INSEAD The Business School for the World®

Working Paper

2025/31/TOM

Optimizing Service Encounters: A Co-Productive, Experiential, and People-Centric Approach

Guillaume Roels
INSEAD, guillaume.roels@insead.edu

Forthcoming in *Foundations and Trends in Technology, Information, and Operations Management* 19, 4 (2025)

Most value creation in services takes place in service encounters — at the interfaces between customers, employees, and service organizations. However, managers are often unable to effectively optimize them, because their tools either fail to fully comprehend these interfaces, lying therefore on the fringes of the value creation process, or predate digital technologies. This monograph reviews the recent development of three levers for optimizing service encounters, which lie at the core of the service value creation process and are relevant in a digital world. These levers are: leveraging co-production to innovate in service design; delighting customers through experience design; and fostering employee engagement by putting people first. Given today's abundant datasets and short feedback loops enabling scientific experimentation, we argue that the time is ripe for effectively optimizing service encounters.

Keywords: Service Management; Co-production; Experience Design; People-centric Operations; Literature Review; Frameworks

1

Introduction

Service encounters lie at the core of the value creation process in services. As illustrated in the "service triangle" depicted in Figure 1.1, they involve complex interactions between customers, employees, and service organizations. Moreover, digital technologies, such as Information and Communication Technologies (ICTs), have played an increasingly important role in channeling these interactions.

Service encounters can be optimized to be more effective and efficient. This is very much needed, as discussed in §1.1, given the importance of services to the economy, the crises they face, and the variety of new service encounter designs enabled by digital technologies — which creates greater room for optimization.

However, as we argue in §1.2, managers often lack the tools to optimize service encounters due to two reasons. First, traditional approaches to service management too often attempt to mimic approaches developed for manufacturing. Consequently, they fail to fully comprehend the interfaces between customers, employees, and service organizations, wrongly assuming that services can, like manufacturing, operate efficiently through functional division. Second, the few approaches to service management that embrace these interfaces, such as the Service

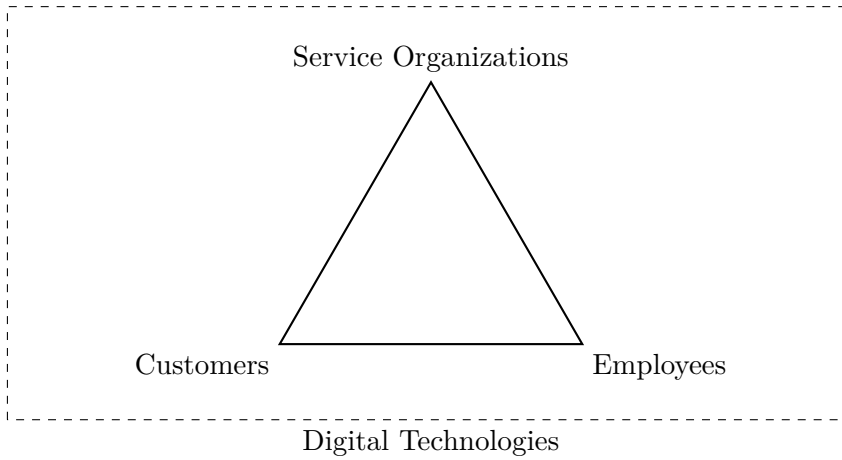


Figure 1.1: The Service Triangle (Adapted from Frei and Morriss (2012)).

Profit Chain (SPC), often predate the digitalization of ICTs, limiting their relevance today.

This monograph introduces three tools for optimizing service encounters, which lie at the interfaces between customers, employees, and the service organization, and are relevant in a world of digital technologies. These tools are:

- leveraging co-production to innovate in service design;
- delighting customers through experience design; and
- fostering employee engagement by putting people first.

We introduce these three levers in §1.3 and argue that the large amounts of data and the short feedback loop generated by digital technologies facilitate the adoption of a scientific method, offering novel opportunities for researchers and managers to experiment with new service designs and receive quick feedback from them. In sum, the time is ripe for optimizing, or even "engineering" (Shostack, 1987), service encounters.

1.1 A Growing Need for Optimizing Service Encounters

There is a growing need for optimizing service encounters given the importance of services to the economy (§1.1.1), the crises affecting many of them (§1.1.2), and the variety of new service encounter designs enabled by digital technologies (§1.1.3).

1.1.1 Importance of Services in Today's Economy

Services account for more than 80% of the GDP in the US and the UK (Statistics Times, 2024). They are by far the largest sector of activity in all advanced economies, as well as in an increasingly larger number of emerging economies. In particular, China, which is often considered to be the world's manufacturing bedrock, generates more than half of its GDP from services. Following this GDP pattern, most jobs today are in services. For instance, in the US, more than 80% of the jobs are service jobs (Ortiz-Ospina and Lippolis, 2017).

These statistics, based on the ISIC industry classification, actually understate the importance of services to the economy. In reality, all activities across a manufacturing value chain heavily depend on various services: in design (*e.g.*, R&D, engineering, design, finance), manufacturing (*e.g.*, product development, quality control, consulting, finance), selling and distribution (*e.g.*, marketing, advertising, transportation, warehousing, retail), and consumption and customization (*e.g.*, servicing, software, training, product support). Bryson and Daniels (2010) refer to the blending of services into the manufacturing economy as a "manuservice" economy.

Pushing this logic further, one could even argue that services are really the fundamental basis of exchange. This is one of the core premises of the Service-Dominant Logic (Vargo and Lusch, 2004), which posits that all economies are service economies. In broad strokes, the Service-Dominant Logic can be described with the following metaphor: "appliance makers really sell holes, not drills." With that logic, any firm, selling goods or services, is into services. That is, services is not a distinct economic sector, it is the entire economy. Along the same lines, Pine and Gilmore (2011) propose a theory of the progression of economic value

from extracting commodities, to making goods, evolving into delivering services, to ultimately competing on staging experiences. While the 19th and first half of the 20th centuries were dominated by manufacturing, the last decades have been dominated by services and experiences. As summed up by Teboul (2006, p. 14), "we are all into services, more or less."

1.1.2 Services Are in Crisis

However important to the economy, services are in crisis. Besides the crises affecting the entire economy (*e.g.*, ageing of populations, immigration, rising inequalities, tariffs and trade barriers) and humanity (*e.g.*, climate change, pandemics, wars), we identify three crises that are specific to services. We articulate these three crises around the three poles of value creation in service encounters, depicted in Figure 1.2: service organizations face rampant costs, customers receive poor experiences, and employees are disengaged or resign from service jobs.

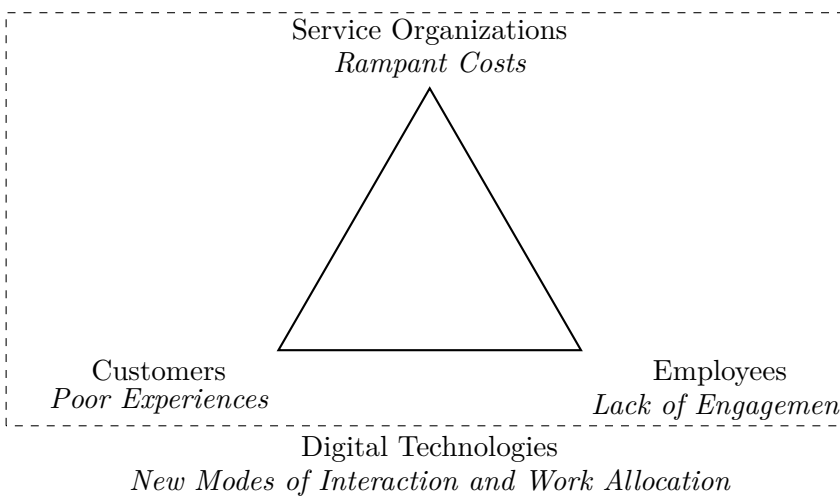


Figure 1.2: Three Crises and One Opportunity.

Service Organizations: Rampant Costs

Many service organizations (*e.g.*, education, healthcare) have recently faced rampant costs, potentially leading to a bubble burst in these sectors; see Figure 1.3. Baumol and Bowen (1965) and Baumol (1993) offer a theory, casually called "Baumol's cost disease," explaining why this may be the case.

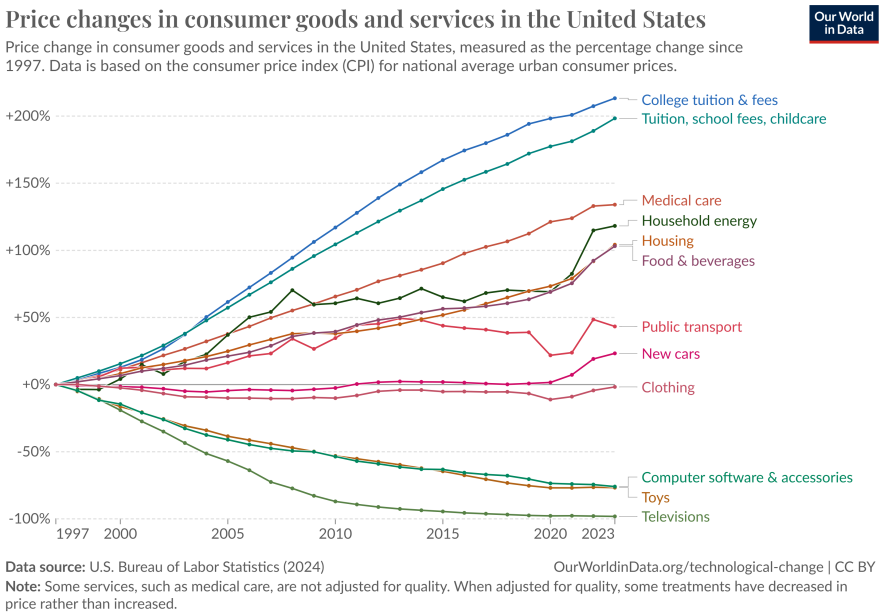


Figure 1.3: Price Changes in Consumer Goods and Services in the United States (Global Change Data Lab, 2024).

Their theory can be outlined as follows: The manufacturing and agricultural sectors, while experiencing large productivity gains through industrialization, can afford paying higher salaries to keep attracting top talent. To remain competitive in the labor market, the service sector, even though it has experienced limited (if any) productivity gains (*e.g.*, symphony orchestras, barber shops), needs to adjust upwards its salaries, in alignment with the salaries in agriculture and manufacturing.

As a result, and somewhat paradoxically, the greater importance of services in today's economies is largely due to their slower productivity

gains relative to agriculture and manufacturing (Karmarkar *et al.*, 2015).

Customers: Poor Experience

A second crisis affecting services is the delivery of poor customer experiences. In 2022, the American Customer Satisfaction Index, which tracks quality of economic output (products and services) in the US (as gauged by the consumers of that output), was lower than in 1994, after a five-year decline (ACSI, 2024). It has recovered since then, but it remains to be seen whether this recovery is sustainable.

This stagnation of customer experience somewhat conflicts with the theory developed by Pine and Gilmore (2011), which predicted that staging experiences would become the ultimate battleground competition. If so, one would have expected a race to the top, with service organizations continuously introducing innovations to delight their customers. While some service standards may have raised (*e.g.*, same-day delivery), customer satisfaction has instead declined in the 1994-2022 time span.

Employees: Lack of Engagement

A third crisis affecting services is the lack of engagement and shortage of service employees. After the COVID-19 pandemic, many of them indeed decided to quit their job, leading to the so-called Great Resignation. In the US alone, 40% of employees were thinking about leaving their jobs within the next three to six months in 2022 (Flynn, 2024). The sectors that have been the most affected were accommodation and food services, leisure and hospitality, and retail — all in service industries.

To add insult to injury, the growth of platforms has led to more precarious working conditions. Given that urban gig work is predominantly carried out by migrant workers (Katta *et al.*, 2024), the deterioration of their working conditions raises fundamental societal questions.

In white-collar offices, the burnout rate has reached unprecedented high levels. According to a 2020 Gallup survey, 76% of employees experience burnout on the job at least sometimes, and 28% say they are burned out “very often” or “always” at work (Wigert, 2020).

1.1.3 Service Design Opportunities Enabled by Digital Technologies

In parallel to these three crises, the development of digital technologies, and in particular ICTs, has offered novel ways to optimize service encounters. Prior to their development, service encounters mostly involved direct synchronous employee-customer interactions in person or over the phone.

In contrast, today's encounters are often indirect, asynchronous, and taking place over digital channels. They may involve one, several, or even communities of customers, employees, and service organizations (*e.g.*, platforms). Because of the low marginal cost of digital technologies, customer experiences can easily be individualized or dynamically adjusted as conditions evolve.

The proliferation of channels has reshaped the role of the customer: On the one hand, their role has been expanded, as extensively discussed by Field (2024); on the other hand, automation has taken away some of their jobs. In the near future, there is no doubt that Artificial Intelligence (AI) agents will also become part of the equation, working with or instead of customers, employees, and service organizations (Sampson, 2021).

In short, digital technologies offer many more degrees of freedom for designing service encounters, creating more room for optimizing them.

1.2 Effective Approaches to Service Encounter Optimization

Despite the great need for optimizing service encounters, managers often lack the tools to do so for two reasons.

First, many approaches to service management tend to lie on the fringes of the core process of service value creation, failing to capture the interfaces between customers, employees, and service organizations. In contrast to manufacturing, which operates efficiently under functional specialization, service encounters often require a truly integrative approach, comprehending these interfaces (§1.2.1); accordingly, effective approaches to service encounter optimization need to be multidisciplinary.

Second, the few service management approaches that truly comprehend these interfaces, such as the celebrated SPC framework, often

predate digital technologies, shedding little light beyond direct, synchronous, in-person service encounters (§1.2.2).

1.2.1 Integrative, Multi-Disciplinary Approaches

Because value creation in services involves customers, employees, and the service organization (Figure 1.1), managing services calls for integrative and multi-disciplinary approaches, blending Operations Management (OM), Marketing, and Organizational Behavior (OB), in contrast to the management of manufacturing activities, which can effectively operate through functional division.

Towards a Grand Theory of Services?

It may be puzzling to compare the extent of knowledge on manufacturing relative to that on services. Knowledge on manufacturing operations is both broad and deep, ranging across scientific management and productivity, inventory theory, facility layout and location, production planning and scheduling, quality management, logistics, and supply chain integration and coordination. Given this large body of work, today's supply chains are extremely lean, factories run at a high level of automation, and digital twins allow managers to anticipate rare events.

In contrast, one may argue that research on services is still in its early days, with some researchers still debating (at least until recently) on their defining nature (Sampson and Froehle, 2006), and many services facing rampant costs instead of productivity gains (§1.1.2).

The reason why it has been more challenging to develop a grand theory of services is that they are much more heterogeneous than manufacturing. In manufacturing activities, the same principles govern the logistics of moving boxes of toys, boxes of electronics, and boxes of garment. What matters is the box, and not what is inside it. In contrast, the world of services encompasses sectors as varied as education, retail, leisure and hospitality, consulting, utilities, transportation and freight, entertainment and media, online consumer services, healthcare, and governmental services. How could such a varied set of industries be governed by a unified set principles?

Could Manufacturing Approaches Apply to Services?

Given the success of manufacturing theories, it could be tempting to apply them to services. While valid in certain cases (*e.g.*, staffing, congestion management, revenue management), one may argue that many issues tackled by manufacturing theories are rather peripheral to the core process of value creation in services, as they do not capture (by design) the interfaces between customers, employees, and service organizations (Figure 1.1).

Applying to services manufacturing theories, without adaptation, may be ineffective because services have traditionally been defined as almost the opposite of manufactured products, namely, as intangible, heterogeneous, inseparable, and perishable (Parasuraman *et al.*, 1985) — IHIP for short — which starkly contrasts with the tangible, uniform, decoupling, and storable nature of boxes.

Yet, some manufacturing approaches may be useful. For instance, lean management approaches, which have been transformational in manufacturing, could also inspire service managers to transform their organizations (Womack and Jones, 2015). Still, they require adaptation to fully embrace the interfaces of the service encounter (Figure 1.1).

Functional Division

The early success of research on manufacturing has led to a functional division of many organizations and research institutions, following the traditional value chain decomposition (Porter, 1985).

This functional division has deeply affected research on service management, preventing it from fully capturing the interfaces between customers, employees, and service organizations, and sometimes creating functional silos. On the one hand, the Marketing community has made important developments in the mapping of customer journeys (Shostack, 1984) and in the measurement of service quality and customer satisfaction (Parasuraman *et al.*, 1985; Parasuraman *et al.*, 1988). On the other hand, the OM community has developed a large body of knowledge on waiting line management and queuing theory (Erlang, 1909), service facility layout and capacity planning, staffing planning and scheduling, and, more recently, pricing and revenue management

(Talluri and Van Ryzin, 2006). Comparing MBA course outlines between Marketing (Wirtz and Lovelock, 2021) and OM (Bordoloi *et al.*, 2022) reveals little overlap in topics.

These functional silos in service management could have in fact been self-inflicted. Indeed, the early research on service management, reviewed in §2.1.1, called for decoupling service operations from customers, accentuating the divide in perspectives between Marketing and OM.

While functional specialization is effective in manufacturing, given that the market boundary separating production from consumption is well defined, it is suboptimal in services, where the market boundary is blurred (Karmarkar and Roels, 2015).

As depicted in Figure 1.1, the core value creation process in services — service encounters — happens at the interfaces between customers, employees, and service organizations. Accordingly, managerial tools for service management can be effective only if they adopt a multi-disciplinary approach.

A Notable Integrative Framework: The Service Profit Chain

A notable exception to the aforementioned functional division in service research is the SPC framework, which cuts across OM, Marketing, and OB (Heskett *et al.*, 1994). The SPC, depicted in Figure 1.4, posits that employee satisfaction drives service quality, which itself drives customer satisfaction, and the latter drives the service organization's profitability and revenue growth.

The SPC framework is multi-disciplinary. On the one hand, its right part falls under the realm of Marketing: Service quality drives customer satisfaction, which then drives their loyalty, and ultimately leads to revenue growth and profitability. Parasuraman *et al.* (1988) and Cronin Jr and Taylor (1992) identify drivers of customer satisfaction as quality gaps, building on the early expectancy-disconfirmation theory (Oliver, 2014).

On the other hand, its left part falls under the realm of OM and OB: Many operational and OB levers, such as workplace design, job design, employee selection and development, employee rewards and recognition,

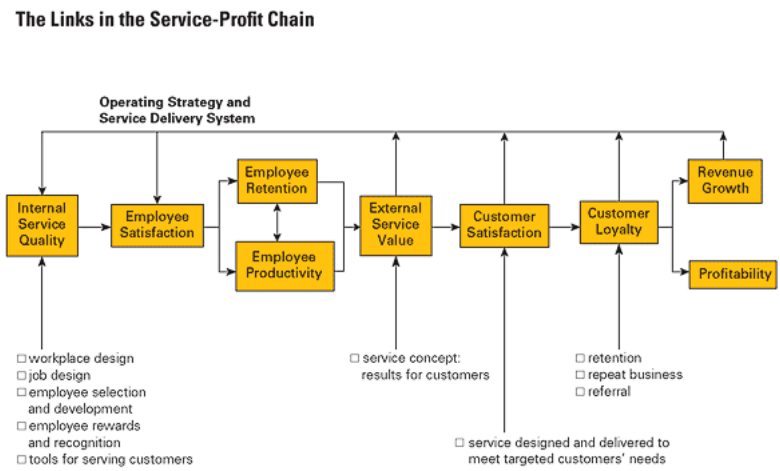


Figure 1.4: The Service-Profit Chain (Heskett *et al.*, 1994).

and tools for serving customers, structure the construct of internal service quality, which is then posited to drive employee satisfaction, or, perhaps more accurately, employee engagement (Heskett and Sasser, 2010) or well-being (Hogreve *et al.*, 2022).

The SPC framework truly lies at the core of the service value creation process as it fully captures the interfaces between customers (external service value), employees (internal service value), and service organizations (profitability and revenue growth); see Field (2024).

An important dimension of the framework is its dynamic nature: The upper backward loop is posited to fuel the internal service quality and employee satisfaction from revenue growth, customer loyalty, customer satisfaction, and external service value. This is truly a "flywheel" effect, epitomized by the classical case studies on Shouldice (Heskett, 2003) and Mercadona (Ton and Harrow, 2010).

1.2.2 Relevance in a Digital World

Because the SPC framework predates the development of digital technologies, its relevance in today's digital world is limited. Indeed, it tends to focus on synchronous employee-customer interactions (in person or over the phone), which were prevailing at the time of its publication.

However, as discussed in §1.1.3, today's service encounters are often indirect, asynchronous, and digital, instead. Consequently, the SPC framework can hardly explain why tech-enabled service companies like Netflix, Zoom, Warby Parker, T-Mobile consistently achieve high Net Promoter Scores (Tessitore, 2023) or the benefit for traditional people-intensive service organizations, such as KFC China (Lin and Zhang, 2020), to offer digital experiences.

Still, we do not believe that the SPC framework should be entirely dismissed: It remains relevant for some people-intensive B2C services in leisure and hospitality, healthcare, and retail, among others. In fact, one could argue that, in reaction to the dehumanization of many service experiences stemming from their process digitalization and the (hyper)scaling of many service organizations, customers are in greater need than ever of authentic experiences, building relationships with empathetic service employees.

However, it is clear that the SPC framework needs to be adapted or new frameworks need to be developed to fully incorporate the deployment of digital technologies in the value creation process. Some recent work has gone in this direction: Hogreve *et al.* (2022) study how digital technologies, and in particular, AI have transformed the SPC, identifying multiple roles for AI such providing support in the left part of the SPC and enhancing customer experience in its right part. Also, Field (2024) discusses how the SPC, and more generally, the value creation process in services, is affected by the following three trends: the rapid pace of technology-enabled service innovations, the expanded role of the customer, and the increasing use of service inventory. These frameworks may keep evolving as new technologies continue to be developed and their use cases continue to be discovered, such as the role of Generative AI in customer care (Blackader *et al.*, 2025).

1.3 Optimizing Service Encounters

Given the importance of services today, the numerous crises they face, and the variety of service encounter designs enabled by digital technologies (§1.1), there is a greater need than ever to optimize service encounters. As discussed in §1.2, effective approaches need to be both

integrative, *i.e.*, lie at the interfaces between customers, employees, and service organizations, and relevant in a digital world. Shostack (1987) already envisioned the possibility to "engineer" services, but the digital technologies available today may now turn her vision into reality.

Today's abundance of data and short feedback loops are offering numerous opportunities for researchers and managers to adopt a scientific method for optimizing service encounters. Indeed, researchers and managers can more easily than ever experiment with new designs of service encounters and collect quick feedback from them. We next classify these opportunities as following either a descriptive or a prescriptive approach (§1.3.1) and then overview the three particular managerial levers studied in this monograph (§1.3.2).

1.3.1 Opportunities for Research Developments

Descriptive Research

From a descriptive standpoint, the large amounts of data generated by digital technologies make it easier to systematically test the causal links posited in service management frameworks, moving beyond the traditional methods of surveys and archival data (Kamakura *et al.*, 2002; Hogreve *et al.*, 2022). One can also more easily measure actual behaviors, moving beyond behavioral intentions, to capture novel relationships, nonlinear effects, and longitudinal effects, as well as identify the boundary conditions of service management frameworks (Hogreve *et al.*, 2017; Hogreve *et al.*, 2022).

Prescriptive Research

From a prescriptive standpoint, the abundance of data and the short feedback loops enabled by digital technologies have offered new ways to adopt data-driven approaches to optimize service delivery at the individual level and unlock value for customers, employees, and service organizations.

We envision an era where service encounters can be optimized for each individual, potentially dynamically — offering new opportunities for value creation. The situation faced by the field of service management

is analogous to what the field of pricing and revenue management experienced at the turn of the century: The science of pricing, which belonged mostly to the marketing and economics fields (Nagle *et al.*, 2023), became an engineering discipline (Talluri and Van Ryzin, 2006) once data became abundant and systems were developed to enable instantaneous price adjustments. For some applications, pricing decisions are now made at the individual level and re-optimized at high frequency, sometimes within milliseconds as in the case of online advertising.

Complementary Approaches

These descriptive and prescriptive approaches are complementary. Although some development of the science needs to precede the engineering of its findings, the development of engineering enables fast experimentation through which new scientific discoveries can be made. Research on service management should alternates between model development, empirical validation, and large-scale deployments.

1.3.2 Three Managerial Levers

This monograph aims to present three managerial levers for optimizing service encounters: co-production, experience design, and employee engagement. These approaches lie at the interfaces between customers, employees, and service organizations (Figure 1.5), and are thus integrative and naturally fit within the SPC (Figure 1.6). Moreover, these approaches fully embrace the development of digital technologies.

Although these three research themes are relatively recent, they build on large bodies of knowledge that preceded them. One of the objectives of this monograph is to make explicit these connections, by discussing the change in perspective from service researchers and managers that have led to the emergence of these themes. Accordingly, our approach will attempt to be comprehensive by offering a historical perspective on these themes and delineating the key lines of research that have shaped this evolution.

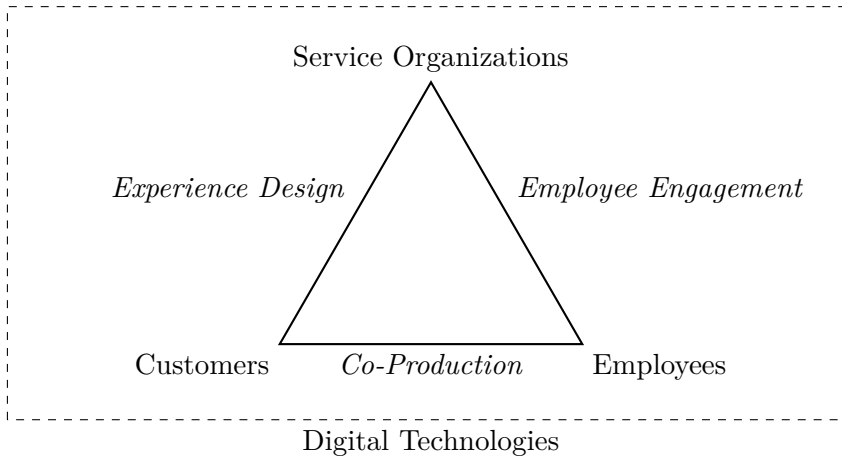


Figure 1.5: Mapping Co-Production, Experience Design, and Employee Engagement on the Service Triangle.

Leveraging Co-Production to Innovate in Service Design

Chapter 2 discusses how leveraging the co-productive nature of many services offers numerous opportunities for service design and innovation. Unlike manufacturing activities, which rely on a single productive resource (namely, the manufacturing organization), service organizations have access to two productive resources, namely, the service provider and the customer. This enables multiples service operating modes such as service factories, collaborative services, and self-services.

The chapter reviews the evolving perspective on the customer's role in service processes. While customers were perceived in the early days of research on service management as a nuisance from which the technical core should be decoupled, they are considered today as an asset that can enhance the effectiveness and efficiency of the service delivery. This chapter also covers different ways co-production has been analytically modeled in the literature, investigates the role of customers as partial employees, and discusses the operational implications of co-production in terms of contracting, capacity management, quality management and transparency, scaling and technology, and servicescape design.

Because the leveraging of co-production fundamentally changes the

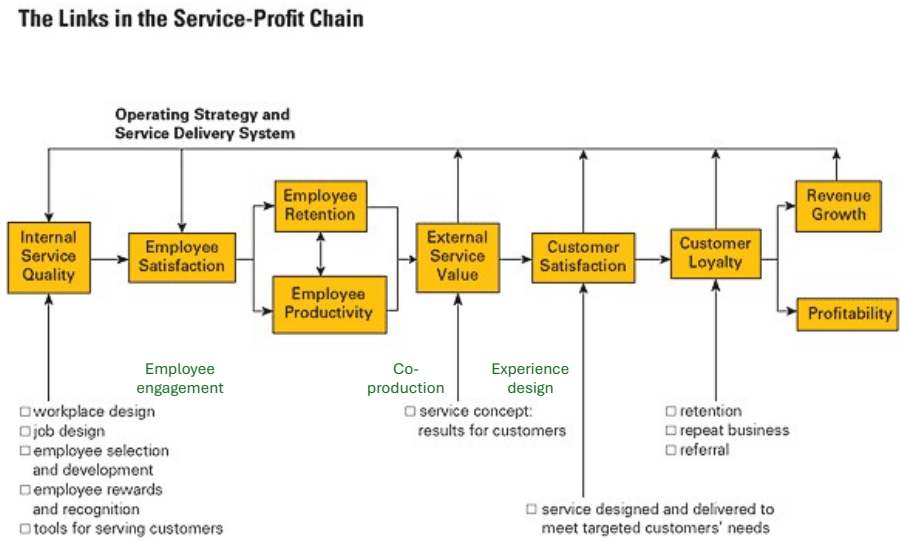


Figure 1.6: Mapping Co-Production, Experience Design, and Employee Engagement on the Service-Profit Chain.

nature of the employee-customer relationship by redefining their roles, we position the theme of co-production as a source of innovation in the design of service encounters along the edge linking employees and customers in Figure 1.5.

As shown in Figure 1.6, the co-productive nature of services lies at the cornerstone of the SPC between the left part focused on employees, and its right part focused on the customer.

We argue that co-productive processes can now be optimized, or even engineered, thanks to the availability of large datasets coming from digital technologies. For instance, Daw *et al.* (2020) and Daw and Yom-Tov (2024) model the dynamics of a conversation between a customer and a service provider (*e.g.*, a call center agent) using a two-sided Hawkes cluster model in which both parties make dynamic contributions. Using data from a company's contact center, consisting of 337,224 service conversations and a total of close to five million messages, they estimate with their model the degree of interaction (measured as synchronicity, *i.e.*, the heterogeneity in rate of processing

information) and work allocation (self-production vs. co-production) between customers and call center agents.

Engineering Experience Design to Delight Customers

Chapter 3 discusses how service experiences can be engineered to maximize customer satisfaction, by leveraging human biases that have been extensively documented in social psychology, such as memory decay and quality contrast. Sequence indeed matters: Would a customer prefer experiencing a low-quality activity before a high-quality one or the contrary? Any sequence is a choice, so not optimizing it is a missed opportunity to delight customers.

This chapter reviews the large literature on experience design focusing first on single-stage encounters, with a strong focus on the management of customer wait, building on queuing theory and the psychology of waiting, and moving then to multi-stage settings to incorporate principles of social psychology, before closing with considerations that go beyond a single encounter, through anticipation, recall, and customer journeys.

We position this theme of research in Figure 1.5 as lying between service organizations, which design experiences, and customers, who are its recipients.

On the SPC, depicted in Figure 1.6, service experience design lies between the provision of external service value and customer satisfaction given that different customers may react differently to the same experiences.

We argue that service experiences can now be engineered. For instance, Deshmane *et al.* (2023) scraped review data from three online platforms across the contexts of movie watching (IMDb), book reading (GoodReads), tourist attractions (TripAdvisor), and eating out (TripAdvisor). In particular, the TripAdvisor attractions data consists of the review histories of 13,426 users who reviewed 92,678 activities. Using this dataset, they calibrate an analytical model of user utility with intertemporal spillovers and offer a dynamic optimization problem to select the sequence of activities that maximizes a consumer's utility.

Fostering Employee Engagement by Putting People First

Chapter 4 identifies ways to foster employee engagement by putting people first. Traditional models of staffing, developed in the context of manufacturing activities, assume that people are interchangeable, have constant productivity, and are linearly scalable. However, employees have unique needs and wants, are subject to fatigue and burnout, and have complementary skills. Given the central role of employees in the service value creation process, as depicted in the SPC framework, it is essential to recognize these behavioral traits to foster their engagement and maximize their well-being along with the profitability of the service organization (Ton, 2014; Corbett, 2024).

This chapter reviews the changes in the research community and practice from treating employees as homogeneous inanimate objects (*e.g.*, human resources, servers in a queuing systems) to treating them as human beings responding to physiological and cognitive stimuli (*e.g.*, financial incentives, workload, values) in a rather predictable way. Specifically, it discusses some classical and more recent research on how people respond to incentives, financial or non-monetary, to motivate them to exert effort; how staffing models have evolved by adopting a more people-centric approach to guide how many and what kind of employees to have; different aspects of job design, such as the trade-off between specialization and variety, employee discretion, and the effect of workload, on employee productivity and engagement; and finally, the role of organizational culture in fostering their engagement.

We position this theme of research in Figure 1.5 as lying between service organizations, which design the employees' jobs and schedules, and employees, who are the recipients thereof.

On the SPC, depicted in Figure 1.6, fostering employee engagement lies between the node of internal service quality and the nodes of employee retention and productivity, as a substitute for employee satisfaction, given that employee engagement has been documented to be more predictive of their productivity (Heskett and Sasser, 2010).

We argue that fostering employee engagement can now be optimized. For instance, Meng *et al.* (2021) study the impact of service facility layout on how service workers organize their tasks in the context of a

hospital's emergency department. To do so, they utilize a unique dataset consisting of infrared nurse location tracking data, patient electronic medical record data, bedside call data, and the architectural floor plan of the emergency department. The location data is generated every six seconds and picked up by one of the 147 receivers located through the emergency department, amounting to four million location records created by more than 200 nurses over five months.

1.3.3 Scope

In the same spirit as the SPC, this monograph focuses on service encounters primarily within Business-to-Consumer (B2C) services. Although we will briefly mention customer journeys, which consist of numerous encounters, our approach tends to be more microscopic than the marketing literature on customer-centricity (Fader, 2020).

Also, we believe that some concepts extend to Business-to-Business (B2B) environments, such as consulting, education, and online services, although we will not venture into B2B topics. Relatedly, we will also ignore the possible interactions between a service and a related product, *i.e.*, Product-Service Systems (Mittal *et al.*, 1999; Mont, 2002), which have led to, on the one hand, servicization (Kowalkowski *et al.*, 2017) and, on the other hand, productization. Moreover, we ignore to a large extent the recent development on platforms (Taylor, 2018; Boehm *et al.*, 2020), which deserve a monograph on its own.

Moreover, our approach will essentially be operational and focused on prescriptions, assuming that a service concept has been defined and focusing on the engineering of its execution. In particular, we will not question the fundamental nature of an exchange as in the Service-Dominant Logic (Vargo and Lusch, 2004).

Finally, given that recent monographs have appeared on the shift of the US economy to information-intensive services (Apte *et al.*, 2012), and, in particular, on service industrialization (Nath *et al.*, 2020), and on worker productivity (Diwas, 2020), we will remain succinct in our review of these research developments and will refer the reader to these studies for further details.

2

Leveraging Co-Production to Innovate in Service Design

This chapter investigates the critical role of the customer in creating value. In contrast to the traditional view that customer's inputs impede service efficiency because of the uncontrolled variability that customers bring with them, this chapter advocates for leveraging customers as an additional productive resource to enhance both the effectiveness and the efficiency of the service delivery.

We first discuss in §2.1 the evolving perception of the customer role in services, from a pure service recipient that needs to be isolated or decoupled from the core production of the service to an active participant, bringing efficiency and/or effectiveness. Section 2.2 then proposes an analytical model of co-production that captures the different participative roles of the customers. Given customers' role as co-producers, they can be seen as partial employees and need to be managed accordingly — an interpretation we investigate in §2.3. Section 2.4 then overviews some operational implications of co-production in terms of contracting, capacity management, quality management and transparency, scaling and technology, and servicscape design. Finally, we offer in §2.5 some future research directions.

Throughout this section, we adopt the standard terminology of

referring to a service provider and a customer, acknowledging that it may not always be clear who really play the roles of providers and recipients of services, given the sometimes blurred market boundary between production and consumption (Karmarkar and Roels, 2015),

2.1 The Evolving Perception of the Customer's Role in Services

This section overviews the evolving perception of the role of customers in service production and delivery in the literature on service management, from a nuisance (§2.1.1) to a source of efficiency (§2.1.2), and then to a source of both efficiency and effectiveness (§2.1.3).

2.1.1 Customer as a Nuisance

Referring to the customer as a "nuisance" is arguably a strong word — in full disclosure, our own choice of word. Although we focus below on the customer contact model to make this point, this analogy has really permeated throughout the entire early literature on service management. This analogy stems from the application of manufacturing frameworks to services: Because variability reduction has demonstrated large productivity gains in manufacturing, the same was deemed to apply to services as well, as reviewed in Appendix §A.

The Customer Contact Model

The customer contact model was motivated by the challenge of dealing with customer variability (Frei, 2006) in services: "Clients ... pose problems for organizations ... by disrupting their routines, ignoring their offers for service, failing to comply with their procedures, making exaggerated demands, and so forth" (Danet, 1981, p. 384). Given that variability reduction had demonstrated large productivity improvements in manufacturing (Lovelock and Young, 1979; Lovelock, 1983), Chase (1978, p. 137) naturally inferred that "the less direct contact the customer has with the service system, the greater the potential of the system to operate at peak efficiency." To be efficient, services should attempt to eliminate customer contact, akin to flow shops in manufacturing (Appendix A).

Accordingly, Chase (1978) proposes to decouple the high-contact operations from the low-contact technical core. This decoupling strategy takes its roots in the seminal work by Thompson (1967), which led to the development of seminal frameworks, such as the product-process matrix (Hayes and Wheelwright, 1979) and the "plant within a plant" concept by Skinner (1974). This notion of decoupling is also very salient in the influential service blueprint framework (Shostack, 1984; Bitner *et al.*, 2008), which is structured around the line of visibility: activities hidden to customers are subject to efficiency, whereas activities that are visible are opportunities for customization. This leads Teboul (2006) to frame activities that involve interaction with customers as "front-stage" and to recommend that they be decoupled from the "back-stage" activities that can be run efficiently.

So, where does the customer fit in a service operation? According to the customer contact model, it should be outside.

Limitations of the Customer Contact Model

The customer contact model has been subject to some criticism. First, many online services, which are more prominent in today's digital world, are often both high touch (high contact) and high tech (low contact) — a contradiction with Chase's original classification. This apparent conundrum has called for a reframing of the notion of customer contact (Sampson and Chase, 2020).

More generally, one could argue that many self-services, some as simple and low-tech as a salad bar, are efficient (as very little labor is involved) while having a high degree of customer contact (Chase, 2010). This contradiction to the theory is blatant in Teboul (2006), who places cafeterias and self-service buffets in the lower-left corner of the product-process matrix (Figure A.1) — this would then imply that they operate suboptimally.

Also, the single-dimension mapping of the customer contact model may be too reductive. Schmenner (1986) indeed argues that hotels, which are categorized as a pure service by Chase (1978) — and are thus posited to be less efficient — are more efficiently run than the lower-contact postal services or repair shops. As a remedy, Schmenner (1986)

proposes to classify services, not on one, but along two dimensions: (i) their degree of labor intensity and (ii) a combined dimension of degree of customer interaction with the service and degree of customization of service for the customer. Later, Schmenner (2004) revised these dimensions as (i) relative throughput time (instead of labor intensity) and (ii) degree of variation (instead of customization and interaction), based on the theory of "swift, even flow" (Schmenner and Swink, 1998). Other frameworks that involve the notion of interaction at the core of the classification include Wemmerlöv (1990) and Kellogg and Nie (1995).

2.1.2 Customer as a Source of Efficiency

In our opinion, the inherent limitation in the customer contact model is that, influenced by the principles of manufacturing operations, it views customers as a nuisance (from which the technical core needs to be decoupled), and not as a productive resource. However, a defining feature of services is that their production process relies on customer inputs (see Figure A.2); see Sampson and Froehle (2006). A service has thus two distinct productive resources: the service provider and the customer.

One of the earliest references of customer participation is Fuchs (1968, pp. 194-195), who states that "in services... the consumer frequently plays an important role in production." The roles can be quite diverse, ranging from passive (sitting at the barber shop) to actual work (laundromat, supermarket), but the quality of the output also depends on the quality of information (medical history) and the customer education (banking, education). This leads Fuchs to conclude that efficiency is not necessarily negatively correlated with the customer's participation (unlike the customer contact model), but that the relationship between customer participation and efficiency is definitely moderated by the customer's "knowledge, experience, and motivation."

Lovelock and Young (1979, p. 177) go one step further, proposing that customers are more than a moderator — they can enhance efficiency: "If customers assume a more active role in the service production and delivery process, they effectively remove some of the labor tasks from

the service organization. There may be benefits for both consumers and service organizations." Similarly, Fitzsimmons (1985, p. 60) notes that "for services, the consumer represents an untapped productive resource. A service delivery system can be designed to permit greater consumer involvement in the production process thereby achieving productivity gains and revising the concept of employment."

2.1.3 Customer as a Source of Efficiency and Effectiveness

It is only in the past two decades that the service managers and researchers have really realized that customers truly lie at the core of the process of value creation. Accordingly, customers can not only bring greater efficiency, but also greater effectiveness.

We first review the early shift of emphasis on effectiveness and then present the Service-Dominant Logic, which posits that the customer is always a co-producer. This will lead us to introduce a three-dimensional product process matrix for services that account for both the role of the customer as providing efficiency and effectiveness. Because this three-dimensional matrix is hard to visualize, we will then project it into two dimensions, discuss the underlying tensions between the choices of operating modes and its implications for the delineation of the market boundary.

A Shift of Focus towards Effectiveness

The earlier perspective on customer participation emphasized efficiency over effectiveness. Indeed, the earlier works on the participatory roles of the customer reviewed in §2.1.2 associate customer involvement with "removing labor tasks" (Lovelock and Young, 1979, p. 177) or "achieving productivity gains" (Fitzsimmons, 1985, p. 60).

However, effectiveness should not be an afterthought. While self-services may be an efficient way to deliver value, they may not always be the most effective. Indeed, Field (2024) notes that, in contrast to ATMs, which are generally deemed to be both efficient and effective, self-checkouts in supermarkets are often a source of frustration, despite being efficient.

Conversely, effectiveness may stem from self-service. Norton *et al.* (2012) identify the so-called "IKEA" effect — an increase in valuation of self-made products. Hence, customers appear to derive greater value from a service if they have been active participants in its delivery — and have successfully completed of the task. In the context of restaurants, Tan and Netessine (2020) report that the introduction of tabletop technology increases sales per minute by more than 10%, due to (as expected) a shorter meal duration, *i.e.*, greater efficiency, but also because the average sales per check increases, *i.e.*, greater effectiveness.

Effectiveness matters even more when there is joint participation of the customer and the provider. In particular, Czerniawska and May (2004, p. 21) note that, based on a survey of award-winning consulting projects, "perhaps the single most common word across the projects that exceeded client expectations is 'together'. In fact, Czerniawska (2006) identifies a radical shift in the consulting industry towards value-based consulting after the turbulent events at the turn of the century (dot-com bubble burst, Enron collapse). In healthcare, clinicians and governments are increasingly advocating for shared decision-making processes, in which doctors and patients work together to choose among treatment options, to ensure that treatment decisions align with each patient's unique needs and preferences (Tuncalp *et al.*, 2023).

Is the Customer Always a Co-Producer?

The co-productive role of customers is a key pillar of the Service-Dominant Logic posited by Vargo and Lusch (2004). This theory envisions that service provision ("sell the hole"), rather than goods ("sell the drill"), is fundamental to economic exchange. The Service-Dominant Logic starkly contrasts with a traditional, goods-based, manufacturing perspective, which has typically separated the role of production and consumption, and not viewed the customer as a co-producer.

Given that the early literature on service management viewed the customer as a nuisance or at best as a source of greater efficiency (but not effectiveness), we agree with the Service-Dominant Logic that it is important to emphasize the co-productive nature of the role of the customer. However, we disagree with the generic statement that the

consumer is always involved in the production of value, at is may imply that production and consumption are always coupled. Instead, as we argue next, a lot of innovations in service designs consist in decoupling operations by reducing the extent of interaction between the producer and the customer, and shifting the work allocation to one or the other.

Hence, while the customer contact model framed the role of the customer as decoupled from the technical core, the Service Dominant Logic, perhaps as a reverse pendulum swing, tends to frame it as always a co-producer. In reality, most service design innovations happen between these two extremes.

A Three-Dimensional Product-Process Matrix

To fully account for the customer as a productive resource, on the same foot as the producer, Roels (2014, Figure 6) proposes to map services processes not along one, but two dimensions: the degree of interaction between the customer and the producer and the degree of task allocation.

For the product dimension, Roels (2014) considers, instead of the classical dimension of product variety/volume (Figure A.1), a dimension of process standardization, which is similar in spirit to the variability dimension proposed by Schmenner (2004). This is because offering a large product variety is challenging only if it entails an inefficient process customization. However, there exist standard processes (*e.g.*, salad bars) that can efficiently offer a wide variety of products.

The proposed three-dimensional matrix is depicted in Figure 2.1. While it is arguably more challenging to visualize than a 2×2 framework, it has the merit of positioning *self-services* (low-interaction task with work shifted towards the customer) as a counterpart to *service factories* (low-interaction task with work shifted towards the producer) to deal with standardized operations, and thus offering similar benefits in terms of short processing time and low cost, *i.e.*, efficiency.

A third operating mode that emerges from this conceptualization is the concept of *collaborative service*, which has a high degree of interaction and more balanced work allocation. There, the focus is not on efficiency, because two resources are involved in the delivery pro-

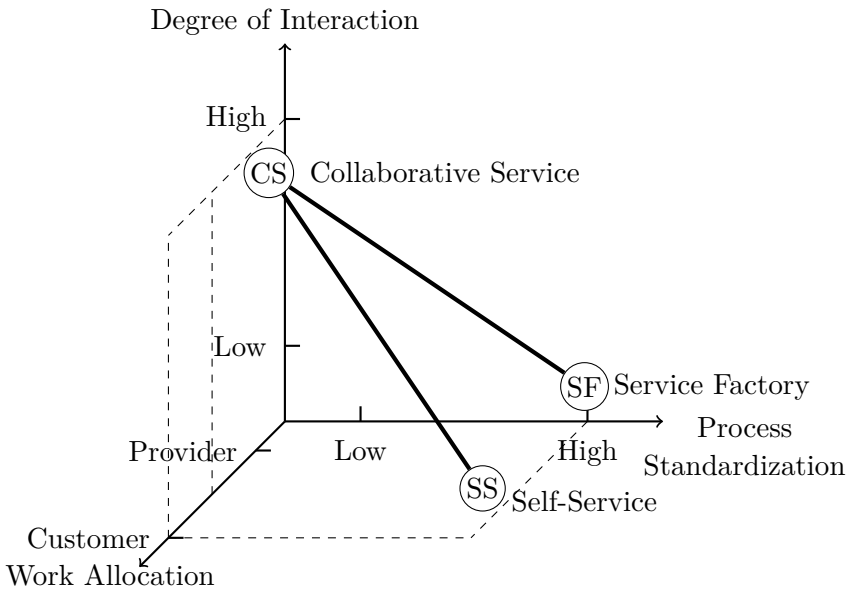


Figure 2.1: A Three-Dimensional Product-Process Matrix for Services (Based on Roels (2014, Figure 6)).

cess, which potentially involves communication and coordination costs. Instead, the focus of this operating mode should be on effectiveness.

This matrix departs from the product-process matrix in manufacturing (Hayes and Wheelwright, 1979) by not contrasting only two operating modes (namely, job shops vs. flow-shops, see Appendix A), but rather by offering a trinity of operating modes: self-services, service factories, and collaborative services.

Very much like the traditional product-process matrix (Figure A.1), this framework prescribes service organizations to match their process type with their desired degree of standardization. However, given the three-dimensionality of the matrix, there are two paths that can be followed, highlighted as thick lines in Figure 2.1.

A Two-Dimensional Projection

Although this three-dimensional matrix is difficult to visualize, its diagonal, which prescribes operating modes, can be projected into one

dimension, depicted in Figure 2.2.

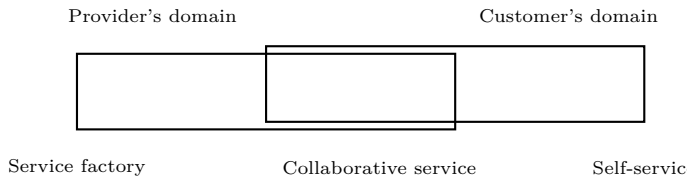


Figure 2.2: A Typology of Co-Productive Services Based on Roels (2014), and Analogous to Grönroos (2011, Figure 2) and Sampson and Chase (2022, Figure 1).

The framework depicts two intersecting domains of operations: the provider's domain (left) and the customer's domain (right). If most of the delivery happens in the provider's domain, the operating mode is a called a *service factory* (adopting the terminology by Schmenner (1986)); if it lies in the customer's domain, it is called a *self-service*; and if it belongs to their intersection, it is a *collaborative service*.

This visualization is in fact similar to some other frameworks presented elsewhere in the literature, such as Grönroos (2011, Figure 2) and Sampson (2012). The latter proposes to map processes according to this dual architecture, distinguishing the provider's domain from the customer's domain, and the intermediate regime of direct interaction. In the framework depicted in Figure 2.2, the intermediate regime of collaborative service does not require a direct interaction, *i.e.*, we do not preclude the co-production from occurring in an asynchronous manner. In another framework, Sampson and Chase (2022, Figure 1) distinguish five operating modes (standardized production, customized production, synchronous production, self-service production, and Do-It-Yourself production) and identify path towards servitization or deservitization. What distinguishes Roels (2014) from these earlier frameworks is that it is mathematically derived, *i.e.*, the existence of the diagonals in Figure 2.1 is proved to be optimal.

Economic Tensions

Sampson and Chase (2022) discuss the tensions that leads to favor one operating mode over another. The first tension is between economies

of scale and customization: Whereas moving towards the provider's domain creates opportunities for economies of scale and offers greater expertise, moving towards the customer's domain creates opportunities for greater customer control and customization potential. For instance, in a museum, customers personalize their experience by choosing a self-direct path based on their interests, which gives them an opportunity to be both cognitively and emotionally immersed and therefore engaged (Minkiewicz *et al.*, 2014; Aouad *et al.*, 2022).

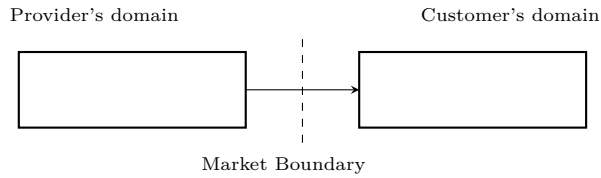
At the core of this first tension is a work allocation model: Who should perform the work? The customer or the provider? Task allocation has been extensively studied in services (Bellos and Kavadias, 2019; Bellos and Kavadias, 2021), and more generally, economics (Itoh, 2001), including models of delegation (Aghion and Tirole, 1997), property rights allocation (Grossman and Hart, 1986), and platforms (Hagiu and Wright, 2019). However, all these models consider a one-on-one relationship, namely between a single provider and a single customer. To fully capture this notion of economies of scale and customization, however, it seems necessary to develop a model with multiple providers and customers, *e.g.*, using a horizontal differentiation model (Hotelling, 1929; Salop, 1979). To the best of our knowledge, this remains to be done.

A second tension is between efficiency and effectiveness: Moving towards the center (collaborative service) offers an opportunity for an enhanced provider-customer relationship, whereas moving towards the extremes offers a greater potential for operating efficiency.

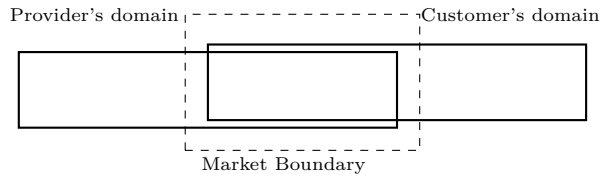
Market Boundary

Although the typology of co-productive services depicted in Figure 2.2 is a projection of the frameworks proposed by Roels (2014) and Sampson and Chase (2022), it has the merit of being analogous to the way supply chain processes are typically mapped, distinguishing production, distribution, and consumption.

In manufacturing, the provider's domain and the customer's domain are non-overlapping, giving rise to a well-defined market boundary (see Figure 2.3a): The provider defines its good or service (*e.g.*, electricity)



(a) Market boundary in manufacturing



(b) Market boundary in services

Figure 2.3: Market Boundaries in Services (Top) and Manufacturing (Bottom) Proposed by Karmarkar and Roels (2015).

as a vector of attributes and the customer pays a price to acquire it, without getting involved in the production of the good or service.

In contrast, with co-productive services (*e.g.*, consulting, education), customers are an integral part of the production process, in line with the Service-Dominant Logic (Vargo and Lusch, 2004), and the market transaction becomes much more challenging to characterize in objective terms, making the market boundary loosely defined as depicted in Figure 2.3b.

2.2 Analytical Models of Co-Production

This section presents some analytical models of co-production. While interacting, the customer and the provider can perform numerous actions, *e.g.*, make different choices and exert various levels of effort. Following the literature on team production (Alchian and Demsetz, 1972; Holmström, 1982) and the principal-agent literature (Holmström, 1979), different models have been proposed to capture the dynamics of interactions not only within the "collaborative service" mode lying at the center of Figure 2.2, but also as potentially shifting towards the extreme operating modes of "service factory" and "self-service."

We first discuss the different roles that the service provider and the customer can take (§2.2.1), then offer a generic representation of co-production (§2.2.2), overview the two most common solution approaches to characterize the existence of an equilibrium (§2.2.3) before investigating more specific analytical forms of co-production function (§2.2.4). We close this section by considering some mathematical conditions on the value function that leads to the emergence of different operating modes depicted in Figure 2.2 (§2.2.5).

2.2.1 Defining Roles

Although Figure 2.2 is symmetric, the relationship between a customer and a service provider is rarely symmetric. Typically, one party, usually the residual claimant of the transaction, initiates the transaction or engagement, hires the other party, sets the rules of collective action, and assigns roles in execution (Roels and Van der Heyden, 2025); in short, "manages or examines the ways to which inputs are used in team production" (Alchian and Demsetz, 1972, p. 782). Following the principal-agent framework (Alchian and Demsetz, 1972; Holmström, 1979), this party is usually called the "principal" and the other party is called the "agent." Often, the role of the principal is to specify the rules of engagement, *i.e.*, a contract.

However, it is not always clear, in a service context, who should play the roles of the principal and the agent, unlike an employment relationship, which has a clear hierarchy. Fortunately, specifying roles is inconsequential if one is only concerned about the total value created provided that the contract is offered in a take-it-or-leave-it fashion, as is common in principal-agent models (Holmström, 1979) and that the contract involves a fixed-fee transfer, which enables the principal to collect the total net surplus and leaves the agent with their reservation profit. Such two-part tariff contracts are second-best optimal under double moral hazard (Bhattacharyya and Lafontaine, 1995; Kim and Wang, 1998).

2.2.2 A Generic Representation of Co-Production

Inputs

For simplicity, suppose that inputs are unidimensional, and let the principal's input x and the agent's input y. For instance, these inputs could represent the time or effort invested in the relationship. With this interpretation, inputs are nonnegative and, given that time is nonfungible, bounded from above. Accordingly, it is often assumed in the literature that $x, y \in [0, 1]$. It is common to assume that these inputs are non-contractible.

Operating Modes

Assuming without loss of generality that the provider is the principal, we have the following operating modes, mapping to Figure 2.2:

- Collaborative Service: $x > 0$ and $y > 0$;
- Service Factory: $x > 0$ and $y = 0$;
- Self-Service: $x = 0$ and $y > 0$.

A final possibility is that $x = 0$ and $y = 0$, in which case no service is delivered.

Value

Let $V(x, y)$ denotes the output produced by combining these inputs. A key feature of co-production is that

$$\frac{\partial^2 V(x,y)}{\partial x \partial y} \neq 0 \tag{2.1}$$

(Alchian and Demsetz, 1972). Moreover, it makes sense to engage in co-production if

$$V(x, y) > V(x, 0) + V(0, y). \tag{2.2}$$

Hence, one way to differentiate services from manufacturing is whether $\partial^2 V(x, y)/(\partial x \partial y) \neq 0$ or not. In supply chain management, it is common to refer to the concept of "added value," which makes

explicit that the production function is additive. Instead, in services, we should refer more to some form of "super-additivity," and, therefore, "super-added value."

Together, Equations (2.1)-(2.2) are the best way to formalize the defining feature of services proposed by Sampson and Froehle (2006) and illustrated in Figure A.2: "in services, customers bring a significant input into the production process."

Sharing Rule

Once value is created, it needs to be shared between the principal and the agent. Considering a two-part tariff contract, let us denote with $\alpha \in (0, 1)$ the principal's share of output and f the fee paid by the agent to the principal. Accordingly, the principal receives $\alpha V(x, y) + f$, whereas the agent receives $(1 - \alpha)V(x, y) - f$. Figure 2.4 presents a simple depiction of the effort contributions and output sharing.

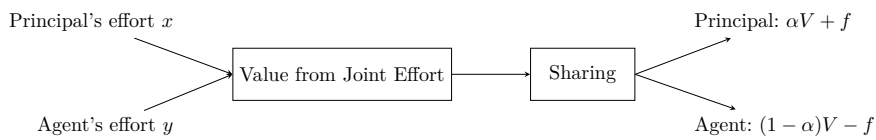


Figure 2.4: An Analytical Framework for Co-Production

The sharing of output is usually assumed to be financial, *e.g.*, through an equity contract or a revenue- or profit-sharing contract (Bhattacharyya and Lafontaine, 1995). However, it does not necessarily have to be. For instance, within an organization, the sharing of output can take the form of sharing authorship on a common piece of work (with perhaps different weights depending on the listed order), such as paper co-authorship or movie credits, or patent ownership (Roels *et al.*, 2024). In a consulting engagement, both parties can derive value from the knowledge acquired through the engagement (Xue and Field, 2008; Özkan-Seely *et al.*, 2015).

Costs of Effort

If the inputs are efforts or time, it is natural to associate a continuously differentiable cost with their exertion.

Let $C_x(x)$ be the principal's cost of effort and $C_y(y)$ be the agent's cost of effort. Naturally, costs are increasing in effort, *i.e.*, $C'_x(x) \geq 0$ and $C'_y(y) \geq 0$, with increasing marginal costs, *i.e.*, $C''_x(x) \geq 0$ and $C''_y(y) \geq 0$. Common functional forms of cost function include linear functions, i.e,

$$C_x(x) = c_x x \text{ and } C_y(y) = c_y y,$$

and quadratic functions, *i.e.*,

$$C_x(x) = c_x x^2 \text{ and } C_y(y) = c_y y^2, \tag{2.3}$$

for some $c_x, c_y > 0$.

It is common to assume that these costs (which may include some utilities as well) are private, and, therefore, non-shareable.

Equilibrium

Effort are often assumed to be chosen simultaneously and non-cooperatively, resulting in an equilibrium (x^*, y^*) solving the following game:

$$\begin{array}{rcl} x^* & = & \arg\max_{x\in[0,1]} \alpha V(x, y^*) - C_x(x) + f \\ y^* & = & \arg\max_{y\in[0,1]} (1-\alpha) V(x^*, y) - C_y(y) - f. \end{array} \tag{2.4}$$

If efforts were chosen instead sequentially, one would consider the Stackelberg equilibrium.

First Best

As a benchmark, we often consider the "first-best" (FB) efforts, which maximize the total value net of the costs of effort:

$$\max_{x,y\in[0,1]} V(x, y) - C_x(x) - C_y(y). \tag{2.5}$$

2.2.3 Solution Approaches

Two approaches are commonly adopted to solve this type of game (2.4).

Supermodularity

The first builds on the theory of supermodularity (Topkis, 1998; Vives, 1999) and relies on Tarski's fixed-point theorem. Inputs are said to be strategic substitutes if $\partial^2 V(x,y)/(\partial x \partial y) < 0$ for all (x,y) and strategic complements if $\partial^2 V(x,y)/(\partial x \partial y) > 0$ for all (x,y). If one of these conditions applies, the game has a Nash equilibrium, even if $V(x,y)$ is not concave in one or any of its arguments. Moreover, monotone comparative statics can be derived.

This is the approach adopted by Roels (2014) to characterize the FB degree of interaction and optimal work allocation in services, considering the process of total value creation (2.5).

Concavity

A second approach, which is the most common in the literature, is to impose some differentiability and concavity requirements to characterize the optimal or equilibrium efforts by solving the first-order optimality conditions. This approach relies on Kakutani's fixed-point theorem.

Specifically, it is common to assume that the value function is continuously differentiable and strictly increasing in the input of each party and exhibits strictly decreasing marginal returns, *i.e.*, $\partial V(x,y)/\partial x > 0$, $\partial V(x,y)/\partial y > 0$, $\partial^2 V(x,y)/\partial x^2 < 0$, and $\partial^2 V(x,y)/\partial y^2 < 0$. Under these assumptions, the equilibrium effort levels (x^*, y^*), if they lie in the interior of the domain $(0,1)$, solve

$$\begin{array}{rcl} \alpha \frac{\partial V(x^*,y^*)}{\partial x} - C'_x(x^*) & = & 0 \\ (1-\alpha) \frac{\partial V(x^*,y^*)}{\partial y} - C'_y(y^*) & = & 0. \end{array} \tag{2.6}$$

Under the aforementioned conditions, the univalence theorem (Gale and Nikaidô, 1965, Theorem 7(ii)) guarantees that there is at most one solution to the following set of optimality conditions such that $x^* \in (0,1)$ and $y^* \in (0,1)$ provided that $\frac{\partial^2 V(x,y)}{\partial x \partial y}$ does not change sign, *i.e.*, efforts are either always strategic complements or strategic substitutes, and the corresponding Jacobian does not vanish, *i.e.*, $\left(\frac{\partial^2 V(x,y)}{\partial x^2} - C''_x(x)\right)\left(\frac{\partial^2 V(x,y)}{\partial y^2} - C''_y(y)\right) \neq \left(\frac{\partial^2 V(x,y)}{\partial x \partial y}\right)^2$.

The equilibrium conditions are in fact very similar to the first-order optimality conditions that characterize the maximization of the net

value $V(x,y) - C_x(x) - C_y(y)$:

$$\begin{array}{rcl} \frac{\partial V(x^*,y^*)}{\partial x} - C'_x(x^*) & = & 0 \\ \frac{\partial V(x^*,y^*)}{\partial y} - C'_y(y^*) & = & 0. \end{array} \tag{2.7}$$

Hence, the equilibrium conditions can be interpreted as maximizing the net value, similar to (2.5), with inflated costs $V(x,y) - C_x(x)/\alpha - C_y(y)/(1-\alpha)$.

2.2.4 Specific Functional Forms

To generate more insights, researchers on co-production have often considered more specific functions. One common approach has been to model $V(x,y)$ as a transformation $F(.)$ of a composite effort function $E(x,y)$, *i.e.*, $V(x,y) = F(E(x,y))$; see Figure 2.5.

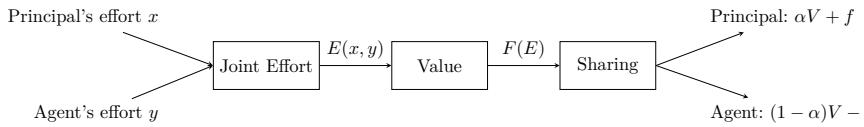


Figure 2.5: An Analytical Framework for Co-Production: Decomposition of Joint Effort and Value

Various forms of composite function have been considered. The simplest, considered by Xue and Field (2008), Bellos and Kavadias (2019), Bellos and Kavadias (2021), Ma *et al.* (2024), and Limon *et al.* (2024) among others, is a linear function:

$$E(x,y) = x + y.$$

Even if $E(x,y)$ is separable, $V(x,y) = F(E)$ may not be so if $F(.)$ is nonlinear (typically, concave). In this case, co-production makes sense if and only if $F(.)$ is super-additive, *i.e.*, $F(x+y) > F(x) + F(y)$.

Another functional form that has been commonly used (Roels *et al.*, 2010; Andritsos and Tang, 2018; Wang *et al.*, 2022; Gupta *et al.*, 2023; Tuncalp *et al.*, 2023) is a Cobb-Douglas function:

$$E(x,y) = x^\alpha y^\beta,$$

with $\alpha, \beta > 0$. To ensure coordinate-wise concavity of $V(x,y)$, it is sufficient to require that $F(.)$ be increasing concave and that $\alpha, \beta < 1$.

With this production function, it does not make sense not to adopt team production because $V(x,0) = V(0,y) = 0$ for all x, y.

Generalizing the linear and Cobb-Douglas functions, some researchers have considered the so-called Constant Elasticity of Substitution (CES) function (Roels, 2014; Rahmani *et al.*, 2018; Chen and Keppo, 2023; Candoğan *et al.*, 2020)

$$E(x,y) = (ax^r + (1-a)y^r)^{\frac{1}{r}},$$

with $0 < a < 1$ and $r \neq 0$. When $r = 1$, $E(x,y) = ax + (1-a)y$; when $r \to 0$, $E(x,y) = x^a y^{(1-a)}$; when $r \downarrow -\infty$, $E(x,y) = \min\{x,y\}$; and when $r \uparrow \infty$, $E(x,y) = \max\{x,y\}$.

Building on the extreme cases of perfect complements ($\min\{x,y\}$) and substitute inputs that cannot be used jointly ($\max\{x,y\}$), Dong *et al.* (2023) consider a convex combination thereof:

$$E(x,y) = \lambda \min\{x,y\} + (1-\lambda)\max\{x,y\},$$

for some $0 < \lambda < 1$.

Another flexible composite effort function, borrowed from decision analysis (Roels *et al.*, 2024), is:

$$E(x,y) = kx + ky + (1-2k)xy,$$

for $k \in [0,1]$. When $k = 1/2$, $E(x,y) = (x+y)/2$; when $k = 0$, $E(x,y) = xy$; and when $k = 1$, $E(x,y) = x + y - xy$. Siggelkow (2002) and de Bettignies (2008) consider a simpler version: $E(x,y) = x+y+kxy$. Efforts are strategic substitutes when $k < 0$ and strategic complements when $k > 0$.

Finally, various choices can be made regarding the concave transformation $F(.)$, such as a power function, *i.e.*, $F(E) = E^b$ with $0 < b < 1$, a logarithmic function, *i.e.*, $F(E) = V_0 + \ln(E)$, or, more generally, $F(E) = V_0 + (E^{1-\rho} - 1)/(1-\rho)$, with $0 < \rho < 1$, for some $V_0 > 1$.

2.2.5 Conditions for Having Different Operating Modes

Modeling joint production is particulary useful to generate insights into the choice of operating mode. To draw the parallel with Figure 2.2, should a service be delivered as a service factory ($x^* > 0$ and $y^* = 0$), as

a self-service ($x^* = 0$ and $y^* > 0$), or as a collaborative service ($x^* > 0$ and $y^* > 0$)?

It turns out that this choice, while relevant to many practical situations, arises mathematically under specific conditions, which we discuss next: The marginal costs of effort at zero must be non-null, efforts need to be strategic substitutes, but their degree of substitutability must be sufficiently low (Roels *et al.*, 2024). While joint concavity is often assumed, it is not necessary.

Non-Null Marginal Costs of Zero Effort

Consider what happens if the marginal cost of zero effort is null, as it is the case with quadratic costs of effort (2.3). Then, setting $x^* > 0$ and $y^* > 0$ would always be optimal (or emerge in equilibrium), *i.e.*, it would never be optimal to operate as a service factory or self-service. Indeed, the derivative of the principal's objective with respect to x would then be equal, when $x = 0$, to $\alpha \frac{\partial V(0,y^*)}{\partial x} - C'_x(0) = \alpha \frac{\partial V(0,y^*)}{\partial x} > 0$ if $\alpha > 0$ and $V(x, y)$ is strictly increasing; and similarly for the derivative of the agent's objective with respect to y.

In the literature on collaborative cost reduction, Corbett *et al.* (2005) show that the equilibrium characterization established by Corbett and DeCroix (2001) with strictly convex costs dramatically changes when costs are restricted to be only (weakly) convex.

Hence, the common assumption of quadratic costs (2.3), perhaps made for convenience, is not innocuous in terms of its implications for service design.

Strategic Substitutes

If efforts were strategic complements, *i.e.*, if $\partial^2 V(x, y)/(\partial x \partial y) > 0$ for all (x, y), then setting $x^* > 0$ and $y^* > 0$ would also always be optimal (or emerge in equilibrium); see Roels *et al.* (2024) for a formal derivation. That is, with strategic complements, the collaborative service is the only equilibrium (or optimal) operating mode.

Low Degree of Substitutability

Finally, if efforts were very substitutable (which we do not precisely define here as it requires a parametrization of the value function; see Roels *et al.* 2024), then one would expect that collaborative production (*i.e.*, $x^* > 0$ and $y^* > 0$) would never be optimal because it would then be optimal to allocate the workload to either the principal or the agent.

Joint Concavity

In the literature, it is often assumed that $V(x, y)$ is jointly concave. This is useful to characterize the FB solution (2.5), because the first-order optimality conditions are then sufficient. However, if efforts are assumed to be chosen in a decentralized fashion, then only coordinate-wise concavity is needed for ensuring that the first-order optimality conditions (2.6) are sufficient. In particular, it may well be that the equilibrium point (x^*, y^*) that solves (2.6) is a saddle point of $V(x, y) - C_x(x) - C_y(y)$ if joint concavity is not required. Accordingly, the assumption of joint concavity, perhaps made for convenience, is not innocuous because it may eliminate this interesting saddle-point case.

2.3 Customers as Employees

The co-productive framework depicted in Figure 2.2 and the analytical modeling of co-production developed in §2.2 suggest a symmetric, dual relationship between employees and customers. As supporting evidence, Schneider *et al.* (1980) report a strong correlation (about 70%) between employees' perceptions of service quality and customer perceptions of service quality in the banking industry. A more recent study of Glassdoor established positive correlation between employee satisfaction and customer satisfaction in people-intensive service industries such as retail; travel and tourism; and restaurants, bars, and food services. In contrast, this correlation was either null or negative in manufacturing-like sectors (Economist, 2019). Going beyond correlation, Heskett *et al.* (1994) postulate a causal relationship between employee satisfaction and customer satisfaction in their SPC framework (Figure 1.4).

Building on this duality, this section discusses this interpretation of customers as employees through their role and the value they provide or derive from the service (§2.3.1) as well as the drivers and measurement of their performance (§2.3.2).

2.3.1 Roles, Value Provision, and Value Obtention

Customers and employees are symmetric in many dimensions: in their roles, in the value they provide to the service, and in the value they derived from the service.

Roles

On the customer and employee roles, Barnard (1940, p. 300) notes, in a critique of his seminal book (Barnard, 1938), that "in the fundamental sociology of business behavior the services of an employee and of a customer when making a purchase are equivalent elements, similar contributions to the *same* organization." Pursuing this line of thought, Solomon *et al.* (1985) propose to view service transactions from a role theory perspective, drawing analogy to dramaturgy. They identify several key elements of the service encounter (reduced there to a human interaction) such as customers' and employees' roles and the service script.

However, Mills and Morris (1986) warns that clients are only "partial" employees, but can be made more productive participants when there is a match between the required production-related skills, knowledge, and attitudes, and the degree of involvement required of the client in service production. The authors propose a multi-stage (process) approach of client involvement in the creation of complex services consisting of the following steps: pre-encounter, initial encounter, and service completion.

Value Provided

Service quality is often framed as a function of the different types of value employees provide (Grönroos, 1984): their technical and functional qualities

Given the symmetry between employees and customers, Kelly *et al.* (1990) apply similar concepts of value creation to customers. Specifically, they define customer technical quality as what the customer provides to the service encounter (*e.g.*, labor, information). In contrast, the customer functional quality is how the service customer behaves during the service provision (*e.g.*, interpersonal aspects, such as friendliness and respect), which can be managed through a process of customer organizational socialization (behavioral and affective).

Value Received

Heskett and Sasser (2010) propose a dual conceptualization of value derived from the service by customers and employees. Specifically, value for customers is defined as the ratio of the sum of the results and quality of the experience over the sum of price and access costs. Symmetrically, value for employees is defined as the sum of their capabilities to deliver results and the quality of their work experience over the sum of the inverse of their total income and their job access costs.

2.3.2 Drivers and Measurement of Performance

Given the duality of roles between customers and employees, customers can be managed similar to the way employees are managed. To effectively perform the tasks they need to do, customers need to have the right capabilities, be motivated, and be licensed to do so, as we discuss next. Similar to employees, who exhibit some heterogeneity in their effectiveness and efficiency, customers are also heterogeneous; accordingly, some service organizations have attempted to measure their efficiency to focus on the most efficient customers.

Customers as Employees

The near symmetry between customers and employees implies that customers can, and perhaps should, be managed like employees. To understand the drivers of employee performance, Blumberg and Pringle (1982) propose a framework along three dimensions: the employees' *capacity* (sometimes, also referred to as their "ability" or "capability"),

their *willingness* (sometimes, also referred to as their "motivation"), and their *opportunity* (sometimes, also referred to as their "license").

A similar framework can be applied to understand the drivers of customer performance. Like employees, customers need to be hired and trained (capabilities), authorized to take action (license), and rewarded or punished (motivation) for their participation in co-production.

Barnard (1940) outlines an entire process to make customers more productive: the bringing of customers into cooperative relationship; the subsequent eliciting of services; the maintenance of customer morale; the maintenance of the scheme of inducements; the maintenance of the scheme of deterrents; supervision and control; inspection; education and training.

In the same spirit as the Service-Dominant Logic (Vargo and Lusch, 2004), which posits that customers are always co-producers, Olsen (2024) proposes "customer lean" principles, requiring that customers participate in organizations' continuous improvement, problem-solving, and decision-making processes.

Customer Efficiency

How do we measure "customer efficiency"? Ideally, it should be measured as the customer's marginal value of effort over their marginal cost. Using the notation of §2.2 (and assuming that the customer is the agent), we would thus ideally measure customer efficiency as the ratio of $\partial V(x,y)/\partial y$ over $C'_y(y)$.

As a proxy for this ideal measure, Xue and Harker (2002) measure customer efficiency as the ratio of total outputs (*e.g.*, number of transactions) over total inputs (*e.g.*, purchase or nonpurchase activity time), *i.e.*, $V(x,y)/y$. They adopt a Data Envelopment Analysis (DEA) approach to account for the multi-dimensionality of both outputs and inputs. The DEA method does not assume *a priori* weights to them, but instead keeps them flexible to build an efficient frontier to identify "efficient" customers, who can combine inputs to obtain a superior combination of outputs.

When customer inputs are only partially observable, customer efficiency can still be measured, but some assumptions need to be made

about which operating mode is the most efficient. For instance, Xue *et al.* (2007) assume that "efficient" customers will conduct more transactions in self-service channels (controlling for other factors that affect self-service choice) because of their relatively lower direct labor and opportunity costs. Under this assumption, efficiency can be measured as the difference between the residuals of an estimation of the value derived under full-service (or collaborative service, to use the same terminology as in Figure 2.2) delivery for a given set of outputs (*e.g.*, $V(x_D, y_D) - \beta_{D,x}x_D - \beta_{D,y}y_D$, for some $\beta_{D,x}, \beta_{D,y}$ estimated through regression) and the residuals of an estimation of the value derived under self-service delivery for a given set of outputs (*e.g.*, $V(0, y_S) - \beta_{S,y}y_S$, for some $\beta_{S,y}$ estimated through regression). Measuring efficiency as a difference in residuals has the benefit of canceling out all non-observable customer inputs, assuming they are the same under both operating modes.

Irrespective of how it is measured, customer efficiency depends on the provider's effort x. Consider an educational service: If a student has not learned the material of a lesson, is it because they have not studied hard enough or because the lesson was poorly explained by the teacher? Hence, a particular customer may be very inefficient for a particular provider, but very efficient for another. For instance, banks and insurance companies have traditionally underserved low-income people, which some other banks, such as Metro Bank, may find very profitable (Lago and Moscoso, 2011).

This notion of customer efficiency or productivity as dependent on the provider's input formalizes the proposition by Mills and Morris (1986) that customers can be made more productive participants when there is a match between the required production-related skills, knowledge, and attitudes, and the degree of involvement required of the client in service production.

Why do we care about customer efficiency? Xue *et al.* (2007) and Xue *et al.* (2011) show that it is a driver profitability. Very much like the generic strategy of customer selection (§A.2.2), service organizations could focus on their most efficient customers.

2.4 Operational Implications

This section discusses some operational implications of the co-production framework in Figure 2.2 in terms of contracting (§2.4.1), capacity management (§2.4.2), quality management (§2.4.3), scaling and technology adoption (§2.4.4), and servicescape design (§2.4.5).

2.4.1 Contracting

In this section, we briefly discuss the two sources of inefficiency (moral hazard) that arise from the co-productive framework and then overview some contracts that are supposed to align incentives to resolve these inefficiencies.

Moral Hazard Within and Across Operating Modes

When efforts are chosen in a decentralized fashion, as in (2.4), and the contract is endogenously set by the principal, two sources of moral hazard lead to inefficient outcomes, relative the FB solution (2.5). First, as is common to standard principal-agent models, the equilibrium efforts *within* an operating mode may be less than the FB efforts. Considering the co-production framework depicted in Figure 2.5, this happens under the self-service option because the buyer, as an agent, receives only a $(1-\alpha)$-share of the value; as well as under the collaborative service operating mode given that both parties receive only a share of the total value, which has the consequence of lowering their incentives, as if their costs of effort were inflated. This phenomenon is referred to as double moral hazard in the literature (Holmström, 1982; Bhattacharyya and Lafontaine, 1995).

A second source of inefficiency arises *across* operating modes. Specifically, the principal may design the contract in such a way that a different operating mode arises in equilibrium than under the FB. In particular, the principal may engage into hoarding, by not partnering enough with the agent (Roels *et al.*, 2024). Consider the following situations. First, suppose the principal role is played by the customer. In this case, hoarding means that customers would engage too much into self-service, *i.e.*, not hire enough service providers who can either be more efficient (if the

FB prescribes that they operate as a service factory) or more effective (if the FB solution prescribes that they operate as a collaborative service). People often complain about lacking time; yet, few really hire service providers to take care of many time-consuming household tasks (*e.g.*, cleaning or gardening). Second, suppose the principal role is played by the service provider. In this case, hoarding means that there is too little self-service in equilibrium, as shown by Xue and Field (2008).

Common Contracts

Common contracts in practice are fixed-fee contracts, time-and-materials contracts, and, to a much smaller extent, performance-based contracts (Karmarkar and Pitbladdo, 1995). When contracts involve a fixed-fee transfers, the sharing of output can be linear without loss of optimality (Bhattacharyya and Lafontaine, 1995; Kim and Wang, 1998).

Ideally, the choice of contract needs to mirror the choice of the operating mode. Accordingly, adopting fixed-fee contracts makes sense in self-services, so that the buyer is the residual claimant and is fully incentivized to exert the required effort. Conversely, adopting time-and-materials makes sense in service factories, to fully motivate the service provider. Finally, performance-based contracts are more suitable for collaborative services. Roels *et al.* (2010) suggest that fixed-fee and time-and-materials contracts can be used for collaborative services as well, but they then require that one party make their efforts observable, *e.g.*, by working on-site or logging hours. Moreover, Bajari and Tadelis (2001) show that time-based contracts, which require observability of a service provider's total cost (though perhaps without precise knowledge of the provider's effort allocation across tasks), tend to dominate fixed-fee contracts when a project is more complex.

In practice, outputs are rarely observable. For instance, how do we measure the value of education? Accordingly, performance-based contracts are rarely used. There are a few notable exceptions, however, in the legal industry (*e.g.*, contingency fees), online advertising (Choi *et al.*, 2020), supply chain collaboration (Corbett and DeCroix, 2001; Corbett *et al.*, 2005; Kim and Netessine, 2013), and servitized products, such as aircraft engines contracted on the basis of "power-by-the-hour" (Kim

et al., 2007; Kim *et al.*, 2022). Although performance-based contracts seem to align incentives (while still being fraught with double moral hazard), they may have unintended consequences when some tasks, albeit important for value creation, may not be contractible (Holmström and Milgrom, 1991), such as investments in reliability (Guajardo *et al.*, 2012; Bakshi *et al.*, 2015).

2.4.2 Capacity Management

Realizing that co-productive service encounter are often unique, recent research on capacity management has departed from the traditional view that service encounter times are fixed (albeit random). In particular, several studies have proposed that the time spent interacting is often discretionary, and that more value is achievable when more time is spent interacting.

However, longer interactions may have negative externalities on other customers through congestion. This has two implications for capacity management: First, service organizations need to determine at which rate to operate, depending on the number of other customers waiting. Second, they need to decide the adequate staffing levels, given this state-dependent changes in service rate.

We next review this literature, distinguishing different substreams depending on whether the task duration is fixed, decided by the provider, by the customer, or whether it is determined through their interaction.

Fixed Task Durations

Co-productive processes involve multiple resources: the service provider and the customer. Even when task times are fixed, managing capacity in a multi-step process with joint resources turns out to be complex because of two reasons: synchronization and coordination.

First, resource allocation across tasks must be synchronized, sometimes resulting in dead time. As a result, the overall capacity of a process that involves joint resources in specific activities may be smaller than the bottleneck capacity (Gurvich and Van Mieghem, 2015), contradicting traditional bottleneck analysis (Goldratt and Cox, 2016). The latter yields the correct network capacity only in certain types of networks,

called networks with a hierarchical collaboration architecture, provided that the tasks that require the most collaboration get scheduling priority (Gurvich and Van Mieghem, 2018).

Second, resources need to be coordinated, which reduces the time available for production. Because no production can happen during coordination meetings, such touchpoints are often perceived to be unproductive, even though they help workers (*e.g.*, a consultancy and a client) resolve issues and be more productive in the future (Roels and Corbett, 2024).

Provider-Chosen Task Durations

Considering a setting in which providers can dynamically adjust the time spent with each customer, Hopp *et al.* (2007) characterize the structure of the provider's time provision policy as a function of the state of the queue, trading off the value provided to the current customer with the waiting cost of the other customers in the queue. They find that, in such discretionary services and unlike the predictions of traditional queueing theory, (i) adding capacity may actually increase congestion, and (ii) task variability in service time can improve system performance.

The "quality–speed trade-off" has also profound implications in terms of contracting and pricing, as studied by Anand *et al.* (2011), Kostami and Rajagopalan (2014), and Debo and Li (2021). A key driver of the type of equilibrium, characterized in terms of price, service speed, demand, and congestion, turns out to be customer sensitivity to the time spent with them (Anand *et al.*, 2011; Debo and Li, 2021). Different contracts are associated with different externalities: In particular, Tong and Rajagopalan (2014) find that fixed-fee contracts leads to a higher demand and thus higher server utilization than time-based contracts.

Customer-Chosen Task Durations

Taking the polar opposite perspective to the provider-chosen task duration, Feldman and Segev (2022) consider a situation where it is the customers who choose their own service times, such as customers securing a parking spot, customers occupying tables in coffee shops, or

gym goers using various pieces of gym equipment. Clearly, the longer customers use the service, the higher the congestion, which has implications on capacity sizing. To help control congestion, setting time limits, *i.e.*, capping the maximum time that customers can spend in service, turns out to be an effective lever.

Interaction-Driven Task Durations

More recently, the literature has explicitly formulated models capturing the dynamics of interaction. Could there be some reciprocity in the customer's choice of speed and the provider's choice of speed, impacting the overall service time with implications on congestion management? Goes *et al.* (2018) find empirically that, in an online customer service chat center, agent multitasking leads to longer in-service delays for customers and lower problem resolution rates. Both result in lower customer satisfaction, although the impact varies across customers.

To investigate further the dynamics of interaction, Daw *et al.* (2020) and Daw and Yom-Tov (2024) model the dynamics of a conversation between a customer and a service provider (*e.g.*, a call center agent) using a two-sided Hawkes cluster model in which both parties make dynamic contributions. They characterize the interaction in terms of both its synchronicity, which measure the heterogeneity in rate of processing information, and its interdependence, measuring the degree of work allocation (self-production vs. co-production), which mirror the two notions of interaction and work allocation depicted in Figure 2.1.

2.4.3 Quality Gaps and Operational Transparency

In services, quality is often measured as a series of gaps, consistent with the expectancy-disconfirmation paradigm (Oliver, 1977; Oliver, 1981; Grönroos, 1984), as we discuss next. Marketing and operational transparency are useful levers to fill these gaps.

Gap Models

Services are fraught with quality gaps. Apte *et al.* (1997) identify three potential gaps. First, there is a potential *performance* gap, which relates

to design, between the customer preferences and the way a service has been specified. Although this could also arise in manufacturing, it is more likely to be more salient in services where customers have heterogenous preferences.

Second, there may be a *conformance* gap, which relates to operations, between the way the service has been specified and the way it is delivered. Although this gap could also arise in manufacturing, it is typically minimized through quality control and assurance, unlike the service environment, which is fraught with customer variability (Frei, 2006).

Finally, there is a potential *communication* gap, which relates marketing, between the customer's expectations and the service specifications. Again, this could also arise with products, but the intangible nature of many services makes it more salient in services than for goods. This last gap could be negative, in case a marketing value proposition underpromises to overdeliver.

This three-gap model by Apte *et al.* (1997) is a simplified version of the seminal gap model proposed by Parasuraman *et al.* (1985), which identifies five gaps. Together, these five gaps shape the customer's gap between their expectation of the service and their perception of the service quality.

The fundamental premise of these gap models is that the greater the gap between perception and expectation, the higher the customer satisfaction, in line with the expectancy-disconfirmation paradigm (Oliver, 1977; Oliver, 1981; Grönroos, 1984). As a follow-up framework, Parasuraman *et al.* (1988) develop a 22-metric scale, the so-called SERVQUAL framework, to assess the customers' perception of service quality along the following five dimensions: tangibles, reliability, responsiveness, assurance, and empathy. See Brown and Swartz (1989) for an application of a version of the gap model to professional services.

The key point of these gap models is that there is a misalignment between the customer's and the provider's domain. Presumably, these misalignments are the strongest when there is little interaction between the customer and the provider. Such misalignments are arguably the most salient when the service involves little ineraction, *i.e.*, operates as a service factory or a self-service (Figure 2.2). Although some gaps could be beneficial to delight the customer (*e.g.*, when a firm underpromises

and overdelivers), they are usually perceived negatively.

Marketing and Operational Transparency

To mitigate the negative impact of such gaps, transparency is key. While the field of Marketing has devoted substantial attention on marketing transparency, attempting to mitigate the negative effect of the performance gap and the marketing gap, it is only recently that the field of OM has looked at operational transparency to mitigate the negative effect of the conformance gap.

For instance, Buell and Norton (2011) discuss that, in self-services, where customers are presumably in charge of most of the service execution but may still depend on some inputs from the provider (*e.g.*, back-end computations to offer listings of flights on an online travel agency platform), engaging in operational transparency by signaling that the provider is exerting effort (*e.g.*, through a spinning wheel) enhance customers' perception of value from the service. In a follow-up piece, Buell *et al.* (2021b) discuss the benefits of operational transparency for governmental services, which are often mistrusted by customers.

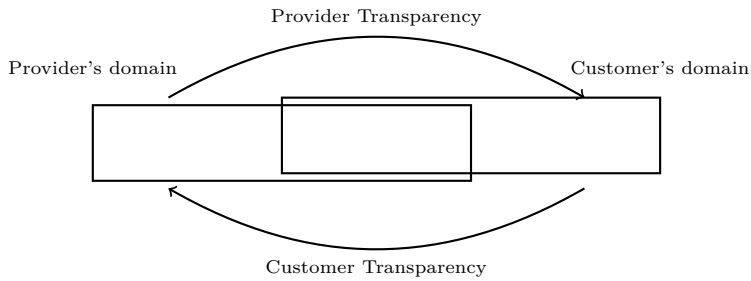


Figure 2.6: A Dual Perspective on Operational Transparency.

Transparency works both ways, as illustrated in Figure 2.6 (Buell, 2019), creating feelings of reciprocity (Buell and Norton, 2011), trust (Buell *et al.*, 2021b), and enhancing value (Buell *et al.*, 2017). In effect, transparency creates a surrogate for the customer or the provider. The service can still be delivered as a service factory or self-service, but benefits, through transparency, from feelings of reciprocity and clear lines of communication, which would otherwise be achieved only through

collaborative service production. Transparency thus enables service delivery to achieve the best of both worlds: Benefiting from the efficiency of the operating modes of service factories or self-services, without compromising the effectiveness of a collaborative service execution.

Transparency could also offer long-term benefits. First, provider transparency can improve customer self-selection into the service, which could drive long-term profitability. For instance, Buell and Choi (2024) reports that a credit card company that is more transparent about the trade-offs involved in its product offering leads to more informed choice by customers, resulting in higher spending and lower cancelation rate, with a lower probability of late payment. Second, customer transparency can help providers better understand their customers' needs and wants, which is especially important if they are evolving, and adapt their service offerings.

However, too much transparency, and more generally, customer-provider interactions, can backfire. Although it is commonly believed that service interactions may help improve customer satisfaction during service failures, Anderson *et al.* (2009) find that the opposite may happen if the customer attributes blame to the service provider. Looking at the various causes of flight delays, they find that customer satisfaction is lower when delays are of internal origin than when they are of external origin or for routine flights. When such delays happen, an important driver of their negative satisfaction is their interactions with employees. Using an economic model of beliefs, Guda *et al.* (2023) also report that, in the context of delay updating information provision, sharing information is beneficial only in the presence of loss aversion and diminishing sensitivity and when low delays are likely. Otherwise, not sharing information is more beneficial.

2.4.4 Scalability and Technology

Services often suffer from low productivity (§1.1.2). We argue this is particularly the case for collaborative services, but less for service factories and self-services. Technology, and in particular AI, can help achieve large productivity gains thanks to service industrialization (Karmarkar, 2004).

Service Inefficiency

One of the key challenges of services is to overcome their notorious low productivity, called Baumol's cost disease discussed in §1.1.2. What makes services inefficient are the numerous frictions induced by the interactions between the provider and the customer. For instance, to have one's hair cut, the process of involving a barber and a customer for (at least) 30 minutes has barely improved over time.

In services, customer satisfaction and employee productivity are indeed negatively correlated, unlike for goods (Anderson *et al.*, 1997). Customer satisfaction and productivity are less likely to be compatible when: 1) customer satisfaction is relatively more dependent on customization — the degree to which the firm's offering is customized to meet heterogeneous customers' needs — as opposed to standardization — the degree to which the firm's offering is reliable, standardized, and free from deficiencies; and 2) when it is difficult (costly) to provide high levels of both customization and standardization simultaneously.

Building on the co-production framework in Figure 2.2, we thus infer that services tend to be inefficient if they operate as collaborative services, but these frictions may disappear if the service delivery evolves towards the service factory operating mode (as promoted by Chase's customer contact model, reviewed in §2.1.1) or towards a self-service operating mode. However, the collaborative service operational mode remains the most effective.

Technology as an Enabler of More Frictionless Operations

Technology could enable service organizations to make that transition towards the less interactive operating modes, especially if they involve transformation of information and not transformation of matter (Apte *et al.*, 2012). As discussed in §A.2.1, service industrialization strategies, such as outsourcing, offshoring, and automation, are especially suitable for information services and gives them scalability (Karmarkar, 2004). Together with transparency (§2.4.3), they can achieve efficiency without compromising on effectiveness.

To support this, consider the following two examples. First, task automation in professional services can (i) augment some tasks, (ii)

deskill some tasks, (iii) move some tasks to self-service technologies, or (iv) centralize tasks to leverage professional workers' distinctive expertise (Sampson, 2021). Second, digital technologies enable telemedicine, which can dramatically increase the outreach of doctors (Delana *et al.*, 2023).

Artificial Intelligence

Among the deployments of service industrial strategies, AI offers great promises to achieve both efficiency and effectiveness. Huang and Rust (2018) identify four types of intelligences that are involved in services and can be provided by AI: mechanical, analytical, intuitive, and empathetic. They postulate that AI is developing in a predictable order, with mechanical mostly preceding analytical, analytical mostly preceding intuitive, and intuitive mostly preceding empathetic intelligence. As an interesting foresight preceding the development of the Large Language Models (LLMs), they assert that AI will eventually be capable of performing even the intuitive and empathetic tasks. Doing so, they argue, will not only enable innovative ways of human–machine integration for providing services, but also result in a fundamental threat for human employment.

One of the challenges for fulfilling this vision is that humans are in general averse to interacting with algorithms, especially if they exhibit erratic behavior (Dietvorst *et al.*, 2015), unless they can override them or modify them (Dietvorst *et al.*, 2018). As a result, disclosing to customers that they interact with a chatbot, however insightful, can be detrimental (Luo *et al.*, 2019). However, enhancing anthropomorphism of chatbots (*e.g.*, by adding interjections and filler words) may potentially counteract the negative effects of disclosing their identity (Xu *et al.*, 2024).

Despite its great promises, AI does not unlock value for everyone, like many other technologies, which do not suit everyone (Field, 2024). Cui *et al.* (2024) report that different user groups derive value from AI in distinct ways: low-experience users derive value from it only passively due to their reliance on AI assistance; in contrast, high-experience users actively enhance the value they get from it by adapting their behaviors. Hence, service providers, if they aim to remain inclusive, need to make sure they keep offering a high-touch channel for their customers who

are less tech-savvy.

2.4.5 Servicescape

A servicescape is the environment in which a service process takes place. The study of servicescapes fits our coproduction framework (Figure 2.2) because it aims to better understand the impact of the environment on customers and employees, building on the Stimulus-Organism-Response psychological model.

In their study of serviscapes, Bitner (1992) distinguishes three types of service environments, which match our typology introduced in Figure 2.2, namely: self-service (involving only the customer), interpersonal services (involving both the customer and an employee), and remote services (involving only employees). Depending on the type of services, the design of the servicescape has different goals. In particular, the servicescape of self-services should aim at enhancing customer satisfaction and support the service strategic positioning and segmentation. In contrast, the servicescape of remote work should aim at fostering employee satisfaction, motivation, and operational efficiency.

Originally, the concept of servicescape was defined within the scope of physical services — and thus, the focus was on the physical environment — but the same question arises for online services. There, servicescape design relates to UX design (Evans, 2017).

While design thinking is relevant to both physical and online environments, there are important differences between these two environments. Specifically, physical servicescapes tend to focus on the the physical appearance of facilities, employees, and equipment, and employees' responsiveness and empathy, which are unobservable online. As a result, trust needs to be conveyed using different artifacts. Considering an online banking environment, Balasubramanian *et al.* (2003) report that perceived trustworthiness of an online broker is a significant antecedent to investors' satisfaction, and that perceived environmental security and perceived operational competence impact the formation of trust.

Although the design of servicescapes is essentially an operations topic, it has received most of the attention from the Marketing research community, and very little from the OM research community, despite

its strong impact on facilities layout management in manufacturing (Kusiak and Heragu, 1987). There is, nevertheless, an emerging OM literature on service facility layout to understand customer paths in self-service facilities such as grocery stores and museums (Aouad *et al.*, 2022) and grocery stores (Moon, 2024) and the drivers of creativity in working spaces (Lee and Sosa, 2024).

2.5 Further Developments

The service dyad in the co-production framework depicted in Figure 2.2 has been increasingly challenged with the development of new technologies and related new business models.

As argued in §2.4.4, *technology* has increasingly contributed to shifting the service delivery away from an interactive, collaborative service operating mode towards *greater labor division*, taking the form of either a self-service or a service factory operating modes. However, with the development of LLMs, AI can now act as a persona on its own, transforming the dyad into a triad — if a service employee is still involved — or replacing the employee altogether. Investigating the deployment of AI in professional services, Sampson (2018) identifies several roles AI could play in this triad, depending on whether it is interacting with the customer, the professional, or a para-professional. Potentially, AI could lead to a *revival of the interactive, collaborative service operating mode*, but this time between customers and AI agents.

Another transformation that has taken place in the last two decades, fueled by technologies, is the development of *platforms* (Hagiu and Wright, 2015; Van Alstyne *et al.*, 2016), shifting the service archetype away from one-on-one interactions to inter-relationships between *communities* (or markets) of customers and communities (or markets) of providers. To be sure, many services (*e.g.*, education, entertainment) have always been delivered to batches of customers, so the fact that multiple customers are involved in the service delivery is not new. What is novel is the scale of these communities, as well as the degree of interaction between members of these communities. In particular, customers can motivate themselves through social interactions (Candogan *et al.*, 2012; Roels and Su, 2014; Zhang *et al.*, 2017; Zeng *et al.*, 2023), which

can lend themselves to innovative business models of service delivery, such as shared appointments (Ramdas and Darzi, 2017). Besides motivating themselves to be engaged service recipients, customers can also turn themselves into service providers through user-content creation (Roels and Fridgeirsdottir, 2009; Zhang *et al.*, 2017), information sharing (Candogan and Drakopoulos, 2020), or peer-to-peer sharing or sales (Benjaafar *et al.*, 2019; Benjaafar and Hu, 2020). In fact, the same person can act both as a customer and a provider of a service. For instance, a customer can produce and watch video content on a social media platform. To illustrate this change, consider the efforts made by KFC China to build communities of users around its various apps (Lin and Zhang, 2020), transforming what was traditionally a physical service into an information-intensive service — a platform, a brand.

The development of these communities create an even greater need to *orchestrate* them. The platform's orchestration role sometimes expands beyond two communities. For instance, online-food delivery platforms are commonly referred to as three-side marketplaces, grouping restaurants, couriers, and households (Chen *et al.*, 2022; Feldman *et al.*, 2023). The orchestration happens through matching of course (Kanoria and Saban, 2021), but also search, assortment optimization, insertion of (time) buffers, information disclosure, content monitoring, *etc.*

Although these developments call for an enrichment of the co-production framework in Figure 2.2, they do not necessarily make it obsolete. For instance, Google and Meta, which are perhaps the epitomic service providers that makes use of AI and leverage communities of users and providers, have been referred to as "data factories" (Thompson, 2018), similar to the service factories in Figure 2.2.

In addition to these model enrichments, we anticipate greater *model calibration*, perhaps at a microscopic level, similar to the work by Daw *et al.* (2020) and Daw and Yom-Tov (2024), leading to finer engineering of service encounters. Another promising area is the extension of the *fundamental operations principles*, *e.g.*, bottleneck theory, to *co-productive environments* (Gurvich and Van Mieghem, 2015). Data-driven experimentation, guided by the insights obtained from modeling, could help service providers identify which parts of the interaction need to happen synchronously, as a direct collaboration, with a concern for effectiveness,

and which parts of the interaction may need to happen with less interaction, perhaps asynchronously, with efficiency being the main concern. And for the latter, whether work needs to be allocated to the service provider operating as a service factory or to the customer in the form of self-service.

Finally, we expect to see applications of the generic models of co-production to specific *service applications*. The sector of education definitely deserves more attention. As shown in Figure 1.3, its costs have rapidly increased recently; yet, PISA test scores have been declining (Economist, 2023) and the value of higher education has been questioned (Economist, 2024). As a potential solution to push the efficiency frontier forward, the digitalization of technologies offers new channels, allowing teachers to "flip the classroom," and learners to engage in lifelong learning. Early experimentation with Massive Open Online Courses (MOOCs) and their integration into blended learning have been received positively (De Moura *et al.*, 2021) and we certainly expect that the development of Generative AI (GenAI), combined with the development of online platforms of course materials (Keppler *et al.*, 2022), would have massive transformational impact on the delivery of educational services. As Patrick Harker, a former Service Science academic and university president, puts it, universities need to transition from being teaching factories to being learner-centric, starting with a revamp of the curriculum to develop relevant skills for today's wold (Harker, 2014). However, Harker acknowledges not knowing exactly how learner-centric curricula and educational delivery processes should really look like — calling for more research in that space. We hope that the service management research community, following the lead of Zhang *et al.* (2017), will keep investigating ways to make them more effective, leveraging, in particular, their co-productive nature.

3

Delighting Customers through Experience Design

This chapter overviews the recent developments in OM and Decision Sciences on service experience design to delight customers. Delighting customers need not be expensive. Small tricks can go a long way. If an experience consists of multiple activities, choosing any sequence is a decision, and not optimizing it could be a missed opportunity. Whether a particular sequence is chosen for historical reasons or explicitly with the goal of delighting the customer is up to the service provider.

Although research on service design experience using social psychology has emerged only in the past two decades, responding to the early calls for research by Chase and Dasu (2001a) and Bitran *et al.* (2008), it connects to and builds upon large bodies of knowledge in arts. Already around 335 BCE, Aristotle identified core principles of story telling, specifically, the rules for the construction of a tragedy. In particular, Aristotle recommends that a plot include a change from good to bad, or from bad to good, with possible reversals and recognitions (Aristotle, 2022). These principles are still deemed relevant today, *e.g.*, for user experience design (IxDF, 2023). These connections are, however, too rare. One exception is Rozin and Rozin (2018) who examine what psychology and music can learn from each other and how this knowl-

edge may be applied to tasting menus and other temporal sequences. With the abundance of data collection characterizing many customer experiences (musical, culinary, education), the time is ripe to make these connections more apparent and leverage humanity's accumulated knowledge in that domain.

This chapter focuses on how to design services to maximize a service's hedonic value. Section 3.1 distinguishes the hedonic value of a service from its utilitarian one and argues why greater emphasis could be put on the former. We then review the literatures on experience management first in single-stage service encounters, with a strong focus on the management of customer wait (§3.2), then in multi-stage settings (§3.3), and finally with considerations beyond a single service encounter, through anticipation, recall, or customer journeys across encounters (§3.4). We conclude with future research directions (§3.5).

3.1 Value in Services: Utilitarian vs. Hedonic

Different perspectives can be adopted on service value: A utilitarian (functional) one, according to which value is determined by usefulness; and an hedonic (experiential) perspective, according to which value is determined by the satisfaction the service provides. The dominant paradigm in economics is the utilitarian perspective (Lancaster, 1966). However, the exclusive application of this utilitarian paradigm to services — without accounting for their hedonic value — has been challenged in two different ways, which we discuss next.

First, according to the Service-Dominant Logic, a service value is not determined by the producer and embedded in the operand resource (goods). Instead, it is perceived and determined by the consumer on the basis of its "value in use" (Vargo and Lusch, 2004). Specifically, a service value results from the beneficial application of operant resources (*e.g.*, knowledge and skills) sometimes transmitted through operand resources (goods). Firms do not determine a service value, they can only make value propositions. Although this Service-Dominant Logic is often cast as a paradigm change, it reflects a long-standing stream of research in Marketing that emphasized the subjective nature of quality of services and which called for a different approach to services than to

goods (Parasuraman *et al.*, 1985; Rust and Oliver, 1993).

A second challenge to applying a purely utilitarian approach to services is the inadequacy of Lancaster's utility representation (*i.e.*, in terms of a vector of attributes) to services. Unlike products, services are processes, which unfold over time. At each point in time during the service process, customers experience an instant utility (Kahneman *et al.*, 1997), which is a measure of hedonic and affective experiences. Moreover, the stimuli and schemata that shape this instant utility are multi-dimensional, including cognitive, affective, sensorial, social, physical, and social dimensions among others (Fließ *et al.*, 2024, Chapter 6). Hence, a proper representation of a service is not a vector of characteristics, unlike what was proposed by Lancaster (1966) for goods, but rather a multivariate, continuous-time function. Integrating this function measures such "temporally extended outcomes" as the "remembered utility" or the "total utility" (Kahneman *et al.*, 1997).

Consequently, adopting a purely utilitarian perspective approach may be unrealistic. Accordingly, recent research has put greater emphasis on the hedonic dimension of services. In fact, the latter dimension has increasingly been a source of service differentiation and way to escape commoditization (Pine and Gilmore, 2011). Even services that have a strong utilitarian value, such as wealth management, can be differentiated on their hedonic dimension (Ponsignon, 2023).

3.2 Single-Stage Experience Management

We first offer a historical perspective on the literature on experience management in single-stage encounters. Given the large literature on this topic, we will not attempt to offer a comprehensive review of the literature, but hope to delineate the key lines of research.

In service OM, experience management was traditionally conceptualized as wait management, with a special focus on call centers (Gans *et al.*, 2003) and hospitals (Armony *et al.*, 2015). Within that stream of research, customers have traditionally been modeled as rational agents, which make rational joining, balking, and reneging decisions, as we review in §3.2.1.

In contrast, recent research developments conceptualize customers

waiting in line as human beings depicting predictable biases, which we review in §3.2.2. Building on prospect theory, this more recent stream of research models satisfaction from a waiting experience as a function of customers' perceptions and expectations of the wait.

3.2.1 Customers as Rational Agents

Traditionally, customers have been modeled as rational agents, trading off the benefit of obtaining a particular service with the disutility of waiting in line. Customers evaluate this trade-off when they decide to join a queue or balk away from it. They might re-evaluate this trade-off while waiting in line, leading to possible reneging behavior. In the following, we first review the literature on joining and balking decisions and then the literature on abandonments or reneging decisions.

These models of rational customer behavior have two merits. First, they depart from the traditional paradigm that customers are inanimate jobs to be processed, and therefore contribute to the literature on people-centric operations (Roels and Staats, 2021). Second, they often serve as a benchmark against which models that account for human biases are evaluated.

Joining and Balking Decisions

One of the earliest papers on people-centric operations is Naor (1969), who explicitly models the customers' joining decisions to a service facility. As in many subsequent works, customers are assumed to have linear utilities; accordingly, they join a queue if their value from the service exceeds the sum of the price and their cost of waiting, estimated as the product of their expected service times and the current queue length. (This linear utility model can in fact be rationalized when customers trade off their time spent on leisure vs. work, provided that they can choose how many hours they work; see Li *et al.* 2024.) Moreover, the firm is assumed to charge the same price to all customers. Consequently, different customers, even if they are homogeneous *ex-ante*, derive different utilities from the service, depending on the state of the queue when they join it.

Naor (1969) compares the firm's optimal pricing strategy, depending on whether it is a monopoly (maximizing revenue) or a social planner (maximizing total welfare). Naor finds that, when the queue is observable, the revenue-maximizing price exceeds the welfare-maximizing price. Therefore, in equilibrium, the queue is shorter for a revenue-maximizing provider than for a welfare-maximizing provider. This result turns out to be quite robust to a variety of model setups (Mendelson, 1985), even though the additive nature of the relationship between service value and waiting cost seems critical (Afeche and Mendelson, 2004).

A second critical assumption is the observability of the queue. Indeed, when the queue is not observable, customers cannot balk, *i.e.*, their joining decision is based on their expectation of the queue length. Considering a setting with an unobservable queue, Edelson and Hilderbrand (1975) show that Naor's seminal result breaks down: the price set by a revenue-maximizing firm is the same as the one set by a welfare-maximizing one. Hassin (1986) studies a dual setting of the problem, where the price is fixed, but a social planner can mandate the service organization to hide or reveal the state of the queue. They find that it may be socially optimal to force a revenue-maximizing firm to reveal the queue length to arriving customers.

A third critical assumption is the linear, and not affine, nature of the pricing scheme. With a two-part tariff, *i.e.*, when the service organization can sell rights to the service in advance, the prices set by a revenue-maximizing and a welfare-maximizing firms are equal (Edelson and Hilderbrand, 1975).

A fourth critical assumption is the homogeneity of customers. When customers have heterogeneous costs of waiting but the firm can neither price-differentiate them nor re-sequence them, the revenue-maximizing price can be smaller than, equal to, or greater then the welfare-maximizing price; still, setting them equal turns out to be quite robust (Edelson and Hilderbrand, 1975). If the firm can resequence customers, it is actually optimal to prioritize them according to the so-called $c\mu$ rule, namely, to prioritize customers that have the highest cost of waiting c and the shortest expected processing time $1/\mu$; see, *e.g.*, Van Mieghem (1995). Pushing this further, Mendelson and Whang (1990) propose an incentive-compatible differentiated pricing scheme, consisting in

charging each customer class the waiting time externality they generate on the system, which turns out to be welfare-maximizing. However, a revenue-maximizing firm quoting prices and (average) throughput times to different classes of customers might prefer adopting a non work-conserving policy, consisting in strategically delaying the work of the lowest-priority class of customers, *i.e.*, remaining idle before commencing service on a waiting job (Afeche, 2013). When the firm cannot price-differentiate its customers, the lowest-priority customers will have an incentive to lie about their types (Snyder *et al.*, 2022), even if it is costly for them (Rodriguez *et al.*, 2024). It is then optimal for the firm to deviate from the welfare-maximizing policy, namely, the $c\mu$ rule, to reduce the extent of lying behavior by "upgrading" some customers who claim to not deserve priority.

While Afeche (2013) considers a system with unobservable queues, Armony and Maglaras (2004) show that offering real-time visibility on the current state of the system upon customer arrival improves the system performance, which aligns with the result by Hassin (1986) in a multi-class setting. Even when the information offered by the service provider may not be truthful, Allon *et al.* (2011) show, using a cheap talk game, that providing information, even if non-verifiable, is optimal for the firm and improves customer utility. Moreover, Yu *et al.* (2018) argue that customers' responses to different delay announcements partly reveals their type to the firm, which constrains the firm's messaging strategy.

Overall, balking can have a severe impact on a firm's revenue. For instance, Lu *et al.* (2013) find that, in the context of a deli store, moderate increases in the number of customers in queue can generate sales reduction equivalent to a 5% price increase.

Reneging Decisions

Abandonments (*i.e.*, queue reneging) are in general suboptimal when customers are fully rational and forward-looking; that is, customers either join the queue or balk upon arrival. This is true even when the queue in unobservable by customers given that any wait experienced so far should be considered as a sunk cost, making it irrelevant for trading

off whether to renege or not.

There are a few reasons where abandonments may occur within a model with rational customers, however. First, when the queue is unobservable and the system experiences random disruptions, abandoning may be optimal because a long wait signals the state of the system to customers (Mandelbaum and Shimkin, 2000). Second, abandonments may arise when the queue is unobservable and customers face strict deadlines after which their valuation for the service drops to zero (Hassin and Haviv, 1995), or when they have nonlinear waiting costs (Haviv and Ritov, 2001; Shimkin and Mandelbaum, 2004). See Hassin and Haviv (2003) for an overview of the assumptions that lead to rational abandonments.

3.2.2 Customers as Human Beings

Departing from the early literature on queue management, which assumed that customers were either inanimate jobs (in line with the origins of queuing theory in electrical engineering (Kleinrock, 1974)) or fully rational (§3.2.1), Maister (2004) raises the attention that the perception of wait time also matters and postulates several principles that make wait feel longer, such as:

- occupied time feels shorter than unoccupied time;
- uncertain waits are longer than known, finite waits;
- unexplained waits are longer than explained waits;
- unfair waits are longer than equitable waits; and
- solo waits feel longer than group waits.

In a follow-up managerial article, Norman (2009) advocates for delivering fair practices, maintaining an open communication, providing feedback, explaining the underlying causes of the wait, and managing perceived fairness by reframing what waiting is. In the spirit of the SPC framework depicted in Figure 1.4, Norman stresses that the goal of service organizations should be to optimize the experience for both

customers and employees, thereby enhancing customer satisfaction and reducing employee stress and turnover.

Closer to the OM research communities, both Larson (1987) and Bitran *et al.* (2008) call for more research on customers' attitudes towards queues. In particular, Larson (1987) raises attention on the potential social injustice in case of deviations from a FIFO policy, the queueing environment (which relate to the servicescape concept discussed in §2.4.5), and the feedback regarding the likely magnitude of the delay. Building on social psychology, Bitran *et al.* (2008) call for more research on the management of the experience, and not the wait, identifying both personal and environmental moderators of the impact of wait duration on the overall evaluation of the service, leading them to build a "waiting-profit chain" framework.

Within the context of congestion management with customers who are subject to biases, different mechanisms have been explored, as we overview next, such as: offering wait-time guarantee, real-time progress feedback, and explaining the causes of the delay, raising attention on queue length vs. wait times, and finally leveraging congestion externalities on quality signaling and consumption.

Wait-Time Guarantee

At the core of many models of wait time management with "irrational" customer is the combination of some form of uncertainty about their wait time, which is resolved through Bayesian updating, and the expectancy-disconfirmation framework (Oliver, 1977). Customer satisfaction is determined by the difference between the their prior estimates and updates of their waiting times. Compounding this comparison is prospect theory, which associates higher weights with losses than with gains (Kahneman and Tversky, 1979), as depicted in Figure 3.1.

Using these components, Kumar *et al.* (1997) build a model and experimentally report that customers' satisfaction increases during the wait if the service time is less than expected, but their satisfaction takes a U-shape if customers observe the services times to be more than expected. Wait time guarantees, if met, appear to increase satisfaction at the end of a wait; however, if violated, they decrease satisfaction.

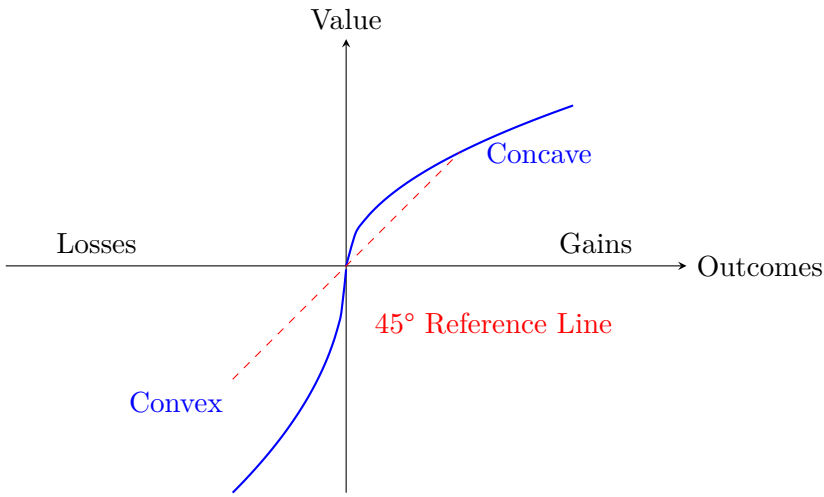


Figure 3.1: Prospect Theory: Customers Are More Sensitive to Losses than to Gains, and Exhibit Increasing (resp., Decreasing) Marginal Benefits Below (resp., Above) a Reference Point.

Wait time guarantee need to be carefully optimized. Indeed, Batt and Terwiesch (2015) find that although faster promises increase sales and profits, they also increase product returns and reduce customer retention. Also, Yu *et al.* (2021) find that customers are loss-averse in time, consistent with prospect theory (Kahneman and Tversky, 1979). Although delay announcements do not alter their loss-averse nature, announcements may affect customers' belief about the offered waiting time and thus, impact their reference points. Consequently, providing conservative wait time estimates may be a way to turn away the most sensitive customers and delight the remaining ones, but too conservative estimates may severely affect revenue.

The accuracy of these wait time guarantees also matters: Pavlov *et al.* (2024) find through experimental studies that both shorter and longer than expected waiting times reduce participant satisfaction with wait. What may explain why shorter waits cause dissatisfaction is that waiting participants may engage in "filler" activities while waiting, which they have to abruptly abandon when their wait finishes.

Real-Time Progress Feedback

In contrast to wait-time guarantees, which affect customers' joining decisions, real-time progress feedback may also affect their reneging behavior. In a large-scale field study with a ride-sharing platform, Yu *et al.* (2022) report that the initial wait-time information and update frequency have a significant impact of customer abandonment. Similarly, Munichor and Rafaeli (2007) report from several experimental studies that fillers in telephone waiting situations can influence queue abandonment and satisfaction. Apologies heard while waiting were found to yield the most negative caller reactions, whereas information about position in the queue produced the most positive reactions. More broadly, they find that displaying a sense of progress improves customer satisfaction — a phenomenon Buell and Norton (2011) cast as operational transparency (also discussed in §2.4.3).

To better understand the dynamics of these empirical phenomena, Guda *et al.* (2023) build a model to investigate whether providing real-time progress information via process trackers improves or hurts consumer satisfaction when customers are delay-sensitive and experience gain-loss utility (again, consistent with prospect theory). They find that sharing information is beneficial only when customers have diminishing sensitivity and low delays are likely. Otherwise, not sharing information is preferred.

In modeling, it is often assumed that delay announcements impact customers' beliefs about the system through a Bayesian updating process, assuming exogenous costs of waiting. However, Yu *et al.* (2017) argue, using a structural estimation model, that delay announcements may also shape customers' waiting costs. Specifically, they find that customers' per-unit waiting cost tends to decrease with the offered waiting times associated with the announcements. That is, customers tend to be more patient if they are provided a longer waiting time estimate, which somewhat contradicts the commonly assumed convex shape of their waiting time cost function (Van Mieghem, 1995).

In physical services, real-time progress feedback is readily available by observing the queue length. Batt and Terwiesch (2015) document the effects of queue length and queue flows on abandonment in the

context of hospital operations. They show that patients are sensitive to being "jumped" in the line and that patients respond differently to people more sick and less sick moving through the system, consistent with the call by Larson (1987) to capture fairness concerns in queues.

Attribution

Consistent with Maister's third principle that unexplained waits are longer than explained waits, customers tend to be more sensitive to congestion when the cause of delays is under the control of the service provider. For instance, Taylor (1994) finds that delays negatively impact satisfaction, but the degree to which the service provider is perceived to have control and the degree to which the delayed customer's time is filled also matter. Similarly, Anderson *et al.* (2009) find that customer satisfaction is lower when delays are attributable to the service provider's actions. As discussed in §2.4.3, operational transparency is often useful, but it could also backfire.

Queue Length vs. Waiting Time

Even though information about queue length is in theory equivalent to information about waiting time (by Little's Law), it appears that customers may base their balking decisions on the former rather than the latter (Lu *et al.*, 2013), without adjusting enough for the speed at which the line moves. Hence, adopting a pooled queue configuration, which is in general considered to be more efficient (Kleinrock, 1974), may backfire because it may lead to more abandonments.

Congestion Externalities

Besides influencing customer satisfaction during the waiting experience, congestion has also externalities on customers' *ex-ante* perception of quality and on their *ex-post* consumption.

Ex-Ante **Perception of Quality** A long queue may act as a signaling device of quality (Tu *et al.*, 2018). Customers thus face a conundrum: On the one hand, they would like to join shorter queues to benefit from

a short waiting time. On the other hand, doing so may also be associated with a lower quality since other customers seem to prefer other service providers. Debo *et al.* (2012) optimize customers' optimal queue joining strategy in this context. They show that the optimal policy has a "hole": a customer should join a queue whose length is either below or above the hole, but not at the hole itself. More broadly, using queues as a signal of quality leads to herding behavior (Veeraraghavan and Debo, 2011): customers may ignore their private signal about the quality of a service provider and herd instead, joining the longer queue. Kremer and Debo (2016) develop and test a theory of observational learning that predict the effect of wait times act as a signal of quality, making low (resp., high) quality products generate shorter (resp., longer) wait times. Although long queues might signal quality in some cases, overemphasizing this signaling mechanism leads to wasted time, lower overall welfare, and misallocation of resources.

Ex-Post **Consumption** Congestion could also have spillover effects on consumption behavior, potentially leading to shorter or longer service times. In particular, longer waits increase consumption. For instance, Kim *et al.* (2018) find that customers who have waited for service for a long time choose to use the service for a long time when it becomes their turn. Ülkü *et al.* (2020) observe a similar pattern and identify mental accounting as the underlying mechanism that drives this behavior: A larger purchase allows customers to offset the long wait suffered. As a result, common managerial practices to make the wait more enjoyable — such as distracting customers while they are waiting — can backfire in terms of their actual spending, even though their satisfaction goes up.

A second set of spillover effect stems from the social interactions in the queues. Specifically, Ülkü *et al.* (2022) report that, due to a concern for other customers, customers tend to accelerate their own service time, and in doing so, sacrifice their own consumption utility. However, the effect is diminished when they have themselves waited, as it is perceived as fair to let others wait if one also had to wait. Similarly, Kim *et al.* (2018) report that customers tend to mimic the previous customers' service usage behavior. That is, customers feel entitled to use a self-service slowly if other did so previously.

3.3 Multi-Stage Experience Design

In the past two decades, different streams of research in OM, decision analysis, and economics have adopted a broader scope for experience design, moving from a single-stage to a multi-stage (or process) approach and considering other stimuli than congestion.

We first review some early developments in research, calling for a multi-stage approach to experiential service design (§3.3.1). We then review the key decisions that are typically made (§3.3.2) and common applications in this stream of research (§3.3.3). Section 3.3.4 presents an overview of our analytical framework, which consists of four components: the utility formation (§3.3.5), the reference point adaptation (§3.3.6), the stock accumulation (§3.3.7), and the aggregation of utilities (§3.3.8). We then overview some key managerial insights about the optimal structure of experiences in §3.3.10. Section 3.3.11 discusses the implications of these policies for resource and congestion management, leading to strategic idling and buffering.

3.3.1 Early Developments

Early on, Shostack (1987) called for the application of "structural process design" to "engineer" services on a more scientific, rational basis, and proposed blueprinting service processes to document all process steps and points of divergence. Later, Chase and Dasu (2001b) proposed to leverage social psychology to maximize customer satisfaction from an experiential service process; see also Fliess *et al.* (2024). Zomerdijk and Voss (2010) build on the literature on services and experience design to derive propositions that reflect design principles for experience-centric services. Their theoretical background builds on the drama metaphor, sometimes referred to as "theatricalization," associating the physical environment with a stage, service employees with actors, the service delivery process with a script, fellow customers as the audience, and the back-office support as backstage.

3.3.2 Decisions

Typical decisions involved in multi-stage experience design are activity sequencing and/or activity duration sizing (Das Gupta *et al.*, 2015), information provision (Ely *et al.*, 2015; Chen *et al.*, 2024a), effort exertion (Bellos and Kavadias, 2019), budget allocation among various operational drivers such as staffing, training, and task assignment (Soteriou and Chase, 2000), and insertion of breaks (Baucells and Zhao, 2019).

3.3.3 Applications

Potential applications of multi-stage experience design include various scheduling decisions, such as songs in a concert (Baucells and Zhao, 2020), operas in a concert season (Dixon and Verma, 2013), levels in a video game balancing reward and difficulty (Li *et al.*, 2023), attractions visits in a theme park (Tsai and Chung, 2012) or in a city trip (Dixon and Thompson, 2019; Deshmane *et al.*, 2023), or practice sessions to maximize performance (Baucells and Zhao, 2019; Roels, 2020); layout decisions such as positioning paintings in a museum to balance the visitors' desire to see the museum's highlights while avoiding congestion (Aouad *et al.*, 2022); effort allocation across multiple activities to maximize a patient's experience in their hospital journey (Soteriou and Chase, 2000); pricing (Popescu and Wu, 2007); or assortment rotation in retail (Caro and Martínez-de-Albéniz, 2012).

Most applications focus on provider-routed experiences. In customer-routed experiences, there may be a gap between the provider's intended experience and the customer's choice of experience (Ponsignon *et al.*, 2017), even though it may be reduced with technology such as audioguides (Aouad *et al.*, 2022) or RFID bands (Tsai and Chung, 2012).

3.3.4 Framework

As highlighted by Chase and Dasu (2001b), what aggregates instantaneous utilities into satisfaction are the sequence of stimuli, the effects of emotions on encoding through schemata (*i.e.*, the cognitive frameworks for the interpretation and organization of new information), which might give more weight on certain elements than others, depending on the

context, and the perceptual distortion over time, which may alter the remembered utility over time (Palmer, 2010).

At the risk of oversimplifying these behavioral mechanisms, Figure 3.2 displays a framework to structure how customer satisfaction is derived from various stimuli. For simplicity, we adopt a discrete-time setting. We consider an experience lasting T periods.

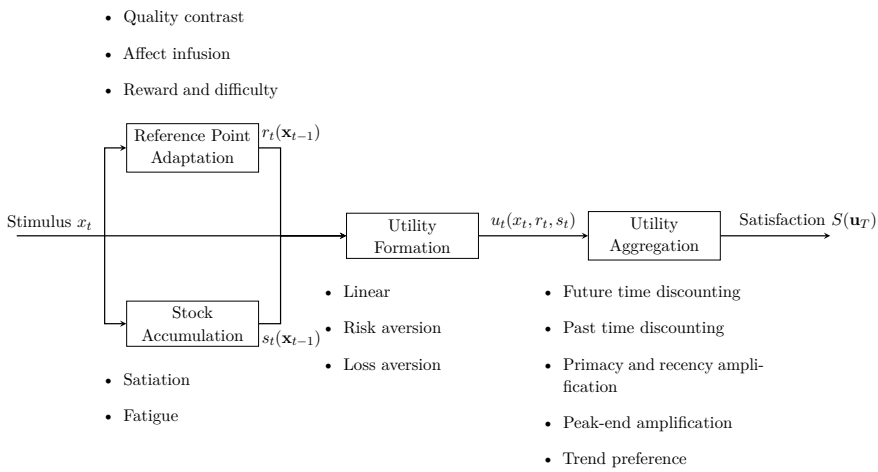


Figure 3.2: An Analytical Framework for Utility Formation and Aggregation

In period t, the customer experiences a stimulus (a/k/a service level) x_t, which we assume, for simplicity to be uni-dimensional. This stimulus, together with a reference point r_t and a stock variable s_t (*e.g.*, a stock of fatigue or satiation) shapes the current period's utility u_t; accordingly, we write the instantaneous utility as $u_t(x_t, r_t, s_t)$

Both the reference point and the stock variables may depend on past stimuli, represented as a vector $\mathbf{x}_{t-1} = (x_1, \ldots, x_{t-1})$. Hence, the current stimulus x_t does not change the current reference point and stock variables, but it alters future ones. Unlike the stock variable, which accumulates over time, the reference point is comparable to the stimulus x_t, and thus acts more like a flow.

Finally, at the end of the experience, the customer derives satisfaction S from all T instantaneous utilities, $\mathbf{u}_T = (u_1, \ldots, u_T)$.

Illustration

To illustrate the framework, consider the model of inter-temporal consumption with adaptation and satiation proposed by Baucells and Sarin (2010). A multi-dimensional version of the model applies, among others, to daily choices of food, *e.g.*, between American, Indian, and Chinese food. In each period t, utility depends on the stimulus x_t, relative to a reference point r_t, and given an accumulated stock of consumption s_t. Because the comparison to the reference point is assumed to occur only in case of consumption, the incremental utility is the difference between a nonlinear function $v(.)$ evaluated at two different points, namely, $x_t - r_t + s_t$ (which involves consumption) and s_t (as if there was no consumption). Hence, $u_t(x_t, s_t) = v(x_t - r_t + s_t) - v(s_t)$. For instance, the value of a Chinese meal (x_t) is naturally contrasted to the quality of the past Chinese meals one may have had in the past (r_t); a second component affecting the evaluation of the meal is presumably the cumulative number of times one may have had this meal in the past (s_t); *e.g.*, someone having the same Chinese meal every day may be "satiated" and not enjoy the current meal as much as someone who is having it for the first time.

The reference point r_t is assumed to adapt to past stimuli, *i.e.*, $r_{t+1} = \alpha x_t + (1 - \alpha) r_t$, for all t, in which $0 < \alpha < 1$ denotes the speed of adaptation. The stocks s_t accumulates over time and decays at rate $\gamma \in (0, 1)$ if it is not replenished, *i.e.*, $s_{t+1} = \gamma s_t + x_t$ for all t.

According to the discounted utility model (Koopmans, 1960), a customer may discount their future stream of utilities at a rate $\delta < 1$, *i.e.*, $S(\mathbf{u}_T) = \sum_{t=1}^{T} \delta^t u_t(x_t, r_t, s_t)$. For a given set of activities with stimuli $\{x_1, \ldots, x_T\}$, perhaps subject to some precedence constraints, the problem then consists in choosing the sequence that maximizes $S(\mathbf{u}_T)$.

We next discuss the four components of the model: the utility formation process (which also applies to single-stage experience design discussed in §3.2), the reference point adaptation, the stock accumulation, and the aggregation of utilities. As we introduce different versions of the model, we label equations to later map them (in Figure 3.3) to the framework proposed in Figure 3.2.

3.3.5 Utility Formation

We first discuss functional forms of the utility function as a function of the stimulus only; we then separately introduce the relationship between the stimulus and a reference point and the relationship between the stimulus and a stock; we finally incorporate all three dimensions — stimulus, reference point, and stock — into a single model.

Functions of the Stimulus Only

In our review of functions of the stimulus only, we first discuss linear models of utility and then nonlinear ones — first monotone and then non-monotone.

Linear Models. The most basic model of utility is a linear model, *i.e.*,

$$u_t(x_t) = \mathbb{E}[x_t + \epsilon_t] = x_t, \tag{3.1}$$

in which ϵ_t is a noise with zero mean and given variance associated with the perception of stimulus x_t. This model, which often comes from an Ordinary Least Squares regression, can also include customer fixed effects and various co-variates.

Nonlinear, Monotone Models. Some nonlinearities can be introduced, such as concavity or convexity. In particular, a risk-averse customer would have concave preferences, whereas a risk-seeking customer would have convex preferences. Specifically,

$$u_t(x_t) = \mathbb{E}[v(x_t + \epsilon_t)], \tag{3.2}$$

in which $v(x)$ is an increasing function, concave if the customer is risk-averse and converse if the customer is risk-seeking. A common function for risk aversion is:

$$v(x) = \frac{x^b - 1}{b}, \tag{3.3}$$

with $0 < b < 1$ (Baucells and Sarin, 2007). This function tends to a logarithmic function when $b \downarrow 0$ and a linear function when $b \uparrow 1$. Implicit to this transformation is the assumption that the support of ϵ_t is such that $x_t + \epsilon_t$ never becomes negative.

Nonlinear, Non-monotone Models. Additional non-linearities could be introduced when utility is non-monotone in the stimulus. For instance, the utility from heat indoors probably peaks at around 20°C. In this case, $v(.)$ could be a polynomial function with a local maximum.

Functions of the Stimulus and Reference Point

We first introduce the most common way reference points shape utilities — through prospect theory, and then discuss variations of this model — including modulation of contrast effects, non-monotone effects, assimilation, and multiplicity of reference points.

Reference Points and Prospect Theory. Departing from the expected utility paradigm, prospect theory posits that individuals make decisions in relativity and not in absolutes. Let $r_t(\mathbf{x}_{t-1})$ be a reference point derived from past stimuli. As depicted in Figure 3.1, customers may be risk-averse when faced with risky choices leading to gains, but risk-seeking when faced with risky choices leading to losses. The stimulus x_t is then evaluated in comparison to the reference point r_t. Popescu and Wu (2007) formalize three properties of prospect theory: reference dependence, loss aversion, and diminishing sensitivity. Assuming again that more is better, a formalization of prospect theory, in the spirit of the model by Kőszegi and Rabin (2006), could be:

$$u_t(x_t, r_t) = v(x_t - r_t + \epsilon_t),$$

in which

$$v(z) = \begin{cases} z^{b_g} & \text{if } z \geq 0, \\ \lambda \cdot (-z)^{b_l} & \text{if } z < 0, \end{cases} \tag{3.4}$$

with $0 < b_g \leq 1 \leq b_l$ and $\lambda > 1$. Chen *et al.* (2024a) use a simpler version, assuming that $b_g = b_l = 1$.

Modulation of Contrast Effect. A more general version of the model could modulate the intensity of the contrast with the reference point. For instance,

$$u_t(x_t, r_t) = v(x_t - \beta r_t + \epsilon_t),$$

for some $\beta > 0$, which could be estimated empirically. This is the approach pursued by Deshmane *et al.* (2023).

This approach is in fact consistent with the separability of drivers of utility as utilitarian and hedonic, as in Bellos and Kavadias (2019), Bellos and Kavadias (2021), and Guo *et al.* (2024). Indeed, the argument in this utility function can be decomposed as a functional term $(1-\beta)x_t$ and an experiential one $\beta(x_t - r_t)$.

Non-Monotone Effects. Most models assume that the larger the gap $x_t - r_t$, the higher the utility. However, this may not always be the case. For instance, Yerkes and Dodson (1908) report an inverted U-shape relationship between arousal and performance. That is, performance increases with physiological or mental arousal, but only up to a point. When levels of arousal become too high, performance decreases. Also, Ely *et al.* (2015) posit that, in entertainment, customers enjoy suspense and/or surprise, which can be both measured as squared deviations from a reference point (acting as a prior belief of a particular outcome).

Assimilation. A dual version of prospect theory, which is based on the disconfirmacy assumption, assimilates (*i.e.*, adds) the reference point to the stimulus, instead of contrasted the stimulus to the reference point (*i.e.*, taking their difference). That is, the higher the reference point, the higher the utility. For instance,

$$u_t(x_t, r_t) = v(x_t + \beta r_t + \epsilon_t),$$

for some $\beta > 0$. For instance, Deshmane *et al.* (2023) model affect assimilation that way. When the customer's affect (which is a generic term that encompasses mood and emotion) is high, the customer perceives any stimulus positively; whereas when the affect is low, all stimulus perceptions tend to be negative.

Multiplicity of Reference Points. Customers could also have multiple reference points, in which case r_t needs to be modeled as multivariate. For instance, Tereyağoğlu *et al.* (2018) document, in the context of pricing theater seats, loss aversion of customers wit respect to both

prices and seats sold: Consumers incur significant utility loss when prices are above their references or when the actual seat sales are lower than their reference points. In the context of project management, Baucells *et al.* (2024) consider reference points defined in terms of project cost and scope. In the context of video game design, Li *et al.* (2023) and Guo *et al.* (2024) consider the interplay between a reward reference point and a difficulty reference point. The balance of reward and difficulty aim to approximate the state of "flow" described by Csikszentmihalyi (2013). To remain in a state of flow, one needs to be constantly challenged, although gradually so as to match the individual's level of skills and motivation.

Functions of the Stimulus and Stocks

When stock variables are introduced (such as satiation, fatigue, difficulty, or reward), utility is usually defined as the incremental gain with or without stimuli (Baucells and Sarin, 2007):

$$u_t(x_t, s_t) = \mathbb{E}[v(x_t + s_t + \epsilon_t) - v(s_t)], \tag{3.5}$$

in which $v(.)$ is given by either (3.3) or (3.4).

As with reference points, there could be multiple stock variables to track, such as fatigue and fitness (Roels *et al.*, 2010). Each of these stock variables can have a positive or negative effect on utility.

Functions of the Stimulus, Reference Point, and Stocks

A more comprehensive model would account for all three variables (x_t, r_t, s_t). Different versions of the model arise, depending on what enters in the comparison. In the context of satiation and habit formation, Baucells and Sarin (2010) define satisfaction as

$$u_t(x_t, s_t) = \mathbb{E}[v(x_t - r_t + s_t + \epsilon_t) - v(s_t)],$$

desirability as

$$u_t(x_t, s_t) = \mathbb{E}[v(x_t - r_t + s_t + \epsilon_t) - v(s_t - r_t)],$$

and withdrawal (associated with no consumption) as

$$u_t(x_t, s_t) = v(s_t) - v(s_t - r_t).$$

3.3.6 Reference Point Adaptation

We first introduce the most common model of reference adaptation — exponential smoothing — and then discuss variations of the model (asymmetric adaptation, Bayesian updating of beliefs) and then non-recursive models of adaptation. We conclude by noting that reference point can be defined in terms of stimulus (as is commonly done), as well as in terms of reactions to stimulus (such as utility from them).

Exponential Smoothing. In most cases studied in the literature, the reference point adapts to past stimuli (Helson, 1964). Because of adaptation, there is a decreased response to repeated stimuli. A common form of adaptation mechanism is the exponential smoothing (Constantinides, 1990; Wathieu, 1997): For $t > 0$,

$$r_{t+1} = \alpha x_t + (1 - \alpha) r_t, \tag{3.6}$$

in which $0 < \alpha < 1$ denotes the speed of adaptation. In a continuous-time setting, this takes the form of Newton's law of cooling: $r'(t) = \alpha(x(t) - r(t))$, in which $x(t)$ is the stimulus rate at time t and $r(t)$ is the reference point at time t. This form of adaptation has been considered in Popescu and Wu (2007), Baucells and Sarin (2010), Das Gupta *et al.* (2015), Li *et al.* (2022), Li *et al.* (2023), Deshmane *et al.* (2023), and Guo *et al.* (2024) among others.

Asymmetric Adaptation. Aflaki and Popescu (2013) embeds prospect theory into the adaptation process. In the spirit of Figure 3.1 and similar to (3.4), the reference point adapts as in (3.6) when the service level improves over the reference point, *i.e.*, when $x_t - r_t > 0$, and it adapts more quickly in case the service level is less than the reference point, *i.e.*, when $x_t - r_t < 0$; specifically, they assume that $r_{t+1} = \lambda\alpha x_t + (1 - \lambda\alpha) r_t$ with $\lambda \in (1, 1/\alpha)$.

Bayesian Updating of Beliefs. When r_t denotes a customer's belief about a particular outcome, as in Ely *et al.* (2015), some Bayesian updating process may need to take place. If x_t denotes the signal and r_{t+1} denotes the posterior probability distribution after observing it,

then $r_{t+1} = f(x_t, r_t)$ for some function $f(.,.)$. Although this is different from (3.6), the components of the updating mechanism are similar. If customers are irrational, they may not engage into a Bayesian updating process but still learn using an exponential smoothing process (3.6); see, *e.g.*, Gaur and Park (2007).

Non-Recursive or Non-Additive Adaptation. All updating processes discussed so far are recursive. However, Baucells *et al.* (2011) do not find support for this recursive character. In particular, they report from an experiment that the reference point is often determined as a combination of the first and the last values of a time series, with intermediate values receiving smaller and nondecaying weights.

Similarly, Nasiry and Popescu (2011b) embed the celebrated peak-end rule into theadaptation process. Specifically, they assume that customers remember the peak and the end of a sequence, *i.e.*,

$$r_{t+1} = \alpha x_t + (1 - \alpha) m_t,$$

in which m_t is defined recursively as $m_t = \max\{m_{t-1}, x_t\}$. In a similar fashion, Roels (2020) considers generalized means (Hardy *et al.*, 1952), such as a multiplicative mean or a maximum. For instance, $r_{t+1} = \max\{x_t, r_t\}$ whenever some experiences leave a strong imprint, crossing a threshold according to which all future experiences will be compared.

Stimulus vs. Utility Adaptation. Whenever the service process involves a random component, which we modeled as ϵ_t above, the reference point may not be based on the stimulus itself x_t, but rather on the reaction to it, *e.g.*, $x_t + \epsilon_t$. For instance, the reference point could be shaped by past utilities, *i.e.*, $r_{t+1} = f(u_t(x_t, r_t), r_t)$ for some function $f(.,.)$, as in Deshmane *et al.* (2023) who attempt to capture the affect infusion (either assimilation or contrast) effect documented by Forgas (1995) and Schwarz and Clore (1983) among others. For instance,

$$r_{t+1} = \alpha u_t + (1 - \alpha) r_t. \tag{3.7}$$

In practice, utilities are hard to observe directly, but a service provider could survey its customers and obtain satisfaction ratings from the

different activities they experienced, *e.g.*, using customer reviews on online platforms (*e.g.*, TripAdvisor).

3.3.7 Stock Accumulation

In contrast to the reference point, which is of the same magnitude as the stimuli x_t, stocks accumulate over time and are, therefore, unbounded in principle. If the stock is not replenished by x_t, it decays at a certain rate, say $\gamma < 1$. Hence, a common model of stock transition is the following:

$$s_{t+1} = \gamma s_t + x_t, \tag{3.8}$$

in which $0 < \gamma < 1$. In a continuous-time setting, the stock accumulation takes the form of an integral: $s(t) = s_0 e^{-\gamma t} \int_0^t e^{-\gamma(t-s)} x(s) ds$, in which s_0 is a constant, $x(t)$ is the stimulus rate at time t, and $s(t)$ is the stock level at time t (Baucells and Zhao, 2019; Baucells and Zhao, 2020). This form of accumulation has been considered to model satiation (Baucells and Sarin, 2007; Baucells and Sarin, 2010; Caro and Martínez-de-Albéniz, 2012; Baucells and Zhao, 2020), fatigue (Baucells and Zhao, 2019; Roels, 2020), and fitness (Roels, 2020).

3.3.8 Aggregation of Utilities

The aggregation of utilities can be of two types: prospective or retrospective.

Prospective Evaluation

The main paradigm in decision analysis is the discounted utility model (Koopmans, 1960), according to which the prospective utility from an experience is the discounted sum of expected utilities, *i.e.*,

$$S(\mathbf{u}_T) = \sum_{t=1}^{T} \delta^t u_t(x_t, r_t, s_t), \tag{3.9}$$

with $0 < \delta < 1$.

Economic theory posits that people make decisions to maximize their discounted sum of utilities. However, O'Donoghue and Rabin (1999) point out that decisions are often time-inconsistent, suggesting a present

bias, leading to modeling the discount factor as hyperbolic. Plambeck and Wang (2013) discuss the operational implications of hyperbolic discounting for pricing and scheduling unpleasant services that may generate future benefits.

Retrospective Evaluation

Kahneman *et al.* (1997) point out that the remembered utility (or satisfaction) is often inconsistent with the decision utility, which is a measure of total utility inferred from choices, presumably based on the prospective evaluation of an experience. We review different ways to account for discounting utilities when forming retrospective evaluations, such as memory decay, primacy-recency effects, the peak-end rule, and a preference for upwards trends.

Memory Decay. A common bias characterizing retrospective evaluations is memory decay, *i.e.*, people put more weight on more recent events (Ebbinghaus, 1913). For instance, in package tracking services, customers punish early idleness less than late idleness, leaving higher delivery service scores when track-package activities cluster toward the end of the shipping horizon (Bray, 2023). Forgetting has also been extensively studied in OM as a dual of learning; see Lapré and Nembhard (2011). Accordingly, satisfaction could be modeled as follows:

$$S(\mathbf{u}_T) = \sum_{t=1}^{T} \delta^{T-t} u_t(x_t, r_t, s_t), \tag{3.10}$$

with $0 < \delta < 1$. This simple model has been adopted by Das Gupta *et al.*, 2015, Li *et al.*, 2023, and Deshmane *et al.* (2023) among others. Note that satisfaction (3.10) can be formulated as (3.9) with $\delta > 1$, up to a multiplying constant. Hence, the same model (3.10) can be used for prospective evaluations if $\delta < 1$ and retrospective evaluation if $\delta > 1$. Loewenstein and Prelec (1991) call this backward time discounting process "negative discounting."

Clearly, discounting the past would create a preference for a happy ending (Ross and Simonson, 1991), consistent with the peak-end rule (Kahneman *et al.*, 1993; Redelmeier and Kahneman, 1996).

Loewenstein and Prelec (1993) report that people express, in their prospective evaluation of sequences, a preference for increasing sequences, in contrast to the discounted utility paradigm, which would prescribe a preference for decreasing sequences. It may be because people express preference for sequences based on their anticipated *ex-post* satisfaction, *i.e.*, by assessing (3.10) as opposed to (3.9).

Primacy-Recency Effects. Some activities may leave a stronger imprint in memories than others. Ebbinghaus (1913) already noted the so-called serial-order effects, *i.e.*, the fact that people tend to remember the most the beginning and end of a sequence — also known as the primacy and recency biases. From a series of experiments, Kahneman *et al.* (1993) and Redelmeier and Kahneman (1996) conclude that the peak and the end of an experience matter more. Model (3.10) can easily be adapted to account for these effects. For instance, the primacy and recency effects, although they are not always salient depending on the valence of the event (Garnefeld and Steinhoff, 2013), can be captured by associated different weights with the first and last elements of the experience, *i.e.*,

$$S(\mathbf{u}_T) = \delta_b u_1(x_1, r_1, s_1) + \sum_{t=2}^{T-1} \delta^{T-t} u_t(x_t, r_t, s_t) + \delta_e u_T(x_T, r_T, s_T), \tag{3.11}$$

in which δ_b and δ_e are the discount factors corresponding to the beginning and end of the sequence. This is in fact quite similar to hyperbolic discounting, which discounts more the first activity in a sequence.

Peak-End Rule. Inspired by the peak-end rule, Li *et al.* (2022) associate a higher discount factor with the peak activity. They define the peak activity as the one that is associated with the highest service level. An alternative model could have been to define it as the one with the highest utility u_t. Following their approach, we thus have

$$S(\mathbf{u}_T) = \sum_{t=1}^{T} \Big(\mathbb{1}[x_t < \max_\tau x_\tau]\delta + \mathbb{1}[x_t = \max_\tau x_\tau]\delta_p \Big)^{T-t} u_t(x_t, r_t, s_t), \tag{3.12}$$

in which $\mathbb{1}[X] = 1$ if X is true and 0 otherwise, and $\delta_p \geq \delta$. While they focus on a situation where the peak activity is still discounted ($\delta_p < 1$), it may be that it is actually amplified ($\delta_p > 1$). Dixon and Verma (2013) reports evidence of peak-end effect in an opera's season concert schedule and Dixon and Thompson (2016) propose a scheduling algorithm that accounts for this bias. Given the larger imprint of the peak on memories, it makes sense to schedule the peak towards the end of an experience — an insight that is confirmed by Dixon *et al.* (2017) in a series of experiments.

Preference for Upwards Trends. Another aspect that may shape satisfaction is a preference for upwards trends (Varey and Kahneman, 1992; Ariely and Carmon, 2000). Dixon and Verma (2013) propose to measure this as the slope of a regression line passing through the different utilities, *i.e.*, $\sum_t (u_t - \overline{u})(t - T/2)/(\sum_t (u_t - \overline{u})^2)$ in which $\overline{u} = \sum_t u_t/T$. Dixon and Thompson (2016) posits that customer satisfaction is a weighted combination of the discounted utility model (3.10), perhaps with higher weights on some activities (peak, beginning, end), and the slope. With constant weights on activities, we would then have:

$$S(\mathbf{u}_T) = \sum_{t=1}^{T} \delta^{T-t} u_t(x_t, r_t, s_t) + \omega \frac{\sum_t (u_t - \overline{u})(t - T/2)}{\sum_t (u_t - \overline{u})^2}, \tag{3.13}$$

for some $\omega > 0$. However, it is important to note that customer preference for upward trends in stimuli is already captured in models of reference points with adaptation, with or without prospect theory (3.4), so before adding a term to (3.10), researchers need to investigate further whether the preference for upward trends needs to be defined in terms of utilities or in terms of stimuli.

3.3.9 Summary

Figure 3.3 puts together all the different versions of the models of utility formation and aggregation into the framework presented in Figure 3.2.

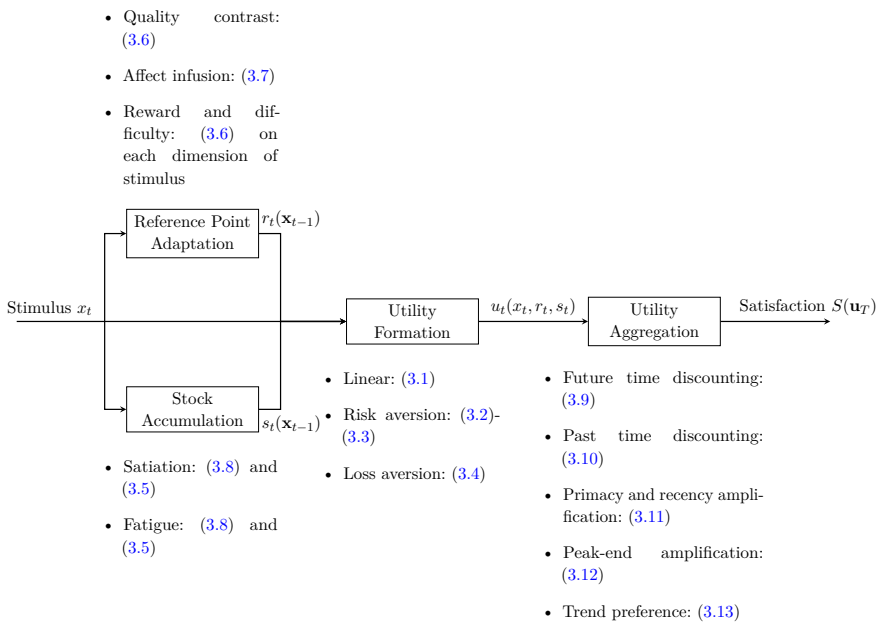


Figure 3.3: An Analytical Framework for Utility Formation and Aggregation

3.3.10 Properties of Optimal Sequencing Policies

For many service experiences, sequence matters. Yet, Palmer (2010, p.201) notes that the "contribution of models of service quality to a customer experience is limited by the neglect of order effects in most studies of service quality."

The optimal sequencing policy can be solved in closed-loop, *i.e.*, dynamically adjusted to changing states, or in open-loop, *i.e.*, decided once and for all at time 0. When there is no uncertainty or when its effects wash out when taking the expectation of utilities, the closed-loop policy and the open-loop policy are identical. For instance, there is no uncertainty in the mechanisms of memory decay, quality contrast, and satiation.

Uncertainty matters when customers are uncertain about a particular outcome and express a preference for suspense or surprise (Ely *et al.*, 2015). In this case, adopting an open-loop policy would eliminate any suspense or surprise. Uncertainty also matters when references are based

on utilities — *e.g.*, in the form of emotions, moods, or affects (Deshmane *et al.*, 2023).

We next highlight a few key design principles, which hold when there is no precedence constraint among activities:

- Under memory decay alone, crescendos are optimal. Because customers remember more the end, higher service levels need to be scheduled later.

- Under forward-time discounting, decrescendos are optimal. This is because customers are impatient to consume.

- Under quality contrast, with a reference point that adapts over time (3.6), and no time discounting, crescendos are optimal. Because customers derive positive utility from positive increments in service levels, and negative utility from negative increments, the latter should be avoided as much as possible.

- Under both memory decay and adaptation, a U-shape sequence is optimal (Das Gupta *et al.*, 2015). (The U-shape may degenerate into a crescendo.). The intuition is that customers put more weight at the end (due to memory decay) and derive positive utility from positive increments (due to adaptation and quality contrast). It is thus optimal to have a large positive increment at the end. To accentuate this last increment, a negative increment may need to be experienced at the beginning of the sequence. Even though this negative increment yields negative utility, it will be quickly forgotten.

- Under forward-time discounting and adaptation, a U-shape sequence is optimal (Wathieu, 1997) — this time, to trade off customer's preference for positive increments with customer's impatience.

- With memory decay, adaptation, and a higher memory associated with the peak service level, an interior peak may be optimal (Li *et al.*, 2022). This happens typically when the discount rate associated with the peak is neither too small (otherwise the sequence

degenerates into a U-shape) or too large (in which case it is always optimal to put the peak at the end).

- With forward-time-discounting and satiation or fatigue, a U-shape sequence is optimal (Baucells and Sarin, 2007; Baucells and Zhao, 2019). This pattern is driven by the customer's impatience to consume and the following need to de-satiate or rest.

- Non-linear utilities, such as those associated with prospect theory (3.4) can induce local peaks (Chen *et al.*, 2024a), perhaps as an attempt to approximate a constant service level throughout (Nasiry and Popescu, 2011b).

- Affect assimilation may also lead to the placement of local peaks (Deshmane *et al.*, 2023). This can be done proactively to ensure that the customer is in good spirits to enjoy future activities, enabling the service provider to "ride the wave" by investing in a peak and benefiting from its positive spillovers subsequently. But this can also be done reactively after a service failure to recover from it and avoid facing a situation where a customer's negative spirits may taint all subsequent evaluations of activities.

These principles complement, and enrich in some cases, the principles outlined by Chase and Dasu (2001b):

- Finish strong;

- Get the bad experience out of the way early;

- Segment the pleasure, combine the pain;

- Build commitment through choice;

- Give people rituals, and stick to them.

In our opinion, the analytical literature on service experience design has refined the first three principles, but has largely ignored the latter two. One type of refinement is that the "finishing strong" recommendation should really be understood as "finishing steep" since, in the presence

of adaptation, the customer cares mostly about the steepness of the gradient than the actual service level.

Further developments of service experience design could also revolve around the incorporation of other behavioral biases that have not been considered so far (Karmarkar and Karmarkar, 2014). In particular, there appears to be little work on customer engagement, unless it is a derivative of customer satisfaction.

In general, optimizing a service experience design is a complex problem for three reasons. First, it may involve nonlinearities, due to the nonlinear nature of the utility function as in (3.4) or the nonlinear nature of the adaptation process if it involves taking a maximum function as in Nasiry and Popescu (2011a) and Roels (2020). Second, when it involves sequencing activities or selecting activities, it is a combinatorial problem involving $T!$ possible permutations. Third, if uncertainties are not washed away when taking expectations as in Deshmane *et al.* (2023), it involves solving a stochastic dynamic optimization problem. Dixon and Thompson (2016), Baucells and Zhao (2020), and Li and Qi (2022) propose different algorithms to solve various versions of the problem.

3.3.11 Implications for Resource Management

Although service levels are often assumed to be exogenous, they can sometimes be optimized in practice by allocating more or less resources at the different stages of the service encounter (Soteriou and Chase, 2000). For instance, Ahmadi (1997) and Rajaram and Ahmadi (2003) consider how to optimize the ride capacity of a theme park to minimize congestion and maximize merchandise sales. In the spirit of co-production reviewed in Chapter 2, Bellos and Kavadias (2019) and Bellos and Kavadias (2021) study which stages of a process should involve the active participation of customers to maximize the sum of utilitarian and hedonic values from the service.

Rather than adding capacity to process customers faster, which is expensive, Baron *et al.* (2014) propose to reduce the workload at some stages by strategically idling servers upstream. This policy, which is not work-conserving, leads to longer throughput times overall, but may help reduce the peak waits at some stations, which may result

in higher customer satisfaction. Baron *et al.* (2017) propose to embed strategic idling into a dynamic scheduling policy in an open job shop. In their model, the service organization needs to decide when to dispatch customers from one station to another (respecting some precedence constraints) and may ask servers to strategically delay their service starting time so as to avoid inducing congestion on their downstream stages. Chen *et al.* (2024b) apply a version of this policy to a healthcare facility, actively maintaining a pool of customers to dispatch from one station to the other to avoid peak congestion.

Overall, these works bring the spirit of service experience design into the core OM issue of congestion management. While the key performance metrics used in these applications revolve around the total throughput time and the peak time at any given stage, one could envision expanding these models to account for satiation or quality contrast, among other factors.

3.4 Beyond the Service Encounter

In this section, we briefly touch upon two aspects of service experience design that fall outside a single encounter, namely the anticipation and the recall of the encounter (§3.4.1) and the management of customer journeys, operating across encounters (§3.4.2).

3.4.1 Before and After the Encounter: Anticipation and Recall

Customers may derive satisfaction anticipating the experience as well as recalling the experience after it has happened. On anticipation, Oliver (2014) posits that customers form preconsumption expectancies, observe product (attribute) performance, compare performance with expectations, form disconfirmation perceptions, combine these perceptions with expectation levels, and form satisfaction judgments.

One interpretation of the expectation disconfirmacy framework can be cast in terms of the price customers pay to obtain the service. In most services, contracting happens before the service delivery (Karmarkar and Pitbladdo, 1995). Accordingly, customers will evaluate whether they are willing to obtain the service if their expected utility from it,

say, $\mathbb{E}[u(x+\epsilon)]$, exceeds the price charged for it, denoted as p. If they are willing to pay for it, customers obtain a realized utility $u(x+\epsilon)$ from the service. Their net utility is therefore $u(x+\epsilon)-p$. The marginal customer, whose anticipation of the service value is exactly equal to $\mathbb{E}[u(x+\epsilon)]$, thus obtains $u(x+\epsilon)-\mathbb{E}[u(x+\epsilon)]$. Anderson and Sullivan (1993) posit that this gap is a key driver of customer satisfaction and repurchase intentions. This gap can be further decomposed into various quality gaps, in the spirit of §2.4.3 (Karmarkar and Roels, 2015).

On recall, Cowley (2014) highlight the salience of the peaks and troughs in customers' memories of experience, consistent with the celebrated peak-end rule (Kahneman *et al.*, 1993). However, new peaks may emerge during the phase of recall itself.

Combining anticipation and recall, Baucells and Bellezza (2017) offer a fairly comprehensive model that accounts for the conceptual consumption (Ariely and Norton, 2009), defined as psychological consumption that is temporally dissociated from physical consumption; adaptation (Helson, 1964; Wathieu, 1997) as in (3.6); and the time distance through time discounting as in (3.9) or (3.10).

3.4.2 Across Encounters: Customer Journeys

While most of the literature on service experience design has focused on a single service encounter, a more holistic approach should consider the entire customer's journey with a service organization, consisting of multiple encounters, with potential churns in between. Lemon and Verhoef (2016) conceptualize the total customer experience as a dynamic process. The customer experience process flows through different phases of prepurchase (including search) to purchase and then to postpurchase; it is iterative and dynamic. This process incorporates past experiences (including previous purchases) as well as external factors. In each stage, customers experience touch points, only some of which are under the firm's control.

Popescu and Wu (2007) and Nasiry and Popescu (2011a) also consider a multi-encounter setting. Within that context, they study dynamic pricing strategies for an aggregate population that is behaviorally biased. Aggregate models of population churn are typically modeled using

Markov chains (Heitz *et al.*, 2011; Ascarza and Hardie, 2013).

Considering customer journeys at the individual level, Aflaki and Popescu (2013) and Guo *et al.* (2024) explicitly model the churn and retention probability. While Aflaki and Popescu (2013) model a customer's retention probability as a function of their anticipated utility from future encounters (shaped by their past experience), Guo *et al.* (2024) model it as a function of their recalled satisfaction. This is where recall and anticipation, discussed in §3.4.1, get mixed.

In such models, the service provider could act strategically across encounters to shape customers' expectation and maximize retention. Using a multi-armed bandit framework, Kanoria *et al.* (2023) model a service firm as pulling one of two arms, namely, a safe mode and a risky mode, to keep a customer satisfied and prevent churn. They find that the firm should use the safe service mode when the customer is marginally satisfied and the risky service mode when the customer is marginally unsatisfied. If the service provider is starting its business and has thus unknown capability (a situation many restaurants face), we conjecture they would want to over-deliver first to build their reputation, in the spirit of the career concerns in organizations (Holmström, 1999).

Customers could also be captive, but modulate their consumption over time, across encounters, with significant consequences for resource management and congestion. In subscription services, customers indeed pay a flat fee and then decide the frequency of usage of the service. Arkes and Blumer (1985) report that, presumably because of sunk cost fallacy, customers who had initially paid more for a season subscription to a theater series attended more plays during the next six months. Building on that phenomenon, Bhaskaran *et al.* (2022) propose that firms should control consumption by lowering prices, in contrast to the classical recommendation to pursue admission control by raising prices. An alternative would be to impost time limits on usage (Feldman and Segev, 2022). On the other hand, subscription services are also fraught with overconfidence about self-control and efficiency, leading people to use less a service under a membership than under a pay-per-use scheme (Della Vigna and Malmendier, 2006).

3.5 Further Developments

By and large, the literature on service experience design can be classified as either documenting phenomena through experimentation (and, thereby, relating to social psychology or behavioral economics) or making prescriptions with analytical models (and, thereby, relating to decision analysis or operations research).

Unfortunately, these research communities turn out to often be distinct, resulting in perhaps overly simplistic prescriptions that emerge from the lab (such as the "peak-end" rule, which remains silent about the overall optimal sequence) and models that may be hard to implement in practice.

We thus foresee a large research opportunity to attempt to *bridge these two communities*, using a data-driven prescriptive approach to service experience design, in same lines as the ones developed by Deshmane *et al.* (2023). Experiences could be individualized and adjusted dynamically to changes of states. What the abundance of data generated by online platforms and the flexibility of service delivery empowered by AI, there is a huge opportunity to develop a new field of research on service experience optimization.

While new services (*e.g.*, online food delivery platforms) generate tremendous amounts of data, we should not dismiss the accumulated experience humanity has built in experience design through storytelling. In particular, Rozin and Rozin (2018) acknowledge the experience accumulated in classical music and, more recently, multi-course tasting menus, connecting some best practices with findings from social psychology, and making improvement suggestions for enhancing the experiential character of multi-course menus. Although their approach involves small data (*e.g.*, surveys), the development of LLMs offers a novel opportunity to *revisit this stock of accumulated knowledge* to identify what made some stories more popular than others, going through books, movies, plays, and music, among others.

New applications may also emerge. For instance, the focus of this stream of research has so far been provider-routed experiences (*e.g.*, shows, concerts), but one could also investigate experience design for *customer-routed services*, such as theme parks (Ahmadi, 1997) or grocery

stores (Moon, 2024), and the power of behavioral nudges to customers to follow a path that suits them the most. Also, while most of the standard applications of service experience design have focused on physical experiences, such as a concerts (Baucells and Zhao, 2020) and touristic tours (Deshmane *et al.*, 2023), different design principles might be relevant for *digital experiences*, such as online games (Li *et al.*, 2023). Investigating the similarities and differences between physical and online experiences is another promising line of research.

4

Fostering Employee Engagement by Putting People First

In the early days of research on service management — in particular, service operations given OM's focus on what happens within organizations — employees were often conceptualized as unhumanistic servers, solely characterized by a constant service rate. Indeed, labor was typically perceived to be a costly resource to be minimized subject to a service constraint. This conceptualization, while still prevalent in many practical settings, could have dangerous implications for the long-term viability of a business as it leads to low investment in people development, resulting in a lack of engagement and high turnover.

The SPC, presented in Figure 1.4, takes a different stance: Instead of cutting down on labor expenses, it suggests investing in people development (the "internal service quality") to drive their productivity and retention and achieve a high external service quality. We first review this argument in §4.1.

We then review research on employee management that has contributed to moving away from this unhumanistic perspective towards a people-centric approach. This literature builds on economics and social psychology to account, on one hand, for people's strategic behavior and, on the other hand, their predictable biases. We consider the following

aspects: employees' motivation and reward (§4.2), staffing decisions (§4.3), job design (§4.4), and organizational culture (§4.5). We conclude with some directions for further research in §4.6.

Overall, this chapter advocates for adopting a more humanistic perspective on labor, following the call by Ton (2014) to invest in "good jobs" and the call by Corbett (2024) to leverage the accumulated knowledge of OM to foster well-being in the workplace.

4.1 The Good Jobs Strategy

4.1.1 An Early, Unhumanistic Perspective on Employees

Early research in service management typically considered employees as inanimate objects. Starting from the early work by Erlang (1909) on queuing theory, service employees were often modeled as exogenous and identically distributed random variables — the parameter μ in queuing parlance. In particular, there was absolutely no consideration of either people's individual traits or their capabilities, motivation, and opportunities. A typical staffing model often consisted of balancing the costs of waiting experienced by customers with the costs of labor, viewing the latter as a scalable commodity. For instance, Holt *et al.* (1956) treat workforce as just another variable, along with inventory, production, and shipments.

Sadly, this unhumanistic view has also permeated many practical settings. The name of many organizations' people department — "Human Resources" (HR) — is quite telling, drawing an immediate parallel between people and physical or financial resources. People are assets; yet, not assets we invest in, but rather assets that tie in cash.

This financial perspective on labor probably stems from the fact that, in many people-intensive services, labor constitutes a large portion of the service provider's cost. It is, therefore, the first thing that comes to mind when one attempts to be margin-conscious. For instance, in retail, labor accounts for 40%-60% of a retailer's cost (HSO, 2023). Given that retailers' net profit margin has been declining to hover around 1-2% (Calcbench, 2023), it seems paramount to save on personnel costs.

4.1.2 The Service Profit Chain: Emphasizing Employee Engagement

Contrary to this unhumanistic view, the SPC, presented in Figure 1.4, posits that investing in people development will pay off in the long-run. The argument works as follows: Fostering employee satisfaction — which was later revisited as their "engagement" (Heskett and Sasser, 2010) — boosts their productivity and retention. The latter would then drive external service quality, from which is derived customer satisfaction. Satisfied customers are posited to be loyal and induce referrals, which would then lead to the service organization's revenue growth and profitability. The SPC's statement on the essential role of employees in driving service quality is consistent with, and perhaps inspired by, the early evidence of correlation between employee satisfaction and customer satisfaction observed in banking by Schneider *et al.* (1980).

Evidence of the Key Role of Employees in Driving Service Value

The left-part of the SPC has received some empirical support, traditionally based on surveys and case studies. See Hogreve *et al.* (2017) for a meta-study. We next review studies across various industries — namely, healthcare, retail, and call centers.

In the context of healthcare, Goldstein (2003) reports clear evidence of SPC linkages from employee development (which includes systems for work and job design, training and development, and attention to employee well-being) to both employee productivity and employee satisfaction; and from the latter to customer satisfaction. However, only some SPC linkages to financial performance turned out to be significant.

In retail, Maxham III *et al.* (2008) report statistical significance of the linkages between employees' job perceptions (conscientiousness, perceived organizational justice, and organizational identification), their job performance (in-role performance, extra-role performance toward customers, and extra-role performance toward the organization), customers' evaluations (satisfaction, purchase intent, loyalty, and word-of-mouth composite), and finally, store performance (customer spending and comparable store sales growth). Also in retail, Fisher *et al.* (2006) report that customer ratings of employee knowledge have a direct effect on customer satisfaction, as well as an indirect one — through customers'

perception of in-stock availability. Briggs *et al.* (2020) find that retailers' service orientation directly influences employee satisfaction and customer relationship performance; yet, they do not obtain support for a direct relationship between employee satisfaction and customer relationship performance when controlling for the effects of service orientation.

For call centers, Chicu *et al.* (2019) find that job design and job discretion affect employee satisfaction, whereas training affects employee productivity, which in turn, relates to employee satisfaction. Employee satisfaction leads to higher retention, but does not necessarily encourage productivity. Subsequently, employee retention is a key mediator in the relationships between employee satisfaction and customer satisfaction and between employee performance and customer satisfaction.

From these numerous pieces of evidence across different contexts, it is clear that employees lie at the core of the process of value creation and should, therefore, not be relegated as an expendable resource. Building on this argument, Ton (2014) advocates to invest in a "good jobs" strategy. Many retailers, she claims, are caught in a vicious circle: They have limited labor budgets, which leads them to under-invest in people — either having too little staff or having them underskilled. This underinvestment in people leads to poor operational execution of the stores, such as misplaced or stocked-out items, lack of assistance, long waiting lines, or filthiness. In the long run, this hurts store sales and profitability, making labor budgets even tighter. Ton (2014) argues that the vicious circle can be reversed and turned into a virtuous circle: By investing in good jobs, store execution improves, which leads to higher sales and profits, and offers an opportunity to invest more in people development.

4.2 Motivation and Reward

One of the earliest developments towards a more people-centric perspective to labor stemmed from the realization that people respond to incentives and need to be motivated to exert effort. These studies originally built on the principal-agent literature in economics, but have increasingly accounted for social preferences and behavioral biases. We

structure this discussion in two parts, reviewing first the monetary incentives (§4.2.1) and then the non-monetary ones (§4.2.2). Given the breadth of the field of personnel economics (Lazear and Gibbs, 2014), our review will be partial only.

4.2.1 Monetary Rewards

This section investigates the motivational effects of monetary rewards. We first review the benefits and challenges of rewarding employees for their performance, not only at the individual level but also in terms of social comparisons, and then discuss the challenges of raising salaries whenever this comes from a regulatory mandate.

Pay per Performance

Building on agency theory (Jensen and Meckling, 1976; Holmström, 1979), a large body of work in both Marketing and OM has studied compensation of service employees — particularly salesforce — though not always with a particular service focus; see Basu *et al.* (1985), John and Weitz (1989), and Albers (1996).

Pay-per-performance contracts should be designed according to a certain "information principle" (Holmström, 1979): Any information about employees' actions, however imperfect, should be used when assessing their performance. According to this information principle, combining team-level and individual-level incentives helps triangulate the information received from different sources and lower their noise. Moreover, such portfolios of incentives help induce cooperative behavior, *i.e.*, encourage employees to account for the synergies in their respective tasks, help each other, or share knowledge (Siemsen *et al.*, 2007).

In the context of queuing theory (the dominant model of service operations), the consideration of servers as people who respond to incentives (as opposed to unhumanistic random variables) flips a classical result in queuing theory: While a pooled facility (*i.e.*, single line) reduces variability in servers' idle times, and is thus traditionally associated with better operational performance (Kleinrock, 1974), the competitive dynamics that arise from high-powered (*i.e.*, pay-per-performance) incentives in dedicated facilities (*i.e.*, separate lines) may overturn the

benefits of pooling (Kalai *et al.*, 1992; Gilbert and Weng, 1998). This is because paying servers per customer served induces competition among them, making them work faster. Given that competition is more intense in dedicated facilities than in a pooled one, the classical queuing theory result, showing the dominance of pooling, may flip.

Given their motivational effect, why are pay-per-performance not ubiquitous in practice? While they are common in salesforce or franchise management, many organizations still use fixed wages. To explain why such high-powered incentives contracts are only adopted selectively, Holmström and Milgrom (1994) propose, with a multi-task model (Holmström and Milgrom, 1991), that asset ownership and job restrictions are linked to the nature of incentives provided. In particular, employment contracts (*i.e.*, within organizations) are characterized with weak incentives for maintaining asset values, weak incentives for narrowly measured performance, and significant restrictions on worker freedom.

Moreover, pay-per-performance contracts may be associated with high psychological costs such as overconfidence and social comparison (Larkin *et al.*, 2012). They can also be associated with long-term and serious mental health problems, leading to a greater usage of antidepressant and antianxiety medication, especially among low-performing and older workers (Dahl and Pierce, 2020).

Hence, one should not be fooled by the theoretically presumed benefit of pay-per-performance contracts. While they are effective in certain situations, the world may be more complex than what is captured in simple models (*e.g.*, multiple tasks may need to be attended) and one should not forget that employees are human beings subject to fatigue and burnout. Moreover, fixed wages can still be motivating, through the prospect of career advancement (Holmström, 1999).

Salary-Based Social Comparisons

Rewarding employees differently, in line with their differences in individual performance, inevitably induces social comparisons whenever salaries are made public, leading to positive or negative feelings and affecting employee's incentives to help each other (Long and Nasiry, 2020). In particular, Card *et al.* (2012) report from a natural experiment

that workers with salaries below the median for their pay unit and occupation report lower pay and job satisfaction and a significant increase in the likelihood of looking for a new job, whereas above-median earners are unaffected about pay transparency. Chan *et al.* (2014a) also report peer effects in department stores. In line with the theoretical model of Siemsen *et al.* (2007), they find that compensation systems influence worker incentives to help and compete with peers within the same firm.

Raising Salaries

Why do service organizations not pay higher wages and break the vicious circle of the "bad jobs" strategy described by Ton (2014)? Rahmandad and Ton (2020) posit the existence of different equilibria, one characterized by a high employee compensation and rich job design (the "good jobs" strategy) and another characterized by a low employee compensation and narrow job design (the "bad jobs" strategy). Using a system dynamics model, they explore how firms discover, move to, and remain at the "good jobs" equilibrium. They uncover several challenges, one of which is how to cope with demand variability with limited inventory buffer — a perennial problem in retail. As a result, they conclude, efforts to adjust labor supply to highly variable demand in services often lead to unstable schedules given with short notice. This schedule unpredictability, in turn, drives high-quality employees away and compromises the good jobs strategy.

A recent line of empirical work has studied the impact of government mandates to raise minimum wages. The general finding is that it hurts service organizations (though with some heterogeneity) and might hurt their employees as well. Specifically, Agarwal *et al.* (2024) report that increasing the minimum wage in the hotel industry could reduce average hotel revenues and occupancy rates. However, the response to an increase in minimum wage is heterogeneous. Specifically, hotels in the middle-end of the market are potentially the most severely affected by a raise in minimum wages because of two reasons. First, they face price-sensitive customers and have therefore limited ability to pass through the cost increase to their customers. Second, they might suffer the most from losses in occupancy rates and revenues due to the downgrading of their

quality as they cope with higher costs.

A naive view would conclude that higher wages are beneficial for employees. However, this would not account for the firms' response to these higher wages. Yu *et al.* (2023) study the change in firms' scheduling practices as a result of higher wages and find that their net effect is to lead to more precarious working conditions. Specifically, they obtain that a $1-increase in the minimum wage, although having a negligible impact on the total labor hours used by the stores, leads to a 27.7% increase in the number of workers scheduled per week, but a 19.4% reduction in weekly hours per worker.

In sum, it appears that, to be effective, a good jobs strategy needs to be adopted by the service organizations themselves and be an integral part of their business strategy; if not, it could backfire through the firms' responses in service design and personnel management.

4.2.2 Non-Monetary Rewards

We next discuss employees' non-monetary incentives, revolving first around the meaningfulness of their work and second around peer effects.

Meaningfulness of the Work

Besides financial incentives, employees of a service organization can be intrinsically motivated to deliver excellent service because they find meaning in their work, due to the reciprocal nature of their interaction with customers (see Chapter 2). Indeed, in one of the earliest studies on co-production, Fuchs (1968, p. 189) notes that "the direct confrontation between consumer and worker that occurs frequently in services creates the possibility of a more completely human and satisfying work experience... at their best many service occupations are extremely rewarding, and the line between "work" and "leisure" activity is often difficult to draw."

Thanks to the abundant datasets stemming from digital technologies, this customer-employee reciprocity can now be studied rigorously. For instance, Altman *et al.* (2021) use micro-level data on call center interactions to report that the emotional load created by negative customer emotions increases agents' response time (*i.e.*, the elapsed time

between each customer message and the agent's response), the length of agents' messages, and the required number of messages needed to complete a service request. Conversely, a long agent response time and a high number of messages produce more negative customer emotions.

In the context of queuing theory, fostering this feeling of reciprocity between customers and servers — by developing a greater sense of "customer ownership" — appears to induce servers to work harder. Considering an emergency department that switched from assigning patients to the next available doctor (as is done in a pooled queuing configuration) to assigning them to doctors upon check-in in a round-robin fashion (which is consistent with a dedicated queuing configuration), Song *et al.* (2020) report shorter waiting times and lengths of stay, without adversarial effects on service quality — in contrast to classical queuing theory that posits that a pooled queuing configuration dominates a dedicated one in terms of operational performance (Kleinrock, 1974). Based on extensive interviews of the doctors involved in this change of configuration, Song *et al.* (2020) suggest that a dedicated queuing configuration is associated with heightened feelings of customer ownership. Similar evidence has been reported in the assignment of client portfolios to account representatives in banks (Jouini *et al.*, 2008). This result can be partly, but not fully, explained by the fact that the sharing of customers in a pooled setting leads to a dilution of customer ownership, resulting in less incentives to work hard to process the queue quickly (Armony *et al.*, 2021). Hence, the level of customer ownership must be much higher in a dedicated configuration than in a pooled one, exceeding the uniform splitting induced by sharing responsibilities.

To complement these observational studies and analytical models, several lab experiment studies have attempted to identify the underlying mechanisms at work. Whereas Shunko *et al.* (2018) stresses the importance of making the line visible, Song *et al.* (2024) report that it may not be enough — awareness of the queue seems to be what matters. Both studies report a higher operational performance of dedicated queues when servers have discretion over their task.

In sum, dedicated queues are not only associated with a higher degree of competition among servers under pay-per-performance contracts (Gilbert and Weng, 1998), they also appear to be associated with a higher

degree of customer ownership. When combined, these two effects could reverse the traditional thinking — based on a unhumanistic perspective of servers — that a pooled queuing configuration always leads to higher operational performance. Studying the queues of competing checkout counter clerks at a grocery store, Wang and Zhou (2018) disentangle the effects of the sharing of responsibilities, a/k/a social loafing or free riding, in a pooled setting; the competition among servers to capture market share; and the negative effect queue length on service times — presumably due to customer ownership.

The discussion so far has revolved around a noble goal — feelings of reciprocity or customer ownership — but perhaps a more basic need that employees attempt to fulfill is a preference for breaks. Considering such a preference for idleness, Gopalakrishnan *et al.* (2016) and Zhong *et al.* (2023) characterize the optimal staffing and routing in a pooled queuing system with strategic servers.

In sum, workers can be motivated by monetary incentives as well as by the intrinsic nature of their work. However, these two sources of incentives could potentially conflict — the economic incentives potentially undermining the employees' intrinsic motivation (Bénabou and Tirole, 2003). Hence it is paramount to find the right balance between monetary incentives and intrinsic motivation.

Peer Effects

Similar to the peer effects that may arise when workers compare their salaries (as discussed in §4.2.1), peer effects may also prevail when workers observe their respective efforts and operational performance. As argued by Mas and Moretti (2009), employee heterogeneity may have dual effects. On the one hand, the introduction of a high productivity worker could lower the effort of incumbent workers because of free riding. On the other hand, it could increase the effort of incumbent workers because of peer effects induced by social norms, social pressure, or learning. We discuss in turn these two sides of peer effect.

Positive Spillovers. In the context of a grocery chain, Mas and Moretti (2009) find strong evidence of positive productivity spillovers from the

introduction of highly productive personnel into a shift: A 10% increase in average co-worker permanent productivity is associated with 1.7% increase in a worker's effort. Using an analytical model, Roels and Su (2014) posit that social comparisons always heighten performance. However, the distribution of performance depends on whether people are ahead-seeking (*i.e.*, deriving positive utility from being ahead of others) or behind averse (*i.e.*, deriving negative utility from falling behind others): Whereas ahead-seeking behavior drives polarization of outcomes, behind-averse behavior drives clustering.

Peer effects could also offer long-term benefits in terms of learning. In particular, Chan *et al.* (2014b) report, in the context of a department store, that working with high-ability peers substantially increases the long-term productivity growth of new salespeople, not only through direct teaching but also because it enables inexperienced employees to observe successful sales techniques. In healthcare, Song *et al.* (2018) find that publicly disclosing the relative performance of doctors along operational performance metrics (such as length of stay), which tend to be overlooked by many doctors, offers an opportunity to identify their top-performing coworkers, which in turn helps identify and validate best work practices. This relative performance feedback not only leads to an increase in the average productivity, it also reduces the variation in productivity across providers, stemming from bottom-ranked workers exhibiting differentially large improvements in productivity.

These positive spillovers may not only affect the sharing of good practices, however: Bad practices can also be shared through peer effects. In the context of restaurants, Chan *et al.* (2021) report that servers are more likely to steal when working with high-theft peers.

Negative Spillovers. Consider next the negative spillovers. First, the presence of a high-ability workers may lead others to free-ride (Alchian and Demsetz, 1972; Holmström, 1982) — a practice more formally called social loafing (Latané, 1981). Peer effects could also diminish productivity when feelings of reciprocity come into the way of financial incentives, especially when working alongside friends (Bandiera *et al.*, 2005).

Peer effects could also bias the perspective of a common asses-

sor. Cassar and Ko (2023) report evidence of peer effects in teaching evaluations. They find evidence of quality contrast (as in §3.3.5), *i.e.*, subjective performance ratings are lower for employees with higher-quality peer groups. Although they do not measure the impact of these evaluations on motivation, we can certainly infer the negative impact high-ability co-workers could have on their peers' motivation.

Mixed Effects. The combination of positive and negative spillovers can result in non-monotone effects overall. Studying heterogeneous teams of restaurant waiters, Tan and Netessine (2019) find that the presence of a high-ability waiter has a positive spillover when the other waiters' ability is low, prompting them to redouble both upselling and cross-selling efforts, but a negative spillover when the the other waiters' ability is high, prompting them to reduce sales efforts.

4.3 Staffing

Staffing decisions consist of two dimensions: a quantity dimension, *i.e.*, how many employees to hire (§4.3.1); and a quality dimension, *i.e.*, what skill level to hire (§4.3.2). Although these two dimensions are interconnected, we explore them separately.

4.3.1 Quantity

While traditional approaches to staffing have typically considered workers as inanimate objects, more recent approaches have attempted to account for their human elements, first as exogenous factors (e.g., no-show probability, deterministic improvements in productivity) and then as endogenous factors (e.g., impact of workload on absenteeism or on productivity), leading to the development of a more people-centric approach to labor scheduling.

Traditional Approaches to Staffing

Labor has traditionally been considered as a resource, along with physical (*e.g.*, inventory, capacity) and financial resources. In particular,

staffing models often reduce labor to a continuous variable with linear costs. Due to the inherently stochastic nature of services (which production is typically "pulled" by customers, consistent with Figure A.2), demand is typically modeled as a random variable. This inanimate perspective on labor reduces staffing to a newsvendor problem (Harrison and Zeevi, 2005): hire too much, and your staff will be idle; hire too little, and your staff will be overworked.

In the presence of congestion, *i.e.*, when customers create externalities among each other, the capacity requirements typically stem from a queuing model. Two types of uncertainty arise in this context: uncertainty about the overall demand rate and, for a given demand rate, the inherent stochastic variability arising in the queuing system. In a quality and efficiency-driven (QED) regime characterized by Halfin and Whitt (1981), a square-root law of staffing is optimal (Gans *et al.*, 2003), consistent with the newsvendor analogy applied to a Poisson random variable. Specifically, in a system with an arrival rate λ and a per-server service rate μ, the total number of servers should be about $(\lambda/\mu) + \beta\sqrt{\lambda/\mu}$, in which β is chosen to guarantee a particular service level. This square-root safety staffing logic is no longer valid when demand uncertainty dominates stochastic variability, but simple capacity prescriptions can still be derived using a suitable newsvendor problem (Bassamboo *et al.*, 2010).

When employees have different skills and customers different needs, the staffing problem embeds an assignment problem (namely, which type of employee should serve each type of customer), offering risk-pooling opportunities when a particular type of employee can serve multiple types of customers. A typical approach consists of solving the problem in two steps, by first identifying the required safety buffers from queuing models; and by then optimizing the scheduling of staff subject to various constraints, such as the requirement for working complete shifts, using linear optimization. See Ernst *et al.* (2004) for a review of staffing and scheduling models in specific service industries. More integrative methods combine linear optimization with either queuing analysis (Harrison and Zeevi, 2005) or simulation (Cezik and L'Ecuyer, 2008).

Uncertain and Dynamic Supply

Moving beyond the (unrealistic) assumption that labor supply is abundant and easily adjustable, some staffing models incorporate the fact that employees could fail to show up or could leave the company; see, *e.g.*, Whitt (2006) and Berenguer *et al.* (2024). The dimension of supply uncertainty is especially salient in on-demand platforms, on which workers self-select to work or not, with no long-term financial consequences if they opt to not work on a given day (Gurvich *et al.*, 2019).

In addition to the potential shortage of supply, learning effects could make employees more efficient over time, as studied by Gans and Zhou (2002).

Externalities

Although the early works on staffing consider staffing and demand as independent, there are connected in two ways. On the one hand, a high workload may lead to a high rate of absenteeism, as reported by Green *et al.* (2013). Failing to incorporate absenteeism as an endogenous variable results in understaffing. To mitigate the negative effects of absenteeism, a multi-unit service organization, such as a hospital, could cross-train staff across departments (Ryu and Jiang, 2025; Yuan, 2025)

On the other hand, the staffing level may have a nonlinear impact on sales (beyond the newsvendor model's kinked functional form, which assumes that sales are the minimum between demand and capacity). Specifically, Perdikaki *et al.* (2012) report that retail store sales volume exhibits diminishing returns to scale with respect to traffic, and that labor moderates the impact of traffic on sales. Similarly, Mani *et al.* (2015) argue that eliminating understaffing in retail stores can result in a significant increase in sales and profitability. Building on these insights, Chuang *et al.* (2016) model a sales response function based on labor adequacy (the labor to traffic ratio) and they embed it into a labor-planning model of an apparel retail chain. Along similar lines, Fisher *et al.* (2021b) estimate how revenue varies with the staffing level at each store of a retail chain and propose a novel staffing model, tailored per store.

A People-Centric Approach to Labor Scheduling

Staff schedules are often subject to last-minute changes and, consequently, quite erratic. We posit that it is because they are often determined as outputs of linear optimization problems (or versions thereof), which get anchored on corner solutions, similar to Manufacturing Requirement Planning (MRP) systems, which are known to respond "nervously" to small changes in demand (Blackburn *et al.*, 1986). While managers who perceive their employees as inanimate human resources may have never considered this erratic nature of schedules to be a problem, a recent stream of research and practice on people-centric operations has studied its negative consequences and how to remedy it. We separately consider the perspectives of employees and contractors, given that the latter may inherently prefer scheduling flexibility.

Employees. For employees, the empirical literature makes a clear case for the detrimental effects of employees' erratic schedules on firm profitability. Fortunately, a little bit of advance notice appears to go a long way. As an example of such studies, Kamalahmadi *et al.* (2021) report that, for a large restaurant chain, short-notice (one-day notice) schedules do not harm server productivity overall, but real-time schedules (same-day notice) do so by 4.4%. The authors identify the following underlying mechanism: Servers may be poorly motivated when working on real-time schedules, which may make them reluctant to check on customers during their meal, thus missing opportunities to sell additional items (cross-selling opportunities) and making them less likely to make additional efforts to sell more expensive items (up-selling opportunities).

Short-term scheduling also leads to inconsistent schedules. That is, employees may be scheduled to work on different days in different weeks, creating hassles to make personal plans (*e.g.*, securing daycare). In line with the effects of schedule unpredictability reported above, the literature finds a detrimental impact of inconsistent schedules on firms' profitability. In particular, Lu *et al.* (2022) find that, in the context of a grocery retail chain, on average, hour-of-the-day consistency and day-of-the-week consistency increase cashier productivity by 0.95% and 1.63%, respectively. Moreover, these effects are much stronger

for inexperienced cashiers. Similarly, Kesavan *et al.* (2022) report that implementing responsible scheduling practices in a retail chain increased store productivity by 5.1%, a result of increasing sales (by 3.3%) and decreasing labor (by 1.8%).

As with the case of the raise of the minimum wage discussed in §4.2.1, firm-initiated efforts to improve scheduling practices seem more effective than regulatory ones. In particular, Kwon and Raman (2023) find that Fair Workweek Laws, which penalize employers for making certain kinds of unilateral changes to work schedules without sufficient advance notice, significantly increase advance notice for covered employees, consistent with their objective of improving work schedule predictability, but are ineffective in terms of increasing work schedule stability.

Overall, it appears that service organizations, when determining their employees' shift schedules, need to account for the long-term consequences of their proposed schedules' unpredictability and lack of consistency, especially for inexperienced employees. Doing so would not only improve their employees' well-being, but could also improve their own bottom line.

Contractors. In some ways, the "gig economy" associated with the development of online platforms appears to solve the problem of scheduling flexibility: Employers, instead of controlling the number of employee hours required to serve their demand, simply need to adjust upwards or downwards the wage they offer to their contractors per utilized hour to perfectly match supply and demand (Lobel *et al.*, 2024). This market solution seems to be very beneficial for employers: Lobel *et al.* (2024) shows that the contractor model is more flexible and outperforms the employee model, especially if demand is very uncertain.

This perspective relies on the assumption of labor supply being completely commoditized and is entrenched in the tradition of considering workers as inanimate objects. However, this assumption, while perhaps valid in the early days of ride-hailing platforms, has been increasingly challenged in the recent years, now that many platforms have realized that supply was not as unlimited — as they hoped for — and, to add insult to injury, started facing competition, encouraging workers to "multi-home." In a recent article, Uber's CEO said that "I think that

the industry as a whole, to some extent, has taken drivers for granted," noting that drivers had always been in abundant supply, but that the pandemic-fueled labor shortage forced a company-wide introspection to "re-examine every single assumption that we've made" (Wall Street Journal, 2023).

To encourage more driver loyalty, Besbes *et al.* (2023) recommend that online platforms allow workers to express their temporal preferences before assigning jobs to them. In their assignment, they should prioritize "full-time" workers over "part-time" workers to nurture their loyal pool of workers by increasing their effective wages.

Along this line, we anticipate a growing body of research on adopting people-centric approaches to scheduling contractors. Even though the market-based solution of adjusting wages to control supply seems to benefit contractors, leading them to rationally decide when to work and when to engage in leisure activities, this logic has failed to anticipate the precarization of the working conditions for this pool of low-wage workers with virtually no benefits (Van Doorn *et al.*, 2023).

Similar to our discussion of the impact of low wages and erratic schedules on firm performance (§4.2.1, making the work more precarious can also backfire. Indeed, Wiengarten *et al.* (2021) report a U-shaped relationship between the adoption of precarious work and flexibility and financial performance: While low levels of precarious work improve flexibility and return on assets, high levels of precarious work harm both. Moreover — and perhaps paradoxically — costs appear to monotonically increase as work becomes more precarious.

The issue of the precarious working conditions of gig workers is completely entangled with the societal debate on the integration of migrants, given that platforms are essentially the only channel to offer migrants much-needed opportunities to improve their livelihoods (Van Doorn *et al.*, 2023). In that sense, reclassifying gig workers as employees will most likely not be sufficient to improve migrants' living conditions. Hence, research on people-centric operational practices for gig workers need to factor in how migration — as a lived experience and object of governance — intersects with the gig economy (Katta *et al.*, 2024).

4.3.2 Quality

We next explore the quality of the staff, namely, both the breadth and depth of their skills.

Breadth of Skills

Cross-training employees offers the benefit of allocating resources to areas that experience a shortage of resources. As shown by Jordan and Graves (1995) in their seminal work on flexibility, a little bit of flexibility can go a long way. Although their work was motivated by multi-plant manufacturing networks, the same insights apply to service organizations that cater to multiple markers, such as a consultancy offering a variety of services. Adopting this principle, each employee can limit themselves to be specialized in only a few areas of expertise for the service organization to appear fully flexible, provided that the graph matching people with skills exhibits a "chain." See Ryu and Jiang (2025) and Yuan (2025) for applications of these flexibility principles to hospital operations in the presence of absenteeism or staff shortage.

Skills can be ranked horizontally (*i.e.*, different areas of expertise) or vertically (*i.e.*, different levels of expertise). Flexibility offers an opportunity for revenue management, where the most skilled resources are used to fulfill different demand streams (*e.g.*, basic and specialized requests); see Netessine *et al.* (2002). Call centers extensively practice skill-based routings, being structured, among the typical canonical designs, as "I-shaped," "V-shaped," "N-shaped," "X-shaped," "W-shaped," or "M-shaped" (Gans *et al.*, 2003), depending on their servers' degree of (horizontal) flexibility. Some call centers are also structured as two-level hierarchies, with a gatekeeper processing simple requests and relaying more specialized requests to more expert agents (Shumsky and Pinker, 2003).

Building on this principle of flexibility, Hopp and Van Oyen (2004) propose a framework for cross-training and coordination. Two key decisions are involved: (i) which skill(s) are strategically most desirable for workers to gain and (ii) how to coordinate these workers to respond dynamically to congestion. Although some results regarding the best cross-training strategy apply to assembly lines (Hopp *et al.*, 2004),

the underlying chaining flexibility principle establishing their good performance is fairly universal.

Depth of Skills

A large portion of many service organizations' employee development revolves around skill acquisition. In the context of an online training program in a retail chain, Fisher *et al.* (2021a) report that sales associates who engaged in training saw an increase in about 2% for every online module taken, which was higher than the direct or indirect costs associated with this training.

Training can help employees cope with a higher workload, and it can heighten their motivation, leading to higher performance (Bendoly and Prietula, 2008), but only if they are challenged, consistent with the theory of flow proposed by Csikszentmihalyi (2013). Indeed, in the absence of additional workload challenges, increases in skill may in fact significantly limit and in some cases actually degrade overall motivation, as well as objective performance (Bendoly and Prietula, 2008).

Learning could of course be on the job, coming from repeated practice, consistent with the classical "learning curve" studied at the organizational level (Lapré and Nembhard, 2011) and incorporated in staffing models by Gans and Zhou (2002). However, it could also be more deliberate, through training programs, as studied by Fisher *et al.* (2021a).

Given the financial cost and the time investment associated with training, it can be beneficial to conceptualize it as a process that can be optimized to deliver maximum performance (Roels, 2020). In particular, learning theories suggest that, in addition to repeated practice, interleaving, *i.e.*, mixing practice on several related skills together, often offers long-term benefits for learning and transfer of skills acquired in a particular context to other contexts (Brown *et al.*, 2014). In an experiment simulating an organizational setting, Schilling *et al.* (2003) report that the learning rate under conditions of related variation is significantly greater than under conditions of specialization or unrelated variation.

4.4 Job design

Labor practices are inherently associated with the characteristics of the jobs offered to workers. MacDuffie (1995) suggest a complementary relationship between labor practices (job rotation, labor division, training, hiring) and operational practices (flexibility), in line with the required alignment between strategy, technology, and organization Milgrom and Roberts (1995). Although MacDuffie's study takes place in a manufacturing context, we expect the notion of alignment to be applicable to services.

This posited complementary relationship calls for more research on the drivers of worker productivity through elements of the design of their job. Our review here will be relatively succinct, referring the reader to the extensive review by Diwas (2020). In particular, we will focus on a few elements, consisting of the trade-off between task specialization and variety (§4.4.1), employee discretion (§4.4.2), and the impact of workload on their motivation and productivity (§4.4.3).

Altogether, this literature on "what people do" and "how they do it" highlights that employees are far from inanimate objects, but respond in a rather predictable way to various physiological or cognitive stimuli. Still, what matters is perhaps less the "regressed mean", but rather the residuals, which should not necessarily be considered as a nuisance to eliminate, but rather as valuable sources of information (Corbett, 2024). Understanding human behavior is thus key to enhance their job and achieve higher organizational performance.

4.4.1 Specialization and Variety

In the past decades, many organizations, and in particular, service organizations, have transformed themselves from a "Tayloristic" organization (characterized by specialization by tasks) to a "holistic" organization (featuring job rotation, integration of tasks, and learning across tasks) (Lindbeck and Snower, 2000). Consistent with the framework proposed by MacDuffie (1995), many elements of these different organizational designs turn out to be complementary. In particular, Lindbeck and Snower (2000) examine four driving forces behind this restructuring

process: advances in production technologies promoting technological task complementarities, advances in information technologies promoting informational task complementarities, changes in worker preferences in favor of versatile work, and advances in human capital that make workers more versatile. All these forces are especially prevalent in information services, which offer large opportunities for industrialization (Karmarkar, 2004).

Worker motivation, and the overall organizational performance, is often enhanced when jobs mix specialization with variety. (Here, specialization means experience in a role, and not years with the organization. As noted by Huckman *et al.* (2009), the former is a better predictor of performance than the latter.) This ideal balance between specialization and variety is consistent with our discussion of learning processes in §4.3.2, which posit that learning is enhanced through both repetition and interleaving. In particular, Narayanan *et al.* (2009) find that, in a large offshore software development company, achieving a proper balance between specialization and exposure to variety leads to the highest productivity, being cognizant that too much variety may impede learning. Moreover, Staats and Gino (2012) find that specialization has short-term benefits through accumulated experience, whereas variety has long-term benefits in terms of worker motivation. Accordingly, workers can focus on specialized tasks over the course of a single day, but alternate between tasks across days.

This trade-off between specialization and variety occurs not only at the task level, but also at the level of the team co-workers — the so-called "team familiarity" (Huckman *et al.*, 2009). Specifically, employees could improve their productivity by either working with the same team mates or instead being exposed to different ways of working — and thus potentially acquiring useful knowledge — by changing team mates. In a healthcare context, Akşin *et al.* (2021) find that a high partner exposure, *i.e.*, frequently rotating teams, is beneficial in non-standardized processes, as well as in standardized processes but only whenever the workers have already acquired substantial individual experience. Very much like the conclusions reached by Staats and Gino (2012), this suggests that specialization should precede variety.

Moreover, the trade-off between task specialization vs. variety in-

teracts with the trade-off between team familiarity vs. rotation. In particular, Huckman and Staats (2011) find that *intra*personal team diversity (i.e., whether individuals on the team are more or less specialized) helps a team cope with challenges such as task changes; whereas, *inter*personal team diversity (i.e., differences in experience across the entire team) tends to hurt the team in coping with challenges.

4.4.2 Discretion

In addition to inducing greater learning, task variety helps reduce boredom and enhance worker motivation (Staats and Gino, 2012). Another motivating element is to give workers discretion on the execution of their tasks. In some contexts, such as highly complex tasks, task discretion happens *de facto* given that managers are less knowledgeable about what needs to be done than front-line workers (Hopp *et al.*, 2007).

Two aspects of discretion have been studied in the analytical literature on service management: discretion over the time spent with customers and discretion over the sequence of tasks. When front-line employees choose how much time to spend with their customers, a fundamental trade-off, presented in as an operational implication of co-production, discussed in §2.4.2, arises between, on one hand, the value they generate with customers in service and, on the other hand, the congestion large service times entail. The literature on this topic can be classified according to whether employees care about the monetary implications of their behavior (as reviewed in §4.2.1), see, *e.g.*, Hopp *et al.* (2007), Gilbert and Weng (1998), and Wang and Zhou (2018); or its non-monetary implications (as reviewed in §4.2.2), see, *e.g.*, Song *et al.* (2024) and Gopalakrishnan *et al.* (2016).

When workers have discretion over the sequence of their tasks, how do they tend to prioritize them? Several pieces of evidence report a preference for early task completion, with some variation around that heuristic. In particular, Ibanez *et al.* (2018) show that medical doctors prioritize similar tasks (batching) and those tasks they expect to complete faster (shortest expected processing time). Moreover, they exercise more discretion as they accumulate experience. Similarly, KC *et al.* (2020) report a preference for early task completion. However,

prioritizing the easy tasks can hurt performance, resulting in a lower throughput overall and lower learning (KC *et al.*, 2020).

Perhaps a blind spot in the research literature on employee discretion is the lack of studies on the discretion over which task to engage in. In a case study, Buell *et al.* (2015) describe an initiative developed by Oberoi Hotels to give their employees discretion over ways to delight customers, provided that it stays within preset budget limits. Although these discretionary tasks remain marginal (both in terms of time spent and budget) relative to the overall operations of running a hotel, they have a significant impact on motivating employees and delighting customers, and they should, therefore, not be dismissed.

4.4.3 Workload

People respond to workload in ambivalent ways: If the workload is too little, they get bored and lose motivation. If the workload is too high, they get stressed out and may burn out. Although most of the queuing literature assumes that service times are independent of workload, KC and Terwiesch (2009) show, in the context of hospital operations, workers work faster as their load increases. However, this acceleration may not be sustainable: Long periods of increased load (overwork) have the effect of slowing them down.

In the context of a restaurant chain, which might presumably operate at lower levels of workload than hospitals, Tan and Netessine (2014) report a service time speedup effect when the workload increases, similar to KC and Terwiesch (2009), but only at high levels of workload. At low levels of workload, they report instead a service slowdown effect because in that regime, servers expend more sales efforts with the increase in workload at a cost of slower service speed. Similarly, in the context of banking, Xu *et al.* (2022) report a U-shaped impact of the workload on operational risk error rate. More specifically, as workload increases, the error rate of operational risk events first decreases and then increases. In the context of rail traffic control, Men *et al.* (2024) report a monotone effect between workload and train delays, which is partly mitigated by the operator's experience overriding the system.

While the dependence of workload on service times has obvious

implications for staffing, as discussed in §4.3.1, one should not neglect its impact on worker motivation and the quality of service as well.

4.5 Organizational Culture

Although rarely studied in the field of management science and operations research, organizational culture can have highly motivating effects on employees. Building on a frameworks for collective action proposed by Goudsmet and Van der Heyden (2023), Roels and Van der Heyden (2025) suggest that research in OM may have overlooked the importance of values in driving collective action. We believe the same applies to research on service management.

Metters *et al.* (2019) review and call for more research on the effect of culture on OM, focusing mostly on differences in culture across the globe given the global span of companies. A notable study is the comparison of the notion of "trust" across different countries (Özer *et al.*, 2014). The classical case studies of Euro Disney (Loveman *et al.*, 1992) and Four Seasons (Hallowell *et al.*, 2002) appearing in many service management syllabi indicate the relevance of the topic of dealing across international cultures for service managers, despite the shortage of service management research on it.

At the local level, organizational culture can be shaped by operational initiatives, consistent with the argument made by MacDuffie (1995) about the complementarity of operational processes and human resource practices. In a study of retail store performance, Ton and Huckman (2008) report that a high degree of a store's process conformance — the extent to which managers aim to reduce variation in store operations in accordance with a set of prescribed standards for task performance — mitigates the negative impact of employee turnover on store performance. Technology could also help shape culture. In particular, Pierce *et al.* (2015) report that implementing a theft-monitoring information technology may have led to greater fairness concerns among employees.

While operations can impact human resource practices, the opposite is also true: Gubler *et al.* (2018) report productivity improvements after an organization implemented a corporate wellness program. By improving their diet and taking on more exercise, employees saw their

productivity increase.

At the team level, organizational culture can have profound impact on the degree of psychological safety. Psychological safety is a shared belief held by members of a team that the team is safe for interpersonal risk-taking (Edmondson, 1999). Teams that are characterized with a high degree of psychological safety tends to outperform and learn more over time. Despite the large evidence of psychological safety on performance, few operations scholars have really considered this impact, with the exception of Siemsen *et al.* (2009). We hope that future research will study the drivers and boundary conditions of psychological safety from an operational lens.

4.6 Further Developments

The notion of employment in people-intensive services has been recently challenged in numerous ways: first, by the development of platforms and the rise of the gig economy; then, by the COVID-19 pandemic, which has led to structural shifts in demand, more home-working, and the Great Resignation,; and, more recently, the development of GenAI, perhaps making many information workers redundant.

These changes have certainly offered numerous opportunities in terms of service design innovation by experimenting with new operating modes leveraging the co-productive nature of services (Chapter 2); as well as in terms of service experience engineering (Chapter 3). However, they may have severe consequences for service employees — and more generally, given that most jobs today are service jobs, for our societies. The lure of deploying technology across the board may lead to a possible service de-humanization, lacking authenticity. It might also lead managers to fall into the trap — yet again — that labor is an expendable resource, leading to a precarization of the working conditions. Finally, long-term investments in technology deployment may divert resources that were allocated to employee development.

Academic research is well positioned to adopt an objective perspective on *anticipating the negative long-term consequences of these developments for our societies*. In the spirit of engaging in responsible research (Netessine, 2022), we urge academics to adopt a people-centric

approach to service employee management and foster their well-being at work (Corbett, 2024) and engagement by offering them "good jobs" (Ton, 2014).

Although the study of employee management typically relates to the field of OB, we believe that researchers in service management across disciplines have a unique perspective to share on the topic to enrich the discussion. First, as discussed throughout this chapter, many service design decisions impact employee engagement and *vice versa*. Ignoring this connection often leads to short-sighted decisions, favoring short-term profitability over long-term growth. Second, the availability of data in the workplace enabled by digital technologies and the short feedback loops offer novel opportunities for adopting scientific methods to better understand what people do and what drives their engagement and performance and to experiment with new ways of operating. A good example of this scientific approach, aimed at optimizing service encounters, is Meng *et al.* (2021) who capture microscopic data on nurses on a hospital floor, enabling them to make tour and layout recommendations.

Perhaps paradoxically, *emerging economies* offer a fertile ground for the development of research on employee engagement aimed at optimizing service encounters. This is perhaps paradoxical because we may (perhaps, wrongly) expect that these economies have less access to digital technologies. However, some emerging economies have large population densities, so even a small-scale data collection per individual could generate large data sets, when aggregated across individuals or transactions. For instance, Aouad *et al.* (2024) capture half a million food transactions made by 23,717 consumers with point-of-sales data in only 39 micro-retail stores in India. Although they take the perspective of customers, one could imagine that similar scale of data applied to service employees as well, such as Mumbai's famed "dabbawalas" (Thomke, 2012). As most emerging economies are increasingly becoming service economies, there is indeed a pressing need to study the role of employees in these contexts as well.

Finally, the service management literature often considers short-term performance metrics (*e.g.*, store profitability, patient length of stay) as the main dependent variables, but as this chapter covered,

inter-temporal trade-offs may be at work, *i.e.*, boosting short-term performance may be at the cost of long-term performance. If this is the case, the challenge is, of course, to establish causality between short-term drivers and long-term performance. To resolve that tension, service management researchers need to identify antecedents of long-term performance. Based on the literature reviewed in this chapter, it appears that employee engagement (or conversely, employees' mental health) may be a good predictor for long-term performance. If this is indeed confirmed, we thus encourage researchers to *use a measure of employee engagement as a main dependent variable*, perhaps in addition to the usual short-term performance metrics, to capture this trade-off and measure the long-term implications of various managerial actions.

5

Conclusions

Services are everywhere, and their importance to the economy has been growing, across all economies. However, they have been fraught with three crises: rampant costs, poor customer experiences, and lack of engagement or even shortage of service employees.

To tackle these challenges, we posit that service encounters, which lie at the core of the service value creation process, need to be optimized, fulfilling the vision by Shostack (1987) that they can be "engineered." This need has in fact been strengthened with the development of digital technologies, which have unlocked many new ways to structure service encounters — and thus offer more degrees of freedom along which to optimize.

Unfortunately, managers often lack the tools to effectively optimize service encounters, primarily for two reasons. First, many of the traditional approaches to service management have attempted to mimic the approaches developed for manufacturing. Accordingly, they have been entrenched in functional disciplines, even though service value creation processes, which lie at the interface between customers, employees, and service organizations (Figure 1.1), call for multi-disciplinary approaches. Second, the few integrative frameworks that are available, such as the

SPC (Figure 1.6), which cuts across Marketing, OM, and OB, predate the development of digital technologies and thus remain silent about service encounters that takes place in indirect, asynchronous, and digital channels.

This monograph introduces three managerial levers that both lie at the core of the service value creation process (and are thus multidisciplinary) and embrace the development of digital technologies:

- Leveraging co-production to innovate in service design;
- Delighting customers through experience design;
- Fostering employee engagement by putting people first.

Using these levers, and leveraging the large data sets and short feedback loops offered by digital technologies, we believe the time is ripe for truly optimizing service encounters to achieve higher efficiency and effectiveness.

We hope this monograph will help future researchers to "see the wood for the trees" in service management, aspire to contribute to the development of an integrative and contemporary theory of service management, and investigate innovative ways to design services to deal with their crises and bring prosperity, delight, and well-being.

Acknowledgements

I thank the Editor in Chief, Panos Kouvelis, for inviting me to write a monograph on service management and Zac Rolnic for his technical support along the way. I also thank an anonymous reviewer who offered thoughtful and constructive feedback and the participants of the EURO Working Group (EWG) on Retail Operations for their feedback on an early presentation of the ideas of this monograph. I also thank IESE for graciously hosting me during my sabbatical leave.

Appendices

A

Breaking the Trade-Off between Experience and Efficiency

This appendix studies in the context of services the fundamental trade-off between variability and efficiency. We first map services (as a generic industry, disregarding its heterogeneity) into the classical product-process matrix (Hayes and Wheelwright, 1979), positioning them as job shops (§A.1). Doing so explains why queuing systems have traditionally been the dominant representation of service systems.

Using then a queuing-theoretic framework, we re-interpret in §A.2 three core strategies for efficiently delivering an outstanding service experience proposed by Frei (2006) and Buell (2020).

A.1 Services as Job Shops

One of the most fundamental trade-offs in OM is balancing efficiency and variability (variety). The product-process matrix developed for manufacturing organizations (Hayes and Wheelwright, 1979), depicted in Figure A.1, calls for matching the right process with the right product. Specifically, a manufacturing firm offering a wide variety of one-of-a-kind products should adopt a *job shop*, which is a flexible process with general-purpose resources, typically laid out by functions, in which the flow is jumbled. In contrast, a manufacturing firm offering a small

variety of standard, large-volume products should adopt a *flow shop*, which is a rigid process with specialized resources, typically laid out in the same sequence as the different operations than need to take place to manufacture the product, and in which the flow is structured.

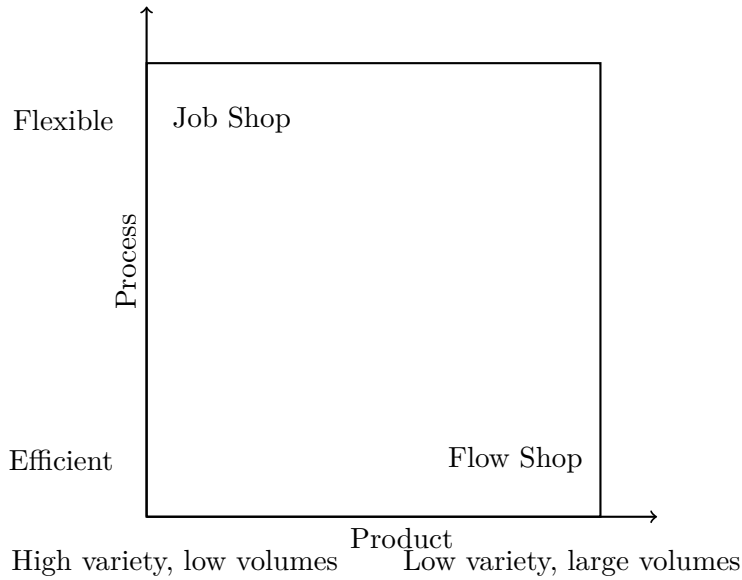


Figure A.1: Matrix of Product vs. Process

Compared to many manufacturing organizations, services (as a whole, and disregarding their heterogeneity) tend to be more flexible, and therefore, to operate as job shops. One could in fact argue the converse: Job shops are effectively operating as service organizations whenever production is triggered by customers (Sampson, 2010).

A.1.1 Variability and Scaling

Drawing the parallel between services and job shops offers two important insights. First, services often need to learn to deal with variability, *i.e.*, they can rarely fully eliminate it. Frei (2006) identifies five sources of variability associated with customers, depicted in Figure A.2: variability in arrivals, variability in service requests, variability in capability, variability in effort, and variability in subjective preferences. Unlike man-

ufacturing organizations, which can certify suppliers (quality assurance) or inspect raw materials (quality control) to eliminate variability, service organizations often have to learn to deal with this customer-induced variability.

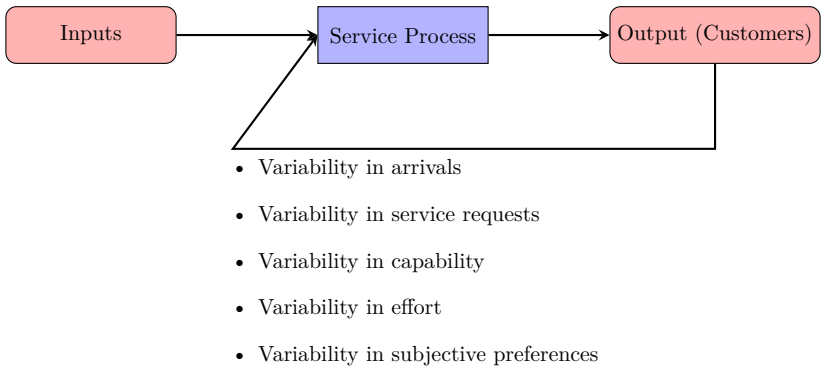


Figure A.2: In Service Processes, Customers Bring a Significant Input (Sampson and Froehle, 2006), but It Is Assorted with Various Kinds of Variability (Frei, 2006)

Second, it may be quite elusive to seek to improve efficiency in services and scale their operations, unless their value proposition to accommodate variety is fundamentally changed. The lack of scalability of many services is indeed why the service industry suffers from low productivity improvement, a/k/a Baumol's cost disease (§1.1.2).

A.1.2 Services as Queuing Systems

Given their inherent variability, job shop operations are often characterized as queuing systems (Buzacott and Shanthikumar, 1993). The association of services with job shops therefore creates a natural representation of services as a queuing system (Mandelbaum and Zeltyn, 2010; Mandelbaum, 2011). In fact, for many years, the research community on service operations was mostly constituted of queuing theorists.

However, service operations involve broader issues than the management of queuing systems. In many services (*e.g.*, education, media and entertainment), congestion is of second order, if relevant at all. Hence, reducing the scope of service operations to queuing systems creates too many blind spots and unnecessarily limits the scope of the OM field to

tackle relevant questions.

Nevertheless, associating services with queuing systems is useful to identify of generic strategies for efficiently delivering outstanding service experiences, as we discuss next.

A.2 Generic Strategies for Efficiently Delivering Outstanding Service Experiences

How can services deliver an outstanding service experience in an efficient manner? Using queuing theory as a framework, we can distinguish three strategies (Buell, 2020), two of which are extensively described by Frei (2006): low-cost variability accommodation, customer selection, and uncompromised variability reduction.

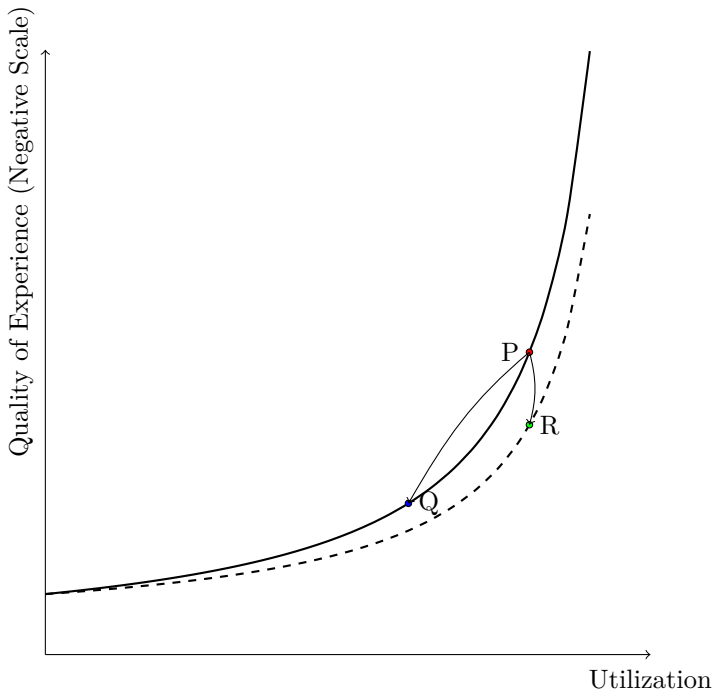


Figure A.3: Low-Cost Service Accommodation and/or Customer Selection (from P to Q) and Uncompromised Variability Reduction (from P to R)

To visualize this, Figure A.3 depicts the classical exponential re-

lationship between utilization (horizontal axis) and throughput time (vertical axis), which we take as a proxy for (the negative of) customer experience. From a baseline situation P, a service provider could enhance customer service experience by reducing either utilization (point Q) or variability (point R). Reducing utilization can be achieved in two ways: either by increasing capacity or by reducing demand. We next discuss these three generic strategies.

A.2.1 Low-Cost Variability Accommodation

A service provider can accommodate variability by *increasing its service capacity*, but this strategy is appealing only if can be achieved at low cost. To do so, various service industrialization strategies can be deployed.

First, a service provider can *outsource* parts of its service process to a third-party supplier that offers the same service to multiple service providers and benefits of economies of scale. This practice is ubiquitous in helpdesk centers, *e.g.*, call centers, giving rise to moral hazard and adverse selection (Hasija *et al.*, 2008; Ren and Zhou, 2008; Aksin *et al.*, 2008).

Second, a service provider can *offshore* parts of its service to take advantage of low-wage labor or pooling. Here, we use the term offshoring as relocation of service activities, without necessarily implying that it moves overseas. For instance, the recent development of cloud (also known as dark or ghost) kitchens has enabled many restaurateurs to decouple the eating process, which is often located in an accessible, and thus expensive urban area, from the cooking process, which can be done in a cheaper industrial area (Roy *et al.*, 2022; Epstein *et al.*, 2023).

Third, a service provider can *automate* parts of its process, using robotics and artificial intelligence. In services, automation offers many opportunities for self-service, such as the self-check-in counters at the airport or self-check-out counters in retail (Gao and Su, 2018; Field, 2024).

Information-intensive services are particularly prone to such service industrialization strategies because "[digital] information can be standardized, built to order, assembled from components, picked, packed, stored, and shipped" (Karmarkar, 2004, p. 102). Although the case

for efficiency and scalability is obvious for many online services (*e.g.*, search, entertainment, social networks), it is only emerging in other sectors of information services, such as education or professional services (Karmarkar, 2010). For instance, Sampson and Santos (2023) identify, through a calibrated simulation model, great opportunities for reengineering professional services through increased automation, offshore outsourcing, and task delegation. Accordingly, one of the most useful categorization of services (which comprise more than 80% of the economy in many advanced economies) is to distinguish whether they transform information or atoms (Apte *et al.*, 2012).

A.2.2 Customer Selection

Instead of increasing its service capacity, a service provider can accommodate variability by *lowering its demand*, by carefully selecting its customers. This customer selection strategy, although it is missing from Frei (2006), is extensively discussed by Buell (2020). It is a natural dual to the low-cost variability accommodation given that it has a similar effect on utilization. The motto is to "work less, but work better."

Besides reducing utilization, customer selection also offers an opportunity for service providers to reduce variability in demand. For instance, hospitals can use deliberate admission controls to smooth out emergency departments' workloads (Helm *et al.*, 2011).

Beyond congestion, adopting a more focused offering could also help improve customer satisfaction. Buell *et al.* (2021a) indeed note that banks that have more focused customer bases tend to achieve higher satisfaction scores and longer customer tenure.

In some B2C services, service providers cannot reject customers, but they can always better convey realistic expectations of the service provision through operational transparency to reduce the so-called marketing gap (Parasuraman *et al.*, 1988; Apte *et al.*, 1997) and facilitate the match between customer's expectations and the service delivery (Buell and Choi, 2024) — a strategy explored in greater detail in §2.4.3.

A.2.3 Uncompromised Variability Reduction

The third strategy to resolve the trade-off between customer experience and efficiency consists in *reducing variability*, either in demand or in service requests, without compromising the service experience. For instance, Benihana has mastered the art of reduction in arrivals (using their bar facilities as a buffer before seating customers), in party sizes (by grouping different parties at the same table), and in service times (by giving the control of the timing of the service to the chef), while delighting customers with a cooking experience (Sasser, 2004).

This strategy is perhaps less obvious than the former two because it tackles variability and not utilization (which is often perceived to be of second order). Because services are analogous to job shops, designed to accommodate variability (see §A.1), this variability reduction strategy effectively suggests migrating the service organization towards a flow shop, and therefore, to focus on high-volume, low-cost specialization. The epitome of this strategy is McDonald's, which is, according to Levitt (1972, p. 44), a "supreme example of the application of manufacturing and technological brilliance to problems that must ultimately be viewed as marketing problems." When praising this approach, Levitt (1972) makes an obvious case for reducing variability, using words such as "carefully controlled execution" and "systematic substitution of equipment for people,". Levitt (1976, pp. 66-67) further identifies industrialization strategies for services, which make the connection to flow shops tighter.

While variability reduction has had large successes in manufacturing (*e.g.*, Six Sigma, Total Quality Management), this strategy has certainly influenced the perception, in the early research on service management (§2.1.1), of the customer's role as hurting efficiency.

References

ACSI. (2024). "U.S. overall customer satisfaction". URL: https://theacsi.org/the-acsi-difference/us-overall-customer-satisfaction/.

Afeche, P. (2013). "Incentive-compatible revenue management in queueing systems: Optimal strategic delay". *Manufacturing & Service Operations Management*. 15(3): 423–443.

Afeche, P. and H. Mendelson. (2004). "Pricing and priority auctions in queueing systems with a generalized delay cost structure". *Management Science*. 50(7): 869–882.

Aflaki, S. and I. Popescu. (2013). "Managing retention in service relationships." *Management Science*. 60(2): 415–433.

Agarwal, S., M. Ayyagari, and R. Kosová. (2024). "Minimum wage increases and employer performance: role of employer heterogeneity". *Management Science*. 70(1): 225–254.

Aghion, P. and J. Tirole. (1997). "Formal and real authority in organizations". *Journal of Political Economy*. 105(1): 1–29.

Ahmadi, R. H. (1997). "Managing capacity and flow at theme parks". *Operations Research*. 45(1): 1–13.

Aksin, O. Z., F. de Vericourt, and F. Karaesmen. (2008). "Call center outsourcing contract analysis and choice". *Management Science*. 54(2): 354–368.

Akşin, Z., S. Deo, J. O. Jónasson, and K. Ramdas. (2021). "Learning from many: Partner exposure and team familiarity in fluid teams". *Management Science*. 67(2): 854–874.

Albers, S. (1996). "Optimization models for salesforce compensation". *European Journal of Operational Research*. 89(1): 1–17.

Alchian, A. A. and H. Demsetz. (1972). "Production, information costs, and economic organization". *The American Economic Review*. 62(5): 777–795.

Allon, G., A. Bassamboo, and I. Gurvich. (2011). ""We will be right with you": Managing customer expectations with vague promises and cheap talk". *Operations Research*. 59(6): 1382–1394.

Altman, D., G. B. Yom-Tov, M. Olivares, S. Ashtar, and A. Rafaeli. (2021). "Do customer emotions affect agent speed? An empirical study of emotional load in online customer contact centers". *Manufacturing & Service Operations Management*. 23(4): 854–875.

Anand, K. S., M. F. Paç, and S. Veeraraghavan. (2011). "Quality–Speed conundrum: Trade-offs in customer-intensive services". *Management Science*. 57(1): 40–56.

Anderson, E. W., C. Fornell, and R. T. Rust. (1997). "Customer satisfaction, productivity, and profitability: Differences between goods and services". *Marketing Science*. 16(2): 129–145.

Anderson, E. W. and M. W. Sullivan. (1993). "The antecedents and consequences of customer satisfaction for firms". *Marketing Science*. 12(2): 125–143.

Anderson, S. W., L. S. Baggett, and S. K. Widener. (2009). "The impact of service operations failures on customer satisfaction: Evidence on how failures and their source affect what matters to customers". *Manufacturing & Service Operations Management*. 11(1): 52–69.

Andritsos, D. A. and C. S. Tang. (2018). "Incentive programs for reducing readmissions when patient care is co-produced". *Production and Operations Management*. 27(6): 999–1020.

Aouad, A., A. Deshmane, and V. Martinez-de-Albeniz. (2022). "Designing layouts for sequential experiences: Application to cultural institutions". *Tech. rep.* IESE. URL: https://ssrn.com/abstract=4158587.

Aouad, A., K. Ramdas, and A. Sungu. (2024). "Digitized Indian micro-grocery transactions reveal that grain subsidies reduce junk food buying by low-income shoppers". *Tech. rep.* London Business School. URL: https://dx.doi.org/10.2139/ssrn.4847728.

Apte, U., U. Karmarkar, and H. Nath. (2012). "The US information economy: Value, employment, industry structure, and trade". *Foundations and Trends® in Technology, Information and Operations Management*. 6(1): 1–87.

Apte, U. M., U. S. Karmarkar, and R. Pitbladdo. (1997). "Quality management in services: Analysis and measurement". In: *The Practice of Quality Management*. Ed. by P. Lederer and U. Karmarkar. Springer. 167–193.

Ariely, D. and Z. Carmon. (2000). "Gestalt characteristics of experiences: The defining features of summarized events". *Journal of Behavioral Decision Making*. 13(2): 191–201.

Ariely, D. and M. I. Norton. (2009). "Conceptual consumption". *Annual Review of Psychology*. 60(1): 475–499.

Aristotle. (2022). *How to tell a story: An ancient guide to the art of storytelling for writers and readers*. Princeton University Press.

Arkes, H. R. and C. Blumer. (1985). "The psychology of sunk cost". *Organizational Behavior and Human Decision Processes*. 35(1): 124–140.

Armony, M., S. Israelit, A. Mandelbaum, Y. N. Marmor, Y. Tseytlin, and G. B. Yom-Tov. (2015). "On patient flow in hospitals: A data-based queueing-science perspective". *Stochastic Systems*. 5(1): 146–194.

Armony, M. and C. Maglaras. (2004). "Contact centers with a call-back option and real-time delay information". *Operations Research*. 52(4): 527–545.

Armony, M., G. Roels, and H. Song. (2021). "Pooling queues with strategic servers: The effects of customer ownership". *Operations Research*. 69(1): 13–29.

Ascarza, E. and B. G. Hardie. (2013). "A joint model of usage and churn in contractual settings". *Marketing Science*. 32(4): 570–590.

Bajari, P. and S. Tadelis. (2001). "Incentives versus transaction costs: A theory of procurement contracts". *RAND Journal of Economics*: 387–407.

Bakshi, N., S.-H. Kim, and N. Savva. (2015). "Signaling new product reliability with after-sales service contracts". *Management Science*. 61(8): 1812–1829.

Balasubramanian, S., P. Konana, and N. M. Menon. (2003). "Customer satisfaction in virtual environments: A study of online investing". *Management Science*. 49(7): 871–889.

Bandiera, O., I. Barankay, and I. Rasul. (2005). "Social preferences and the response to incentives: Evidence from personnel data". *The Quarterly Journal of Economics*. 120(3): 917–962.

Barnard, C. I. (1938). *The Functions of the Executive*. Cambridge, MA: Harvard University Press.

Barnard, C. I. (1940). "Comments on the job of the executive". *Harvard Business Review*. 18(3): 295–308.

Baron, O., O. Berman, D. Krass, and J. Wang. (2014). "Using strategic idleness to improve customer service experience in service networks". *Operations Research*. 62(1): 123–140.

Baron, O., O. Berman, D. Krass, and J. Wang. (2017). "Strategic idleness and dynamic scheduling in an open-shop service network: Case study and analysis". *Manufacturing & Service Operations Management*. 19(1): 52–71.

Bassamboo, A., R. S. Randhawa, and A. Zeevi. (2010). "Capacity sizing under parameter uncertainty: Safety staffing principles revisited". *Management Science*. 56(10): 1668–1686.

Basu, A. K., R. Lal, V. Srinivasan, and R. Staelin. (1985). "Salesforce compensation plans: An agency theoretic perspective". *Marketing science*. 4(4): 267–291.

Batt, R. J. and C. Terwiesch. (2015). "Waiting patiently: An empirical study of queue abandonment in an emergency department". *Management Science*. 61(1): 39–59.

Baucells, M. and R. K. Sarin. (2007). "Satiation in discounted utility". *Operations Research*. 55(1): 170–181.

Baucells, M. and R. K. Sarin. (2010). "Predicting utility under satiation and habit formation". *Management Science*. 56(2): 286–301.

Baucells, M., M. Weber, and F. Welfens. (2011). "Reference-point formation and updating". *Management Science*. 57(3): 506–519.

Baucells, M. and L. Zhao. (2019). "It is time to get some rest". *Management Science*. 65(4): 1455–1947.

Baucells, M. and S. Bellezza. (2017). "Temporal profiles of instant utility during anticipation, event, and recall". *Management Science*. 63(3): 729–748.

Baucells, M., Y. Grushka-Cockayne, and W. Hwang. (2024). "Managerial mental accounting and downstream project decisions". *Management Science*.

Baucells, M. and L. Zhao. (2020). "Everything in moderation: Foundations and applications of the satiation model". *Management Science*. 66(12): 5701–5719.

Baumol, W. J. (1993). "Health care, education and the cost disease: A looming crisis for public choice". In: *The Next Twenty-Five Years of Public Choice*. Ed. by C. K. Rowley, F. Schneider, and R. D. Tollison. Springer. 17–28.

Baumol, W. J. and W. G. Bowen. (1965). "On the performing arts: The anatomy of their economic problems". *The American Economic Review*. 55(1/2): 495–502.

Bellos, I. and S. Kavadias. (2019). "When should customers control service delivery? Implications for service design". *Production and Operations Management*. 28(4): 890–907.

Bellos, I. and S. Kavadias. (2021). "Service design for a holistic customer experience: A process framework". *Management Science*. 67(3): 1718–1736.

Bénabou, R. and J. Tirole. (2003). "Intrinsic and extrinsic motivation". *The Review of Economic Studies*. 70(3): 489–520.

Bendoly, E. and M. Prietula. (2008). "In "the zone": the role of evolving skill and transitional workload on motivation and realized performance in operational tasks". *International Journal of Operations & Production Management*. 28(12): 1130–1152.

Benjaafar, S. and M. Hu. (2020). "Operations management in the age of the sharing economy: What is old and what is new?" *Manufacturing & Service Operations Management*. 22(1): 93–101.

Benjaafar, S., G. Kong, X. Li, and C. Courcoubetis. (2019). "Peer-to-peer product sharing: Implications for ownership, usage, and social welfare in the sharing economy". *Management Science*. 65(2): 477–493.

Berenguer, G., W. B. Haskell, and L. Li. (2024). "Managing volunteers and paid workers in a nonprofit operation". *Management Science*. 70(8): 5298–5316.

Besbes, O., V. Goyal, G. Iyengar, and R. Singal. (2023). "Workforce scheduling with heterogeneous time preferences: Effective wages and workers' supply". *Manufacturing & Service Operations Management*. URL: http://dx.doi.org/10.2139/ssrn.4202484.

Bhaskaran, S. R., S. Erat, and R. Mukherjee. (2022). "Getting your money's worth: Capacity planning through admission control vs. consumption control". *Tech. rep.* SMU Cox. URL: https://ssrn.com/abstract=4287225.

Bhattacharyya, S. and F. Lafontaine. (1995). "Double-Sided Moral Hazard and the Nature of Share Contracts". *The RAND Journal of Economics*. 26(4): 761–781.

Bitner, M. J., A. Otrom, and F. N. Morgan. (2008). "Service Blueprinting: A Practical Technique for Service Innovation". *California Management Review*. 50(3): 65–94.

Bitner, M. J. (1992). "Servicescapes: The impact of physical surroundings on customers and employees". *Journal of Marketing*. 56(2): 57–71.

Bitran, G. R., J.-C. Ferrer, and P. Rocha e Oliveira. (2008). "OM Forum—Managing customer experiences: perspectives on the temporal aspects of service encounters". *Manufacturing & Service Operations Management*. 10(1): 61–83.

Blackader, B., E. Buesing, J. Amar, J. Raabe, M. Mehnidratta, and V. Gupta. (2025). "The Contact Center Crossroads: Finding the Right Mix of Humans and AI". URL: https://www.mckinsey.com/capabilities/operations/our-insights/the-contact-center-crossroads-finding-the-right-mix-of-humans-and-ai.

Blackburn, J. D., D. H. Kropp, and R. A. Millen. (1986). "A comparison of strategies to dampen nervousness in MRP systems". *Management Science*. 32(4): 413–429.

Blumberg, M. and C. D. Pringle. (1982). "The missing opportunity in organizational research: Some implications for a theory of work performance". *Academy of Management Review*. 7(4): 560–569.

Boehm, J., H. K. Bhargava, and G. G. Parker. (2020). "The business of electric vehicles: a platform perspective". *Foundations and Trends® in Technology, Information and Operations Management*. 14(3): 203–323.

Bordoloi, S., J. A. Fitzsimmons, and M. J. Fitzsimmons. (2022). *Service Management: Operations, Strategy, Information Technology*. McGraw-Hill.

Bray, R. L. (2023). "Operational transparency: Showing when work gets done". *Manufacturing & Service Operations Management*. 25(3): 812–826.

Briggs, E., S. Deretti, and H. T. Kato. (2020). "Linking organizational service orientation to retailer profitability: Insights from the service-profit chain". *Journal of Business Research*. 107: 271–278.

Brown, P. C., H. L. Roediger, and M. A. McDaniel. (2014). *Make It Stick: The Science of Succesful Learning*. Cambridge, MA: Harvard University Press.

Brown, S. W. and T. A. Swartz. (1989). "A gap analysis of professional service quality". *Journal of Marketing*. 53(2): 92–98.

Bryson, J. R. and P. W. Daniels. (2010). "Service worlds: The 'services duality' and the rise of the 'manuservice' economy". In: *Handbook of Service Science*. Ed. by P. P. Maglio, C. A. Kieliszewski, and J. C. Spohrer. Springer. 79–104.

Buell, R. and M. Choi. (2024). "Improving customer compatibility with tradeoff transparency". *Management Science*.

Buell, R. W. (2019). "Operational transparency: Make your processes visible to customers and your customers visible to employees". *Harvard Business Review*. 97(4): 102–113.

Buell, R. W. (2020). "Transforming customer engagement in service operations. Module note for instructors." *Tech. rep.* Harvard Business School.

Buell, R. W., D. Campbell, and F. X. Frei. (2021a). "The customer may not always be right: Customer compatibility and service performance". *Management Science*. 67(3): 1468–1488.

Buell, R. W., T. Kim, and C.-J. Tsay. (2017). "Creating reciprocal value through operational transparency". *Management Science*. 63(6): 1673–1695.

Buell, R. W. and M. I. Norton. (2011). "The labor illusion: How operational transparency increases perceived value". *Management Science*. 57(9): 1564–1579.

Buell, R. W., E. Porter, and M. I. Norton. (2021b). "Surfacing the submerged state: Operational transparency increases trust in and engagement with government". *Manufacturing & Service Operations Management*. 23(4): 781–802.

Buell, R. W., A. Raman, and V. Muthuram. (2015). " Oberoi Hotels: Train Whistle in the Tiger Reserve". *Tech. rep.* No. 9-615-043. Harvard Business School.

Buzacott, J. and J. G. Shanthikumar. (1993). *Stochastic Models of Manufacturing Systems*. Ed. by P. Hall.

Calcbench. (2023). "Are retailers weathering inflation?" URL: https://www.calcbench.com/blog/post/7175741043892224 00/are-retailers-weathering-inflation.

Candogan, O., K. Bimpikis, and A. Ozdaglar. (2012). "Optimal pricing in networks with externalities". *Operations Research*. 60(4): 883–905.

Candogan, O. and K. Drakopoulos. (2020). "Optimal signaling of content accuracy: Engagement vs. misinformation". *Operations Research*. 68(2): 497–515.

Candoğan, S. T., C. G. Korpeoglu, and C. S. Tang. (2020). "Team collaboration in innovation contests". *Tech. rep.* NUS Working Paper. URL: https://dx.doi.org/10.2139/ssrn.3607769.

Card, D., A. Mas, E. Moretti, and E. Saez. (2012). "Inequality at work: The effect of peer salaries on job satisfaction". *American Economic Review*. 102(6): 2981–3003.

Caro, F. and V. Martínez-de-Albéniz. (2012). "Product and price competition with satiation effects". *Management Science*. 58(7): 1357–1373.

Cassar, G. and T. Ko. (2023). "Peer effects in subjective performance evaluation". *Contemporary Accounting Research*. 40(3): 1704–1732.

Cezik, M. T. and P. L'Ecuyer. (2008). "Staffing multiskill call centers via linear programming and simulation". *Management Science*. 54(2): 310–323.

Chan, T. Y., Y. Chen, L. Pierce, and D. Snow. (2021). "The influence of peers in worker misconduct: Evidence from restaurant theft". *Manufacturing & Service Operations Management*. 23(4): 952–973.

Chan, T. Y., J. Li, and L. Pierce. (2014a). "Compensation and peer effects in competing sales teams". *Management Science*. 60(8): 1965–1984.

Chan, T. Y., J. Li, and L. Pierce. (2014b). "Learning from peers: Knowledge transfer and sales force productivity growth". *Marketing Science*. 33(4): 463–484.

Chase, R. B. (1978). "Where does the customer fit in a service operation?" *Harvard Business Review*. 56(6): 137–142.

Chase, R. B. (2010). "Revisiting "Where does the customer fit in a service operation?" Background and future development of contact theory". In: *Handbook of Service Science*. Ed. by P. P. Maglio, C. A. Kieliszewski, and J. C. Spohrer. Springer. 11–17.

Chase, R. B. and S. Dasu. (2001a). "Want to perfect your company's service? Use behavioral science". *Harvard Business Review*: 78–84.

Chase, R. B. and S. Dasu. (2001b). "Want to perfect your company's service? Use behavioral science". *Harvard Business Review*. 79(6): 78–84.

Chen, H., M. Hu, J. Liu, and Y. Ravid. (2024a). "Ups and downs in experience design". *Tech. rep.* No. 9. 1895–1911.

Chen, M., O. Baron, A. Mandelbaum, J. Wang, G. B. Yom-Tov, and N. Arber. (2024b). "Frontiers in Operations: Waiting experience in open-shop service networks: Improvements via flow analytics and automation". *Manufacturing & Service Operations Management*.

Chen, M., M. Hu, and J. Wang. (2022). "Food delivery service and restaurant: Friend or foe?" *Management Science*. 68(9): 6539–6551.

Chen, Z. and J. Keppo. (2023). "R&D data sharing in new product development". *Tech. rep.* NUS. URL: https://dx.doi.org/10.2139/ssrn.3915253.

Chicu, D., M. del Mar Pàmies, G. Ryan, and C. Cross. (2019). "Exploring the influence of the human factor on customer satisfaction in call centres". *BRQ Business Research Quarterly*. 22(2): 83–95.

Choi, H., C. F. Mela, S. R. Balseiro, and A. Leary. (2020). "Online display advertising markets: A literature review and future directions". *Information Systems Research*. 31(2): 556–575.

Chuang, H. H.-C., R. Oliva, and O. Perdikaki. (2016). "Traffic-based labor planning in retail stores". *Production and Operations Management*. 25(1): 96–113.

Constantinides, G. M. (1990). "Habit formation: A resolution of the equity premium puzzle". *Journal of Political Economy*. 98(3): 519–543.

Corbett, C. J. and G. A. DeCroix. (2001). "Shared-Savings Contracts for Indirect Materials in Supply Chains: Channel Profits and Environmental Impacts". *Management Science*. 47(7): 881–893.

Corbett, C. J., G. A. DeCroix, and A. Y. Ha. (2005). "Optimal shared-savings contracts in supply chains: Linear contracts and double moral hazard". *European Journal of Operational Research*. 163(3): 653–667.

Corbett, C. J. (2024). "OM forum—The operations of well-being: An operational take on happiness, equity, and sustainability". *Manufacturing & Service Operations Management*. 26(2): 409–430.

Cowley, E. (2014). "Consumers telling consumption stories: word-of-mouth and retrospective evaluations". *Journal of Business Research*. 67(7): 1522–1529.

Cronin Jr, J. J. and S. A. Taylor. (1992). "Measuring service quality: a reexamination and extension". *Journal of Marketing*. 56(3): 55–68.

Csikszentmihalyi, M. (2013). *Flow: The Psychology of Happiness*. Random House.

Cui, Y., Z. Jiang, and Q. Li. (2024). "Unlocking the value of real-time AI assistance: Who benefits, and why?" *Tech. rep.* Cornell University Working Paper. URL: https://dx.doi.org/10.2139/ssrn.4497014.

Czerniawska, F. and P. May. (2004). *Management Consulting in Practice*. London, UK: Management Consultancies Association.

Czerniawska, F. (2006). *The Trusted Firm: How Consulting Firms Build Successful Client Relationships*. John Wiley & Sons.

Dahl, M. S. and L. Pierce. (2020). "Pay-for-performance and employee mental health: Large sample evidence using employee prescription drug usage". *Academy of Management Discoveries*. 6(1): 12–38.

Danet, B. (1981). "Client-organization interfaces". In: *Handbook of Organization Design*. Ed. by P. C. Nystrom and W. H. Starbuck. Vol. 2. Oxford University Press New York, NY. 382–428.

Das Gupta, A., U. S. Karmarkar, and G. Roels. (2015). "The design of experiential services with acclimation and memory decay: Optimal sequence and duration". *Management Science*. 62(5): 1278–1296.

Daw, A., A. Castellanos, G. B. Yom-Tov, J. Pender, and L. Gruendlinger. (2020). "The co-production of service: Modeling service times in contact centers using Hawkes processes". *Management Science*. URL: arXiv%20preprint%20arXiv:2004.07861.

Daw, A. and G. B. Yom-Tov. (2024). "Asymmetries of service: Interdependence and synchronicity". *Tech. rep.* USC Working Paper. URL: https://arxiv.org/pdf/2402.15533.

de Bettignies, J.-E. (2008). "Financing the entrepreneurial venture". *Management Science*. 54(1): 151–166.

De Moura, V. F., C. A. de Souza, and A. B. N. Viana. (2021). "The use of Massive Open Online Courses (MOOCs) in blended learning courses and the functional value perceived by students". *Computers & Education*. 161: 104077.

Debo, L. and C. Li. (2021). "Design and pricing of discretionary service lines". *Management Science*. 67(4): 2251–2271.

Debo, L. G., C. Parlour, and U. Rajan. (2012). "Signaling quality via queues". *Management Science*. 58(5): 876–891.

Delana, K., S. Deo, K. Ramdas, G.-B. B. Subburaman, and T. Ravilla. (2023). "Multichannel delivery in healthcare: the impact of telemedicine centers in southern India". *Management Science*. 69(5): 2568–2586.

Della Vigna, S. and U. Malmendier. (2006). "Paying not to go to the gym". *American Economic Review*. 96(3): 694–719.

Deshmane, A., V. Martinez-de-Albeniz, and G. Roels. (2023). "Intertemporal spillovers in consumer experiences: Empirical evidence and service design implications". *Tech. rep.* INSEAD Working Paper. URL: https://dx.doi.org/10.2139/ssrn.4507191.

Dietvorst, B. J., J. P. Simmons, and C. Massey. (2015). "Algorithm aversion: people erroneously avoid algorithms after seeing them err." *Journal of Experimental Psychology: General*. 144(1): 114.

Dietvorst, B. J., J. P. Simmons, and C. Massey. (2018). "Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them". *Management Science*. 64(3): 1155–1170.

Diwas, K. (2020). "Worker productivity in operations management". *Foundations and Trends® in Technology, Information and Operations Management*. 13(3): 151–249.

Dixon, M. and R. Verma. (2013). "Sequence effects in service bundles: Implications for service design and scheduling". *Journal of Operations Management*. 31(3): 138–152.

Dixon, M. J. and G. M. Thompson. (2016). "Bundling and scheduling service packages with customer behavior: Model and heuristic". *Production and Operations Management*. 25(1): 36–55.

Dixon, M. J. and G. M. Thompson. (2019). "The impact of timing and bundling flexibility on affect-based service package design". *Decision Sciences*. 50(5): 948–984.

Dixon, M. J., L. Victorino, R. J. Kwortnik, and R. Verma. (2017). "Surprise, anticipation, and sequence effects in the design of experiential services". *Production and Operations Management*. 26(5): 945–960.

Dong, M., T. Mayskaya, V. Smirnov, O. Taylor, and A. Wait. (2023). "Diversity in teams". *Tech. rep.* University of Sydney.

Ebbinghaus, H. (1913). *Memory: A Contribution to Experimental Psychology*. New York, NY: Teachers College, Columbia University.

Economist, ((2019). "Why firms should treat their employees well". *The Economist*. URL: https://www.economist.com/graphic-detail/2019/08/28/why-firms-should-treat-their-employees-well.

Economist, ((2023). "Covid-19 was a disaster for the world's schoolchildren". *The Economist*. URL: https://www.economist.com/leaders/2023/12/05/covid-19-was-a-disaster-for-the-worlds-schoolchildren.

Economist, ((2024). "Is your master's degree useless?" *The Economist*. URL: https://www.economist.com/international/2024/11/18/is-your-masters-degree-useless.

Edelson, N. M. and D. K. Hilderbrand. (1975). "Congestion tolls for Poisson queuing processes". *Econometrica: Journal of the Econometric Society*: 81–92.

Edmondson, A. (1999). "Psychological safety and learning behavior in work teams". *Administrative Science Quarterly*. 44(2): 350–383.

Ely, J., A. Frankel, and E. Kamenica. (2015). "Suspense and surprise". *Journal of Political Economy*. 123(1): 215–260.

Epstein, N., S. Gallino, and A. Moreno. (2023). "Operational consequences of customer interaction design: Evidence from last-mile delivery services". *Tech. rep.* Harvard University.

Erlang, A. K. (1909). "The theory of probabilities and telephone conversations". *Nyt. Tidsskr. Mat. Ser. B*. 20: 33–39.

Ernst, A. T., H. Jiang, M. Krishnamoorthy, and D. Sier. (2004). "Staff scheduling and rostering: A review of applications, methods and models". *European Journal of Operational Research*. 153(1): 3–27.

Evans, D. C. (2017). *Bottlenecks: Aligning UX Design with User Psychology*. Apress.

Fader, P. (2020). *Customer Centricity: Focus on the Right Customers for Strategic Advantage*. University of Pennsylvania Press.

Feldman, P., A. E. Frazelle, and R. Swinney. (2023). "Managing relationships between restaurants and food delivery platforms: Conflict, contracts, and coordination". *Management Science*. 69(2): 812–823.

Feldman, P. and E. Segev. (2022). "The important role of time limits when consumers choose their time in service". *Management Science*. 68(9): 6666–6686.

Field, J. M. (2024). *Designing Service Processes to Unlock Value*. Business Expert Press.

Fisher, M., S. Gallino, and S. Netessine. (2021a). "Does online training work in retail?" *Manufacturing & Service Operations Management*. 23(4): 876–894.

Fisher, M., S. Gallino, and S. Netessine. (2021b). "Setting retail staffing levels: A methodology validated with implementation". *Manufacturing & Service Operations Management*. 23(6): 1562–1579.

Fisher, M., J. Krishnan, and S. Netessine. (2006). "Retail store execution: An empirical study". *Tech. rep.* The Wharton School, University of Pennsylvania.

Fitzsimmons, J. A. (1985). "Consumer Participation and Productivity in Service Operations". *Interfaces*. 15(3): 60–67.

Fließ, S., S. Dyck, and M. Volkers. (2024). *Management von Dienstleistungsprozessen. Service Co-Creation - Service Experience - Service Value*. Springer Link.

Fliess, S., S. Dyck, and M. Volkers. (2024). "Understanding the structure of service processes from a customer perspective—An event segmentation approach". *Tech. rep.* University of Hagen. URL: https://www.researchgate.net/publication/339359253.

Flynn, J. (2024). "20 stunning Great Resignation statistics [2023]: How many people quit their jobs in 2022". URL: https://www.zippia.com/advice/great-resignation-statistics/.

Forgas, J. P. (1995). "Mood and judgment: the affect infusion model (AIM)." *Psychological Bulletin*. 117(1): 39.

Frei, F. X. (2006). "Breaking the trade-off between efficiency and service". *Harvard Business Review*: 1–12.

Frei, F. X. and A. Morriss. (2012). *Uncommon Service: How to Win by Putting Customers at the Core of Your Business*. Harvard Business Press.

Fuchs, V. R. (1968). *The Service Economy*. New York, NY: National Bureau of Economic Research.

Gale, D. and H. Nikaidô. (1965). "The Jacobian matrix and global univalence of mappings". *Mathematische Annalen*. 159(2): 81–93.

Gans, N., G. Koole, and A. Mandelbaum. (2003). "Telephone call centers: Tutorial, review, and research prospects". *Manufacturing & Service Operations Management*. 5(2): 79–141.

Gans, N. and Y.-P. Zhou. (2002). "Managing learning and turnover in employee staffing". *Operations Research*. 50(6): 991–1006.

Gao, F. and X. Su. (2018). "Omnichannel service operations with online and offline self-order technologies". *Management Science*. 64(8): 3595–3608.

Garnefeld, I. and L. Steinhoff. (2013). "Primacy versus recency effects in extended service encounters". *Journal of Service Management*. 24(1): 64–81.

Gaur, V. and Y.-H. Park. (2007). "Asymmetric consumer learning and inventory competition". *Management Science*. 53(2): 227–240.

Gilbert, S. M. and Z. K. Weng. (1998). "Incentive effects favor nonconsolidating queues in a service system: The principal-agent perspective". *Management Science*. 44(12-part-1): 1662–1669.

Global Change Data Lab. (2024). "Price changes in consumer goods and services in the United States". URL: https://ourworldindata.org/grapher/price-changes-consumer-goods-services-united-states.

Goes, P. B., N. Ilk, M. Lin, and J. L. Zhao. (2018). "When more is less: Field evidence on unintended consequences of multitasking". *Management Science*. 64(7): 3033–3054.

Goldratt, E. M. and J. Cox. (2016). *The Goal: A Process of Ongoing Improvement*. Routledge.

Goldstein, S. M. (2003). "Employee development: an examination of service strategy in a high-contact service environment". *Production and Operations Management*. 12(2): 186–203.

Gopalakrishnan, R., S. Doroudi, A. R. Ward, and A. Wierman. (2016). "Routing and staffing when servers are strategic". *Operations research*. 64(4): 1033–1050.

Goudsmet, A. and L. Van der Heyden. (2023). "The six dimensions of winning teams". *Tech. rep.* INSEAD. URL: https://ssrn.com/abstract=4670713.

Green, L. V., S. Savin, and N. Savva. (2013). ""Nursevendor problem": Personnel staffing in the presence of endogenous absenteeism". *Management Science*. 59(10): 2237–2256.

Grönroos, C. (1984). "A service quality model and its marketing implications". *European Journal of Marketing*. 18(4): 36–44.

Grönroos, C. (2011). "Value co-creation in service logic: A critical analysis". *Marketing Theory*. 11(3): 279–301.

Grossman, S. J. and O. D. Hart. (1986). "The costs and benefits of ownership: A theory of vertical and lateral integration". *Journal of Political Economy*. 94(4): 691–719.

Guajardo, J. A., M. A. Cohen, S.-H. Kim, and S. Netessine. (2012). "Impact of performance-based contracting on product reliability: An empirical analysis". *Management Science*. 58(5): 961–979.

Gubler, T., I. Larkin, and L. Pierce. (2018). "Doing well by making well: The impact of corporate wellness programs on employee productivity". *Management Science*. 64(11): 4967–4987.

Guda, H., M. Dawande, and G. Janakiraman. (2023). "The economics of process transparency". *Production and Operations Management*. 32(6): 1812–1829.

Guo, Q., Y. Li, L. Liu, and L. Sheng. (2024). "Retention optimization in maintenance training programs". *Tech. rep.* University of Science and Technology of China. URL: https://ssrn.com/abstract=4811915.

Gupta, S., A. Roy, S. Kumar, and R. Mudambi. (2023). "When worse is better: Strategic choice of vendors with differentiated capabilities in a complex cocreation environment". *Management Science*. 69(5): 2833–2851.

Gurvich, I., M. Lariviere, and A. Moreno. (2019). "Operations in the on-demand economy: Staffing services with self-scheduling capacity". In: *Sharing Economy. Making Supply Meet Demand*. Ed. by M. Hu. Springer. 249–278.

Gurvich, I. and J. A. Van Mieghem. (2015). "Collaboration and multitasking in networks: Architectures, bottlenecks, and capacity". *Manufacturing & Service Operations Management*. 17(1): 16–33.

Gurvich, I. and J. A. Van Mieghem. (2018). "Collaboration and multitasking in networks: Prioritization and achievable capacity". *Management Science*. 64(5): 2390–2406.

Hagiu, A. and J. Wright. (2015). "Multi-sided platforms". *International Journal of Industrial Organization*. 43: 162–174.

Hagiu, A. and J. Wright. (2019). "Controlling vs. enabling". *Management Science*. 65(2): 577–595.

Halfin, S. and W. Whitt. (1981). "Heavy-traffic limits for queues with many exponential servers". *Operations Research*. 29(3): 567–588.

Hallowell, R., D. Bowen, and C.-I. Knoop. (2002). " Four Seasons Goes to Paris: "53 Properties, 24 Countires, 1 Philosophy"". *Tech. rep.* No. 9-803-069. Harvard Business School.

Hardy, G., J. E. Littlewood, and G. Pólya. (1952). *Inequalities*. 2nd. New York, NY: Cambridge University Press.

Harker, P. T. (2014). "Commentary—making sense of higher education's future: an economics and operations perspective". *Service Science*. 6(4): 207–216.

Harrison, J. M. and A. Zeevi. (2005). "A method for staffing large call centers based on stochastic fluid models". *Manufacturing & Service Operations Management*. 7(1): 20–36.

Hasija, S., E. J. Pinker, and R. A. Shumsky. (2008). "Call center outsourcing contracts under information asymmetry". *Management Science*. 54(4): 793–807.

Hassin, R. (1986). "Consumer information in markets with random product quality: The case of queues and balking". *Econometrica: Journal of the Econometric Society*: 1185–1195.

Hassin, R. and M. Haviv. (1995). "Equilibrium strategies for queues with impatient customers". *Operations Research Letters*. 17(1): 41–45.

Hassin, R. and M. Haviv. (2003). *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems*. Vol. 59. Springer Science & Business Media.

Haviv, M. and Y. Ritov. (2001). "Homogeneous customers renege from invisible queues at random times under deteriorating waiting conditions". *Queueing Systems*. 38: 495–508.

Hayes, R. H. and S. C. Wheelwright. (1979). "Linking manufacturing process and product life cycles". *Harvard Business Review*. 57(1): 133–140.

Heitz, C., M. Dettling, and A. Ruckstuhl. (2011). "Modelling customer lifetime value in contractual settings". *International Journal of Services Technology and Management*. 16(2): 172–190.

Helm, J. E., S. AhmadBeygi, and M. P. Van Oyen. (2011). "Design and analysis of hospital admission control for operational effectiveness". *Production and Operations Management*. 20(3): 359–374.

Helson, H. (1964). *Adaptation-Level Theory: An Experimental and Systematic Approach to Behavior*. New York: Harper & Row.

Heskett, J. (2003). " Shouldice Hospital Limited". *Tech. rep.* No. 9-683-068. Harvard Business School.

Heskett, J. L., T. O. Jones, G. W. Loveman, W. E. Sasser Jr., and L. A. Schlesinger. (1994). "Putting the service-profit chain to work". *Harvard Business Review*. 72(2): 164–170.

Heskett, J. L. and W. E. Sasser. (2010). "The service profit chain: From satisfaction to ownership". In: *Handbook of Service Science*. Ed. by P. P. Maglio, C. A. Kieliszewski, and J. C. Spohrer. Springer. 19–29.

Hogreve, J., A. Iseke, and K. Derfuss. (2022). "The service-profit chain: reflections, revisions, and reimaginations". *Journal of Service Research*. 25(3): 460–477.

Hogreve, J., A. Iseke, K. Derfuss, and T. Eller. (2017). "The service–profit chain: A meta-analytic test of a comprehensive theoretical framework". *Journal of Marketing*. 81(3): 41–61.

Holmström, B. (1982). "Moral hazard in teams". *The Bell Journal of Economics*. 13(2): 324–340.

Holmström, B. (1979). "Moral hazard and observability". *The Bell Journal of Economics*. 10(1): 74–91.

Holmström, B. (1999). "Managerial incentive problems: A dynamic perspective". *The Review of Economic Studies*. 66(1): 169–182.

Holmström, B. and P. Milgrom. (1991). "Multitask principal–agent analyses: Incentive contracts, asset ownership, and job design". *The Journal of Law, Economics, and Organization*. 7(Special Issue): 24–52.

Holmström, B. and P. Milgrom. (1994). "The firm as an incentive system". *The American Economic Review*: 972–991.

Holt, C. C., F. Modigliani, and J. F. Muth. (1956). "Derivation of a linear decision rule for production and employment". *Management Science*. 2(2): 159–177.

Hopp, W. J., S. M. R. Iravani, and G. Y. Yuen. (2007). "Operations systems with discretionary task completion". *Management Science*. 53(1): 61–77.

Hopp, W. J., E. Tekin, and M. P. Van Oyen. (2004). "Benefits of skill chaining in serial production lines with cross-trained workers". *Management Science*. 50(1): 83–98.

Hopp, W. J. and M. P. Van Oyen. (2004). "Agile workforce evaluation: a framework for cross-training and coordination". *IIE Transactions*. 36(10): 919–940.

Hotelling, H. (1929). "Stability in competition". *The Economic Journal*. 39(153): 41–57.

HSO. (2023). "What are the four biggest expense costs in retail?" URL: https://www.hso.com/blog/what-are-the-four-biggest-expense-costs-in-retail.

Huang, M.-H. and R. T. Rust. (2018). “Artificial intelligence in service”. *Journal of Service Research*. 21(2): 155–172.

Huckman, R. S. and B. R. Staats. (2011). “Fluid tasks and fluid teams: The impact of diversity in experience and team familiarity on team performance”. *Manufacturing & Service Operations Management*. 13(3): 310–328.

Huckman, R. S., B. R. Staats, and D. M. Upton. (2009). “Team familiarity, role experience, and performance: Evidence from Indian software services”. *Management Science*. 55(1): 85–100.

Ibanez, M. R., J. R. Clark, R. S. Huckman, and B. R. Staats. (2018). “Discretionary task ordering: Queue management in radiological services”. *Management Science*. 64(9): 4389–4407.

Itoh, H. (2001). “Job design and incentives in hierarchies with team production”. *Hitotsubashi Journal of Commerce and Management*. 36(1): 1–17.

IxDF, I. D. F. .-. (2023). “Aristotle on storytelling in user experience”. URL: https://www.interaction-design.org/literature/article/aristotle-on-storytelling-in-user-experience.

Jensen, M. C. and W. H. Meckling. (1976). “Theory of the firm: Managerial behavior, agency costs and ownership structure”. *Journal of Financial Economics*. 11(4): 5–50.

John, G. and B. Weitz. (1989). “Salesforce compensation: An empirical investigation of factors related to use of salary versus incentive compensation”. *Journal of Marketing Research*. 26(1): 1–14.

Jordan, W. C. and S. C. Graves. (1995). “Principles on the benefits of manufacturing process flexibility”. *Management Science*. 41(4): 577–594.

Jouini, O., Y. Dallery, and R. Nait-Abdallah. (2008). “Analysis of the impact of team-based organizations in call center management”. *Management Science*. 54(2): 400–414.

Kahneman, D., B. L. Fredrickson, C. A. Schreiber, and D. A. Redelmeier. (1993). “When more pain is preferred to less: Adding a better end”. *Psychological Science*. 4(6): 401–405.

Kahneman, D. and A. Tversky. (1979). “Prospect theory: An analysis of decision under risk”. *Econometrica*. 47(2): 263–292.

Kahneman, D., P. P. Wakker, and R. Sarin. (1997). "Back to Bentham? Explorations of experienced utility". *The Quarterly Journal of Economics*. 112(2): 375–406.

Kalai, E., M. I. Kamien, and M. Rubinovitch. (1992). "Optimal service speeds in a competitive environment". *Management Science*. 38(8): 1154–1163.

Kamakura, W. A., V. Mittal, F. De Rosa, and J. A. Mazzon. (2002). "Assessing the service-profit chain". *Marketing Science*. 21(3): 294–317.

Kamalahmadi, M., Q. Yu, and Y.-P. Zhou. (2021). "Call to duty: Just-in-time scheduling in a restaurant chain". *Management Science*. 67(11): 6751–6781.

Kanoria, Y., I. Lobel, and J. Lu. (2023). "Managing customer churn via service mode control". *Mathematics of Operations Research*.

Kanoria, Y. and D. Saban. (2021). "Facilitating the search for partners on matching platforms". *Management Science*. 67(10): 5990–6029.

Karmarkar, U. S. (2004). "Will uou survive the services revolution?" *Harvard Business Review*. 82(6): 100–107.

Karmarkar, U. S. and U. R. Karmarkar. (2014). "Customer experience and service design". In: *Managing Consumer Services. Factory or Theater?* Ed. by E. Baglieri and U. S. Karmarkar. Springer. Chap. 7. 109–130.

Karmarkar, U. S. and R. Pitbladdo. (1995). "Service markets and competition". *Journal of Operations Management*. 12(3-4): 397–411.

Karmarkar, U. S. (2010). "The industrialization of Information Services". In: *Handbook of Service Science*. Ed. by P. P. Maglio, C. A. Kieliszewski, and J. C. Spohrer. Springer. 419–435.

Karmarkar, U. S., K. Kim, and H. Rhim. (2015). "Industrialization, productivity and the shift to services and information". *Production and Operations Management*. 24(11): 1675–1695.

Karmarkar, U. S. and G. Roels. (2015). "An analytical framework for value co-production in services". *Service Science*. 7(3): 163–180.

Katta, S., F. Ferrari, N. van Doorn, and M. Graham. (2024). "Migration, migrant work (ers) and the gig economy". *Environment and Planning A: Economy and Space*: 0308518X241250168.

KC, D. S., B. R. Staats, M. Kouchaki, and F. Gino. (2020). "Task selection and workload: A focus on completing easy tasks hurts performance". *Management Science*. 66(10): 4397–4416.

KC, D. S. and C. Terwiesch. (2009). "Impact of workload on service time and patient safety: An econometric analysis of hospital operations". *Management Science*. 55(9): 1486–1498.

Kellogg, D. L. and W. Nie. (1995). "A framework for strategic service management". *Journal of Operations Management*. 13(4): 323–337.

Kelly, S. W., J. H. Donnelly Jr, and S. J. Skinner. (1990). "Customer participation in service production and delivery". *Journal of Retailing*. 66(3): 315–336.

Keppler, S. M., J. Li, and D. Wu. (2022). "Crowdfunding the front lines: An empirical study of teacher-driven school improvement". *Management Science*. 68(12): 8809–8828.

Kesavan, S., S. J. Lambert, J. C. Williams, and P. K. Pendem. (2022). "Doing well by doing good: Improving retail store performance with responsible scheduling practices at the Gap, Inc." *Management Science*. 68(11): 7818–7836.

Kim, H., Y. S. Lee, and K. S. Park. (2018). "The psychology of queuing for self-service: Reciprocity and social pressure". *Administrative Sciences*. 8(4): 75.

Kim, S. H. and S. Netessine. (2013). "Collaborative cost reduction and component procurement under information asymmetry". *Management Science*. 59(1): 189–206.

Kim, S.-H., M. A. Cohen, and S. Netessine. (2007). "Performance contracting in after-sales service supply chains". *Management Science*. 53(12): 1843–1858.

Kim, S.-H., J. A. Guajardo, and S. Netessine. (2022). "Performance-based contracting: Past, present, and future". In: *Creating Values with Operations and Analytics: A Tribute to the Contributions of Professor Morris Cohen*. Springer. 85–103.

Kim, S. K. and S. Wang. (1998). "Linear contracts and the double moral-hazard". *J. Econom. Theor.* 82(2): 342–378.

Kleinrock, L. (1974). *Queueing Systems: Theory*. wiley.

Koopmans, T. C. (1960). "Stationary ordinal utility and impatience". *Econometrica: Journal of the Econometric Society*: 287–309.

Kostami, V. and S. Rajagopalan. (2014). "Speed–quality trade-offs in a dynamic model". *Manufacturing & Service Operations Management*. 16(1): 104–118.

Kőszegi, B. and M. Rabin. (2006). "A model of reference-dependent preferences". *The Quarterly Journal of Economics*. 121(4): 1133–1165.

Kowalkowski, C., H. Gebauer, B. Kamp, and G. Parry. (2017). "Servitization and deservitization: Overview, concepts, and definitions". *Industrial Marketing Management*. 60: 4–10.

Kremer, M. and L. Debo. (2016). "Inferring quality from wait time". *Management Science*. 62(10): 3023–3038.

Kumar, P., M. U. Kalwani, and M. Dada. (1997). "The impact of waiting time guarantees on customers' waiting experiences". *Marketing Science*. 16(4): 295–314.

Kusiak, A. and S. S. Heragu. (1987). "The facility layout problem". *European Journal of Operational Research*. 29(3): 229–251.

Kwon, C. and A. Raman. (2023). "The real effects of Fair Workweek Laws on work schedules: Evidence from Chicago, Los Angeles, and Philadelphia". *Tech. rep.* Harvard Business School. URL: http://dx.doi.org/10.2139/ssrn.4609755.

Lago, A. and P. G. Moscoso. (2011). "Metro Bank: The British Banking Revolution Begins". *Tech. rep.* No. P-1112-E. IESE.

Lancaster, K. J. (1966). "A new approach to consumer theory". *Journal of Political Economy*. 74(2): 132–157.

Lapré, M. A. and I. M. Nembhard. (2011). "Inside the organizational learning curve: Understanding the organizational learning process". *Foundations and Trends® in Technology, Information and Operations Management*. 4(1): 1–103.

Larkin, I., L. Pierce, and F. Gino. (2012). "The psychological costs of pay-for-performance: Implications for the strategic compensation of employees". *Strategic Management Journal*. 33(10): 1194–1214.

Larson, R. C. (1987). "OR Forum—Perspectives on queues: Social justice and the psychology of queueing". *Operations Research*. 35(6): 895–905.

Latané, B. (1981). "The psychology of social impact". *American Psychologist*. 36(4): 343–356.

Lazear, E. P. and M. Gibbs. (2014). *Personnel Economics in Practice*. John Wiley & Sons.

Lee, S. and M. Sosa. (2024). "Spaces for creativity: Unconventional workspaces and divergent thinking".

Lemon, K. N. and P. C. Verhoef. (2016). "Understanding customer experience throughout the customer journey". *Journal of Marketing*. 80(6): 69–96.

Levitt, T. (1972). "Production line approach to services". *Harvard Business Review*. 50(5): 41–52.

Levitt, T. (1976). "The industrialization of service". *Harvard Business Review*. 54(5): 63–74.

Li, S., M. A. Lariviere, and A. Bassamboo. (2024). "Is full price the full story when consumers have time and budget constraints?" *Manufacturing & Service Operations Management*. 26(1): 370–388.

Li, Y., T. Dai, and X. Qi. (2022). "A theory of interior peaks: Activity sequencing and selection for service design". *Manufacturing & Service Operations Management*. 24(2): 993–1001.

Li, Y. and X. Qi. (2022). "A geometric branch-and-bound algorithm for the service bundle design problem". *European Journal of Operational Research*. 303(3): 1044–1056.

Li, Y., C. T. Ryan, and L. Sheng. (2023). "Optimal sequencing in single-player games". *Management Science*. 69(10): 6057–6075.

Limon, Y., T. Martagan, and A. Krishnamurthy. (2024). "Contracts for biopharmaceutical manufacturing based on production cost and capabilities". *International Journal of Production Research*. 62(7): 2640–2662.

Lin, C. and C. Zhang. (2020). "KFC China: Building Competitive Advantages through Digitilization". *Tech. rep.* No. CB0031. CEIBS.

Lindbeck, A. and D. J. Snower. (2000). "Multitask learning and the reorganization of work: From Tayloristic to holistic organization". *Journal of Labor Economics*. 18(3): 353–376.

Lobel, I., S. Martin, and H. Song. (2024). "Frontiers in Operations: Employees vs. contractors: An operational perspective". *Manufacturing & Service Operations Management*.

Loewenstein, G. and D. Prelec. (1991). "Negative time preference". *The American Economic Review*. 81(2): 347–352.

Loewenstein, G. F. and D. Prelec. (1993). "Preferences for sequences of outcomes." *Psychological Review*. 100(1): 91.

Long, X. and J. Nasiry. (2020). "Wage transparency and social comparison in sales force compensation". *Management Science*. 66(11): 5290–5315.

Lovelock, C. H. (1983). "Classifying services to gain strategic marketing insights". *Journal of Marketing*. 47(3): 9–20.

Lovelock, C. H. and R. F. Young. (1979). "Look to consumers to increase productivity". *Harvard Business Review*. 57(3): 168–178.

Loveman, G. W., L. A. Schlesinger, and R. T. Anthony. (1992). " Euro Disney: The First 100 Days". *Tech. rep*. No. 9-693-013. Harvard Business School.

Lu, G., R. Y. Du, and X. Peng. (2022). "The impact of schedule consistency on shift worker productivity: An empirical investigation". *Manufacturing & Service Operations Management*. 24(5): 2780–2796.

Lu, Y., A. Musalem, M. Olivares, and A. Schilkrut. (2013). "Measuring the effect of queues on customer purchases". *Management Science*. 59(8): 1743–1763.

Luo, X., S. Tong, Z. Fang, and Z. Qu. (2019). "Frontiers: Machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases". *Marketing Science*. 38(6): 937–947.

Ma, L., F. Yang, M. Lin, and W. Xue. (2024). "Pricing and quality decisions for standardized and collaborative services in a home health care service platform". *Transportation Research Part E: Logistics and Transportation Review*. 181(103366): 1–31.

MacDuffie, J. P. (1995). "Human resource bundles and manufacturing performance: Organizational logic and flexible production systems in the world auto industry". *ILR Review*. 48(2): 197–221.

Maister, D. H. (2004). "The Anatomy of a Consulting Firm". In: *The Advice Business: Essential Tools and Models for Managing Consulting*. Ed. by C. J. Fombrun and M. D. Nevis. Pearson Prentice-Hall.

Mandelbaum, A. (2011). "Service engineering: Data-based science in support of service management, or empirical adventures in call centers and hospitals". URL: https://repository.gatech.edu/entities/publication/de0b9ad2-ad17-4f67-9b9b-dc2139ec97be.

Mandelbaum, A. and N. Shimkin. (2000). "A model for rational abandonments from invisible queues". *Queueing Systems*. 36: 141–173.

Mandelbaum, A. and S. Zeltyn. (2010). "Service engineering: Data-based course development and teaching". *INFORMS Transactions on Education*. 11(1): 3–19.

Mani, V., S. Kesavan, and J. M. Swaminathan. (2015). "Estimating the impact of understaffing on sales and profitability in retail stores". *Production and Operations Management*. 24(2): 201–218.

Mas, A. and E. Moretti. (2009). "Peers at work". *American Economic Review*. 99(1): 112–145.

Maxham III, J. G., R. G. Netemeyer, and D. R. Lichtenstein. (2008). "The retail value chain: linking employee perceptions to employee performance, customer evaluations, and store performance". *Marketing Science*. 27(2): 147–167.

Men, C., M. Van den Broeke, M. Verschelde, and B. Roets. (2024). "Human-Machine Interaction in Digital Control Rooms: The Dual Impact on Workload and Operational Performance". *Tech. rep.* KUL. URL: https://ssrn.com/abstract=4891595.

Mendelson, H. (1985). "Pricing computer services: Queueing effects". *Communications of the ACM*. 28(3): 312–321.

Mendelson, H. and S. Whang. (1990). "Optimal incentive-compatible priority pricing for the M/M/1 queue". *Operations Research*. 38(5): 870–883.

Meng, L., R. J. Batt, and C. Terwiesch. (2021). "The impact of facility layout on service worker behavior: An empirical study of nurses in the emergency department". *Manufacturing & Service Operations Management*. 23(4): 819–834.

Metters, R., D. Marshall, and M. Pagell. (2019). "Cultural research in the production and operations management field". *Foundations and Trends® in Technology, Information and Operations Management*. 13(1-2): 1–150.

Milgrom, P. and J. Roberts. (1995). "Complementarities and fit. Strategy, structure, and organizational change in manufacturing". *Journal of Accounting and Economics*. 19(2-3): 179–208.

Mills, P. K. and J. H. Morris. (1986). "Clients as "partial" employees of service organizations: Role development in client participation". *Academy of Management Review*. 11(4): 726–735.

Minkiewicz, J., J. Evans, and K. Bridson. (2014). "How do consumers co-create their experiences? An exploration in the heritage sector". *Journal of Marketing Management*. 30(1-2): 30–59.

Mittal, V., P. Kumar, and M. Tsiros. (1999). "Attribute-level performance, satisfaction, and behavioral intentions over time: a consumption-system approach". *Journal of Marketing*. 63(2): 88–101.

Mont, O. K. (2002). "Clarifying the concept of product–service system". *Journal of Cleaner Production*. 10(3): 237–245.

Moon, K. (2024). "(Machine) learning preferences from complex choice sets: An application to service networks". *Tech. rep.* University of Pennsylvania. URL: http://dx.doi.org/10.2139/ssrn.3819117.

Munichor, N. and A. Rafaeli. (2007). "Numbers or apologies? Customer reactions to telephone waiting time fillers." *Journal of Applied Psychology*. 92(2): 511.

Nagle, T. T., G. Müller, and E. Gruyaert. (2023). *The Strategy and Tactics of Pricing: A Guide to Growing More Profitably*. Routledge.

Naor, P. (1969). "The regulation of queue size by levying tolls". *Econometrica: journal of the Econometric Society*: 15–24.

Narayanan, S., S. Balasubramanian, and J. M. Swaminathan. (2009). "A matter of balance: Specialization, task variety, and individual learning in a software maintenance environment". *Management Science*. 55(11): 1861–1876.

Nasiry, J. and I. Popescu. (2011a). "Dynamic pricing with loss-averse consumers and peak-end anchoring". *Operations Research*. 56(6): 1361–1368.

Nasiry, J. and I. Popescu. (2011b). "Dynamic pricing with loss-averse consumers and peak-end anchoring". *Operations Research*. 59(6): 1361–1368.

Nath, H., U. Apte, and U. Karmarkar. (2020). "Service industrialization, employment and wages in the us information economy". *Foundations and Trends® in Technology, Information and Operations Management*. 13(4): 250–343.

Netessine, S. (2022). "OM forum—A vision of responsible research in operations management". *Manufacturing & Service Operations Management*. 24(6): 2799–2808.

Netessine, S., G. Dobson, and R. A. Shumsky. (2002). "Flexible service capacity: Optimal investment and the impact of demand correlation". *Operations Research*. 50(2): 375–388.

Norman, D. A. (2009). "Designing waits that work". *MIT Sloan Management Review*. 50(4): 23–28.

Norton, M. I., D. Mochon, and D. Ariely. (2012). "The IKEA effect: When labor leads to love". *Journal of Consumer Psychology*. 22(3): 453–460.

O'Donoghue, T. and M. Rabin. (1999). "Doing it now or later". *American Economic Review*. 89(1): 103–124.

Oliver, R. L. (1977). "Effect of expectation and disconfirmation on postexposure product evaluations: An alternative interpretation." *Journal of Applied Psychology*. 62(4): 480.

Oliver, R. L. (1981). "Measurement and evaluation of satisfaction processes in retail settings". *Journal of Retailing*.

Oliver, R. L. (2014). *Satisfaction: A Behavioral Perspective on the Consumer*. Routledge.

Olsen, E. O. (2024). "Customer Lean. Moving beyond customer engagement. Unleashing the power of customer knowledge and creativity". URL: https://www.purpose-ccl.org/customerlean.

Ortiz-Ospina, E. and N. Lippolis. (2017). "Structural transformation: how did today's rich countries become 'deindustrialized'?" URL: https://ourworldindata.org/structural-transformation-and-deindustrialization-evidence-from-todays-rich-countries.

Özer, Ö., Y. Zheng, and Y. Ren. (2014). "Trust, trustworthiness, and information sharing in supply chains bridging China and the United States". *Management Science*. 60(10): 2435–2460.

Özkan-Seely, G. F., C. Gaimon, and S. Kavadias. (2015). "Dynamic knowledge transfer and knowledge development for product and process design teams". *Manufacturing & Service Operations Management*. 17(2): 177–190.

Palmer, A. (2010). "Customer experience management: a critical review of an emerging idea". *Journal of Services Marketing*. 24(3): 196–208.

Parasuraman, A., V. A. Zeithaml, and L. L. Berry. (1988). "SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality". *Journal of Retailing*. 64(1): 12.

Parasuraman, A., V. A. Zeithaml, and L. L. Berry. (1985). "A conceptual model of service quality and its implications for future research". *Journal of Marketing*. 49(4): 41–50.

Pavlov, V., R. Mogre, and T. Lennon Olsen. (2024). "The sooner the better but not too soon, please." *Tech. rep.* University of Auckland. URL: https://dx.doi.org/10.2139/ssrn.4744367.

Perdikaki, O., S. Kesavan, and J. M. Swaminathan. (2012). "Effect of traffic on sales and conversion rates of retail stores". *Manufacturing & Service Operations Management*. 14(1): 145–162.

Pierce, L., D. C. Snow, and A. McAfee. (2015). "Cleaning house: The impact of information technology monitoring on employee theft and productivity". *Management Science*. 61(10): 2299–2319.

Pine, B. J. and J. H. Gilmore. (2011). *The Experience Economy*. Harvard Business Press.

Plambeck, E. L. and Q. Wang. (2013). "Implications of hyperbolic discounting for optimal pricing and scheduling of unpleasant services that generate future benefits". *Management Science*. 59(8): 1927–1946.

Ponsignon, F. (2023). "Making the customer experience journey more hedonic in a traditionally utilitarian service context: a case study". *Journal of Service Management*. 34(2): 294–315.

Ponsignon, F., F. Durrieu, and T. Bouzdine-Chameeva. (2017). "Customer experience design: a case study in the cultural sector". *Journal of Service Management*. 28(4): 763–787.

Popescu, I. and Y. Wu. (2007). "Dynamic pricing strategies with reference effects". *Operations Research*. 55(3): 413–429.

Porter, M. E. (1985). *Competitive Advantage: Creating and Sustaining Superior Performance*. New York, NY: Simon and Schuster.

Rahmandad, H. and Z. Ton. (2020). "If higher pay is profitable, why is it so rare? Modeling competing strategies in mass market services". *Organization Science*. 31(5): 1053–1071.

Rahmani, M., G. Roels, and U. S. Karmarkar. (2018). "Team leadership and performance: Combining the roles of direction and contribution". *Management Science*. 64(11): 5234–5249.

Rajaram, K. and R. Ahmadi. (2003). "Flow management to optimize retail profits at theme parks". *Operations Research*. 51(2): 175–184.

Ramdas, K. and A. Darzi. (2017). "Adopting innovations in care delivery-the case of shared medical appointments". *New England Journal of Medicine*. 376: 1105–1107.

Redelmeier, D. A. and D. Kahneman. (1996). "Patients' memories of painful medical treatments: Real-time and retrospective evaluations of two minimally invasive procedures". *Pain*. 66(1): 3–8.

Ren, Z. J. and Y.-P. Zhou. (2008). "Call center outsourcing: Coordinating staffing level and service quality". *Management Science*. 54(2): 369–383.

Rodriguez, A. E., R. Ibrahim, and D. Zhan. (2024). "On customer (dis) honesty in priority queues: The role of lying aversion". *Management Science*.

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

Roels, G. and L. Van der Heyden. (2025). "Modeling Collective Action in POM: The Strategy-Operations-People ($S_3O_2P_1$) Framework". *Tech. rep.* INSEAD. URL: https://dx.doi.org/10.2139/ssrn.5197018.

Roels, G. (2014). "Optimal design of coproductive services: Interaction and work allocation". *Manufacturing & Service Operations Management*. 16(4): 578–594.

Roels, G. (2020). "High-performance practice processes". *Management Science*. 66(4): 1509–1526.

Roels, G. and C. J. Corbett. (2024). "Too many meetings? Scheduling rules for team coordination". *Management Science*.

Roels, G. and K. Fridgeirsdottir. (2009). "Dynamic revenue management for online display advertising". *Journal of Revenue and Pricing Management*. 8(5): 452–466.

Roels, G., V. Smirnov, I. Tsetlin, and A. Wait. (2024). "You, me, or we? Co-productive principal-agent dynamics". *Tech. rep.* INSEAD. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3892218.

Roels, G. and B. R. Staats. (2021). "OM forum—people-centric operations: Achievements and future research directions". *Manufacturing & Service Operations Management*. 23(4): 745–757.

Roels, G. and X. Su. (2014). "Optimal design of social comparison effects: Setting reference groups and reference points". *Management Science*. 60(3): 606–627.

Ross, W. T. and I. Simonson. (1991). "Evaluations of pairs of experiences: A preference for happy endings". *Journal of Behavioral Decision Making*. 4(4): 273–282.

Roy, D., E. Spiliotopoulou, and J. de Vries. (2022). "Restaurant analytics: Emerging practice and research opportunities". *Production and Operations Management*. 31(10): 3687–3709.

Rozin, P. and A. Rozin. (2018). "Advancing understanding of the aesthetics of temporal sequences by combining some principles and practices in music and cuisine with psychology". *Perspectives on Psychological Science*. 13(5): 598–617.

Rust, R. T. and R. L. Oliver. (1993). *Service Quality: New Directions in Theory and Practice*. Sage Publications.

Ryu, M. and R. Jiang. (2025). "Nurse staffing under absenteeism: A distributionally robust optimization approach". *Manufacturing & Service Operations Management*.

Salop, S. C. (1979). "Monopolistic competition with outside goods". *The Bell Journal of Economics*: 141–156.

Sampson, S. E. (2010). "The Unified Service Theory: A paradigm for Service Science". In: *Handbook of Service Science*. Ed. by P. P. Maglio, C. A. Kieliszewski, and J. C. Spohrer. Springer. 107–131.

Sampson, S. E. (2012). "Visualizing service operations". *Journal of Service Research*. 15(2): 182–198.

Sampson, S. E. and C. M. Froehle. (2006). "Foundations and implications of a proposed unified services theory". *Production and Operations Management*. 15(2): 329–343.

Sampson, S. E. (2018). "Professional service jobs: Highly paid but subject to disruption?" *Service Science*. 10(4): 457–475.

Sampson, S. E. (2021). "A strategic framework for task automation in professional services". *Journal of Service Research*. 24(1): 122–140.

Sampson, S. E. and R. B. Chase. (2020). "Customer contact in a digital world". *Journal of Service Management*. 31(6): 1061–1069.

Sampson, S. E. and R. B. Chase. (2022). "Optimizing customer involvement: how close should you be to your customers?" *California Management Review*. 65(1): 119–146.

Sampson, S. E. and R. P. dos Santos. (2023). "Reengineering professional services through automation, remote outsourcing, and task delegation". *Journal of Operations Management*. 69(6): 911–940.

Sasser, W. E. (2004). " Benihana of Tokyo". *Tech. rep.* No. 9-673-057. Harvard Business School.

Schilling, M. A., P. Vidal, R. E. Ployhart, and A. Marangoni. (2003). "Learning by doing something else: Variation, relatedness, and the learning curve". *Management Science*. 49(1): 39–56.

Schmenner, R. W. (1986). "How can service businesses survive and prosper". *Sloan Management Review*. 27(3): 21–32.

Schmenner, R. W. (2004). "Service businesses and productivity". *Decision Sciences*. 35(3): 333–347.

Schmenner, R. W. and M. L. Swink. (1998). "On theory in operations management". *Journal of Operations Management*. 17(1): 97–113.

Schneider, B., J. J. Parkington, and V. M. Buxton. (1980). "Employee and customer perceptions of service in banks". *Administrative Science Quarterly*: 252–267.

Schwarz, N. and G. L. Clore. (1983). "Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states." *Journal of Personality and Social Psychology*. 45(3): 513.

Shimkin, N. and A. Mandelbaum. (2004). "Rational abandonment from tele-queues: Nonlinear waiting costs with heterogeneous preferences". *Queueing Systems*. 47: 117–146.

Shostack, G. L. (1984). "Designing services that deliver". *Harvard Business Review*: 132–139.

Shostack, G. L. (1987). "Service positioning through structural change". *Journal of Marketing*. 51(1): 34–43.

Shumsky, R. A. and E. J. Pinker. (2003). "Gatekeepers and referrals in services". *Management Science*. 49(7): 839–856.

Shunko, M., J. Niederhoff, and Y. Rosokha. (2018). "Humans are not machines: The behavioral impact of queueing design on service time". *Management Science*. 64(1): 453–473.

Siemsen, E., S. Balasubramanian, and A. V. Roth. (2007). "Incentives that induce task-related effort, helping, and knowledge sharing in workgroups". *Management Science*. 53(10): 1533–1550.

Siemsen, E., A. V. Roth, S. Balasubramanian, and G. Anand. (2009). "The influence of psychological safety and confidence in knowledge on employee knowledge sharing". *Manufacturing & service operations management*. 11(3): 429–447.

Siggelkow, N. (2002). "Misperceiving interactions among complements and substitutes: Organizational consequences". *Management Science*. 48(7): 900–916.

Skinner, W. (1974). "The focused factory". *Harvard Business Review*. 52(3): 113–121.

Snyder, H., L. Witell, A. Gustafsson, and J. R. McColl-Kennedy. (2022). "Consumer lying behavior in service encounters". *Journal of Business Research*. 141: 755–769.

Solomon, M. R., C. Surprenant, J. A. Czepiel, and E. G. Gutman. (1985). "A role theory perspective on dyadic interactions: The service encounter". *Journal of Marketing*. 49(1): 99–111.

Song, H., M. Armony, and G. Roels. (2024). "Queue configurations and operational performance: An interplay between customer ownership and queue length awareness". *Manufacturing & Service Operations Management*. 26(6): 2284–2304.

Song, H., A. L. Tucker, R. Graue, S. Moravick, and J. J. Yang. (2020). "Capacity pooling in hospitals: The hidden consequences of off-service placement". *Management Science*. 66(9): 3825–3842.

Song, H., A. L. Tucker, K. L. Murrell, and D. R. Vinson. (2018). "Closing the productivity gap: Improving worker productivity through public relative performance feedback and validation of best practices". *Management Science*. 64(6): 2628–2649.

Soteriou, A. C. and R. B. Chase. (2000). "A robust optimization approach for improving service quality". *Manufacturing & Service Operations Management*. 2(3): 264–286.

Staats, B. R. and F. Gino. (2012). "Specialization and variety in repetitive tasks: Evidence from a Japanese bank". *Management Science*. 58(6): 1141–1159.

Statistics Times. (2024). "List of countries by GDP sector composition". URL: http://statisticstimes.com/economy/countries-by-gdp-sector-composition.php.

Talluri, K. T. and G. J. Van Ryzin. (2006). *The Theory and Practice of Revenue Management*. Vol. 68. Springer Science & Business Media.

Tan, T. F. and S. Netessine. (2014). "When does the devil make work? An empirical study of the impact of workload on worker productivity". *Management Science*. 60(6): 1574–1593.

Tan, T. F. and S. Netessine. (2019). "When you work with a superman, will you also fly? An empirical study of the impact of coworkers on performance". *Management Science*. 65(8): 3495–3517.

Tan, T. F. and S. Netessine. (2020). "At your service on the table: Impact of tabletop technology on restaurant performance". *Management Science*. 66(10): 4496–4515.

Taylor, S. (1994). "Waiting for service: the relationship between delays and evaluations of service". *Journal of Marketing*. 58(2): 56–69.

Taylor, T. A. (2018). "On-demand service platforms". *Manufacturing & Service Operations Management*. 20(4): 704–720.

Teboul, J. (2006). *Service Is Front-Stage*. New York, NY: INSEAD Business Press, Palgrave McMillan.

Tereyağoğlu, N., P. S. Fader, and S. Veeraraghavan. (2018). "Multiattribute loss aversion and reference dependence: Evidence from the performing arts industry". *Management Science*. 64(1): 421–436.

Tessitore, S. (2023). "Highest NPS scores: Best NPS scores from top companies in 2023". URL: https://customergauge.com/benchmarks/blog/top-highest-nps-scores.

Thomke, S. (2012). "Mumbai's models of service excellence". *Harvard Business Review*. 90(11): 121–126.

Thompson, B. (2018). "Data factories". URL: https://stratechery.com/2018/data-factories/.

Thompson, J. D. (1967). *Organizations in Action: Social Science Bases of Administrative Theory*. New-York: McGraw-Hill.

Ton, Z. and S. Harrow. (2010). " Mercadona". *Tech. rep.* No. 9-610-089. Harvard Business School.

Ton, Z. (2014). *The Good Jobs Strategy: How the Smartest Companies Invest in Employees to Lower Costs and Boost Profits*. Houghton Mifflin Harcourt.

Ton, Z. and R. S. Huckman. (2008). "Managing the impact of employee turnover on performance: The role of process conformance". *Organization Science*. 19(1): 56–68.

Tong, C. and S. Rajagopalan. (2014). "Pricing and operational performance in discretionary services". *Production and Operations Management*. 23(4): 689–703.

Topkis, D. M. (1998). *Supermodularity and Complementarity*. Princeton, NJ: Princeton University Press.

Tsai, C.-Y. and S.-H. Chung. (2012). "A personalized route recommendation service for theme parks using RFID information and tourist behavior". *Decision Support Systems*. 52(2): 514–527.

Tu, R., W. Feng, C. Lin, and P. Tu. (2018). "Read into the lines: the positive effects of queues". *Journal of Service Theory and Practice*. 28(5): 661–681.

Tuncalp, F., R. Ibrahim, S.-H. Kim, and J. Tong. (2023). "When should doctors and patients use shared decision-making under bounded rationality?" *Tech. rep.* Bilkent University.

Ülkü, S., C. Hydock, and S. Cui. (2020). "Making the wait worthwhile: Experiments on the effect of queueing on consumption". *Management Science*. 66(3): 1149–1171.

Ülkü, S., C. Hydock, and S. Cui. (2022). "Social queues (cues): Impact of others' waiting in line on one's service time". *Management Science*. 68(11): 7958–7976.

Van Alstyne, M. W., G. G. Parker, and S. P. Choudary. (2016). "Pipelines, platforms, and the new rules of strategy". *Harvard Business Review*. 94(4): 54–62.

Van Doorn, N., F. Ferrari, and M. Graham. (2023). "Migration and migrant labour in the gig economy: An intervention". *Work, Employment and Society*. 37(4): 1099–1111.

Van Mieghem, J. A. (1995). "Dynamic scheduling with convex delay costs: The generalized c| mu rule". *The Annals of Applied Probability*: 809–833.

Varey, C. and D. Kahneman. (1992). "Experiences extended across time: Evaluation of moments and episodes". *Journal of Behavioral Decision Making*. 5(3): 169–185.

Vargo, S. L. and R. F. Lusch. (2004). "Evolving to a new dominant logic for Marketing". *Journal of Marketing*. 68(1): 1–17.

Veeraraghavan, S. K. and L. G. Debo. (2011). "Herding in queues with waiting costs: Rationality and regret". *Manufacturing & Service Operations Management*. 13(3): 329–346.

Vives, X. (1999). *Oligopoly Pricing: Old Ideas and New Tools*. MIT Press.

Wall Street Journal. (2023). "What happened when Uber's CEO started driving for Uber". URL: https://www.wsj.com/articles/uber-ceo-started-driving-for-uber-5bef5023.

Wang, J., L. Ma, W. Xue, and Y.-H. Kuo. (2022). "Impact of self-service technology in designing a service delivery system". *Production and Operations Management*.

Wang, J. and Y.-P. Zhou. (2018). "Impact of queue configuration on service time: Evidence from a supermarket". *Management Science*. 64(7): 3055–3075.

Wathieu, L. (1997). "Habits and the anomalies in intertemporal choice". *Management Science*. 43(11): 1552–1563.

Wemmerlöv, U. (1990). "A taxonomy for service processes and its implications for service design". *Int. J. Service Ind. Management*. 1(1): 20–40.

Whitt, W. (2006). "Staffing a call center with uncertain arrival rate and absenteeism". *Production and Operations Management*. 15(1): 88–102.

Wiengarten, F., M. Pagell, C. F. Durach, and P. Humphreys. (2021). "Exploring the performance implications of precarious work". *Journal of Operations Management*. 67(8): 926–963.

Wigert, B. (2020). "Employee burnout: The biggest myth". URL: https://www.gallup.com/workplace/288539/employee-burnout-biggest-myth.aspx.

Wirtz, J. and C. Lovelock. (2021). *Services Marketing: People, Technology, Strategy*. World Scientific.

Womack, J. P. and D. T. Jones. (2015). *Lean Solutions: How Companies and Customers Can Create Value and Wealth Together*. Simon and Schuster.

Xu, Y., H. Dai, and W. Yan. (2024). "Identity disclosure and anthropomorphism in voice chatbot design: A field experiment". *Management Science*.

Xu, Y., T. F. Tan, and S. Netessine. (2022). "The impact of workload on operational risk: Evidence from a commercial bank". *Management Science*. 68(4): 2668–2693.

Xue, M. and J. M. Field. (2008). "Service coproduction with information stickiness and incomplete contracts: Implications for consulting services design". *Production and Operations Management*. 17(3): 357–372.

Xue, M. and P. T. Harker. (2002). "Customer efficiency concept and its impact on e-business management". *Journal of Service Research*. 4(4): 253–267.

Xue, M., L. M. Hitt, and P. T. Harker. (2007). "Customer efficiency, channel usage, and firm performance in retail banking". *Manufacturing & Service Oper. Management*. 9(4): 535–558.

Xue, M., L. M. Hitt, and P.-y. Chen. (2011). "Determinants and outcomes of internet banking adoption". *Management Science*. 57(2): 291–307.

Yerkes, R. M. and J. D. Dodson. (1908). "The relation of strength of stimulus to rapidity of habit-formation". *Journal of Comparative Neurology and Psychology*. 18: 459–482.

Yu, Q., G. Allon, and A. Bassamboo. (2017). "How do delay announcements shape customer behavior? An empirical study". *Management Science*. 63(1): 1–20.

Yu, Q., G. Allon, and A. Bassamboo. (2021). "The reference effect of delay announcements: A field experiment". *Management Science*. 67(12): 7417–7437.

Yu, Q., G. Allon, A. Bassamboo, and S. Iravani. (2018). "Managing customer expectations and priorities in service systems". *Management Science*. 64(8): 3942–3970.

Yu, Q., S. Mankad, and M. Shunko. (2023). "Evidence of the unintended labor scheduling implications of the minimum wage". *Manufacturing & Service Operations Management*. 25(5): 1947–1965.

Yu, Q., Y. Zhang, and Y.-P. Zhou. (2022). "Delay information in virtual queues: A large-scale field experiment on a major ride-sharing platform". *Management Science*. 68(8): 5745–5757.

Yuan, Y. (2025). "Managing Flexible Capacity in Service Systems with Worker Shortages". *Manufacturing & Service Operations Management*.

Zeng, Z., H. Dai, D. J. Zhang, H. Zhang, R. Zhang, Z. Xu, and Z.-J. M. Shen. (2023). "The impact of social nudges on user-generated content for social network platforms". *Management Science*. 69(9): 5189–5208.

Zhang, D. J., G. Allon, and J. A. Van Mieghem. (2017). "Does social interaction improve learning outcomes? Evidence from field experiments on massive open online courses". *Manufacturing & Service Operations Management*. 19(3): 347–367.

Zhong, Y., R. Gopalakrishnan, and A. R. Ward. (2023). "Behavior-aware queueing: The finite-buffer setting with many strategic servers". *Operations Research*.

Zomerdijk, L. G. and C. A. Voss. (2010). "Service design for experience-centric services". *Journal of Service Research*. 13(1): 67–82.