frac{G^{r}+H^{r}}{2H^{r}}, \tag{A.362}$$

at which

$$\frac{\left(\sigma^B\left(h^{\mathrm{d}}\right)\right)^2}{h^{\mathrm{d}}} = G^r + \frac{(G^r + H^r)^2}{4H^r}. \tag{A.363}$$

Furthermore, when $p^{\mathrm{a}}\left(h^{\mathrm{d}}\right) = 0$,

$$\frac{\left(\sigma^B\left(h^{\mathrm{d}}\right)\right)^2}{h^{\mathrm{d}}} = G^r, \tag{A.364}$$

and when $p^{\mathrm{a}}\left(h^{\mathrm{d}}\right) = 1$,

$$\frac{\left(\sigma^B\left(h^{\mathrm{d}}\right)\right)^2}{h^{\mathrm{d}}} = 2G^r, \tag{A.365}$$

Note that

$$G^r + \frac{(G^r + H^r)^2}{4H^r} \geq 2G^r \iff (G^r - H^r)^2 \geq 0. \tag{A.366}$$

Then, the minimum value of (A.359) is G^r, and the maximum value is $G^r + \frac{(G^r+H^r)^2}{4H^r}$. (A.353) follows. $\square$

Now we prove the results of the Proposition. In general, note that when $\hat{a}^e = x$, from (2.50),

$$\begin{aligned}\hat{h}^{\mathrm{d}} &= \underset{h^{\mathrm{d}} \in [h^{\min},1]}{\arg\max} \left((x+\mu^u)\left(R - \mathcal{V}\left(h^{\mathrm{d}}\right)\right) - \left((x+\mu^u)^2 + (\sigma^u)^2\right)\mathcal{M}\left(h^{\mathrm{d}}\right)\right) \\ &= \underset{h^{\mathrm{d}} \in [h^{\min},1]}{\arg\min} \left((x+\mu^u)\,\mathcal{V}\left(h^{\mathrm{d}}\right) + \left((x+\mu^u)^2 + (\sigma^u)^2\right)\mathcal{M}\left(h^{\mathrm{d}}\right)\right). \end{aligned} \tag{A.367}$$

Let $Z\left(h^{\mathrm{d}}\right)$ represent the expression being minimized. Using the results from Proposition 7,

$$\begin{aligned}
Z\left(h^{\mathrm{d}}\right) =& \left(x+\mu^{u}\right)\mathcal{V}\left(h^{\mathrm{d}}\right)+\left(\left(x+\mu^{u}\right)^{2}+\left(\sigma^{u}\right)^{2}\right)\mathcal{M}\left(h^{\mathrm{d}}\right) \\
=& \left(x+\mu^{u}\right)\left(c_{F}\left(\sigma^{F}\left(h^{\mathrm{d}}\right)\right)^{2}+c_{B}\left(\sigma^{B}\left(h^{\mathrm{d}}\right)\right)^{2}\right) \\
&+\left(\left(x+\mu^{u}\right)^{2}+\left(\sigma^{u}\right)^{2}\right)\left(c_{F}\left(\mu^{F}\left(h^{\mathrm{d}}\right)\right)^{2}+c_{B}\left(\mu^{B}\left(h^{\mathrm{d}}\right)\right)^{2}\right) \\
=& c_{F}\left(x+\mu^{u}\right)\left(\left(1+3p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\right)\mathrm{Var}\left[S\right]+p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\left(1-p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\right)\left(\mathrm{E}\left[S\right]\right)^{2}\right) \\
&+c_{B}\left(x+\mu^{u}\right)\left(h^{\mathrm{d}}\left(\left(1+p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\right)G^{r}+p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\left(1-p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\right)H^{r}\right)\right) \\
&+c_{F}\left(\left(x+\mu^{u}\right)^{2}+\left(\sigma^{u}\right)^{2}\right)\left(\mathrm{E}\left[S\right]\left(1+p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\right)\right)^{2} \\
&+c_{B}\left(\left(x+\mu^{u}\right)^{2}+\left(\sigma^{u}\right)^{2}\right)\left(h^{\mathrm{d}}\left(1+p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\right)\mathrm{E}\left[\frac{1}{r}\right]\right)^{2}. \qquad \text{(A.368)}
\end{aligned}$$

Then,

$$\begin{aligned}
\frac{dZ}{dh^{\mathrm{d}}} =& c_F\left(x+\mu^u\right)\left(\left(3\frac{dp^{\mathrm{a}}}{dh^{\mathrm{d}}}\right)\mathrm{Var}\left[S\right]+\frac{dp^{\mathrm{a}}}{dh^{\mathrm{d}}}\left(1-2p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\right)\left(\mathrm{E}\left[S\right]\right)^2\right)\\
&+c_B\left(x+\mu^u\right)\left(\left(\left(1+p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\right)+h^{\mathrm{d}}\left(\frac{dp^{\mathrm{a}}}{dh^{\mathrm{d}}}\right)\right)G^r\right.\\
&\left.+\left(p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\left(1-p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\right)+h^{\mathrm{d}}\frac{dp^{\mathrm{a}}}{dh^{\mathrm{d}}}\left(1-2p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\right)\right)H^r\right)\\
&+c_F\left(\left(x+\mu^u\right)^2+\left(\sigma^u\right)^2\right)\left(\mathrm{E}\left[S\right]\right)^2 2\left(1+p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\right)\frac{dp^{\mathrm{a}}}{dh^{\mathrm{d}}}\\
&+c_B\left(\left(x+\mu^u\right)^2+\left(\sigma^u\right)^2\right)\left(\mathrm{E}\left[\frac{1}{r}\right]\right)^2 2\left(h^{\mathrm{d}}\left(1+p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\right)\right)\\
&\left(\left(1+p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\right)+h^{\mathrm{d}}\frac{dp^{\mathrm{a}}}{dh^{\mathrm{d}}}\right)\\
=&-c_F\left|\frac{dp^{\mathrm{a}}}{dh^{\mathrm{d}}}\right|\left(3\left(x+\mu^u\right)\mathrm{Var}\left[S\right]+\left(\mathrm{E}\left[S\right]\right)^2\left(\left(x+\mu^u\right)\left(1-2p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\right)\right.\right.\\
&\left.\left.+2\left(\left(x+\mu^u\right)^2+\left(\sigma^u\right)^2\right)\left(1+p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\right)\right)\right)\\
&+c_B\left(\left(x+\mu^u\right)\left(\left(1+p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)-h^{\mathrm{d}}\left|\frac{dp^{\mathrm{a}}}{dh^{\mathrm{d}}}\right|\right)G^r\right.\right.\\
&\left.\left.+\left(p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\left(1-p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\right)-h^{\mathrm{d}}\left|\frac{dp^{\mathrm{a}}}{dh^{\mathrm{d}}}\right|\left(1-2p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\right)\right)H^r\right)\right.\\
&\left.+2\left(\left(x+\mu^u\right)^2+\left(\sigma^u\right)^2\right)\left(\mathrm{E}\left[\frac{1}{r}\right]\right)^2\right.\\
&\left.\times\left(h^{\mathrm{d}}\left(1+p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\right)^2-\left(h^{\mathrm{d}}\right)^2\left(1+p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\right)\left|\frac{dp^{\mathrm{a}}}{dh^{\mathrm{d}}}\right|\right)\right), \qquad \text{(A.369)}
\end{aligned}$$

since $\frac{dp^{\mathrm{a}}}{dh^{\mathrm{d}}} \leq 0$. Finally, for any x, if $\frac{dZ\left(h^{\mathrm{d}}\right)}{dh^{\mathrm{d}}} \leq 0$, then $\hat{h}^{\mathrm{d}} = 1$. On the other hand, if $\frac{dZ\left(h^{\mathrm{d}}\right)}{dh^{\mathrm{d}}} \geq 0$, then $\hat{h}^{\mathrm{d}} = h^{\min}$. We determine sufficient conditions to ensure each case.

To begin, consider when $\frac{dZ\left(h^{\mathrm{d}}\right)}{dh^{\mathrm{d}}} \leq 0$. First, note that

$$\begin{aligned}
&3\left(x+\mu^u\right)\mathrm{Var}\left[S\right]\\
&+\left(\mathrm{E}\left[S\right]\right)^2\left(\left(x+\mu^u\right)\left(1-2p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\right)+2\left(\left(x+\mu^u\right)^2+\left(\sigma^u\right)^2\right)\left(1+p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\right)\right)\\
&\geq 3\min\left(x+\mu^u,\left(x+\mu^u\right)^2+\left(\sigma^u\right)^2\right)\left(\mathrm{Var}\left[S\right]+\left(\mathrm{E}\left[S\right]\right)^2\right)\geq 0. \qquad \text{(A.370)}
\end{aligned}$$

Furthermore, if $p^{\mathrm{a}}\left(h^{\mathrm{d}}\right) \leq 0.5$,

$$p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\left(1-p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\right)-h^{\mathrm{d}}\left|\frac{dp^{\mathrm{a}}}{dh^{\mathrm{d}}}\right|\left(1-2p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\right) \leq p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\left(1-p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\right) \leq 0.25, \tag{A.371}$$

and if $p^{\mathrm{a}}\left(h^{\mathrm{d}}\right) > 0.5$,

$$\begin{aligned} p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\left(1-p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\right)-h^{\mathrm{d}}\left|\frac{dp^{\mathrm{a}}}{dh^{\mathrm{d}}}\right|\left(1-2p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\right) &\leq p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\left(1-p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\right)+\gamma^{h} \\ &\leq 0.25+\gamma^{h}, \end{aligned} \tag{A.372}$$

since

$$\max_{x\in[0,1]} x\left(1-x\right) = 0.25. \tag{A.373}$$

Thus,

$$\begin{aligned} &\left(1+p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)-h^{\mathrm{d}}\left|\frac{dp^{\mathrm{a}}}{dh^{\mathrm{d}}}\right|\right)G^{r}+\left(p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\left(1-p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\right)-h^{\mathrm{d}}\left|\frac{dp^{\mathrm{a}}}{dh^{\mathrm{d}}}\right|\left(1-2p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\right)\right)H^{r} \\ &\leq 2G^{r}+\left(0.25+\gamma^{h}\right)H^{r}. \end{aligned} \tag{A.374}$$

Finally,

$$h^{\mathrm{d}}\left(1+p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\right)^{2}-\left(h^{\mathrm{d}}\right)^{2}\left(1+p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\right)\left|\frac{dp^{\mathrm{a}}}{dh^{\mathrm{d}}}\right| \leq 4. \tag{A.375}$$

Combining this,

$$\begin{aligned} \frac{dZ}{dh^{\mathrm{d}}} &\leq -c_{F}\gamma^{l}\left(3\min\left(x+\mu^{u},(x+\mu^{u})^{2}+(\sigma^{u})^{2}\right)\left(\mathrm{Var}\left[S\right]+\left(\mathrm{E}\left[S\right]\right)^{2}\right)\right) \\ &\quad + c_{B}\left(\left(x+\mu^{u}\right)\left(2G^{r}+\left(0.25+\gamma^{h}\right)H^{r}\right)+8\left((x+\mu^{u})^{2}+(\sigma^{u})^{2}\right)\left(\mathrm{E}\left[\frac{1}{r}\right]\right)^{2}\right) \leq 0. \end{aligned} \tag{A.376}$$

Therefore, $\hat{h}^{\mathrm{d}} = 1$ if

$$c_B \leq c_F \gamma^l \left(\frac{3 \min \left(x + \mu^u, (x+\mu^u)^2 + (\sigma^u)^2 \right) \left(\mathrm{Var}\left[S\right] + \left(\mathrm{E}\left[S\right]\right)^2 \right)}{\left(x+\mu^u\right)\left(2G^r + \left(0.25 + \gamma^h\right) H^r\right) + 8 \left(\left(x+\mu^u\right)^2 + \left(\sigma^u\right)^2 \right) \left(\mathrm{E}\left[\frac{1}{r}\right] \right)^2} \right). \tag{A.377}$$

Next, we determine sufficient conditions to ensure $\frac{dZ\left(h^{\mathrm{d}}\right)}{dh^{\mathrm{d}}} \geq 0$. First, note that

$$\begin{aligned}
&3\left(x+\mu^u\right)\mathrm{Var}\left[S\right] \\
&+ \left(\mathrm{E}\left[S\right]\right)^2 \left(\left(x+\mu^u\right)\left(1 - 2p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\right) + 2\left(\left(x+\mu^u\right)^2 + \left(\sigma^u\right)^2 \right)\left(1 + p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\right)\right) \\
&\leq 3\left(x+\mu^u\right)\mathrm{Var}\left[S\right] + \left(\mathrm{E}\left[S\right]\right)^2 \left(\left(x+\mu^u\right) + 4\left(\left(x+\mu^u\right)^2 + \left(\sigma^u\right)^2 \right)\right).
\end{aligned} \tag{A.378}$$

Furthermore,

$$\begin{aligned}
&\left(1 + p^{\mathrm{a}}\left(h^{\mathrm{d}}\right) - h^{\mathrm{d}} \left| \frac{dp^{\mathrm{a}}}{dh^{\mathrm{d}}} \right| \right) G^r + \left(p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\left(1 - p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\right) - h^{\mathrm{d}} \left| \frac{dp^{\mathrm{a}}}{dh^{\mathrm{d}}} \right| \left(1 - 2p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\right)\right) H^r \\
&\geq \left(1 - \gamma^h\right) G^r - \gamma^h H^r,
\end{aligned} \tag{A.379}$$

and

$$\begin{aligned}
&2\left(\left(x+\mu^u\right)^2 + \left(\sigma^u\right)^2 \right)\left(\mathrm{E}\left[\frac{1}{r}\right] \right)^2 \left(h^{\mathrm{d}}\left(1 + p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\right)^2 - \left(h^{\mathrm{d}}\right)^2 \left(1 + p^{\mathrm{a}}\left(h^{\mathrm{d}}\right)\right) \left| \frac{dp^{\mathrm{a}}}{dh^{\mathrm{d}}} \right| \right) \\
&\geq -4\left(\left(x+\mu^u\right)^2 + \left(\sigma^u\right)^2 \right)\left(\mathrm{E}\left[\frac{1}{r}\right] \right)^2 \gamma^h.
\end{aligned} \tag{A.380}$$

Thus,

$$\begin{aligned}
\frac{dZ}{dh^{\mathrm{d}}} \geq & - c_F \gamma^h \left(3\left(x+\mu^u\right)\mathrm{Var}\left[S\right] + \left(\mathrm{E}\left[S\right]\right)^2 \left(\left(x+\mu^u\right) + 4\left(\left(x+\mu^u\right)^2 + \left(\sigma^u\right)^2 \right)\right)\right) \\
& + c_B \left(\left(x+\mu^u\right)\left(\left(1-\gamma^h\right) G^r - \gamma^h H^r \right) - 4\left(\left(x+\mu^u\right)^2 + \left(\sigma^u\right)^2 \right)\left(\mathrm{E}\left[\frac{1}{r}\right] \right)^2 \gamma^h \right) \\
\geq & 0,
\end{aligned} \tag{A.381}$$

and therefore $\hat{h}^{\mathrm{d}} = h^{\min}$, if

$$G^r \geq \frac{\gamma^h}{1-\gamma^h}\left(4\left(\frac{(x+\mu^u)^2+(\sigma^u)^2}{(x+\mu^u)}\right)\left(\mathrm{E}\left[\frac{1}{r}\right]\right)^2 + H^r\right) \tag{A.382}$$

and

$$c_B \geq c_F\left(\frac{3\mathrm{Var}\left[S\right] + \left(\mathrm{E}\left[S\right]\right)^2\left(1+4\left(\frac{(x+\mu^u)^2+(\sigma^u)^2}{(x+\mu^u)}\right)\right)}{\left(\frac{1-\gamma^h}{\gamma^h}\right)G^r - H^r - 4\left(\frac{(x+\mu^u)^2+(\sigma^u)^2}{(x+\mu^u)}\right)\left(\mathrm{E}\left[\frac{1}{r}\right]\right)^2}\right). \tag{A.383}$$

Using this result, we consider two extreme cases: $\hat{a}^e = 0$ and $\hat{a}^e = E$.

a) From (2.49), $\hat{a}^e = 0$ if

$$\min\left(\frac{R-\mathcal{V}\left(h^{\mathrm{d}}\right)}{2\mathcal{M}\left(h^{\mathrm{d}}\right)} - \mu^u, E\right) \leq 0. \tag{A.384}$$

This holds if

$$\begin{aligned}\frac{R-\mathcal{V}\left(h^{\mathrm{d}}\right)}{2\mathcal{M}\left(h^{\mathrm{d}}\right)} - \mu^u \leq 0 \iff R \leq & 2\mu^u\mathcal{M}\left(h^{\mathrm{d}}\right) + \mathcal{V}\left(h^{\mathrm{d}}\right) \\ = & 2\mu^u\left(c_F\left(\mu^F\left(h^{\mathrm{d}}\right)\right)^2 + c_B\left(\mu^B\left(h^{\mathrm{d}}\right)\right)^2\right) \\ & + c_F\left(\sigma^F\left(h^{\mathrm{d}}\right)\right)^2 + c_B\left(\sigma^B\left(h^{\mathrm{d}}\right)\right)^2.\end{aligned} \tag{A.385}$$

Applying Lemma 11,

$$\begin{aligned}& 2\mu^u\left(c_F\left(\mu^F\left(h^{\mathrm{d}}\right)\right)^2 + c_B\left(\mu^B\left(h^{\mathrm{d}}\right)\right)^2\right) + \left(c_F\left(\sigma^F\left(h^{\mathrm{d}}\right)\right)^2 + c_B\left(\sigma^B\left(h^{\mathrm{d}}\right)\right)^2\right) \\ & \geq c_F\left(2\mu^u\left(\mathrm{E}\left[S\right]\right)^2 + \mathrm{Var}\left[S\right]\right) + c_B\left(2\mu^u\left(h^{\mathrm{d}}\left(\mathrm{E}\left[\frac{1}{r}\right]\right)\right)^2 + h^{\mathrm{d}}G^r\right) \\ & \geq c_F\left(2\mu^u\left(\mathrm{E}\left[S\right]\right)^2 + \mathrm{Var}\left[S\right]\right).\end{aligned} \tag{A.386}$$

Then,

$$R \leq c_F \left(2\mu^u \left(\mathrm{E}\left[S\right]\right)^2 + \mathrm{Var}\left[S\right]\right) \tag{A.387}$$

ensures that $\hat{a}^e = 0$.

In this case, from (A.377), $\hat{h}^{\mathrm{d}} = 1$, if

$$c_B \leq c_F \gamma^l \left(\frac{3\min\left(\mu^u, \left(\mu^u\right)^2 + \left(\sigma^u\right)^2\right)\left(\mathrm{Var}\left[S\right] + \left(\mathrm{E}\left[S\right]\right)^2\right)}{\mu^u\left(2G^r + \left(0.25 + \gamma^h\right)H^r\right) + 8\left(\left(\mu^u\right)^2 + \left(\sigma^u\right)^2\right)\left(\mathrm{E}\left[\frac{1}{r}\right]\right)^2}\right). \tag{A.388}$$

Additionally, from (A.382) and (A.383), $\hat{h}^{\mathrm{d}} = h^{\min}$, if

$$G^r \geq \frac{\gamma^h}{1-\gamma^h}\left(4\left(\frac{\left(\mu^u\right)^2 + \left(\sigma^u\right)^2}{\left(\mu^u\right)}\right)\left(\mathrm{E}\left[\frac{1}{r}\right]\right)^2 + H^r\right) \tag{A.389}$$

and

$$c_B \geq c_F \left(\frac{3\mathrm{Var}\left[S\right] + \left(\mathrm{E}\left[S\right]\right)^2\left(1 + 4\left(\frac{\left(\mu^u\right)^2 + \left(\sigma^u\right)^2}{\left(\mu^u\right)}\right)\right)}{\left(\frac{1-\gamma^h}{\gamma^h}\right)G^r - H^r - 4\left(\frac{\left(\mu^u\right)^2 + \left(\sigma^u\right)^2}{\left(\mu^u\right)}\right)\left(\mathrm{E}\left[\frac{1}{r}\right]\right)^2}\right) \tag{A.390}$$

b) From (2.49), $\hat{a}^e = E$ if

$$\min\left(\frac{R - \mathcal{V}\left(h^{\mathrm{d}}\right)}{2\mathcal{M}\left(h^{\mathrm{d}}\right)} - \mu^u, E\right) = E. \tag{A.391}$$

Then,

$$\begin{aligned}\frac{R - \mathcal{V}\left(h^{\mathrm{d}}\right)}{2\mathcal{M}\left(h^{\mathrm{d}}\right)} - \mu^u \geq E \iff R \geq & 2\mathcal{M}\left(h^{\mathrm{d}}\right)\left(E + \mu^u\right) + \mathcal{V}\left(h^{\mathrm{d}}\right) \\ = & 2\left(c_F\left(\mu^F\left(h^{\mathrm{d}}\right)\right)^2 + c_B\left(\mu^B\left(h^{\mathrm{d}}\right)\right)^2\right)\left(E + \mu^u\right) \\ & + c_F\left(\sigma^F\left(h^{\mathrm{d}}\right)\right)^2 + c_B\left(\sigma^B\left(h^{\mathrm{d}}\right)\right)^2.\end{aligned} \tag{A.392}$$

Applying Lemma 11,

$$
\begin{aligned}
&2\left(c_F\left(\mu^F\left(h^{\mathrm{d}}\right)\right)^2+c_B\left(\mu^B\left(h^{\mathrm{d}}\right)\right)^2\right)(E+\mu^u)+c_F\left(\sigma^F\left(h^{\mathrm{d}}\right)\right)^2+c_B\left(\sigma^B\left(h^{\mathrm{d}}\right)\right)^2\\
=&c_F\left(2\left(\mu^F\left(h^{\mathrm{d}}\right)\right)^2(E+\mu^u)+\left(\sigma^F\left(h^{\mathrm{d}}\right)\right)^2\right)\\
&+c_B\left(2\left(\mu^B\left(h^{\mathrm{d}}\right)\right)^2(E+\mu^u)+\left(\sigma^B\left(h^{\mathrm{d}}\right)\right)^2\right)\\
\leq&c_F\left(8\left(\mathrm{E}\left[S\right]\right)^2(E+\mu^u)+\mathrm{Var}\left[S\right]+\left(\frac{3\mathrm{Var}\left[S\right]+\left(\mathrm{E}\left[S\right]\right)^2}{2\mathrm{E}\left[S\right]}\right)^2\right)\\
&+c_B\left(8\left(\mathrm{E}\left[\frac{1}{r}\right]\right)^2(E+\mu^u)+G^r+\frac{\left(G^r+H^r\right)^2}{4H^r}\right).
\end{aligned}
\tag{A.393}
$$

Then, $\hat{a}^e=E$ if

$$
\begin{aligned}
R\geq&c_F\left(8\left(\mathrm{E}\left[S\right]\right)^2(E+\mu^u)+\mathrm{Var}\left[S\right]+\left(\frac{3\mathrm{Var}\left[S\right]+\left(\mathrm{E}\left[S\right]\right)^2}{2\mathrm{E}\left[S\right]}\right)^2\right)\\
&+c_B\left(8\left(\mathrm{E}\left[\frac{1}{r}\right]\right)^2(E+\mu^u)+G^r+\frac{\left(G^r+H^r\right)^2}{4H^r}\right),
\end{aligned}
\tag{A.394}
$$

In this case, from (A.377), $\hat{h}^{\mathrm{d}}=1$, if $E\geq 1$ and

$$
c_B\leq c_F\gamma^l\left(\frac{3\left(\mathrm{Var}\left[S\right]+\left(\mathrm{E}\left[S\right]\right)^2\right)}{2G^r+\left(0.25+\gamma^h\right)H^r+8\left(\frac{(E+\mu^u)^2+(\sigma^u)^2}{E+\mu^u}\right)\left(\mathrm{E}\left[\frac{1}{r}\right]\right)^2}\right),
\tag{A.395}
$$

since

$$
\min\left(E+\mu^u,(E+\mu^u)^2+(\sigma^u)^2\right)=E+\mu^u
\tag{A.396}
$$

if $E\geq 1$. Additionally, from (A.382) and (A.383), $\hat{h}^{\mathrm{d}}=h^{\min}$, if

$$
G^r\geq\frac{\gamma^h}{1-\gamma^h}\left(4\left(\frac{(E+\mu^u)^2+(\sigma^u)^2}{(E+\mu^u)}\right)\left(\mathrm{E}\left[\frac{1}{r}\right]\right)^2+H^r\right)
\tag{A.397}
$$

and

$$c_B \geq c_F \left(\frac{3\text{Var}\left[S\right] + \left(\text{E}\left[S\right]\right)^2 \left(1 + 4\left(\frac{(E+\mu^u)^2+(\sigma^u)^2}{(E+\mu^u)}\right)\right)}{\left(\frac{1-\gamma^h}{\gamma^h}\right) G^r - H^r - 4\left(\frac{(E+\mu^u)^2+(\sigma^u)^2}{(E+\mu^u)}\right)\left(\text{E}\left[\frac{1}{r}\right]\right)^2} \right). \tag{A.398}$$

□

Proof of Proposition 10

From Lemma 2, the objective function of the hospital in the absence of backroom costs is

$$\Pi_{\text{A}}^{\text{FE}}\left(\mathbf{a}^e, \mathbf{h}^{\text{d}}\right) = \sum_{i=1}^{N} a_i^e \mathcal{A}_i^{\text{FE}}\left(\mathbf{h}^{\text{d}}\right) - c_F \left(\sum_{i=1}^{N} a_i^e \mu_i^F\left(h_i^{\text{d}}\right)\right)^2 + \sum_{i=1}^{N} \mathcal{B}_i^{\text{FE}}\left(\mathbf{h}^{\text{d}}\right), \tag{A.399}$$

where

$$\mathcal{A}_i^{\text{FE}}\left(\mathbf{h}^{\text{d}}\right) = R_i - c_F \left(2\mu_i^F\left(h_i^{\text{d}}\right)\sum_{j=1}^{N} \mu_j^u \mu_j^F\left(h_j^{\text{d}}\right) + \left(\sigma_i^F\left(h_i^{\text{d}}\right)\right)^2\right), \tag{A.400}$$

$$\begin{aligned}\mathcal{B}_i^{\text{FE}}\left(\mathbf{h}^{\text{d}}\right) =& \mu_i^u R_i \\ &- c_F \left(\left((\sigma_i^u)^2 + (\mu_i^u)^2\right)\left(\mu_i^F\left(h_i^{\text{d}}\right)\right)^2 + \mu_i^u \left(\sigma_i^F\left(h_i^{\text{d}}\right)\right)^2 \right. \\ &\left. + \sum_{j\neq i} \left(\rho_{ij}^u \sigma_i^u \sigma_j^u + \mu_i^u \mu_j^u\right) \mu_i^F\left(h_i^{\text{d}}\right) \mu_j^F\left(h_j^{\text{d}}\right) \right).\end{aligned} \tag{A.401}$$

Substituting (2.29) and (2.30), we can rewrite the above expressions as functions of $\mathbf{p}^{\mathrm{a}}$ instead of $\mathbf{h}^{\mathrm{d}}$:

$$\begin{aligned}\mathcal{A}_i^{\mathrm{FE}}\left(\mathbf{p}^{\mathrm{a}}\right) =& R_i - c_F \left(2\mathrm{E}\left[S_i\right]\left(1+p_i^{\mathrm{a}}\right)\sum_{j=1}^{N}\left(\mu_j^u \mathrm{E}\left[S_j\right]\left(1+p_j^{\mathrm{a}}\right)\right)\right. \\ &\left. + \left(1+3p_i^{\mathrm{a}}\right)\mathrm{Var}\left[S_i\right] + p_i^{\mathrm{a}}\left(1-p_i^{\mathrm{a}}\right)\left(\mathrm{E}\left[S_i\right]\right)^2 \right), \end{aligned} \tag{A.402}$$

$$\begin{aligned}\mathcal{B}_i^{\mathrm{FE}}\left(\mathbf{p}^{\mathrm{a}}\right) =& \mu_i^u R_i - c_F \left(\left(\left(\sigma_i^u\right)^2 + \left(\mu_i^u\right)^2\right)\left(\mathrm{E}\left[S_i\right]\left(1+p_i^{\mathrm{a}}\right)\right)^2 \right. \\ &+ \mu_i^u\left(\left(1+3p_i^{\mathrm{a}}\right)\mathrm{Var}\left[S_i\right] + p_i^{\mathrm{a}}\left(1-p_i^{\mathrm{a}}\right)\left(\mathrm{E}\left[S_i\right]\right)^2\right) \\ &\left. + \sum_{j\neq i}\left(\left(\rho_{ij}^u\sigma_i^u\sigma_j^u + \mu_i^u\mu_j^u\right)\mathrm{E}\left[S_i\right]\mathrm{E}\left[S_j\right]\left(1+p_i^{\mathrm{a}}\right)\left(1+p_j^{\mathrm{a}}\right)\right)\right). \end{aligned} \tag{A.403}$$

Plugging this in gives,

$$\begin{aligned}\Pi_{\mathrm{A}}^{\mathrm{FE}}\left(\mathbf{a}^e,\mathbf{p}^{\mathrm{a}}\right) =& \sum_{i=1}^{N}\left(a_i^e R_i\right) - c_F\left(\sum_{i=1}^{N}\left(a_i^e 2\left(\mathrm{E}\left[S_i\right]\left(1+p_i^{\mathrm{a}}\right)\right)^2\mu_i^u\right)\right. \\ &+ \sum_{i=1}^{N}\left(a_i^e 2\mathrm{E}\left[S_i\right]\left(1+p_i^{\mathrm{a}}\right)\sum_{j\neq i}\left(\mu_j^u\mathrm{E}\left[S_j\right]\left(1+p_j^{\mathrm{a}}\right)\right)\right) \\ &\left. + \sum_{i=1}^{N}\left(a_i^e\left(\left(1+3p_i^{\mathrm{a}}\right)\mathrm{Var}\left[S_i\right] + p_i^{\mathrm{a}}\left(1-p_i^{\mathrm{a}}\right)\left(\mathrm{E}\left[S_i\right]\right)^2\right)\right)\right) \\ &- c_F\left(\sum_{i=1}^{N}a_i^e\mathrm{E}\left[S_i\right]\left(1+p_i^{\mathrm{a}}\right)\right)^2 \\ &+ \sum_{i=1}^{N}\left(\mu_i^u R_i\right) - c_F\left(\sum_{i=1}^{N}\left(\left(\left(\sigma_i^u\right)^2 + \left(\mu_i^u\right)^2\right)\left(\mathrm{E}\left[S_i\right]\left(1+p_i^{\mathrm{a}}\right)\right)^2\right)\right. \\ &+ \sum_{i=1}^{N}\left(\mu_i^u\left(\left(1+3p_i^{\mathrm{a}}\right)\mathrm{Var}\left[S_i\right] + p_i^{\mathrm{a}}\left(1-p_i^{\mathrm{a}}\right)\left(\mathrm{E}\left[S_i\right]\right)^2\right)\right) \\ &\left. + \sum_{i=1}^{N}\sum_{j\neq i}\left(\left(\rho_{ij}^u\sigma_i^u\sigma_j^u + \mu_i^u\mu_j^u\right)\mathrm{E}\left[S_i\right]\mathrm{E}\left[S_j\right]\left(1+p_i^{\mathrm{a}}\right)\left(1+p_j^{\mathrm{a}}\right)\right)\right) \end{aligned} \tag{A.404}$$

Then, we can optimize the objective function over $\mathbf{p}^{\mathrm{a}}$ and $\mathbf{a}^{e}$. We first hold the latter fixed and consider the optimal value of $\mathbf{p}^{\mathrm{a}}$. Taking the partial derivative of the objective function (A.399) with respect to p_k^{a}, we get

$$
\begin{aligned}
\frac{\partial \Pi_{\mathrm{A}}^{\mathrm{FE}}}{\partial p_k^{\mathrm{a}}} = -c_F \Bigg(& 4a_k^e \left(\mathrm{E}\left[S_k\right]\right)^2 \left(1+p_k^{\mathrm{a}}\right) \mu_k^u + 2a_k^e \mathrm{E}\left[S_k\right] \sum_{j \neq k} \left(\mu_j^u \mathrm{E}\left[S_j\right] \left(1+p_j^{\mathrm{a}}\right)\right) \\
& + 2\mu_k^u \mathrm{E}\left[S_k\right] \sum_{i \neq k} \left(a_i^e \mathrm{E}\left[S_i\right] \left(1+p_i^{\mathrm{a}}\right)\right) \\
& + a_k^e \left(3\mathrm{Var}\left[S_k\right] + \left(1-2p_k^{\mathrm{a}}\right) \left(\mathrm{E}\left[S_k\right]\right)^2\right) \\
& + 2a_k^e \mathrm{E}\left[S_k\right] \sum_{i=1}^{N} \left(a_i^e \mathrm{E}\left[S_i\right] \left(1+p_i^{\mathrm{a}}\right)\right) \\
& + 2\left(\left(\sigma_k^u\right)^2 + \left(\mu_k^u\right)^2\right) \left(\mathrm{E}\left[S_k\right]\right)^2 \left(1+p_k^{\mathrm{a}}\right) \\
& + \mu_k^u \left(3\mathrm{Var}\left[S_k\right] + \left(1-2p_k^{\mathrm{a}}\right) \left(\mathrm{E}\left[S_k\right]\right)^2\right) \\
& + 2\mathrm{E}\left[S_k\right] \sum_{i \neq k} \left(\left(\rho_{ik}^u \sigma_i^u \sigma_k^u + \mu_i^u \mu_k^u\right) \mathrm{E}\left[S_i\right] \left(1+p_i^{\mathrm{a}}\right)\right) \Bigg) \\
= -c_F \Bigg(& 2\left(\mathrm{E}\left[S_k\right]\right)^2 \left(1+p_k^{\mathrm{a}}\right) \left(2a_k^e \mu_k^u + \left(\sigma_k^u\right)^2 + \left(\mu_k^u\right)^2\right) \\
& + 2a_k^e \mathrm{E}\left[S_k\right] \left(\sum_{j \neq k} \left(\mu_j^u \mathrm{E}\left[S_j\right] \left(1+p_j^{\mathrm{a}}\right)\right) + \sum_{i=1}^{N} \left(a_i^e \mathrm{E}\left[S_i\right] \left(1+p_i^{\mathrm{a}}\right)\right)\right) \\
& + 2\mu_k^u \mathrm{E}\left[S_k\right] \sum_{i \neq k} \left(a_i^e \mathrm{E}\left[S_i\right] \left(1+p_i^{\mathrm{a}}\right)\right) \\
& + \left(a_k^e + \mu_k^u\right) \left(3\mathrm{Var}\left[S_k\right] + \left(1-2p_k^{\mathrm{a}}\right) \left(\mathrm{E}\left[S_k\right]\right)^2\right) \\
& + 2\mathrm{E}\left[S_k\right] \sum_{i \neq k} \left(\left(\rho_{ik}^u \sigma_i^u \sigma_k^u + \mu_i^u \mu_k^u\right) \mathrm{E}\left[S_i\right] \left(1+p_i^{\mathrm{a}}\right)\right) \Bigg),
\end{aligned}
\tag{A.405}
$$

which is a linear function of p_k^{a}. Further note that at $p_k^{\mathrm{a}} = 0$,

$$\begin{aligned}
\frac{\partial \Pi_{\mathrm{A}}^{\mathrm{FE}}}{\partial p_k^{\mathrm{a}}} = & -c_F \left(2 \left(\mathrm{E}\left[S_k \right] \right)^2 \left(2 a_k^e \mu_k^u + \left(\sigma_k^u \right)^2 + \left(\mu_k^u \right)^2 \right) \right. \\
& + 2 a_k^e \mathrm{E}\left[S_k \right] \left(\sum_{j \neq k} \left(\mu_j^u \mathrm{E}\left[S_j \right] \left(1 + p_j^{\mathrm{a}} \right) \right) + \sum_{i=1}^{N} \left(a_i^e \mathrm{E}\left[S_i \right] \left(1 + p_i^{\mathrm{a}} \right) \right) \right) \\
& + 2 \mu_k^u \mathrm{E}\left[S_k \right] \sum_{i \neq k} \left(a_i^e \mathrm{E}\left[S_i \right] \left(1 + p_i^{\mathrm{a}} \right) \right) \\
& + \left(a_k^e + \mu_k^u \right) \left(3 \mathrm{Var}\left[S_k \right] + \left(\mathrm{E}\left[S_k \right] \right)^2 \right) \\
& \left. + 2 \mathrm{E}\left[S_k \right] \sum_{i \neq k} \left(\left(\rho_{ik}^u \sigma_i^u \sigma_k^u + \mu_i^u \mu_k^u \right) \mathrm{E}\left[S_i \right] \left(1 + p_i^{\mathrm{a}} \right) \right) \right) \\
& \leq 0. \qquad \text{(A.406)}
\end{aligned}$$

Then, there are two possibilities. First, if (A.405) is negative for all $p_k^{\mathrm{a}} \in [0, 1]$, then the objective function is maximized at $p_k^{\mathrm{a}} = 0$. On the other hand, if (A.405) switches from negative to positive for some $p_k^{\mathrm{a}} \in (0, 1)$, then the objective function is a convex function of p_k^{a} and maximized at either $p_k^{\mathrm{a}} = 0$ or $p_k^{\mathrm{a}} = 1$.

Suppose the latter case is true, and consider the value of the objective function at the end points. In both cases, any term containing $p_k^{\mathrm{a}} \left(1 - p_k^{\mathrm{a}} \right)$ is equal to 0, and the remaining terms are decreasing in p_k^{a}, the objective function is maximized at $p_k^{\mathrm{a}} = 0$. Therefore, for any $\mathbf{a}^e$, the global maximizer of (A.399) is $\mathbf{p}^{\mathrm{a}} = (0, \ldots, 0)$, or equivalently, $\mathbf{h}^{\mathrm{d}} = (1, \ldots, 1)$.

We can now consider the optimal value of $\mathbf{a}^e$. Plugging the above result into (A.399) gives

$$\bar{\Pi}_{\mathrm{A}}^{\mathrm{FE}} \left(\mathbf{a}^e \right) = \sum_{i=1}^{N} a_i^e \bar{\mathcal{A}}_i^{\mathrm{FE}} - c_F \left(\sum_{i=1}^{N} a_i^e \mathrm{E}\left[S_i \right] \right)^2 + \sum_{i=1}^{N} \bar{\mathcal{B}}_i^{\mathrm{FE}}, \qquad \text{(A.407)}$$

where

$$\bar{\mathcal{A}}_i^{\mathrm{FE}} = R_i - c_F \left(\mathrm{Var}\left[S_i\right] + 2\mathrm{E}\left[S_i\right] \left(\sum_{j=1}^{N} \mu_j^u \mathrm{E}\left[S_j\right] \right) \right), \tag{A.408}$$

$$\begin{aligned} \bar{\mathcal{B}}_i^{\mathrm{FE}} =& \mu_i^u R_i - c_F \Bigg(\left((\sigma_i^u)^2 + (\mu_i^u)^2 \right) \left(\mathrm{E}\left[S_i\right]\right)^2 + \mu_i^u \left(\mathrm{Var}\left[S_i\right]\right) \\ & + \sum_{j \neq i} \left(\left(\rho_{ij}^u \sigma_i^u \sigma_j^u + \mu_i^u \mu_j^u \right) \mathrm{E}\left[S_i\right] \mathrm{E}\left[S_j\right] \right) \Bigg). \end{aligned} \tag{A.409}$$

Note that the Hessian matrix $\mathbf{H}$ of $\bar{\Pi}^{\mathrm{FE}}$ is given by

$$\mathbf{H} = -2c_F \left(M M^T \right), \tag{A.410}$$

where

$$M = \left[\mathrm{E}\left[S_1\right] \cdots \mathrm{E}\left[S_N\right] \right]. \tag{A.411}$$

The Hessian is clearly negative definite, so the objective function is concave in $\mathbf{a}_{\mathrm{FE}}^e$ and a global maximum exists. Then, the Lagrangean is given by

$$L(\mathbf{a}^e, \lambda, \nu) = \bar{\Pi}^{\mathrm{FE}}\left(\mathbf{a}^e\right) + \sum_{i=1}^{N} \lambda_i a_i^e + \sum_{i=1}^{N} \nu_i \left(E_i - a_i^e\right), \tag{A.412}$$

and the optimal solution $\hat{\mathbf{a}}^e$ is a critical point of the Lagrangean satisfying the following equations:

$$\begin{aligned} \frac{\bar{\mathcal{A}}_i^{\mathrm{FE}} + \hat{\lambda}_i - \hat{\nu}_i}{2c_F \mathrm{E}\left[S_i\right]} &= \sum_{j=1}^{N} \left(\hat{a}_{\mathrm{FE}}^e\right)_j \mathrm{E}\left[S_j\right], \quad i = 1, \ldots, N \\ \hat{\lambda}_i \left(\hat{a}_{\mathrm{FE}}^e\right)_i &= 0, \quad i = 1, \ldots, N \\ \hat{\nu}_i \left(E_i - \left(\hat{a}_{\mathrm{FE}}^e\right)_i\right) &= 0, \quad i = 1, \ldots, N, \\ \lambda_i, \nu_i &\geq 0, \quad i = 1, \ldots, N. \end{aligned} \tag{A.413}$$

Without loss of generality, assume that

$$\frac{R_1 - c_F \text{Var}\left[S_1\right]}{\text{E}\left[S_1\right]} \geq \cdots \geq \frac{R_N - c_F \text{Var}\left[S_N\right]}{\text{E}\left[S_N\right]}, \tag{A.414}$$

and define

$$I_+ = \left\{ i \in \{1, \ldots, N\} \,\middle|\, \frac{\bar{\mathcal{A}}_i^{\text{FE}}}{2c_F \text{E}\left[S_i\right]} > \sum_{j=1}^{N} \left(\hat{a}_{\text{FE}}^e\right)_j \text{E}\left[S_j\right] \right\}, \tag{A.415}$$

$$I = \left\{ i \in \{1, \ldots, N\} \,\middle|\, \frac{\bar{\mathcal{A}}_i^{\text{FE}}}{2c_F \text{E}\left[S_i\right]} = \sum_{j=1}^{N} \left(\hat{a}_{\text{FE}}^e\right)_j \text{E}\left[S_j\right] \right\}, \tag{A.416}$$

$$I_- = \left\{ i \in \{1, \ldots, N\} \,\middle|\, \frac{\bar{\mathcal{A}}_i^{\text{FE}}}{2c_F \text{E}\left[S_i\right]} < \sum_{j=1}^{N} \left(\hat{a}_{\text{FE}}^e\right)_j \text{E}\left[S_j\right] \right\}. \tag{A.417}$$

Then,

$$\begin{aligned} \left(\hat{a}_{\text{FE}}^e\right)_i &= E_i \text{ for } i \in I_+, \\ 0 < \left(\hat{a}_{\text{FE}}^e\right)_i &< E_i \text{ for } i \in I, \\ \left(\hat{a}_{\text{FE}}^e\right)_i &= 0 \text{ for } i \in I_-. \end{aligned} \tag{A.418}$$

We consider three possible cases. First, if

$$\frac{\bar{\mathcal{A}}_1^{\text{FE}}}{2c_F \left(\text{E}\left[S_1\right]\right)^2} < E_1, \tag{A.419}$$

then $I = \{1\}$, $I_- = \{2, \ldots, N\}$, and $I_+ = \varnothing$, with

$$\left(\hat{a}_{\text{FE}}^e\right)_1 = \frac{\left(R_1 - c_F \left(\text{Var}\left[S_1\right] + 2\text{E}\left[S_1\right] \left(\sum_{j=1}^{N} \mu_j^u \text{E}\left[S_j\right]\right)\right)\right)^+}{2c_F \left(\text{E}\left[S_1\right]\right)^2}. \tag{A.420}$$

Second, if

$$\frac{\bar{\mathcal{A}}_N^{\mathrm{FE}}}{2c_F \mathrm{E}\left[S_N\right]} > \sum_{j=1}^{N} E_j \mathrm{E}\left[S_j\right], \tag{A.421}$$

then $I = \varnothing$, $I_- = \varnothing$, and $I_+ = \{1, \ldots, N\}$.

Finally, if

$$E_1 \leq \frac{\bar{\mathcal{A}}_1^{\mathrm{FE}}}{2c_F \left(\mathrm{E}\left[S_1\right]\right)^2}, \quad \frac{\bar{\mathcal{A}}_N^{\mathrm{FE}}}{2c_F \mathrm{E}\left[S_N\right]} \leq \sum_{j=1}^{N} E_j \mathrm{E}\left[S_j\right], \tag{A.422}$$

then define

$$i_{\mathrm{FE}}^* = 1 + \max\left(i \in \{1, \ldots, N\} \,\middle|\, \frac{\bar{\mathcal{A}}_i^{\mathrm{FE}}}{2c_F \mathrm{E}\left[S_i\right]} > \sum_{j=1}^{N} \left(\hat{a}_{\mathrm{FE}}^e\right)_j \mathrm{E}\left[S_j\right] \right). \tag{A.423}$$

Therefore, $I = \{i_{\mathrm{FE}}^*\}$, $I_- = \{i_{\mathrm{FE}}^* + 1, \ldots, N\}$, and $I_+ = \{0, \ldots, i_{\mathrm{FE}}^* - 1\}$, with

$$\hat{a}_{i_{\mathrm{FE}}^*}^e = \frac{R_{i_{\mathrm{FE}}^*} - c_F\left(\mathrm{Var}\left[S_{i_{\mathrm{FE}}^*}\right] + 2c_F \mathrm{E}\left[S_{i_{\mathrm{FE}}^*}\right]\left(\sum_{j=1}^{N} \mu_j^u \mathrm{E}\left[S_j\right]\right)\right)}{2c_F\left(\mathrm{E}\left[S_{i_{\mathrm{FE}}^*}\right]\right)^2} - \frac{\sum_{j=1}^{i_{\mathrm{FE}}^*-1} E_j \mathrm{E}\left[S_j\right]}{\mathrm{E}\left[S_{i_{\mathrm{FE}}^*}\right]}. \tag{A.424}$$

□

Proof of Proposition 11

As stated in (2.73), the optimal discharge policy is determined by minimizing the backroom costs under an arbitrary portfolio of elective procedures:

$$\mathbf{h}_r^{\mathrm{d}}\left(\mathbf{a}^e\right) = \underset{\mathbf{h}^{\mathrm{d}}\in[\mathbf{h}^{\min},\mathbf{1}]}{\arg\min} \left(\sum_{i=1}^{N} \left(\left(\sigma_i^B\left(h_i^{\mathrm{d}}\right)\right)^2 + 2\mu_i^B\left(h_i^{\mathrm{d}}\right) \left(\sum_{k=1}^{N} \mu_k^u \mu_k^B\left(h_k^{\mathrm{d}}\right) \right) \right) a_i^e + \left(\sum_{i=1}^{N} a_i^e \mu_i^B\left(h_i^{\mathrm{d}}\right) \right)^2 \right.$$
$$+ \sum_{i=1}^{N} \left(\left(\mu_i^u \left(\sigma_i^B\left(h_i^{\mathrm{d}}\right)\right)^2 + \left(\mu_i^u\right)^2 + \left(\sigma_i^u\right)^2 \right) \left(\mu_i^B\left(h_i^{\mathrm{d}}\right)\right)^2 \right.$$
$$\left. \left. + 2\sum_{j\neq i} \left(\mu_i^u \mu_j^u + \rho_{ij}^u \sigma_i^u \sigma_j^u\right) \mu_i^B\left(h_i^{\mathrm{d}}\right) \mu_j^B\left(h_j^{\mathrm{d}}\right) \right) \right). \tag{A.425}$$

Notice that if $\left(\sigma_i^B\left(h_i^{\mathrm{d}}\right)\right)^2$ and $\mu_i^B\left(h_i^{\mathrm{d}}\right)$ are increasing in h_i^{d}, the expression being minimized is also increasing in h_i^{d}, and therefore $\hat{h}_i^{\mathrm{d}} = h_i^{\min}$. Thus, from Proposition 7, if for $i = 1, \ldots, N$,

$$\gamma_i^h \leq \frac{G_i^r}{G_i^r + H_i^r}, \tag{A.426}$$

then under any portfolio of elective procedures, the hospital will always discharge patients as quickly as possible, i.e.

$$\mathbf{h}_r^{\mathrm{d}} = \left(h_1^{\min}, \ldots, h_N^{\min}\right). \tag{A.427}$$

Define the probability that a patient undergoing procedure i will be readmitted under this policy as

$$p_i^{\max} = p_i^{\mathrm{a}}\left(h_i^{\min}\right). \tag{A.428}$$

Then, the objective function of the hospital under this discharge policy is

$$\bar{\Pi}_{\mathrm{A}}^{\mathrm{SI}}\left(\mathbf{a}^e\right) = \sum_{i=1}^{N} a_i^e \bar{\mathcal{A}}_i^{\mathrm{SI}} - c_F \left(\sum_{i=1}^{N} a_i^e \mathrm{E}\left[S_i\right]\left(1 + p_i^{\max}\right) \right)^2 + \sum_{i=1}^{N} \bar{\mathcal{B}}_i^{\mathrm{SI}}, \tag{A.429}$$

where

$$\bar{\mathcal{A}}_i^{\mathrm{SI}} = R_i - c_F \left(2\mathrm{E}\left[S_i\right]\left(1+p_i^{\max}\right)\sum_{j=1}^{N}\left(\mu_j^u \mathrm{E}\left[S_j\right]\left(1+p_j^{\max}\right)\right) + \left(1+3p_i^{\max}\right)\mathrm{Var}\left[S_i\right] + p_i^{\max}\left(1-p_i^{\max}\right)\left(\mathrm{E}\left[S_i\right]\right)^2 \right), \tag{A.430}$$

$$\bar{\mathcal{B}}_i^{\mathrm{SI}} = \mu_i^u R_i - c_F \left(\left(\left(\sigma_i^u\right)^2 + \left(\mu_i^u\right)^2\right)\left(\mathrm{E}\left[S_i\right]\left(1+p_i^{\max}\right)\right)^2 + \mu_i^u\left(\left(1+3p_i^{\max}\right)\mathrm{Var}\left[S_i\right] + p_i^{\max}\left(1-p_i^{\max}\right)\left(\mathrm{E}\left[S_i\right]\right)^2\right) + \sum_{j\neq i}\left(\left(\rho_{ij}^u\sigma_i^u\sigma_j^u + \mu_i^u\mu_j^u\right)\mathrm{E}\left[S_i\right]\mathrm{E}\left[S_j\right]\left(1+p_i^{\max}\right)\left(1+p_j^{\max}\right)\right) \right). \tag{A.431}$$

Assume without loss of generality that

$$\frac{\bar{\mathcal{A}}_1^{\mathrm{SI}}}{\mathrm{E}\left[S_1\right]\left(1+p_1^{\max}\right)} \geq \cdots \geq \frac{\bar{\mathcal{A}}_N^{\mathrm{SI}}}{\mathrm{E}\left[S_N\right]\left(1+p_N^{\max}\right)}. \tag{A.432}$$

We can apply the results from Proposition 10, substituting $\bar{\mathcal{A}}_i^{\mathrm{SI}}$, $\bar{\mathcal{B}}_i^{\mathrm{SI}}$, and $\mathrm{E}\left[S_i\right]\left(1+p_i^{\max}\right)$ for $\bar{\mathcal{A}}_i^{\mathrm{FE}}$, $\bar{\mathcal{B}}_i^{\mathrm{FE}}$, and $\mathrm{E}\left[S_i\right]$, respectively, for all i. □

Proof of Proposition 12

As given in (3.4), students' best response function is given by

$$e_s^*\left(e_t, \pi_s\right) = \underset{e_j \leq e_s \leq 1}{\arg\max}\left\{\pi_s Pr\left[\beta_1 = P|\beta_0\right] - c_s e_s\right\}. \tag{A.433}$$

The second derivative of the objective function is clearly negative. Therefore, applying (3.1) and the first-order condition, the unconstrained maximum is given by,

$$a\pi_s g\left(\beta_0\right)\left(e_s^*\right)^{a-1}\left(e_t\right)^b - c_s = 0 \iff e_s^* = \left(\frac{a\pi_s g\left(\beta_0\right)\left(e_t\right)^b}{c_s}\right)^{\frac{1}{1-a}}, \tag{A.434}$$

where, for clarity, we omit the designation (e_t, π_s). Applying the constraints on effort, (3.6) follows. □

Proof of Proposition 13

Teachers' maximization problem is given in (3.5). Because teachers anticipate students' response but are uncertain about their cost of exerting effort, this becomes

$$e_t^*(\pi_s, \pi_t) = \underset{e_j \leq e_t \leq 1}{\arg\max} \left\{ \pi_t g(\beta_0) E\left[(e_s^*)^a\right] (e_t)^b - c_t e_t \right\}. \tag{A.435}$$

The second derivative of the objective function is clearly negative. For notational convenience, let e_s^h and e_s^l refer to the unconstrained optimal effort level for students under high and low cost of effort, respectively. That is,

$$e_s^h = \left(\frac{a\pi_s g(\beta_0) (e_t)^b}{c^h} \right)^{\frac{1}{1-a}} \text{ and } e_s^l = \left(\frac{a\pi_s g(\beta_0) (e_t)^b}{c^l} \right)^{\frac{1}{1-a}}. \tag{A.436}$$

Since $c^h > c^l$, then $e_s^h < e_s^l$. Then, there are six possible cases: i) $1 < e_s^h$, ii) $e_j \leq e_s^h \leq 1 < e_s^l$, iii) $e_s^h < e_j \leq 1 < e_s^l$, iv) $e_j \leq e_s^h < e_s^l \leq 1$, v) $e_s^h < e_j \leq e_s^l \leq 1$, and vi) $e_s^l < e_j$. We consider each of these in turn below. Two of these cases do not result in a closed-form solution for teachers' optimal effort level.

Case 1: $1 < e_s^h$**.** Suppose that students exert the maximum effort level, for any possible cost of effort. That is, suppose that

$$1 < \left(\frac{a\pi_s g(\beta_0) (e_t)^b}{c^h} \right)^{\frac{1}{1-a}} \iff \left(\frac{c^h}{a\pi_s g(\beta_0)} \right)^{\frac{1}{b}} < e_t. \tag{A.437}$$

Then, plugging in the students' effort level and applying the first-order condition, the teachers' unconstrained optimal effort decision is

$$e_t^* = \left(\frac{b\pi_t g(\beta_0)}{c_t} \right)^{\frac{1}{1-b}}. \tag{A.438}$$

If (A.438) is between e_j and 1, then, comparing (A.438) to (A.437), this case holds if

and only if

$$\left(\frac{c^h}{a\pi_s}\right)^{1-b}\left(\frac{c_t}{b\pi_t}\right)^{b} < g\left(\beta_0\right). \tag{A.439}$$

Case 2: $e_j \leq e_s^h \leq 1 < e_s^l$**.** Suppose that students exert maximum effort if the cost of effort is low and some effort level between the minimum and maximum possible values if the cost is high. That is, suppose that

$$\begin{aligned} 1 < \left(\frac{a\pi_s g\left(\beta_0\right)\left(e_t\right)^b}{c^l}\right)^{\frac{1}{1-a}} &\iff \left(\frac{c^l}{a\pi_s g\left(\beta_0\right)}\right)^{\frac{1}{b}} < e_t \text{ and} \\ e_j \leq \left(\frac{a\pi_s g\left(\beta_0\right)\left(e_t\right)^b}{c^h}\right)^{\frac{1}{1-a}} \leq 1 &\iff \left(\frac{c^h (e_j)^{1-a}}{a\pi_s g\left(\beta_0\right)}\right)^{\frac{1}{b}} \leq e_t \leq \left(\frac{c^h}{a\pi_s g\left(\beta_0\right)}\right)^{\frac{1}{b}}. \end{aligned} \tag{A.440}$$

Then, plugging in the students' effort level and applying the first-order condition, the teachers' unconstrained optimal effort decision satisfies

$$\left(\frac{c_t}{\pi_t g\left(\beta_0\right)}\right)\left(e_t^*\right)^{\frac{1-a-b}{1-a}} - \left(1-p_s^h\right) b\left(e_t^*\right)^{\frac{-ab}{1-a}} = p_s^h\left(\frac{a\pi_s g\left(\beta_0\right)}{c^h}\right)^{\frac{a}{1-a}}\left(\frac{b}{1-a}\right). \tag{A.441}$$

A closed-form solution cannot be obtained.

Case 3: $e_s^h < e_j \leq 1 < e_s^l$**.** Suppose that students exert maximum effort if the cost of effort is low and minimum effort if the cost of effort is high. That is, suppose that

$$\begin{aligned} 1 < \left(\frac{a\pi_s g\left(\beta_0\right)\left(e_t\right)^b}{c^l}\right)^{\frac{1}{1-a}} &\text{ and } \left(\frac{a\pi_s g\left(\beta_0\right)\left(e_t\right)^b}{c^h}\right)^{\frac{1}{1-a}} < e_j \\ \iff c^l < a\pi_s g\left(\beta_0\right)\left(e_t\right)^b &< c^h (e_j)^{1-a}. \end{aligned} \tag{A.442}$$

Then, plugging in the students' effort level and applying the first-order condition, the

teachers' unconstrained optimal effort decision is

$$e_t^* = \left(\frac{b\pi_t g\left(\beta_0\right)\left(p_s^h (e_j)^a + \left(1 - p_s^h\right)\right)}{c_t} \right)^{\frac{1}{1-b}}. \tag{A.443}$$

If (A.438) is between e_j and 1, then, comparing (A.443) to (A.442), this case holds if and only if

$$c^l < a\pi_s g\left(\beta_0\right) \left(\frac{b\pi_t g\left(\beta_0\right)\left(p_s^h (e_j)^a + \left(1 - p_s^h\right)\right)}{c_t} \right)^{\frac{b}{1-b}} < c^h (e_j)^{1-a}. \tag{A.444}$$

Case 4: $e_j \leq e_s^h < e_s^l \leq 1$**.** Suppose that students exert some effort level between the minimum and maximum possible values, for any possible cost of effort. That is, suppose that

$$\begin{aligned} & e_j \leq \left(\frac{a\pi_s g\left(\beta_0\right)\left(e_t\right)^b}{c^h} \right)^{\frac{1}{1-a}} \text{ and } \left(\frac{a\pi_s g\left(\beta_0\right)\left(e_t\right)^b}{c^l} \right)^{\frac{1}{1-a}} \leq 1 \\ \iff & c^h (e_j)^{1-a} \leq a\pi_s g\left(\beta_0\right)\left(e_t\right)^b \leq c^l. \end{aligned} \tag{A.445}$$

Then, plugging in the students' effort level and applying the first-order condition, the teachers' unconstrained optimal effort decision is

$$e_t^* = \left(g\left(\beta_0\right)\left(\pi_s a\right)^a \left(\left(\frac{b}{1-a} \right) \left(\frac{\pi_t}{c_t} \right) \left(\frac{p_s^h}{\left(c^h\right)^{\frac{a}{1-a}}} + \frac{\left(1 - p_s^h\right)}{\left(c^l\right)^{\frac{a}{1-a}}} \right) \right)^{1-a} \right)^{\frac{1}{1-a-b}} \tag{A.446}$$

If (A.438) is between e_j and 1, then, comparing (A.446) to (A.445), this case holds if

and only if

$$c^h(e_j)^{1-a} \leq$$
$$\left((a\pi_s g(\beta_0))^{1-b}\left(\left(\frac{b}{1-a}\right)\left(\frac{\pi_t g(\beta_0)}{c_t}\right)\left(\frac{p_s^h}{(c^h)^{\frac{a}{1-a}}}+\frac{(1-p_s^h)}{(c^l)^{\frac{a}{1-a}}}\right)\right)^b\right)^{\frac{1-a}{1-a-b}} \leq c^l. \tag{A.447}$$

Case 5: $e_s^h < e_j \leq e_s^l \leq 1$**.** Suppose that students exert minimum effort when the cost of exerting effort is high and some effort level between the minimum and maximum possible values when the cost is low. That is, suppose that

$$\left(\frac{a\pi_s g(\beta_0)(e_t)^b}{c^h}\right)^{\frac{1}{1-a}} < e_j \leq \left(\frac{a\pi_s g(\beta_0)(e_t)^b}{c^l}\right)^{\frac{1}{1-a}} \leq 1$$
$$\iff a\pi_s g(\beta_0)(e_t)^b < c^h(e_j)^{1-a} \text{ and } c^l(e_j)^{1-a} \leq a\pi_s g(\beta_0)(e_t)^b \leq c^l. \tag{A.448}$$

Then, plugging in the students' effort level and applying the first-order condition, the teachers' unconstrained optimal effort decision satisfies

$$bp_s^h(e_j)^a = \left(\frac{c_t}{\pi_t g(\beta_0)}\right)(e_t^*)^{1-b} - \left(\frac{b}{1-a}\right)\left(1-p_s^h\right)\left(\frac{a\pi_s g(\beta_0)}{c^l}\right)^{\frac{a}{1-a}}(e_t^*)^{\frac{ab}{1-a}}. \tag{A.449}$$

A closed-form solution cannot be obtained.

Case 6: $e_s^l < e_j$**.** Suppose that students exert minimal effort, for any possible cost of effort. That is, suppose that

$$\left(\frac{a\pi_s g(\beta_0)(e_t)^b}{c^l}\right)^{\frac{1}{1-a}} < e_j \iff a\pi_s g(\beta_0)(e_t)^b < c^l(e_j)^{1-a}. \tag{A.450}$$

Then, plugging in the students' effort level and applying the first-order condition, the

teachers' unconstrained optimal effort decision is

$$e_t^* = \left(\frac{b\pi_t g\left(\beta_0\right)\left(e_j\right)^a}{c_t} \right)^{\frac{1}{1-b}}. \tag{A.451}$$

If (A.438) is between e_j and 1, then, comparing (A.451) to (A.450), this case holds if and only if

$$\left(\frac{a\pi_s}{c^l} \right)^{1-b} \left(\frac{b\pi_t}{c_t} \right)^b g\left(\beta_0\right) < \left(e_j\right)^{1-(a+b)}. \tag{A.452}$$

□

BIBLIOGRAPHY

I. Adan and J. Vissers. Patient mix optimisation in hospital admission planning: a case study. *International Journal of Operations and Production Management*, 22(4):445–461, 2002.

I. Adan, J. Bekkers, N. Dellaert, J. Vissers, and X. Yu. Patient mix optimisation and stochastic resource requirements: A case study in cardiothoracic surgery planning. *Health Care Management Science*, 12(2):129–141, 2009.

D. A. Andritsos and C. S. Tang. Incentive programs for reducing readmissions when patient care is co-produced. *Production and Operations Management*, 27(6):999–1020, 2018.

N. Ayvaz and W. T. Huh. Allocation of hospital capacity to multiple types of patients. *Journal of Revenue and Pricing Management*, 9(5):386–398, 2010.

G. R. Baker, P. G. Norton, V. Flintoft, R. Blais, A. Brown, J. Cox, E. Etchells, W. A. Ghali, P. Hébert, S. R. Majumdar, et al. The Canadian Adverse Events Study: the incidence of adverse events among hospital patients in Canada. *Canadian Medical Association Journal*, 170(11):1678–1686, 2004.

G. Barlevy and D. Neal. Pay for percentile. *American Economic Review*, 102(5):1805–31, 2012.

D. P. Baron and R. B. Myerson. Regulating a monopolist with unknown costs. *Econometrica: Journal of the Econometric Society*, pages 911–930, 1982.

H. Bavafa, C. Leys, L. Örmeci, and S. Savin. Managing portfolio of elective surgical procedures: A multidimensional inverse newsvendor problem. *Operations Research*, 2019. (forthcoming).

M. Begen and M. Queyranne. Appointment scheduling with discrete random durations. *Mathematics of Operations Research*, 36(2):240–257, 2011.

J. R. Behrman, S. W. Parker, P. E. Todd, and K. I. Wolpin. Aligning learning incentives of students and teachers: results from a social experiment in mexican high schools. *Journal of Political Economy*, 123(2):325–364, 2015.

J. Beliën and E. Demeulemeester. Building cyclic master surgery schedules with leveled resulting bed occupancy. *European Journal of Operational Research*, 176(2):1185–1204, 2007.

J. Beliën, E. Demeulemeester, and B. Cardoen. A decision support system for cyclic master surgery scheduling with multiple objectives. *Journal of Scheduling*, 12(2):147, 2009.

K. C. Bertin. Minimally invasive outpatient total hip arthroplasty: a financial analysis. *Clinical Orthopaedics and Related Research*, 435:154–163, 2005.

E. P. Bettinger. Paying to learn: The effect of financial incentives on elementary school test scores. *Review of Economics and Statistics*, 94(3):686–698, 2012.

P. Billingsley. *Probability and Measure*. Wiley Series in Probability and Mathematical Statistics. 3rd edition, 1995.

Blue Cross Blue Shield. A study of cost variations for knee and hip replacement surgeries in the U.S., 2015.

K. E. Bulkley, L. N. Oláh, and S. Blanc. Introduction to the special issue on benchmarks for success? interim assessments as a strategy for educational improvement. *Peabody Journal of Education*, 85(2):115–124, 2010.

K. Carey. Measuring the hospital length of stay/readmission cost trade-off under a bundled payment mechanism. *Health economics*, 24(7):790–802, 2015.

K. Carey and M.-Y. Lin. Hospital length of stay and readmission: An early investigation. *Medical Care Research and Review*, 71(1):99–111, 2014.

S. Cavanagh. As McGraw-Hill Education leaves state testing, market thrives for classroom assessment. *Education Week*, 2015. URL `https://www.edweek.org/ew/articles/2015/07/31/as-mcgraw-hill-education-leaves-state-testing-market.html`.

Centers for Medicare and Medicaid Services. CMS manual system, 2008. URL `https://www.cms.gov/Regulations-and-Guidance/Guidance/Transmittals/downloads/R1460CP.pdf`. Retrieved on Mar 05, 2019.

Centers for Medicare and Medicaid Services. Icd-10-cm/pcs ms-drg v34.0 definitions manual, 2019. URL `https://www.cms.gov/icd10manual/version34-fullcode-cms/fullcode_cms/P0188.html`. Retrieved on Mar 19, 2019.

C. W. Chan, V. F. Farias, N. Bambos, and G. J. Escobar. Optimizing intensive care unit discharge decisions with patient readmissions. *Operations Research*, 60(6):1323–1341, 2012.

R. B. Chase. Where does the customer fit in a service operation? *Harvard Business Review*, 56(6):137–142, 1978.

R. B. Chase. The customer contact approach to services: theoretical bases and practical extensions. *Operations Research*, 29(4):698–706, 1981.

H. Chiang, C. Speroni, M. Herrmann, K. Hallgren, P. Burkander, and A. Wellington. Evaluation of the Teacher Incentive Fund: Final Report on Implementation and Impacts of Pay-for-Performance Across Four Years (NCEE 2017-4004). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2017.

L. Y. Chu and D. E. Sappington. Implementing high-powered contracts to motivate intertemporal effort supply. *The Rand Journal of Economics*, 40(2):296–316, 2009.

C. T. Clotfelter and H. F. Ladd. Recognizing and rewarding success in public schools. *Holding schools accountable: Performance-based reform in education*, pages 23–64, 1996.

C. T. Clotfelter, H. F. Ladd, and J. L. Vigdor. Teacher-student matching and the assessment of teacher effectiveness. *Journal of human Resources*, 41(4):778–820, 2006.

J. S. Coleman, E. Q. Campbell, C. J. Hobson, J. McPartland, A. M. Mood, F. D. Weinfeld, and R. L. York. *Equality of Educational Opportunity*, volume 38001. US Department of Health, Education, and Welfare, Office of Education, 1966.

C. J. Corbett and X. De Groote. A supplier's optimal quantity discount policy under asymmetric information. *Management science*, 46(3):444–450, 2000.

J. Crémer, F. Khalil, and J.-C. Rochet. Contracts and productive information gathering. *Games and Economic Behavior*, 25(2):174–193, 1998.

M. L. Davison, Y. S. Seo, E. C. Davenport Jr, D. Butterbaugh, and L. J. Davison. When do children fall behind? What can be done? *Phi Delta Kappan*, 85(10):752, 2004.

L. Dearden, C. Emmerson, C. Frayne, and C. Meghir. Conditional cash transfers and school dropout rates. *Journal of Human Resources*, 44(4):827–857, 2009.

T. S. Dee and J. Wyckoff. Incentives, selection, and teacher performance: Evidence from IMPACT. *Journal of Policy Analysis and Management*, 34(2):267–297, 2015.

A. Diamant, J. Milner, and F. Quereshy. Dynamic patient scheduling for multi-appointment health care programs. *Production and Operations Management*, 27(1):58–79, 2018.

E. Dillon and A. Rotherham. States' evidence: What it means to make 'Adequate Yearly Progress' under NCLB. *Education sector explainers*, July 2007.

C. S. Ding and M. L. Davison. A longitudinal study of math achievement gains for initially low achieving students. *Contemporary Educational Psychology*, 30(1):81–95, 2005.

District of Columbia Office of the State Superintendent of Education. Assessment and Accountability Manual, 2011. URL `https://osse.dc.gov/publication/district-columbia-assessment-and-accountability-manual`.

District of Columbia Public Schools. DCPS Data Set - DC CAS, 2018a. URL `https://dcps.dc.gov/publication/dcps-data-set-dc-cas`.

District of Columbia Public Schools. DCPS at a Glance: Enrollment, 2018b. URL `https://dcps.dc.gov/page/dcps-glance-enrollment`.

District of Columbia Public Schools. Our Schools, 2018c. URL https://dcps.dc.gov/page/our-schools.

R. Eberts, K. Hollenbeck, and J. Stone. Teacher performance incentives and student outcomes. *Journal of Human Resources*, pages 913–927, 2002.

M. Evans. What does knee surgery cost? Few know, and that's a problem. *Wall Street Journal*, (August 21), 2018. URL https://www.wsj.com/articles/what-does-knee-surgery-cost-few-know-and-thats-a-problem-1534865358.

D. N. Figlio and L. W. Kenny. Individual teacher incentives and student performance. *Journal of Public Economics*, 91(5):901–914, 2007.

E. S. Fisher, J. E. Wennberg, T. A. Stukel, and S. M. Sharp. Hospital readmission rates for cohorts of Medicare beneficiaries in Boston and New Haven. *New England Journal of Medicine*, 331(15):989–995, 1994.

A. J. Forster, I. Stiell, G. Wells, A. J. Lee, and C. Van Walraven. The effect of hospital occupancy on emergency department length of stay and patient disposition. *Academic Emergency Medicine*, 10(2):127–133, 2003.

N. Freeman, M. Zhao, and S. Melouk. An iterative approach for case mix planning under uncertainty. *Omega*, 76:160–173, 2018.

R. G. J. Fryer. Financial incentives and student achievement: Evidence from randomized trials. *The Quarterly Journal of Economics*, 126(4):1755–1798, 2011.

R. G. J. Fryer. Teacher incentives and student achievement: Evidence from New York City public schools. *Journal of Labor Economics*, 31(2):373–407, 2013.

V. R. Fuchs. *The Service Economy*. National Bureau of Economic Research, New York, 1968.

D. Fudenberg, B. Holmstrom, and P. Milgrom. Short-term contracts and long-term agency relationships. *Journal of economic theory*, 51(1):1–31, 1990.

P. C. Fuloria and S. A. Zenios. Outcomes-adjusted reimbursement in a health-care delivery system. *Management Science*, 47(6):735–751, 2001.

B. S. Georgopoulos and F. C. Mann. *The community general hospital*. Macmillan, 1962.

U. Gneezy, S. Meier, and P. Rey-Biel. When and why incentives (don't) work to modify behavior. *Journal of Economic Perspectives*, 25(4):191–210, 2011.

A. Y. Ha. Supplier-buyer contracting: Asymmetric cost information and cutoff level policy for buyer participation. *Naval Research Logistics (NRL)*, 48(1):41–64, 2001.

E. Hanushek. The economics of schooling: Participation and performance. *The Journal of Economic Literature*, 24(3):1141–77, 1986.

E. A. Hanushek. *The education of Negroes and whites*. PhD thesis, Massachusetts Institute of Technology, 1968.

E. A. Hanushek. Conceptual and empirical issues in the estimation of educational production functions. *Journal of human Resources*, pages 351–388, 1979.

E. A. Hanushek. Education production functions. In *The Economics of Education*, pages 161–170. Elsevier, 2020.

Healthcare Cost and Utilization Project. Overview of the Nationwide Readmissions Database (NRD), 2016. URL `https://www.hcup-us.ahrq.gov/nrdoverview.jsp`. Retrieved on Feb 27, 2019.

J. Helm and M. Van Oyen. Design and optimization methods for elective hospital admissions. *Operations Research*, 62(6):1265–1282, 2014.

J. Helm, S. AhmadBeygi, and M. Van Oyen. Design and analysis of hospital admission control for operational effectiveness. *Production and Operations Management*, 20(3):359–374, 2011. ISSN 1937-5956.

S. Hof, A. Fügener, J. Schoenfelder, and J. O. Brunner. Case mix planning in hospitals: a review and future agenda. *Health Care Management Science*, 20(2):207–220, 2017.

C. K. Jackson. A little now for a lot later a look at a texas advanced placement incentive program. *Journal of Human Resources*, 45(3):591–639, 2010.

S. M. Johnson. Merit pay for teachers: A poor prescription for reform. *Harvard Educational Review*, 54(2):175–186, 1984.

U. S. Karmarkar and R. Pitbladdo. Service markets and competition. *Journal of Operations Management*, 12(3-4):397–411, 1995.

D. S. Kc and C. Terwiesch. An econometric analysis of patient flows in the cardiac intensive care unit. *Manufacturing & Service Operations Management*, 14(1):50–65, 2012.

J. C. King, P. A. Manner, D. L. Stamper, D. C. Schaad, and S. S. Leopold. Is minimally invasive total knee arthroplasty associated with lower costs than traditional TKA? *Clinical Orthopaedics and Related Research*, 469(6):1716–1720, 2011.

A. Klein. No child left behind: An overview. *Education Week*, April 10, 2015. URL `http://www.edweek.org/ew/section/multimedia/no-child-left-behind-overview-definition-summary.html`.

R. J. Kusters and P. M. Groot. Modelling resource availability in general hospitals de-

sign and implementation of a decision support model. *European Journal of Operational Research*, 88(3):428–445, 1996.

K. L. Leonard and J. G. Zivin. Outcome versus service based payments in health care: lessons from african traditional healers. *Health economics*, 14(6):575–593, 2005.

H. M. Levin. Measuring efficiency in educational production. *Public Finance Quarterly*, 2 (1):3–24, 1974.

S. D. Levitt, J. A. List, and S. Sadoff. The effect of performance-based incentives on educational achievement: Evidence from a randomized experiment. Technical report, National Bureau of Economic Research, 2016.

T. R. Lewis and D. E. Sappington. Information management in incentive problems. *Journal of political Economy*, 105(4):796–821, 1997.

R. L. Linn. Assessments and accountability. *Educational researcher*, 29(2):4–16, 2000.

N. Liu, V.-A. Truong, X. Wang, and B. Anderson. Integrated scheduling and capacity planning with considerations for patients' length-of-stays. *Production and Operations Management*, 0(ja), 2019.

E. F. Long and K. S. Mathews. The boarding patient: Effects of ICU and hospital occupancy surges on patient flow. *Production and Operations Management*, Forthcoming, 2017.

H. Lutze and Ö. Özer. Promised lead-time contracts under asymmetric information. *Operations Research*, 56(4):898–915, 2008.

G. Ma and E. Demeulemeester. A multilevel integrative approach to hospital case mix and capacity planning. *Computers & Operations Research*, 40(9):2198–2207, 2013.

J. Matos and P. P. Rodrigues. Modeling decisions for hospital bed management. In *Proceedings of the International Conference on Health Informatics (Healthinf-2011)*, pages 504–507, 2011.

P. H. Mitchell and S. M. Shortell. Adverse outcomes and variations in organization of care delivery. *Medical Care*, pages NS19–NS32, 1997.

R. Murnane and D. Cohen. Merit pay and the evaluation problem: Why most merit pay plans fail and a few survive. *Harvard Educational Review*, 56(1):1–18, 1986.

K. Nagpal, S. Arora, A. Vats, H. W. Wong, N. Sevdalis, C. Vincent, and K. Moorthy. Failures in communication and information transfer across the surgical care pathway: interview study. *BMJ Quality & Safety*, 21(10):843–849, 2012.

D. M. Nair, J. J. Fitzpatrick, R. McNulty, E. R. Click, and M. M. Glembocki. Frequency of nurse–physician collaborative behaviors in an acute care hospital. *Journal of Interprofessional Care*, 26(2):115–120, 2012.

D. Neal and D. W. Schanzenbach. Left behind by design: Proficiency counts and test-based accountability. *The Review of Economics and Statistics*, 92(2):263–283, 2010.

P. Oreopoulos. Do dropouts drop out too soon? wealth, health and happiness from compulsory schooling. *Journal of public Economics*, 91(11-12):2213–2229, 2007.

J. Patrick, M. L. Puterman, and M. Queyranne. Dynamic multipriority patient scheduling for a diagnostic resource. *Operations Research*, 56(6):1507–1525, 2008.

M. Perie, S. Marion, B. Gong, and J. Wurtzel. The role of interim assessments in a comprehensive assessment system. *Washington, DC: The Aspen Institute*, November 2007.

E. L. Plambeck and S. A. Zenios. Performance-based incentives in a dynamic principal-agent model. *Manufacturing & Service Operations Management*, 2(3):240–263, 2000.

S. Rath, K. Rajaram, and A. Mahajan. Integrated anesthesiologist and room scheduling for surgeries: Methodology and application. *Operations Research*, 65(6):1460–1478, 2017.

K. Raugust, K. Mickey, and K. Meaney. PreK-12 testing market forecast 2019-2020, 2019.

J. Riccio, N. Dechausay, C. Miller, S. Nuñez, N. Verma, and E. Yang. Conditional Cash Transfers in New York City: The Continuing Story of the Opportunity NYC-Family Rewards Demonstration., 2013.

D. L. Richter and D. R. Diduch. Cost comparison of outpatient versus inpatient unicompartmental knee arthroplasty. *Orthopaedic Journal of Sports Medicine*, 5(3): 2325967117694352, 2017.

S. G. Rivkin, E. A. Hanushek, and J. F. Kain. Teachers, schools, and academic achievement. *Econometrica*, 73(2):417–458, 2005.

G. Roels. Optimal design of coproductive services: Interaction and work allocation. *Manufacturing & Service Operations Management*, 16(4):578–594, 2014.

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

A. V. Roth and R. Van Dierdonck. Hospital resource planning: concepts, feasibility, and framework. *Production and Operations Management*, 4(1):2–29, 1995.

P. Shi, J. Helm, J. Deglise-Hawkinson, and J. Pan. Timing it right: Balancing inpatient congestion versus readmission risk at discharge, 2019. URL `https://dx.doi.org/10.2139/ssrn.3202975`.

S. Shortell, S. Becker, and D. Neuhauser. The effects of management practices on hospital efficiency and quality of care. *Organizational Research in Hospitals*, ed. SM Shortell and M. Brown. Chicago: Inquiry Books, Blue Cross Association, 1976.

R. A. Shumsky and E. J. Pinker. Gatekeepers and referrals in services. *Management Science*, 49(7):839–856, 2003.

M. G. Springer, D. Ballou, L. Hamilton, V.-N. Le, J. R. Lockwood, D. F. McCaffrey, M. Pepper, and B. M. Stecher. Teacher pay for performance: Experimental evidence from the Project on Incentives in Teaching (POINT). *Society for Research on Educational Effectiveness*, 2011.

T. F. Tan and S. Netessine. When does the devil make work? an empirical study of the impact of workload on worker productivity. *Management Science*, 60(6):1574–1593, 2014.

P. Todd and K. I. Wolpin. Accounting for mathematics performance of high school students in mexico: Estimating a coordination game in the classroom. *Forthcoming Journal of Political Economy*, 2018.

B. Topol, J. Olson, E. Roeber, and P. Hennon. Getting to higher-quality assessments: Evaluating costs, benefits and investment strategies, 2012.

United States General Accounting Office. Title I: Characteristics of Tests Will Influence Expenses; Information Sharing May Help States Realize Efficiencies. (GAO-03-389), May 2003.

U.S. Department of Education. Race to the Top: Executive Summary, November 2009.

G. P. Whitaker. Coproduction: Citizen participation in service delivery. *Public administration review*, pages 240–246, 1980.

M. Xue and J. M. Field. Service coproduction with information stickiness and incomplete contracts: Implications for consulting services design. *Production and Operations Management*, 17(3):357–372, 2008.

M. Xue and P. T. Harker. Customer efficiency: Concept and its impact on e-business management. *Journal of Service Research*, 4(4):253–267, 2002.

M. Xue, L. M. Hitt, and P. T. Harker. Customer efficiency, channel usage, and firm performance in retail banking. *Manufacturing & Service Operations Management*, 9(4):535–558, 2007.

V. M. Young and D. H. Kim. Using assessments for instructional improvement: A literature review. *education policy analysis archives*, 18:19, 2010.

C. Zacharias and M. Pinedo. Managing customer arrivals in service systems with multiple identical servers. *Manufacturing & Service Operations Management*, 19(4):639–656, 2017.

H. Zhang and S. Zenios. A dynamic principal-agent model with hidden information: Sequential optimality through truthful state revelation. *Operations Research*, 56(3):681–696, 2008.