Manufacturing & Service Operations Management

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Optimal Design of Coproductive Services: Interaction and Work Allocation

Guillaume Roels

To cite this article:
Guillaume Roels (2014) Optimal Design of Coproductive Services: Interaction and Work Allocation. Manufacturing & Service Operations Management 16(4):578-594. http://dx.doi.org/10.1287/msom.2014.0495

M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 16, No. 4, Fall 2014, pp. 578–594
ISSN 1523-4614 (print) | ISSN 1526-5498 (online)

informs

http://dx.doi.org/10.1287/msom.2014.0495

Optimal Design of Coproductive Services: Interaction and Work Allocation

Guillaume Roels
UCLA Anderson School of Management, University of California, Los Angeles, Los Angeles, California 90095, groels@anderson.ucla.edu

In services, customers provide significant inputs into the production process. In particular, these inputs may be the customers themselves participating in the service delivery. Although many service firms have explored different ways of involving customers in their production process, there is no clear guideline for the design of such coproductive systems. In this paper, we develop an analytical model of joint production between a service provider and a customer and characterize how a service firm should design its coproductive system. We show that, as a task becomes more standard, it is desirable to decrease the degree of interaction between the provider and the customer by making their efforts more substitutable and to allocate most of the work to whoever is the most efficient. Conversely, as a task becomes less standard, it is optimal to increase interaction by making efforts more complementary and to balance the work allocation. Our analysis gives rise to a service-process framework with three archetypes of coproductive services: collaborative services, service factories, and self-services. We discuss the implications of our results for service process reengineering.

Keywords: operations strategy; service operations; joint production; monotone comparative statics
History: Received: June 17, 2013; accepted: May 31, 2014. Published online in *Articles in Advance* August 18, 2014.

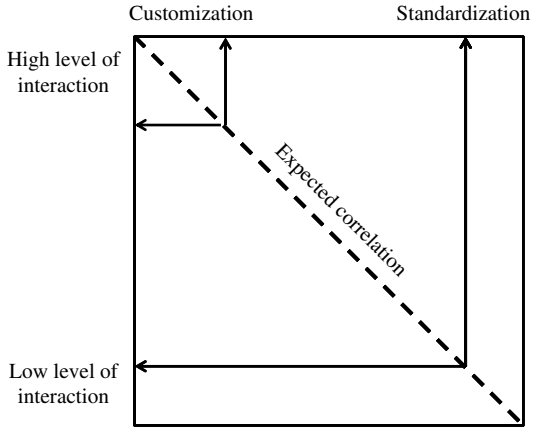
1. Introduction

In services, customers provide significant inputs into the production process (Fuchs 1968, Vargo and Lusch 2004, Sampson and Froehle 2006, Spohrer and Maglio 2010). In particular in *coproductive* services, these inputs are the customers themselves, who jointly participate with the provider in the service delivery. For instance, the value generated through a consulting engagement may depend on both the client sharing contextual knowledge and data and the consultant bringing methodologies and frameworks. Similarly, the quality of a tutoring session depends on both the tutor's effort at explaining a concept clearly and the student's effort at trying to assimilate it.

Many service firms are currently facing an unprecedented set of reengineering opportunities (Karmarkar 2004). The development of information technologies and the enhanced access to global markets have indeed enabled service firms to explore novel ways to involve their customers in their production process. Consider for instance the impact of information technologies on higher education. Should a university embrace information technology to "flip the classroom" and use class time to truly engage with its students (*The Economist* 2011)? Should a university specialize in publishing massive open online courses (MOOCs), perhaps with only distant interaction with its students? Should a university act as a content platform, offering students access to third-party generated course modules, talks, and internships? Each of these three delivery models is based on a different value proposition, which may need to be fully embraced to yield results. As Sally Blount, dean of the Kellogg School of Management, notes, "Our industry is about to transform itself. And you have to decide whether you are in or out of face-to-face education" (*The Economist* 2014). As this quote illustrates, service firms must often make strategic decisions regarding the design of their coproductive processes, especially regarding their degree of interaction and work allocation.

Despite the need for frameworks in these turbulent times, there are few guidelines for the design of coproductive service processes. Most existing frameworks are, moreover, conceptual, which makes them open to subjective interpretation. Consider for instance the service-process matrix proposed by Teboul (2006), depicted in Figure 1. This matrix suggests that service firms should match the degree of interaction of their service delivery process to the degree of standardization of their products. Although this may seem to be a natural match, not every service fits that prescription. For instance, salad bars, discount brokers (e.g., Charles Schwab), and online retailers (e.g., Amazon.com) may offer highly customized services with low levels of interaction (Teboul 2006, pp. 43–47). One may thus question the prescriptive value of the matrix (or at least of that particular interpretation) if it allows for exceptions, which moreover seem to be increasingly common

Figure 1 A Product-Process Matrix for Services



Source. Reproduced from Teboul (2006) with permission of Palgrave Macmillan.

with the development of information technologies. Our goal here is to provide analytical foundations for such conceptual frameworks, refine them, and expand them, in the same way Milgrom and Roberts (1990) and de Groote (1994) have provided a theoretical base for the celebrated product-process matrix for manufacturing activities (Hayes and Wheelwright 1979).

In this paper, we develop an analytical model of joint-production between a customer and a service provider and characterize how a service firm should design its joint-production system for a particular task as a function of its degree of standardization. We model the value from a particular service encounter as a function of the joint effort between the customer and the provider. Within this framework, a task is *standard* if it can be consistently and efficiently completed across all encounters, i.e., if it is associated with predictably high marginal returns to effort.

Focusing on service execution occurring after an initial diagnostic stage, we assume that efforts are discretionary, i.e., can be adapted to the type of encounter (Hopp et al. 2007), and that the parties are symmetric, notwithstanding differences in costs of effort. We model the joint effort as a generalized mean of the provider's and the customer's efforts (Hardy et al. 1952), using the constant elasticity of substitution (CES) function (Arrow et al. 1961). A particular design of a coproductive system will, accordingly, be characterized by its degree of *effort complementarity*, measuring interaction, and its *work allocation* between the provider and the customer.

We find that, as a task becomes more standard, efforts should be more substitutable and the work allocation should be more unbalanced. Conversely, as a task becomes less standard, efforts should be more complementary and the work allocation should be more balanced. These monotone comparative statics give rise to a product-process matrix for coproductive services with three extreme operating modes: *collaborative services*, involving direct interaction between the customer and the provider; *service factories*, involving independent processing by the provider; and *self-services*, involving independent processing by the customer. We discuss the implications for service process reengineering and assess its sensitivity to cost reductions.

This paper is organized as follows. We first review the related literature in §2. We introduce the model in §3 and characterize its comparative statics in §4. These results lead to the development of a product-process matrix for coproductive services in §5. In §6, we study how the framework is affected by endogenous choices of a task's degree of standardization and by changes in cost of effort. We conclude in §7. We present additional sensitivity results in Appendices A and B. All proofs are gathered in an electronic companion (available as supplemental material at http://dx.doi.org/10.1287/msom.2014.0495).

2. Literature Review

We review first the literature on the concept of customer contact in services, then different analytical models of joint production, and finally some analytical frameworks for strategic fit.

2.1. Customer Contact and Product-Process Frameworks

Chase (1978; 1981, p. 700) argues that a service system's potential operating efficiency is a function of "the degree to which the customer is in direct contact with the service facility relative to the total service creation time for the customer." Accordingly, he suggests, high-contact service activities must be managed differently than, and decoupled from, low-contact activities.

Since Chase's early papers, the notion of customer contact has evolved, moving away from physical contact to capture instead the extent of interaction and customization (Schmenner 1986), measured as person-to-person contact (Karmarkar and Pitbladdo 1995), active contact (Mersha 1990), or direct contact with interaction (Wemmerlöv 1990). In fact, Kellogg and Chase (1995, p. 1736) themselves suggest that "it is not merely time in the system that is being considered [in measuring customer contact] but also the ability to react and customize the service offering."

Despite its evolving interpretation, the concept of customer contact lies at the core of many service classifications (see Silvestro et al. 1992 and Cook et al. 1999 for reviews). In particular, Kellogg and Nie (1995) and Teboul (2006) suggest that service firms should match the level of customer interaction (or customer contact) to the scope of their product offerings; see Figure 1. In a similar vein, Schmenner (1986, p. 23) argues that "for many services, customization and interaction go hand in hand." Moreover, these prescriptions have been (partially) empirically validated. In particular, Huete and Roth (1988) find some, albeit weak, evidence

in the banking industry of a relationship between the industrialization level of a service delivery channel (process dimension) and the service complexity of the product offered through that channel (product dimension). Similarly, Lewis and Brown (2012) empirically validate the existence of an "efficient" diagonal in the legal industry after asking partners of a legal practice to rate their service offering, on a 1–10 scale, in terms of relative throughput time (process dimension) and degree of variation (product dimension). However, many of these conceptual frameworks lack analytical foundations and may therefore be open to subjective interpretations. For instance, Sampson and Froehle (2006) argue that Schmenner's (1986) joint measure of customer interaction and customization is difficult to operationalize because customization may occur independently of interaction. Similarly, Wemmerlöv (1990) and Silvestro et al. (1992) argue that Shostack's (1987) concept of process complexity is ambiguous. Kellogg and Nie (1995) propose a product-process-type service classification based on the degree of customization and the level of customer influence, but they broadly define the latter dimension as encompassing the notions of customer contact, interaction, participation, and impact on the service process.

We contribute to the literature on service operations by offering an analytical model of joint production and by formalizing the notion of customer contact, or interaction, as degree of effort complementarity. Our mathematical framework is aimed at supporting and refining the aforementioned conceptual studies with greater formalism and expanding the service process matrix depicted in Figure 1 along a third dimension, capturing the degree of work allocation.

2.2. Analytical Models of Joint Production

Despite an early emphasis on the customer's active role in the production of services (Fuchs 1968), there appears to be little analytical work on joint production in services before Karmarkar and Pitbladdo (1995). Building on their work, Xue and Field (2008) model a coproduction environment with information stickiness, Roels et al. (2010) study the optimal choice of contracts with coproduction, and White and Badinelli (2012) propose a model of workforce planning with customer involvement. We extend the ideas of Karmarkar and Pitbladdo (1995) and Roels et al. (2010), who, respectively, considered a linear and a Cobb–Douglas production functions, by considering a CES production function (Arrow et al. 1961).

Although the CES function is commonly used in economics (e.g., Bhattacharyya and Lafontaine 1995, Adams 2006), our focus is on effort allocation between two parties in a transaction, and not on contracting. Similarly, Bellos and Kavadias (2014) study the optimal allocation of control in a service process with multiple touchpoints, trading off efficiency against experiential variability. In their model, each touchpoint involves substitutable efforts; moreover, control is a binary decision, so the work is fully allocated to either the provider or the customer. Focusing on a particular task (or touchpoint), we complement their work by characterizing the shift of efforts from complements to substitutes as well as the gradual shift in work allocation between the provider and the customer.

Value cocreation is obviously not limited to services. Ramirez (1999, p. 54) indeed argues that "services is a framework to think of value creation;" see also Spohrer and Maglio (2010). In general, the notion of effort complementarity is related to the classical concepts of interaction (Marschak and Radner 1972) and interdependence (Thompson 1967); see Chase and Tansik (1983) for similar analogies. Although our structural results could, in principle, apply to any kind of joint production, as arises in white-collar processes (Marschak and Radner 1972, Hopp et al. 2009), including new product development processes (Eppinger et al. 1994, Schaefer 1999), or in public services (Whitaker 1980), their implications for design are the most relevant to service processes.

In operations management, value cocreation arises in manufacturing activities (Buzacott 2004, Gurvich and Van Mieghem 2014), as well as in joint product development (Bhaskaran and Krishnan 2009, Iyer et al. 2005), joint detection of defective products (Baiman et al. 2000), joint quality investment (Wang and Yin 2014), collaborative cost reduction in supply chains (Corbett and DeCroix 2001, Corbett et al. 2005, Kim and Netessine 2013), and repair services (Jain et al. 2013). Our abstract representation of joint production steps back from those context-specific studies, which typically consider efforts to be either substitutes or complements, by studying the shift from one to the other in the design of the coproductive system.

2.3. Complementarities and Strategic Fit

Most service classification frameworks (e.g., Silvestro et al. 1992, Kellogg and Nie 1995, Teboul 2006) propose to match service configurations to the environment characteristics in the spirit of contingency theory (Donaldson 2001), similar to the seminal product-process matrix for manufacturing activities by Hayes and Wheelwright (1979). Building upon this idea, Chase (1978) and Chase and Tansik (1983) propose to match, respectively, operational tactics and organizational structure to the service environment. Most of these classification schemes, however, are conceptual.

Milgrom and Roberts (1995) argue that the tools of supermodularity (Topkis 1998) are well suited to answer strategic questions due to the importance of alignment and complementarity in strategy. Moreover, models based on supermodularity do not require differentiability or convexity assumptions, which would

not be realistic at the level of strategic design decisions. Furthermore, they are robust to uncertainty, decentralized decision making, and incomplete model specification (Milgrom and Roberts 1995). Similar to Milgrom and Roberts (1990) and de Groote (1994), who use supermodularity to build models of strategic fit for manufacturing firms, we use supermodularity to characterize the strategic complementarities that arise in service design.

3. Model

We consider a coproductive service in which, in each service encounter, value is cocreated by a service provider (e.g., a front-line employee) and a patronizing customer. Since service processes may consist of multiple tasks, each involving different degrees of joint production, we focus our analysis on a particular service task, similar to Wemmerlöv (1990) and Lewis and Brown (2012).

We consider a situation in which efforts are not only *joint*, due to the coproductive nature of the service, but also *discretionary*, i.e., depend on the characteristics of the encounter (Hopp et al. 2007, Anand et al. 2011, Tong and Rajagopalan 2014). Similar to this literature, we focus on the *service execution* step occurring after an initial diagnostic step. According to Bitran and Lojo (1993), this step is typically the longest of the encounter, with closer interactions between customers and providers, and has a strong impact on customer satisfaction. In addition, we assume that the provider and the customer, although they may have different costs of effort, are otherwise *symmetric*, and leave for future research the study of asymmetries. As a result of these four conditions, how much effort is spent and who spends effort in an encounter will depend on

- the relative costs of effort;
- the design of the coproductive system, in particular the degree of effort substitutability and the allocation of work; and
- the type of encounter.

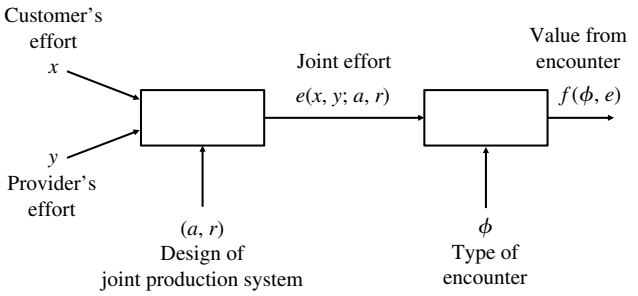
In this setting, two decisions must be made, at different time epochs.

1. At the design stage, the service firm chooses how to configure its coproductive system, specifically its degree of effort substitutability (r) and work allocation (a), which then applies to all service encounters.

2. In each encounter, the provider and a patronizing customer choose how much effort to exert so as to maximize the total value from the encounter. This value is a function of the type of encounter (ϕ) and of the joint effort (e), which is itself a function of individual efforts (x and y) and of the design of the system (a and r). The model is depicted in Figure 2.

Our model of joint production is deliberately stylized to encompass the breadth of service industries. In particular, we do not distinguish business-to-consumer from business-to-business services and generically adopt a provider–customer terminology. Given its level of generality, our model may also apply to other areas of joint production, such as team work in white-collar processes, provided that the coproductive system needs to be *designed*, which is typically the case in services, when a process needs to be established to handle many encounters.

Figure 2 A Model of Joint Production



We first introduce the three components of the model, i.e., the individual efforts, their aggregation into joint effort, and the service value function. We then describe the two decisions, i.e., the choice of effort during each encounter and the coproductive system configuration at the design stage.

3.1. Model Components

3.1.1. Individual Efforts. We assume that, in each encounter between a customer and a provider, inputs from both parties, such as time (Anand et al. 2011) and information (Xue and Field 2008), cocreate value (Fuchs 1968, Vargo and Lusch 2004, Sampson and Froehle 2006, Spohrer and Maglio 2010). Let $x \geq 0$ and $y \geq 0$, respectively, denote the customer's and the provider's inputs.

Providing input is costly. As is common in the literature (e.g., Bhattacharyya and Lafontaine 1995, Xue and Field 2008), we consider convex costs, i.e., $c_x x^p$ and $c_y y^p$, with $p \geq 1$, $c_x, c_y > 0$.

We normalize inputs so that one unit of customer input is equivalent to one unit of provider input. Accordingly, if the customer and the provider have the same cost of time, but the provider is θ times more productive than the customer, then $c_y = c_x/\theta^p$. Hence, input costs account not only for costs of time, but also for differences in skills (through appropriate rescaling), as well as for congestion and fatigue (through the exponentiation). Accordingly, we will refer to inputs as *efforts*, consistent with the literature on joint production (Bhattacharyya and Lafontaine 1995).

We assume identical cost elasticities p so that, if one party has a cost advantage at a given level of effort, it does so at all levels of effort. Effectively, this assumption implies that differences in productivity between the provider and the customer are independent

of the type of job, i.e., of whether it requires high or low effort. In particular, this assumption holds when effort represents time, in which case costs are typically linear (Corbett et al. 2005). This assumption may not always hold in practice, however, and we leave the study of such asymmetries for future research.

3.1.2. Joint Effort. How the customer's and the provider's inputs are combined to cocreate value varies, depending on the context. In certain situations (e.g., investment planning), multiple resources may be needed to complete a task (Gurvich and Van Mieghem 2014). In that case, the total output is a function of the minimum of all efforts. In other situations (e.g., trading stocks), only one resource is needed to complete a task; although one person could do the job, two people could share the workload. In that latter case, the total output is a function of the weighted arithmetic mean of efforts, with weights proportional to the parties' respective value contributions.

More generally, we assume that the service value is a function of the (generalized) mean of x and y (Hardy et al. 1952), i.e., of

$$e(x, y; a, r) = (ax^r + (1-a)y^r)^{1/r},$$

consistent with the literature on joint production (e.g., Bhattacharyya and Lafontaine 1995, Adams 2006). This so-called CES function, which has been introduced for estimating country productivity (Arrow et al. 1961), is particularly amenable to econometric analysis. In particular, Reinhardt (1972) uses a specific form of that function (with $r = 0$) to estimate the productivity of physicians (and nurses, technicians, and office aides) and Xue et al. (2007) do the same to estimate customer efficiency for a retail bank. Specific forms of this function have been widely used in the operations management literature; see Table 1.

In this generalized mean model, the parameter a, $0 < a < 1$, measures the relative weight on the customer's effort in the value generated in the encounter. In particular, the value is a function of only the provider's effort when $a \to 0$ and only the customer's effort when $a \to 1$. As we shall establish in §4.2, the allocation of efforts directly follows from the value of a, and we shall therefore refer to a as *work allocation*.

The parameter r, $r \in \mathbb{R}$, is called the substitution parameter and gives $e(x, y; a, r)$ its versatility:

- When $r \to -\infty$, $e(x, y; a, r) \to \min\{x, y\}$, i.e., efforts are perfect complements.
- When $r \to 0$, $e(x, y; a, r) \to x^a y^{(1-a)}$, which is a Cobb–Douglas function.
- When $r = 1$, $e(x, y; a, r) = ax + (1-a)y$, i.e., efforts are perfect substitutes.
- When $r \to +\infty$, $e(x, y; a, r) \to \max\{x, y\}$, i.e., efforts are redundant.

It turns out that r is an increasing function of the elasticity of substitution between x and y. In particular, efforts are complements when r is low and substitutes otherwise. Because the concept of complementarity (Topkis 1998) is intimately related to the concept of interaction (Marschak and Radner 1972), one may interpret r as an (inverse) measure of *interaction*, i.e., the lower r, the greater the interaction. Since $e(x, y; a, r)$ is increasing in r (Hardy et al. 1952), this interpretation suggests that greater efficiency is achieved at lower levels of interaction, consistent with customer contact theory (Chase 1978, 1981).

3.1.3. Service Value. We assume that higher effort generates more value, but with diminishing returns. Specifically, we denote by $f(\phi, e)$ the value generated with e units of mean effort for an encounter of type ϕ. For notational convenience, we assume that $f(\phi, e)$ is twice-differentiable and denote partial derivatives as subscripts; hence, $f_e > 0$ and $f_{ee} \le 0$.

The type of encounter $\phi \in \Phi \subseteq \mathbb{R}$ depends on both the job content and the customer characteristics. The set Φ is ordered so that the marginal returns on effort decrease in ϕ, i.e., $f_{\phi e} \le 0$. Accordingly, ϕ may represent the difficulty of the task to be executed. For instance, more effort is needed to manage complex financial portfolios. Alternatively, ϕ may represent the degree of causal ambiguity in the task or the customer's lack of

Table 1 Examples of Uses of Specific Forms of the CES Function in the Operations Management Literature

Context	Reference	r	a	$f(e)$
Assembly process	Buzacott (2004)	$-\infty$	1/2	ke
Production with joint resources	Gurvich and Van Mieghem (2014)	$-\infty$	1/2	ke
Joint quality improvement	Wang and Yin (2014)	$-\infty$	1/2	$\max_q \mathbb{E}_D[\min\{e+D, q\}] - cq$
Banking services	Xue et al. (2007)	0		ke^b, $0 < b < 1$
Collaborative services	Roels et al. (2010)	0		ke^b, $0 < b < 1$
Repair services	Jain et al. (2013)	0	1/2	$k - (k+c)/(1+e^2)$, $e > 1$
Joint cost reduction	Kim and Netessine (2013)	0		$\max_q \mathbb{E}_D[\min\{D, q\}] - (c - ke)q$, $e \le 1$
Cell process	Buzacott (2004)	1	1/2	ke
Call center outsourcing	Aksin et al. (2008)	1	1/2	$\mathbb{E}_D[\min\{e, D\}]$
New product development	Bhaskaran and Krishnan (2009)	1	1/2	ke
Joint quality improvement	Wang and Yin (2014)	1	1/2	$\max_q \mathbb{E}_D[\min\{\log(e) + D, q\}] - cq$

Notes. $k, c > 0$ and $D \ge 0$. In all cases but Kim and Netessine (2013), $f(e)$ is concave increasing.

absorptive capacity since these features make information more "sticky" (Szulanski 1996), thereby leading to higher incremental effort for acquiring, transferring, and using information (von Hippel 1994). Note that ϕ affects the individual efforts only through the mean effort e, consistent with our symmetry assumption discussed in §3.1.1.

This setup generalizes several models of discretionary services. In particular, Tong and Rajagopalan (2014) consider a value function $f(\phi, e) = \sqrt{(e - e_0)/\phi}$ if $e \geq e_0$ and zero otherwise, in which e is the total time the provider spends interacting with a customer and e_0 represents the minimum time required to deliver the service, including the initial diagnostic. In their model, ϕ is the "effectiveness of spending an incremental amount of time in delivering value" and captures the degree of complexity or difficulty of a task as well as the degree of customer expectations about the service. For example, they suggest, if a boiler is dirtier and more complex and difficult to clean, it will take more time to achieve the same level of cleanliness or operating efficiency.

As another example, Hopp et al. (2007) consider $f(\phi, e) = 1 - \exp(-\phi e)$ when $e \geq e_0$ and zero otherwise. If $e_0 \geq 1/\phi$, then $f_{\phi e} \leq 0$ for all $e \geq e_0$. In their model, ϕ corresponds to the processing rate and is a characteristic of the task and the customer. For instance, they argue, in a call center with upselling, customers may have a similar revenue potential but, due to different personalities and time constraints, may require different amounts of conversation time to achieve this potential. Similarly, in an engineering group, warranty repair problems may have similar cost saving potential but different levels of complexity. Therefore, the problems may require different amounts of diagnostic time to achieve a certain level of design efficiency.

Similar to Hopp et al. (2007) and Tong and Rajagopalan (2014), different encounters are associated with different values of ϕ. In particular, some encounters are associated with high marginal returns on effort (i.e., are simple) while others have low marginal returns on effort (i.e., are difficult). We denote by $F(\phi; u)$ the distribution of ϕ on Φ, parameterized by $u \in U$, and we assume that $F(\phi; u)$ is stochastically increasing in u, i.e., $F(\phi; u_1) \geq F(\phi; u_2)$ for all ϕ if and only if $u_1 \leq u_2$. Hence, when u is small, most encounters are associated with a high marginal return on effort, and we will refer to such tasks as *standard*. By contrast when u is large, a significant fraction of encounters involves low marginal returns on effort, and we will refer to such tasks as *nonstandard*.

We initially assume that u is exogenous, similar to Hopp et al. (2007) and Tong and Rajagopalan (2014), in the same spirit as contingency theory (Donaldson 2001), and we later relax this assumption in §6 to study the interaction between service design and task standardization.

3.2. Decisions

3.2.1. Effort Exertion. Given a particular design of a coproductive system (a, r), efforts in particular encounter of type ϕ are chosen to maximize the total value $f(\phi, e)$ net of costs. In practice, several factors may prevent the parties from maximizing the total surplus including information asymmetries and moral hazard. Moreover, the whole notion of total surplus may be elusive since service outputs are often hard to measure (Bitran and Lojo 1993, Karmarkar and Pitbladdo 1995). Nevertheless, total surplus maximization constitutes a useful benchmark to characterize effort exertion, consistent with the literature on joint production (e.g., Bhattacharyya and Lafontaine 1995, Roels et al. 2010, White and Badinelli 2012) and on discretionary services (e.g., Tong and Rajagopalan 2014). Moreover, supermodularity is, in general, robust to decentralization of decision-making (Milgrom and Roberts 1995). Formally,

$$\max_{x,\, y \geq 0} f(\phi, e(x, y, a, r)) - c_x x^p - c_y y^p = v^*(\phi, a, r). \quad (1)$$

Let $(x^*(\phi, a, r), y^*(\phi, a, r))$ denote the optimal effort levels. We assume that $r < p$ to guarantee a unique interior optimal solution; see Lemma EC.1 in the electronic companion. In particular when $r \leq 1 < p$, the joint effort function $e(x, y, ; a, r)$ is concave in efforts (x, y) whereas the costs of effort are convex, which are common assumptions in the literature (see Table 1). Without that condition, efforts would be so substitutable that it may be optimal to set one of them to zero, i.e., joint production would degenerate into a single-resource production. Instead of these extreme cases, we are mostly interested here in the gradual changes in coproduction.

3.2.2. Service Design. At the design stage, the firm chooses the configuration of its coproductive system, namely its degree of effort substitutability r and allocation of work a, for all future service encounters. Let $A \times R$ be the set of feasible designs and $N(a)$ and $M(r)$ be the respective costs of design. For instance, $M(r)$ can include costs of coordination (Thompson 1967) and $N(a)$ may include costs of technology and of customer adoption (Campbell et al. 2011).

As argued by Milgrom and Roberts (1995), strategic decisions often involve nonconvexities. In particular, certain designs may be infeasible or more/less costly than other designs. Accordingly, we do not make any specific assumptions about the feasible sets or costs of design, besides assuming that A and R are compact and that $N(a)$ and $M(r)$ are continuous. The sets A and R may thus be nonconvex and the cost functions $N(a)$ and $M(r)$ may be multimodal.

Accordingly, in the design stage, the firm solves

$$\max_{a\in A,\, r\in R} V(a, r, u) \equiv \int_{\Phi} v^*(\phi, a, r)\, dF(\phi; u) - M(r) - N(a). \quad (2)$$

Let $(a^*(u), r^*(u))$ denote the optimal design parameters. We next study the comparative statics of this optimization problem.

4. Analysis

In this section, we first characterize how the optimal design of the coproductive system should be adapted to the degree of task standardization and then show that, at the encounter level, these design changes translate into changes in effort exertion.

4.1. Optimal Design of the Coproductive System

We first study the sensitivity of the optimal design variables $(a^*(u), r^*(u))$ to the degree of task standardization (u). We initially offer a pairwise characterization of the relationships between these three variables and then present a global characterization.

4.1.1. Pairwise Characterization. We characterize the cross derivatives of the optimal value function $v^*(\phi, a, r)$ through three lemmas. We focus here on presenting the technical results and interpret them in §5. First, we find that substitutable efforts are more attractive with higher marginal returns on effort.

Lemma 1. $v^*_{r\phi} \le 0$.

Second, giving more weight to the largest effort is more attractive with higher marginal returns on effort.

Lemma 2. $v^*_{a\phi} \le 0$ *if and only if* $x^*(\phi, a, r) \ge y^*(\phi, a, r)$.

In contrast to these first two lemmas, which hold under full generality, the relationship between a and r may not be monotone in general, unless some restrictions are imposed on their feasible sets and on the effort marginal returns. These assumptions are only needed when the firm has the flexibility to alter both the allocation of work and the degree of effort substitutability, i.e., when A and R are not singletons; otherwise, the relationship between a and r is irrelevant.

We first assume that the marginal returns on joint effort e are bounded from below. In particular, this assumption is satisfied when $f(\phi, e) = g(\phi)e^{b(\phi)}$, for any $g(\phi) > 0$ and $0 < b(\phi) \le 1$, as in Bhaskaran and Krishnan (2009) and Roels et al. (2010).

Assumption 1. $f_e + f_{ee}e \ge 0$.

The second assumption restricts the feasible sets of a and r to avoid extreme work allocations, such as $a = 0$ or $a = 1$, when r is high.

Assumption 2. *For any* $\bar{r}$ *such that* $r \in R \cap (-\infty, \bar{r}]$, a *must belong to* $A \cap [\underline{a}(\bar{r}), \bar{a}(\bar{r})]$, *in which*

$$\underline{a}(\bar{r}) \equiv \min\left\{a \ge \max\left\{\frac{c_x}{c_x + \bar{\xi}(\bar{r})c_y}, h(a, \bar{r})\right\}\right\} \quad \textit{and}$$

$$\bar{a}(\bar{r}) \equiv \max\left\{a \le \min\left\{\frac{c_x}{c_x + c_y\underline{\xi}(\bar{r})}, h(a, \bar{r})\right\}\right\},$$

where

$$h(a, r) \equiv \frac{-(p/(r-p))\ln(\xi(a, r))\xi(a, r) - \xi(a, r) + 1}{(\xi(a, r) - 1)^2}$$

$$\textit{with } \xi(a, r) \equiv \left(\frac{c_x(1-a)}{c_y a}\right)^{r/(r-p)} \quad (3)$$

and $\underline{\xi}(\bar{r}) = 1/150$ *if* $\bar{r} < 0$ *and zero otherwise, and* $\bar{\xi}(\bar{r}) = 150$ *if* $\bar{r} < 0$ *and infinity otherwise.*

For any $\bar{r}$, Assumption 2 restricts r to belong to $R \cap (-\infty, \bar{r}]$ and a to belong to $A \cap [\underline{a}(\bar{r}), \bar{a}(\bar{r})]$, for some functions $\underline{a}(r)$ and $\bar{a}(r)$. It turns out that the size of the interval $[\underline{a}(\bar{r}), \bar{a}(\bar{r})]$ is decreasing in $\bar{r}$, whereas that of $(-\infty, \bar{r}]$ is increasing in $\bar{r}$. Hence, there is a trade-off between a and r: The less constrained r, the more constrained a needs to be to satisfy Assumption 2, and vice versa.

In Appendix B, we show that Assumption 2 mildly restricts the feasible set of A unless costs are very asymmetric or efforts are allowed to be very substitutable (i.e., $\bar{r}/p$ tends to 1). For instance when $c_x = c_y$ and $r \le 0.1p$, Assumption 2 holds as long as $a \in [0.1, 0.9]$, i.e., as long as the total value does not depend on more than 90% of a particular input.

When costs are very asymmetric or when efforts are allowed to be very substitutable, the restrictions put on a are more severe. Nevertheless, these restrictions are without dramatic consequences on the feasible range of efforts, which are ultimately the variables of interest. For instance, when $c_x/c_y = 1/10$ and $\bar{r}/p = 0.9$, Assumption 2 holds when $a \in [0.06, 0.16]$, i.e., when the total value depends on at least 6% but not less than 16% of the customer's input. Although this range of admissible work allocation is very narrow, it still allows the customer to exert as little as 60 times less effort or as much as 693 times more effort than the provider, as shown in Appendix B. Hence, efforts can still vary significantly despite a narrow range of admissible work allocation. This is because efforts are very sensitive to the allocation of work (a) when they are highly substitutable (i.e., when $\bar{r}/p \approx 1$), which is precisely when the range of admissible work allocation is the most restricted. In sum, the restrictions on the range of admissible work allocation under Assumption 2 are either mild (when $\bar{r} \ll p$ and $c_x \approx c_y$) or without dramatic consequences on effort exertion.

Under these two assumptions, the next lemma shows that greater effort substitutability is associated with more unbalanced work allocations.

Lemma 3. *Under Assumptions* 1 *and* 2, $v^*_{ra} \geq 0$ *if and only if* $x^*(\phi, a, r) \geq y^*(\phi, a, r)$.

4.1.2. Global Characterization. Putting these three lemmas together demonstrates a complementary relationship between (i) the extent to which a task is nonstandard (u), (ii) the degree of effort complementarity ($-r$), and (iii) the contribution of the smallest effort to the total value (a or $-a$ depending on whether or not $x^*(\phi, a, r) \leq y^*(\phi, a, r)$):

Proposition 1. *Under Assumption* 1, *for any* $\bar{r}$,

- $V(-a, -r, u)$ *is supermodular on* $A \cap [c_x/(c_x + c_y), \bar{a}(\bar{r})] \times R \cap (-\infty, \bar{r}] \times U$;
- $V(a, -r, u)$ *is supermodular on* $A \cap [\underline{a}(\bar{r}), c_x/(c_x + c_y)] \times R \cap (-\infty, \bar{r}] \times U$.

In particular, Proposition 1 shows that u and r are strategic substitutes, irrespective of whether or not a is large. Specifically, it shows that greater effort substitutability becomes more desirable when a task exhibits a higher frequency of high marginal returns on effort, and vice versa; see the left panel of Figure 3.

In addition, Proposition 1 shows that a and u are strategic complements when a is small and strategic substitutes otherwise. Accordingly, more unbalanced work allocations occur when tasks are more standard; see the right panel of Figure 3.

Unlike the optimal effort substitutability ($r^*(u)$), which is always decreasing in u, the optimal work allocation ($a^*(u)$) may not evolve monotonically as u changes. These nonmonotone design changes may occur because there exist two relationship structures among the variables (a, r, u) depending on whether or not $a^*(u) \geq c_x/(c_x + c_y)$. If the work allocation a is constrained to be always greater or always smaller than $c_x/(c_x + c_y)$, it will evolve monotonically as a task becomes more standard, because only one of the two regimes stated in Proposition 1 would then apply. These trajectories correspond to the solid and dashed lines depicted in Figures 3 and 4. However, when a is not constrained to be greater or smaller than $c_x/(c_x + c_y)$, the optimal work allocation $a^*(u)$ may jump discontinuously from one regime to the other as u changes. See the dotted line on the right panel of Figure 4.

The intuition behind this discontinuous jump is a bottleneck story. When the task is nonstandard and efforts are complementary, the least efficient resource acts as a bottleneck. As a result, the total value heavily depends on its input. By contrast, when the task becomes more standard and efforts more substitutable, the least efficient resource no longer acts as a bottleneck since resources are less coupled, and the total value may then mostly rely on the most efficient resource.

Despite the possible discontinuous jump in the optimal work allocation, we note that it is always profitable, though not necessarily optimal, to (weakly) increase a when $a \geq c_x/(c_x + c_y)$ or to (weakly) decrease it otherwise. That is, the upper or lower paths (dashed and solid lines) depicted in the right panels of Figures 3 and 4 correspond to two local optima. A conservative (but still profitable) strategy could then consist in remaining on one of these two paths as task standardization changes, ignoring the option of flipping the work allocation.

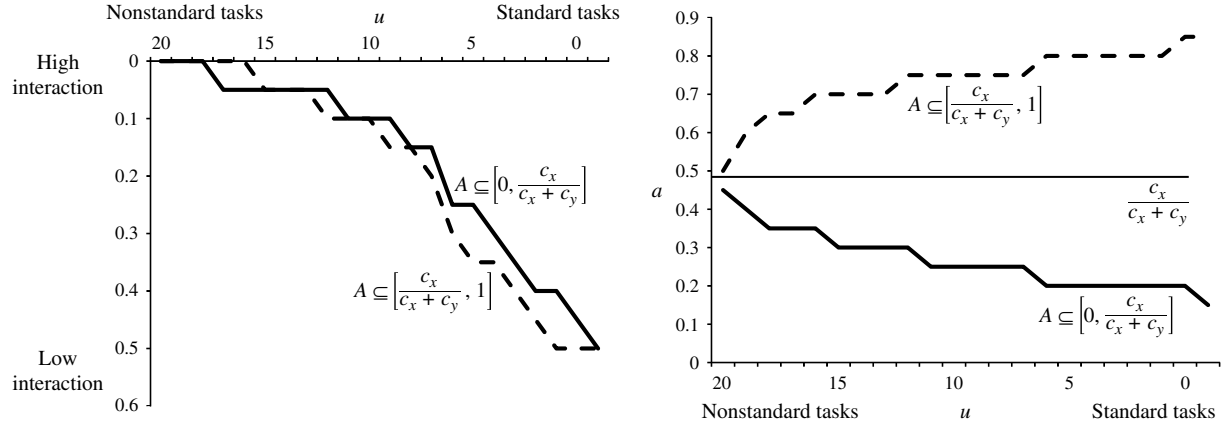
4.2. Optimal Effort Exertion

We next show that the adaptation of the design of the coproductive system to changes in task standardization translates at the encounter stage into changes in effort exertion. Although efforts $(x^*(\phi, a, r), y^*(\phi, a, r))$ may not evolve monotonically with a and r, as shown in Appendix A, the optimal *customer intensity* (White and Badinelli 2012), defined as

$$z^*(\phi, a, r) \equiv \frac{x^*(\phi, a, r)}{y^*(\phi, a, r)},$$

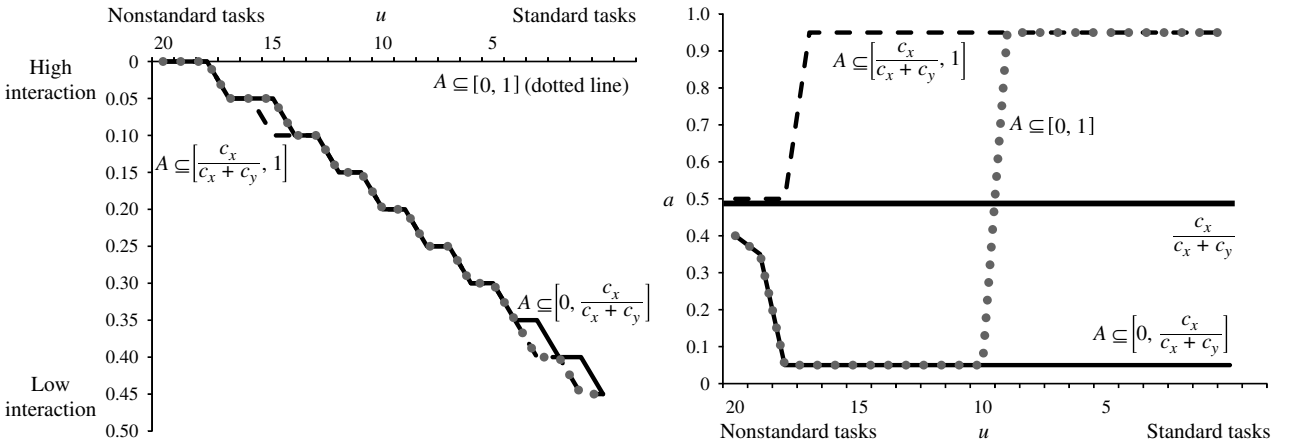
is monotone in the design parameters (a, r), as we show next.

Figure 3 When Tasks Become More Standard, Efforts Become More Substitutable (Left) and the Work Allocation Becomes More Unbalanced (Right)



Note. $F(\phi; u) = 1$ if $\phi \leq u$ and zero otherwise, $f(\phi, e) = (20 - \phi)\sqrt{e}$, $p = 2$, $c_x = 1$, $c_y = 1.05$, $M(r) = r^2$, $N(a) = 3^6(a - 0.5)^6$, $R = \{-0.25, -0.2, \ldots, 0.5\}$, $A = \{0.05, 0.1, \ldots, 0.45\}$ (solid) $A = \{0.5, 0.55, \ldots, 0.95\}$ (dashed).

Figure 4 **When the Work Allocation Is Flexible, the Optimal Work Allocation May Not Evolve Monotonically as a Task Becomes More Standard (Right), Whereas the Degree of Effort Substitutability Is Monotone (Left)**



Note. $F(\phi; u) = 1$ if $\phi \le u$ and zero otherwise, $f(\phi, e) = (18 - 0.9\phi)\sqrt{e}$, $p = 2$, $c_x = 1$, $c_y = 1.05$, $M(r) = r^2$, $N(a) = (a - 0.4)^2$, $R = \{-0.25, -0.2, \ldots, 0.5\}$, $A = \{0.05, 0.1, \ldots, 0.45\}$ (solid), $A = \{0.5, 0.55, \ldots, 0.95\}$ (dashed), $A = \{0.05, 0.1, \ldots, 0.45, 0.5, 0.55, \ldots, 0.95\}$ (dotted line).

Lemma 4. *The optimal customer intensity $z^*(\phi, a, r)$ is independent of ϕ; increasing in a; and increasing in r if and only if $z^*(\phi, a, r) \ge 1$.*

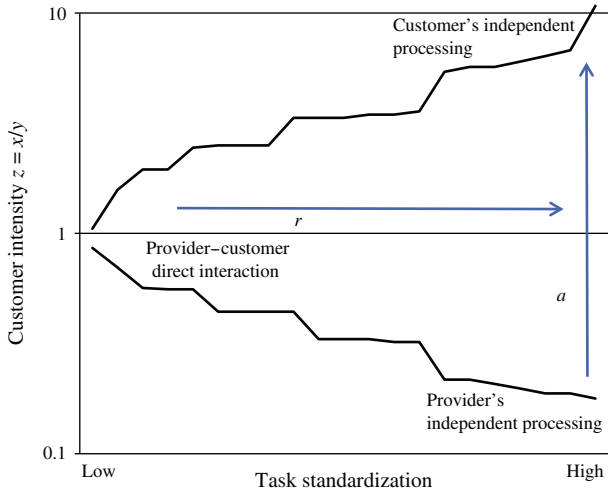
In particular, as a shifts from 0 to 1, the customer should exert proportionally more effort, which validates the interpretation of a as work allocation. Moreover, because z^* increases in r if and only if $z^* \ge 1$, efforts are the most unbalanced ($z \ll 1$ or $z \gg 1$) when they are substitutes (large r), and the most balanced when they are complements. Finally, Lemma 4 shows that, even though efforts may need to be adjusted to the level of task difficulty ϕ, they are adjusted proportionally. This is because, in our model, task difficulty affects both efforts x and y jointly, through the joint effort function $e(x, y; a, r)$, and not individually, and also because efforts have the same cost elasticity p. Without those assumptions, who exerts the highest effort may depend on the type of task. Because $z^*(\phi, a, r)$ is independent of ϕ, we will henceforth denote the customer intensity as $z^*(a, r)$.

The next proposition characterizes the sensitivity of the optimal customer intensity $z^*(a^*(u), r^*(u))$ to task standardization. Because the optimal design is monotone in the degree of task standardization (conditional on being in one out of two regimes) by Proposition 1, and because the customer intensity is monotone in the design variables (in that regime) by Lemma 4, the customer intensity is monotone in the degree of task standardization (in that regime).

Proposition 2. *The optimal customer intensity $z^*(a^*(u), r^*(u))$ decreases in u if and only if $a^*(u) \ge c_x/(c_x + c_y)$.*

Therefore, as a task becomes more standard (i.e., as u decreases), efforts become more substitutable (r increases), more weight is associated with the highest effort, and that effort further increases, at least proportionally. In particular, if $a^*(u) \ge c_x/(c_x + c_y)$, i.e., if $z^* \ge 1$, then a increases, ultimately leading to an increase in customer intensity z^*. Conversely when $a^*(u) \le c_x/(c_x + c_y)$, then a decreases, ultimately leading to a decrease in z^*. Figure 5 shows the evolution of the customer intensity as a result of the adaptation of the service design to changes in task standardization.

Figure 5 **(Color online) With More Standard Tasks, Efforts Become More Substitutable, the Largest Effort Is Weighed More Heavily, and Efforts Become More Unbalanced**



Note. Same example as Figure 3.

5. Discussion

Based on the comparative statics derived in §4, we next derive a product-process matrix for coproductive services. We then illustrate the framework with an example and finally discuss its managerial implications.

5.1. A Product-Process Matrix for Coproductive Services

By demonstrating a complementary relationship between task standardization ($-u$) and effort substitutability (r), Proposition 1 effectively formalizes the service-process matrix shown in Figure 1. In particular, the left panels of Figures 3 and 4, which illustrate the relationship between r and u, mirror the product-process matrix displayed in Figure 1. Our analytical framework thus offers a precise meaning behind the concepts of interaction (namely, effort complementarity) and standardization (namely, high frequency of tasks with high marginal returns to effort).

Beyond semantics, precision of constructs also helps operationalize the matrix. For instance, where would salad bars, discount brokers, and online retailers fit in the matrix of Figure 1? Although Teboul (2006) places them in the lower left corner, arguing that they offer high customization at low degrees of interaction, this seems to contradict the notion of "efficient diagonal" since they can be viable business models. Based on the constructs of effort complementarity and standardization, our model suggests instead that they should be located in the bottom right corner of the matrix, on the diagonal itself. Consider, for instance, the activity of trading stocks. On one hand, trading involves substitutable efforts since the more the provider trades stocks, the less the customer would need to trade them, and vice versa. On the other hand, the work that is involved in trading is relatively constant, irrespective of what stocks are in the portfolio, indicating that the task is mostly standard. Hence, our construct u suggests that the dimension of the horizontal axis of Figure 1 should be less related to *product* standardization, but more to *process* standardization.

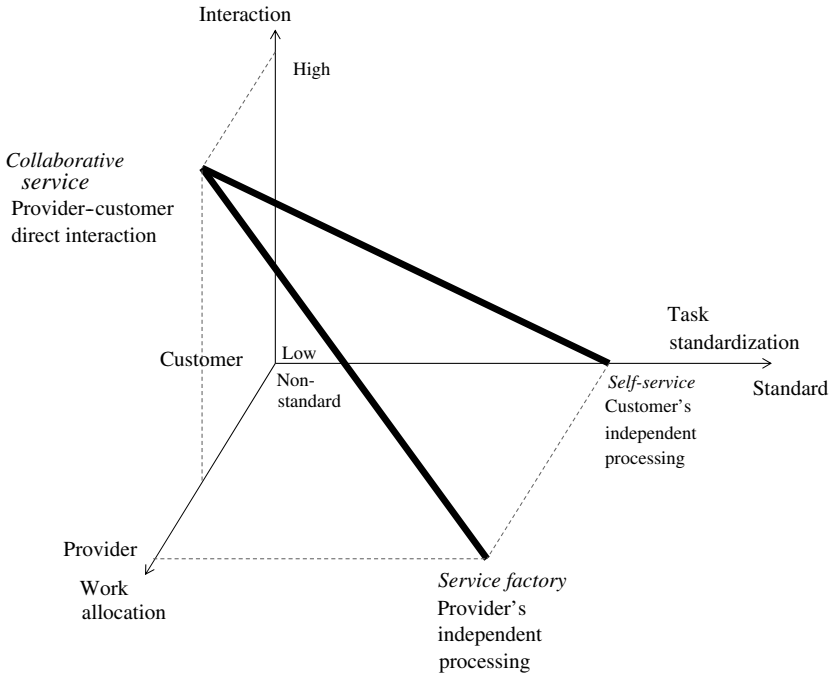
In addition, Proposition 1 shows that the matrix depicted in Figure 1 can be expanded along a third dimension, namely the allocation or work (a). Figure 1 indeed lumps together tasks that are performed independently by customers (e.g., a salad bar) with tasks that are performed independently by the provider (e.g., a backroom kitchen), despite their inherent differences.

Figure 6 depicts the corresponding product-process matrix for coproductive services, refining and expanding the matrix depicted in Figure 1. Unlike other product-process matrices (Hayes and Wheelwright 1979, Kellogg and Nie 1995, Teboul 2006), which map only one type of process to each degree of environmental uncertainty, there are two diagonals, depending on the work allocation, due to the presence of two, and not one, productive resources in coproductive systems.

The complementarity relationship among the three variables $(a, -r, u)$ or $(-a, -r, u)$ implies that more of one of those variables creates a stronger need to increase (or decrease) the other two in a predictable way. For instance, greater effort complementarity leads to more balanced work allocation, which itself leads to greater effort complementarity, and so on. As a result, as a task becomes more standard, the configuration of the coproductive system will converge to one out of the following three service archetypes (inspired by Sampson's (2012) classification and Schmenner's (1986) terminology):

- *service factory*, involving independent processing by the provider (low u, high r, low a);

Figure 6 A Product-Process Matrix for Coproductive Services



- *self-service*, involving independent processing by the customer (low u, high r, high a); and
- *collaborative service*, involving direct interaction between the provider and the customer (high u, low r, intermediate a).

By Proposition 2, the three different design archetypes obtained from Figure 6 are associated with three different modes of operation at the encounter level. In particular, the role of the customer changes as a function of the configuration of the coproductive system (Bitner et al. 1997): For nonstandard tasks, when efforts are complementary and work is evenly allocated, the customer effectively acts as a "partial employee" in the collaborative service configuration (Mills et al. 1983). On the other hand with standard tasks, when efforts are substitutable and concentrated on either the customer or the provider, customers may be viewed either as subcontractors in a self-service configuration (Lovelock and Young 1979) or as an "interference" in a service factory configuration (Frei 2006). Accordingly, service providers should adapt the way they manage their customers to the service system configuration.

The exact configuration of these archetypes is obviously context-specific, since the design feasible sets (A and R), design costs ($M(a)$ and $N(r)$), and effort cost parameters (c_x, c_y, and p) are context-dependent. In particular, operating as a "self-service" for a high-end restaurant may not be the same as for a fast-food restaurant due to differences in market perceptions, competition, and technologies. Nevertheless, as we discuss next, these three archetypes are useful benchmarks for firms belonging to similar market segments.

5.2. Examples

As an illustration, consider the process of preparing a personal tax return. Whereas some clients have simple tax returns (as indicated by the EZ suffix), others have complicated tax returns, including mortgages, home businesses, foreign bank accounts, etc. Depending on the composition of its pool of clients, a tax preparation service business may therefore have predictably high returns on effort, i.e., be associated with a low value of u, if most of its clients are of the EZ type, or a significant fraction of low returns on effort, i.e., be associated with a high value of u, otherwise.

Different coproductive models exist for preparing tax returns. One way is to have the client and a CPA personally meet, exchange information, discuss various tax return items, explore different tax saving opportunities, potentially beyond the current tax return (e.g., retirement, college loans). Under this format, efforts are complementary since the CPA's marginal returns on effort increase in the client's provision of information about her investments and expenses, and conversely, the client's marginal returns on effort at collecting the right information and making the right investments increase in the accountant's effort at probing the client with the right questions.

Alternatively, the client may prepare her tax return using a software program. The software prefills some entries and the client fills in the rest. Sometimes, software companies also offer opportunities for contacting experts to answer questions while preparing the tax return. With no interaction, efforts are substitutable since all entries prefilled by the software need not be filled by the client. With opportunities for interaction, efforts may, at times, be slightly complementary.

If the technology (i.e., $M(r)$ and R) allows so, Proposition 1 indicates that, consistent with Figure 1, it is more desirable to work with a CPA (low r) for complex tax returns (high u), to use a software tool with opportunities for interaction (medium r) for moderately complex tax returns (medium u), and to use a software tool with no interaction (high r) for simple tax returns (low u).

As the content of the tax return becomes simpler, Proposition 1 suggests that the work allocation should be more extreme. Whereas both the CPA and the client actively participate when jointly preparing a tax return, entering information into a software program is often done by either the client or the provider. The allocation of work may depend on the technology in place (i.e., $N(a)$ and A), as well as on the relative unit costs of effort (i.e., c_x and c_y), which include costs of time, adjusted for relative differences in productivity, and costs to access the required information.

More than being extreme, the work allocation involved in preparing a tax return may abruptly shift as the tax return becomes simpler, as discussed in §4.1.2. Whereas moderately simple tax returns are typically prepared by clients, using a tax software program, very simple tax returns can be prepared by the tax preparation service firm (or the government, as is currently the case in Denmark and Sweden); i.e., the service configuration may radically shift from a self-service configuration to a service factory as the frequency of simple tax returns increases. The reason behind this shift is as follows: Because taxpayers are information bottlenecks for moderately complex tax returns, the total output heavily depends on their input. With simpler tax returns, less information is needed, making efforts more substitutable. Consequently, the workload can be shifted toward the most efficient party, which may be the tax preparation firm.

As another example, in the financial industry, consider the portfolio of products offered by Merrill Lynch (Rangan and Bell 2008): (i) the full service involves brokerage services and personalized advice from a financial consultant on a wide range of financial products, such as savings, investments, insurance, and estate planning; (ii) the ML Direct offers online trading without the advice and guidance of a financial consultant; and (iii) the discretionary services assign

portfolio management responsibilities to a financial consultant with little client involvement. Based on our framework, the full service may fit better clients with complex needs and high asset value, whereas the last two products may fit better clients with simpler needs and fewer assets. Among these two, the ML Direct may be preferred by technology-savvy clients with limited investable assets. Interestingly, these different service configurations are associated with different pricing mechanisms (Rangan and Bell 2008), so the design decisions have also marketing implications.

As a third example, in education, a university could use information technology to position itself as a collaborative service by flipping the classroom, or as a service factory by publishing massive open online courses (MOOCs), or as a self-service by giving its students access to a portfolio of third-party generated course modules, talks, and internships. Our framework suggests that the first configuration will be preferred for teaching soft skills, whereas the last two are more suited to teaching hard knowledge. Given the economies of scale in publishing, we anticipate that there will be only a few players active in the MOOC publishing space, and that all other players would then adopt a self-service configuration for teaching hard knowledge.

5.3. Managerial Implications

Overall, Propositions 1 and 2 suggest the following service design principle, mirroring Chase's (1981) decoupling principle and de Groote's (1994) operations strategy principle:

Design Principle. *A shift toward more* (*less*) *standard tasks makes it more desirable to select a configuration that has more substitutable* (*complementary*) *efforts and more unbalanced* (*balanced*) *work allocation. Although it may be optimal to completely flip the work allocation, it is always desirable to allocate more* (*less*) *work to the party that exerts the highest effort as a task becomes more* (*less*) *standard.*

As this design principle reveals, the whole service configuration needs to be adapted to changes in task standardization. In fact, the lack of convexity of the value function $V(a, r, u)$ creates no room for hybrid configurations. In particular, it may not be sustainable to adopt a service configuration that mixes elements of the three coproductive service archetypes. This suggests that a firm aiming at competing on different market segments should adopt multiple "focused" service configurations to efficiently cater to those segments, similar to Skinner's (1974) concept of focused factory.

Practically speaking, this implies that service packages may need to be *unbundled* into specific modules catering to different market segments, as competition on those segments increases. For instance in banking, Merrill Lynch, when it faced increasing competition from Charles Schwab and E*Trade, offered its clients a continuum of products from the fully self-directed to the fully delegated, letting them self-select one or multiple accounts based on their needs and preferences (Rangan and Bell 2008). In consulting, McKinsey started offering consulting software packages in addition to its traditional relationship-based consulting, perhaps as a hedge against disruption from new entrants (Christensen et al. 2013).

In addition, the complementarity relationship between the design variables (a, r, u) reveals that limited value will be attained if one design variable is changed, but not the others. Hence, adapting the service design to environmental changes needs to be *comprehensive*. This suggests that, empirically, one should observe, within the same industry, clusters of service configurations around the three service archetypes.

A third implication, suggested by Milgrom and Roberts (1995), is that such design changes require *strong coordination* and may need to be initiated from the top. If the design decisions are decentralized, then no department responsible for a particular design variable would see the full benefit of adapting the service configuration to changes in the environment. This explains why, in practice, incumbent firms are often slow at embracing new technologies to reconfigure their service delivery model (Christensen et al. 2013).

6. Extensions

In this section, we consider how the optimal design and efforts are affected when the degree of task standardization is endogenously chosen or when the costs of effort can be reduced.

6.1. Endogenous Choice of Degree of Task Standardization

Suppose that, in the design stage, the firm can decrease u by focusing on a narrow market segment, offering a narrow set of services, or streamlining its delivery processes or to increase u by adopting the opposite strategies. Let U be the set of feasible values for u, assumed to be compact, and $P(u)$ be the costs associated with different degrees of standardization, assumed to be continuous. In that case, the firm's design problem (2) becomes

$$\max_{a\in A,\, r\in R,\, u\in U} V(a, r, u) \equiv \int_{\Phi} v^*(\phi, a, r)\, dF(\phi; u) - M(r) - N(a) - P(u).$$

According to Proposition 1, greater effort substitutability leads not only to more unbalanced work allocations, but also to greater task standardization, which then begets further effort substitutability and even more unbalanced work allocation. Conversely, greater effort complementarity leads not only to more balanced work allocation, but also to greater openness

to deal with nonstandard tasks, e.g., tasks that may be difficult, ambiguous, or associated with sticky information. Hence, our results are *amplified* when the degree of task standardization is endogenous. In particular, greater effort complementarity and more balanced work allocation make it more attractive to work on nonstandard tasks, and vice versa.

6.2. Cost Reduction

Reengineering opportunities are often enabled by technological advances leading to cost reductions. Because there exists in general no monotone relationship between efforts (x^* and y^*) and the design variables (a and r), as shown in Appendix A, assessing the consequences of a cost reduction is subtle and requires careful analysis. To see this, suppose that $f(\phi, e) = g(\phi)e^{b(\phi)}$, for $0 < b(\phi) < 1$, and consider the effect of a reduction in c_x. Without changing the service configuration, the customer intensity z should naturally increase. However, if the service is reconfigured, z may not necessarily increase, depending on whether or not $x \geq y$ and on the degree of effort complementarity.

If $x > y$, a decrease in c_x always leads to an increase in x. By Proposition A.1 in Appendix A, x monotonically increases in a and r when $x > y$, so the decrease in c_x leads to an increase in r and an increase in a, which together lead to a further increase in z by Lemma 4.

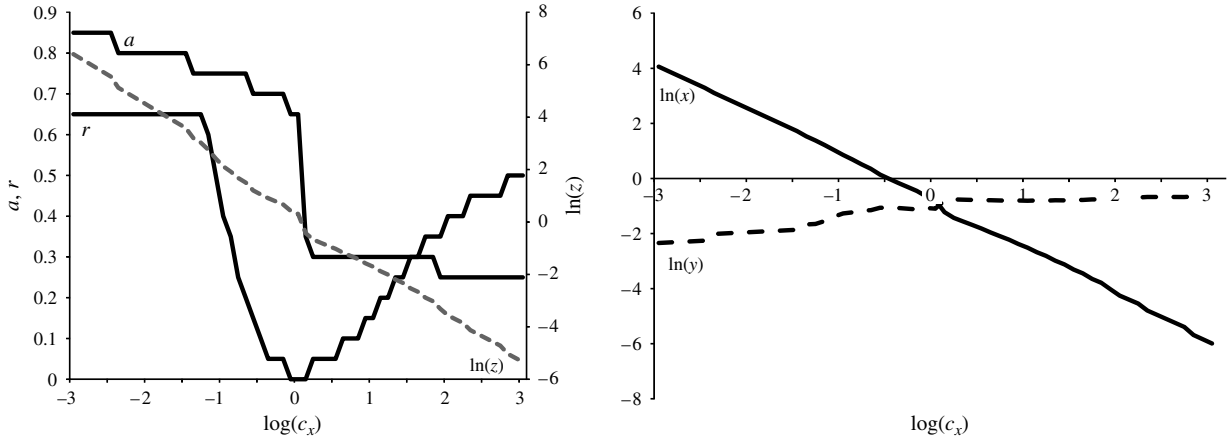
If $x < y$, an increase in x may be associated with a decrease in r and an increase in a (Proposition A.1), as it happens when efforts are substitutes. In that case, the decrease in c_x leads to an increase in z (Lemma 4), similar to the first case, but efforts become more complementary here. For instance, the emergence of online courses, which has reduced the pupils' costs of learning, has been reported to have transformed classroom time into tutoring (*The Economist* 2011).

These two cases are illustrated in Figure 7. When $x > y$, an increase in c_x makes efforts more complementary (r decreases) and allocates more work to the provider (a decreases); see Figure 7, left. As c_x keeps increasing, x keeps decreasing (Figure 7, right), ultimately leading to a structural shift in who exerts the highest effort, i.e., yielding $z < 1$, and then pushing for greater effort substitutability (i.e., higher r). Hence, there may be a complete reversal in design objectives, from more to less interaction, and in who exerts the highest effort as c_x keeps increasing. Moreover, even though the customer intensity (z) always decreases in c_x, the provider's effort (y) may not necessarily be increasing in c_x; for instance in Figure 7, y reaches a local minimum at $c_x = c_y$.

If $x < y$ and efforts are complements, an increase in x could potentially be associated with an increase in r and a decrease in a (Proposition A.1). In that case, a decrease in c_x would result in a *decrease*, and not an increase, in z (Lemma 4). That is, the customer may end up exerting proportionally *less* effort after her cost of effort decreases. See Figure 8 for an example. The intuition is as follows: By reducing the cost of the bottleneck resource (c_x in this case), the productivity of all resources increases (i.e., both x and y increase) due to effort complementarity. The system becomes less coupled (r increases) and the total value depends less on the bottleneck's input (a decreases). For instance, a consulting team's efforts may be impeded by a client's lack of preparation and lack of willingness to share data. When the client is more prepared, the consulting team's return on effort can improve (e.g., it spends less time trying to collect data or it uses better quality data), and roles and responsibilities can be more clearly defined to make better use of everyone's time.

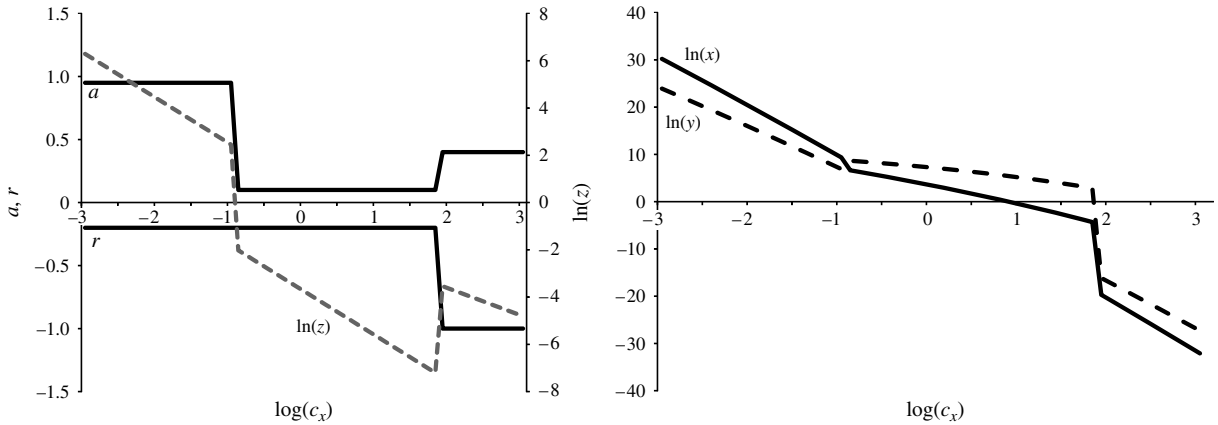
Given that cost reductions may have different outcomes depending on effort complementarity, properly

Figure 7 **When the Cost of Effort of One Party Increases, Less Work May Be Allocated to That Party and Efforts Initially Become More Complementary and May Then Become More Substitutable (Left); If That Is the Case, the Customer Intensity x/y Evolves Monotonically (Right)**



Note. $f(\phi, e) = 3\sqrt{e}$, $p = 2$, $c_y = 1$, $M(r) = r^2$, $N(a) = 3^6(a - 0.5)^6$, $R = \{-0.25, -0.2, \ldots, 0.65\}$, $A = \{0.05, 0.1, \ldots, 0.95\}$.

Figure 8 **With Complementary Efforts, the Customer Intensity x/y May Not Evolve Monotonically with the Cost of Effort of the Party Who Exerts the Least Effort**



Note. $f(\phi, e) = e^{0.8}$, $p = 1$, $c_y = 0.1$, $M(r) = 0.1(|r+1|)^{0.1}$, $N(a) = 3(a-0.4)^6$, $R = \{-1, -0.9, \ldots, -0.2\}$, $A = \{0.1, 0.15, \ldots, 0.95\}$.

mapping service configurations is paramount for assessing these outcomes before reengineering service processes.

7. Conclusions

In this paper, we characterize how the design of a coproductive system, in particular its degree of interaction and its work allocation, should be adapted to the degree of task standardization. We show that as a task becomes more standard, in the sense that its effort marginal returns become higher and more predictable, efforts should be more substitutable and the work allocation should be more unbalanced; as a result of these design changes, effort exertion at the encounter level is also more unbalanced. Conversely, when the effort marginal returns decrease or become less predictable, efforts should be more complementary, the allocation of work should be more balanced, and both parties end up putting in similar efforts. These monotone comparative statics give rise to three extreme configurations for coproductive systems: collaborative services, involving direct interaction between the customer and the provider, service factories, involving independent processing by the provider, and self-services, involving independent processing by the customer.

Our analytical model contributes to the literature on service-process frameworks in two different respects. First, we formalize the concepts of "standardization," measured here as the predictability and magnitude of returns on effort, and "customer contact" or "interaction," measured here as effort complementarity.

Second, we expand the existing service-process frameworks by considering work allocation as a third dimension, thereby leading to a novel product-process matrix for coproductive services; see Figure 6. Unlike other product-process matrices (Hayes and Wheelwright 1979, Kellogg and Nie 1995, Teboul 2006), which map only one type of process to each degree of environmental uncertainty, Figure 6 depicts two diagonals, depending on the work allocation.

Overall, our framework creates a mental map for reengineering service tasks, by questioning the alignment of effort complementarity and work allocation with a task's degree of standardization:

- To what extent is the current degree of effort complementarity aligned with the degree of standardization of the task?
- To what extent is the current work allocation aligned with the relative cost efficiencies of the customer and the provider?
- In case of misalignment, what changes need to be made to realign the degree of interaction and work allocation with the degree of task standardization and the relative costs of effort?

Although simple, these questions may lead to radical reengineering of coproductive systems. The complementary relationship between the design variables suggests that a hybrid strategy, compromising across the three service archetypes, is not sustainable in the long run, leading to an unbundling of the service package into focused modules. Moreover, design changes need to be comprehensive to pay off; that is, changing only the degree of interaction without changing the work allocation (or vice versa) will usually not pay off. Accordingly, design changes need to be strongly coordinated and should perhaps be initiated from the top.

Our model has several limitations, which are direct consequences of our modeling assumptions. Because it considers symmetric parties, it ignores information asymmetries and task-specific differences in expertise. Moreover, we have ignored moral hazard by focusing on the total surplus. Third, our model focuses on interactions between one customer and a provider, and

therefore ignores such externalities as customer interactions (e.g., queuing) and economies of scale. Fourth, our model focuses on service execution and ignores any coproduction in the diagnostic stage. Fifth, the model is defined at the task level and remains silent about possible task interdependencies (e.g., between back and front office). Finally, our model considers the concepts of standardization and interaction as unidimensional. In reality, other dimensions matter such as variability in processing times, employee empathy, responsiveness to unforseen events, and physical proximity, among others.

Nevertheless, our results may be a good starting point to think about service process reengineering. In particular, they justify the recent development of interaction-focused process flowcharting tools (Sampson 2012) and yield a novel classification scheme for coproductive services. Research on joint production is still in its infancy, and we hope that our model will provide a foundation for studying the management of operations of collaborative services, service factories, and self-services.

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/msom.2014.0495.

Acknowledgments

The author thanks Uday Karmarkar for insightful discussions on the topic, and Fred Ponsignon and Alexei Alexandrov for providing useful pointers and interpretations of the results. The author also thanks Steve Graves, the associate editor, and three anonymous referees for their insightful comments, which have significantly improved this paper, as well as the seminar participants at the Kellogg School of Management, Cambridge University, Georgetown University, the 2012 Marketing Science Conference, the 2012 Manufacturing and Service Operations Management Conference, and the 2012 Art and Science of Services Conference for providing useful feedback.

Appendix A. Optimal Efforts

In this appendix, we show that optimal efforts may not be monotone with respect to a and r.

Proposition A.1. *Suppose that* $f(\phi, e) = g(\phi)e^{b(\phi)}$, $0 < b(\phi) < 1$, *and* $r < 1$. *Then the following applies*:

(1) *Sensitivity with respect to* r:

(a) *When* $r \geq b(\phi)$, $x^*(r)$ *is nondecreasing and* $y^*(r)$ *is nonincreasing if and only if* $x^*(r) \geq y^*(r)$.

(b) *When* $r < b(\phi)$: (i) *if* $y^*(r) \geq x^*(r)$, $y^*(r)$ *is nondecreasing; and* (ii) *if* $y^*(r) \leq x^*(r)$, *there exists some* $\bar{r} \leq b(\phi)$ *such that* $y^*(r)$ *is nondecreasing if and only if* $r \leq \bar{r}$.

(2) *Sensitivity with respect to* a:

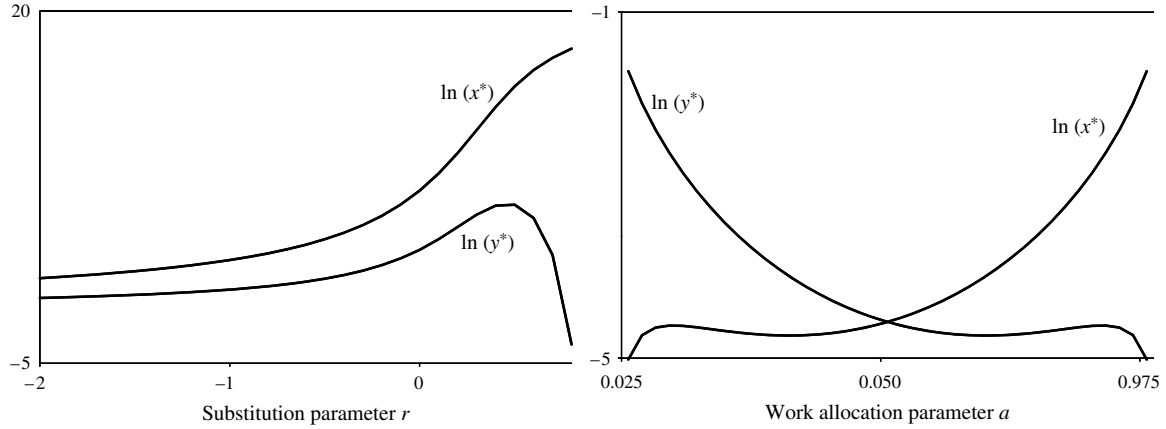
(a) *When* $r \geq b(\phi)$, $y^*(a)$ *is nonincreasing and* $x^*(a)$ *is nondecreasing*.

(b) *When* $r \leq b(\phi)$, $y^*(a)$ *is quasiconvex and then quasiconcave; it is in particular monotone decreasing when* $x^*(a) \leq y^*(a)$; *and* $x^*(a)$ *behaves in an opposite fashion*.

Figure A.1 illustrates the sensitivity result of Proposition A.1. Whereas the highest effort is monotonically increasing in r, the lowest effort is unimodal in r. The unimodal behavior is due to two effects, one direct and one indirect. To see how, suppose that $x^*(r) > y^*(r)$. The direct effect works as follows: as r increases, the marginal returns on x increase due to lower coupling with the bottleneck (y in this case). The marginal returns on y, on the other hand, increase only if $y \approx x$; while the bottleneck initially benefits from lower coupling, i.e., from fewer frictions, making efforts more substitutable ultimately leads to greater specialization. In addition to this direct effect comes an indirect effect: The increase in x^* following an increase in r increases the marginal returns on y when $r < b$ by complementarity, and decreases them when $r > b$ by substitutability. Taken together, these two effects make y^* initially increase and then decrease in r, and the transition happens when $r < b$.

Similarly, a change in the work allocation parameter (a) leads to monotone effects on the highest effort; e.g., an increase in a leads to an increase in $x^*(a)$ when $x^*(a) \geq y^*(a)$. In contrast, the lowest effort evolves nonmonotonically. To see why, suppose that $x^* > y^*$. On one hand, an increase in a reduces the marginal returns on y because more weight is allocated to x. On the other hand, an increase in a increases x^*, which then increases the marginal returns on y when efforts are complements ($r < b$).

Figure A.1 Sensitivity of the Optimal Efforts with Respect to the Substitution Parameter r (Left) and the Work Allocation Parameter a (Right)



Note. Left: $f(\phi, e) = e^{0.8}$, $a = 0.4$, $c_x = .01$, $c_y = p = 1$. Right: $f(\phi, e) = e^{0.8}$, $r = -0.1$, $c_x = c_y = p = 1$.

Table B.1 **Feasible Sets for the Work Allocation Parameter a, i.e., $[\underline{a}(\bar{r}), \bar{a}(\bar{r})]$**

	c_x/c_y				
$\bar{r}/p$	10	4	1	1/4	1/10
−10.0	[0.06, 1.00]	[0.03, 1.00]	[0.01, 0.99]	[0.00, 0.97]	[0.00, 0.94]
−5.0	[0.06, 1.00]	[0.03, 1.00]	[0.01, 0.99]	[0.00, 0.97]	[0.00, 0.94]
−1.0	[0.06, 1.00]	[0.03, 0.99]	[0.02, 0.98]	[0.01, 0.97]	[0.00, 0.94]
−0.1	[0.18, 0.99]	[0.14, 0.97]	[0.07, 0.93]	[0.03, 0.86]	[0.01, 0.82]
0.1	[0.26, 0.99]	[0.20, 0.97]	[0.10, 0.90]	[0.03, 0.80]	[0.01, 0.74]
0.3	[0.38, 0.98]	[0.27, 0.96]	[0.13, 0.87]	[0.04, 0.73]	[0.02, 0.62]
0.5	[0.52, 0.97]	[0.38, 0.94]	[0.18, 0.82]	[0.06, 0.62]	[0.03, 0.48]
0.7	[0.68, 0.96]	[0.52, 0.91]	[0.26, 0.74]	[0.09, 0.48]	[0.04, 0.32]
0.9	[0.84, 0.94]	[0.69, 0.86]	[0.38, 0.62]	[0.14, 0.31]	[0.06, 0.16]

Table B.2 **Feasible Values for the Customer Intensity $\log(z^*(\phi, a, r))$**

	c_x/c_y				
$\bar{r}/p$	10	4	1	1/4	1/10
−10.0	[−0.198, 0.198]	[−0.198, 0.198]	[−0.198, 0.198]	[−0.198, 0.198]	[−0.198, 0.198]
−5.0	[−0.363, 0.363]	[−0.363, 0.363]	[−0.363, 0.363]	[−0.363, 0.363]	[−0.363, 0.363]
−1.0	[−1.088, 0.710]	[−1.042, 0.735]	[−0.832, 0.832]	[−0.735, 1.042]	[−0.710, 1.088]
−0.1	[−1.502, 0.863]	[−1.261, 0.893]	[−1.009, 1.009]	[−0.893, 1.261]	[−0.863, 1.502]
0.1	[−1.605, 0.923]	[−1.348, 0.955]	[−1.079, 1.079]	[−0.955, 1.348]	[−0.923, 1.605]
0.3	[−1.744, 1.003]	[−1.465, 1.038]	[−1.173, 1.173]	[−1.038, 1.465]	[−1.003, 1.744]
0.5	[−1.937, 1.120]	[−1.630, 1.158]	[−1.308, 1.308]	[−1.158, 1.630]	[−1.120, 1.937]
0.7	[−2.233, 1.313]	[−1.893, 1.358]	[−1.529, 1.529]	[−1.358, 1.893]	[−1.313, 2.233]
0.9	[−2.841, 1.784]	[−2.466, 1.839]	[−2.045, 2.045]	[−1.839, 2.466]	[−1.784, 2.841]

Appendix B. Robustness of Assumption 2

In this appendix, we discuss the robustness of Assumption 2. Thresholds $\underline{a}(\bar{r})$ and $\bar{a}(\bar{r})$ can be uniquely characterized in terms of the ratios c_x/c_y and $\bar{r}/p$ (see Lemma EC.4 in the electronic companion). In particular, $\underline{a}(\bar{r})$ and $\bar{a}(\bar{r})$ are, respectively, increasing and decreasing in $\bar{r}/p$. Hence, Assumption 2 can be satisfied by either shrinking R (by lowering the upper bound $\bar{r}$) or by shrinking A (by shrinking the interval $[\underline{a}(\bar{r}), \bar{a}(\bar{r})]$). Moreover, both $\underline{a}(\bar{r})$ and $\bar{a}(\bar{r})$ are increasing in c_x/c_y. Therefore, large (small) values of a are allowed when c_x/c_y is large (small). Table B.1 depicts the sets $[\underline{a}(\bar{r}), \bar{a}(\bar{r})]$ for different values of c_x/c_y and of $\bar{r}/p$.

Overall, Table B.1 shows that the length of the interval $[\underline{a}(\bar{r}), \bar{a}(\bar{r})]$ is the smallest when costs c_x and c_y are asymmetric or when $\bar{r} \approx p$, but is relatively large otherwise. Hence, Assumption 2 is the most stringent when either costs c_x and c_y are asymmetric or when $\bar{r} \approx p$. However, as we argue next, this is without dramatic consequences on the range of feasible efforts, which is the ultimate variable of interest.

Define the customer intensity as $z^*(\phi, a, r) = x^*(\phi, a, r)/y^*(\phi, a, r)$. By Lemma 4, $z^*(\phi, a, r)$ is increasing in a and it is increasing in r if and only if $z^*(\phi, a, r) \geq 1$. Accordingly, there exists a one-to-one mapping between the feasible range $[\underline{a}(\bar{r}), \bar{a}(\bar{r})]$ and the feasible set of relative efforts, for any fixed $\bar{r}/p$ and any fixed c_x/c_y. Table B.2 depicts the feasible set of the *logarithm* (to base 10) of the customer intensity for various cost values of c_x/c_y and $\bar{r}/p$. As $\bar{r}$ increases, the feasible set increases; and as c_x/c_y deviates from 1, the feasible set shifts in favor of the party that has the lowest cost. In particular, the largest feasible sets for $\log(z(\phi, a, r))$ occur for the cases in which the feasible set $[\underline{a}(\bar{r}), \bar{a}(\bar{r})]$ is the smallest. For instance, when $\bar{r}/p = 0.9$ and $c_x/c_y = 0.1$, the feasible range for a is $[0.06, 0.16]$ (Table B.1), which translates into a feasible set for z^* of $[10^{-1.784}, 10^{2.841}] = [0.0164, 693.4]$. Accordingly, the feasible set of efforts is the largest when Assumption 2 constrains the most the feasible set of a, demonstrating that Assumption 2 is either nonstringent or without dramatic consequences on effort.

References

Adams CP (2006) Optimal team incentives with CES production. *Econom. Lett.* 92(1):143–148.

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

Anand KS, Paç MF, Veeraraghavan S (2011) Quality–speed conundrum: Trade-offs in customer-intensive services. *Management Sci.* 57(1):40–56.

Arrow KJ, Chenery HB, Minhas BS, Solow RM (1961) Capital-labor substitution and economic efficiency. *Rev. Econom. Statist.* 43(3):225–250.

Baiman S, Fisher PE, Rajan MV (2000) Information, contracting, and quality costs. *Management Sci.* 48(6):776–789.

Bellos I, Kavadias S (2014) A framework for service design. Working paper, George Mason University, Fairfax, VA. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2476072.

Bhaskaran SR, Krishnan V (2009) Effort, revenue, and cost sharing mechanisms for collaborative new product development. *Management Sci.* 55(7):1152–1169.

Bhattacharyya S, Lafontaine F (1995) Double-sided moral hazard and the nature of share contracts. *RAND J. Econom.* 26(4):761–781.

Bitner M-J, Faranda WT, Hubbert AR, Zeithaml VA (1997) Customer contributions and roles in service delivery. *Internat. J. Service Indust. Management* 8(3):193–205.

Bitran G, Lojo M (1993) A framework for analyzing the quality of the customer interface. *Eur. Management J.* 11(4):385–396.

Buzacott JA (2004) Modelling teams and workgroups in manufacturing. *Ann. Oper. Res.* 126(1–4):215–230.

Campbell CS, Maglio PP, Davis MM (2011) From self-service to super-service: How to shift the boundary between customer and provider. *Inform. Systems and eBusiness Management* 9(2):173–191.

Chase RB (1978) Where does the customer fit in a service operation? *Harvard Bus. Rev.* 56(6):137–142.

Chase RB (1981) The customer contact approach to services: Theoretical bases and practical extensions. *Oper. Res.* 29(4):698–706.

Chase RB, Tansik BA (1983) The customer contact model for organizational design. *Management Sci.* 29(9):1037–1050.

Christensen CM, Wang D, van Bever D (2013) Consulting on the cusp of disruption. *Harvard Bus. Rev.* 91(10):106–114.

Cook DP, Goh C-H, Chung CH (1999) Service typologies: A state of the art survey. *Production Oper. Management* 8(3):318–338.

Corbett CJ, DeCroix GA (2001) Shared-savings contracts for indirect materials in supply chains: Channel profits and environmental impacts. *Management Sci.* 47(7):881–893.

Corbett CJ, DeCroix GA, Ha AY (2005) Optimal shared-savings contracts in supply chains: Linear contracts and double moral hazard. *Eur. J. Oper. Res.* 163(3):653–667.

de Groote X (1994) The flexibility of production processes: A general framework. *Management Sci.* 40(7):933–945.

Donaldson L (2001) *The Contingency Theory of Organizations* (Sage Publications, Thousand Oaks, CA).

Economist, The (2011) Hopes that the internet can improve teaching may at last be bearing fruit. (September 17), http://www.economist.com/node/21529062.

Economist, The (2014) Flipping the classroom—Build it and they may come. (January 18), http://www.economist.com/news/business/21594316-management-schools-are-building-spree-risk-some-build-it-and-they-may-come.

Eppinger SD, Whitney DE, Smith RP, Gebala DA (1994) A model-based method for organizing tasks in product development. *Res. Engrg. Design* 6(1):1–13.

Frei FX (2006) Breaking the trade-off between efficiency and service. *Harvard Bus. Rev.* 84(11):92–101.

Fuchs VR (1968) *The Service Economy* (National Bureau of Economic Research, New York).

Gurvich I, Van Mieghem JA (2014) Collaboration and multitasking in networks: Architectures, bottlenecks, and capacity. *Manufacturing Service Oper. Management*. Forthcoming.

Hardy G, Littlewood JE, Pólya G (1952) *Inequalities*, 2nd ed. (Cambridge University Press, New York).

Hayes RH, Wheelwright SC (1979) Linking manufacturing process and product life cycles. *Harvard Bus. Rev.* 57(1):133–140.

Hopp WJ, Iravani SMR, Liu F (2009) Managing white-collar work: An operations-oriented survey. *Prod. Oper. Management* 18(1):1–32.

Hopp WJ, Iravani SMR, Yuen GY (2007) Operations systems with discretionary task completion. *Management Sci.* 53(1):61–77.

Huete LM, Roth AV (1988) The industrialisation and span of retail banks' delivery systems. *Internat. J. Oper. Prod. Management* 8(3):46–66.

Iyer AV, Schwarz LB, Zenios SA (2005) A principal-agent model for product specification and production. *Management Sci.* 51(1):106–119.

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

Karmarkar US (2004) Will you survive the services revolution? *Harvard Bus. Rev.* 82(6):100–107.

Karmarkar US, Pitbladdo R (1995) Service markets and competition. *J. Oper. Management* 12(3–4):397–411.

Kellogg DL, Chase RB (1995) Constructing an empirically derived measure for customer contact. *Management Sci.* 41(11):1734–1749.

Kellogg DL, Nie W (1995) A framework for strategic service management. *J. Oper. Management* 13(4):323–337.

Kim SH, Netessine S (2013) Collaborative cost reduction and component procurement under information asymmetry. *Management Sci.* 59(1):189–206.

Lewis MA, Brown AD (2012) How different is professional service operations management? *J. Oper. Management* 30(1):1–11.

Lovelock CH, Young RF (1979) Look to consumers to increase productivity. *Harvard Bus. Rev.* 57(3):168–178.

Marschak J, Radner R (1972) *Economic Theory of Teams* (Yale University Press, New Haven, CT).

Mersha T (1990) Enhancing the customer contact model. *J. Oper. Management* 9(3):391–405.

Milgrom P, Roberts J (1990) The economics of modern manufacturing: Technology, strategy, and organization. *Amer. Econom. Rev.* 80(3):511–528.

Milgrom P, Roberts J (1995) Complementarities and fit. Strategy, structure, and organizational change in manufacturing. *J. Accounting Econom.* 19(2–3):179–208.

Mills PK, Chase RB, Margulies N (1983) Motivating the client/employee system as a service production strategy. *Acad. Management* 8(2):301–310.

Ramírez R (1999) Value co-production: Intellectual origins and implications for practice and research. *Strategic Management J.* 20(1):49–65.

Rangan VK, Bell M (2008) Merrill Lynch: Integrated Choice. Technical Report 9-500-090, HBS Case Study, Harvard Business School, Boston.

Reinhardt U (1972) A production function for physician services. *Rev. Econom. Statist.* 54(1):55–66.

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

Sampson SE (2012) Visualizing service operations. *J. Serv. Res.-US* 15(2):182–198.

Sampson SE, Froehle CM (2006) Foundations and implications of a proposed unified services theory. *Production Oper. Management* 15(2):329–343.

Schaefer S (1999) Product design partitions with complementary components. *J. Econom. Behav. Org.* 38(3):311–330.

Schmenner RW (1986) How can service businesses survive and prosper? *Sloan Management Rev.* 27(3):21–32.

Shostack GL (1987) Service positioning through structural change. *J. Marketing* 51(1):34–43.

Silvestro R, Fitzgerald L, Johnston R, Voss C (1992) Towards a classification of service processes. *Internat. J. Service Indust. Management* 3(3):62–75.

Skinner W (1974) The focused factory. *Harvard Bus. Rev.* 52(3):112–120.

Spohrer JC, Maglio PP (2010) Toward a science of service systems. value and symbols. Maglio PP, Kieliszewski CA, Spohrer JC, eds. *Handbook of Service Science* (Springer Science+Business Media, New York), 157–194.

Szulanski G (1996) Exploring internal stickiness: Impediments to the transfer of best practice within the firm. *Strategic Management J.* 17(Winter Special Issue):27–43.

Teboul J (2006) *Service Is Front Stage* (INSEAD Business Press, Palgrave Macmillan, New York).

Thompson JD (1967) *Organizations in Action: Social Science Bases of Administrative Theory* (McGraw-Hill, New York).

Tong C, Rajagopalan S (2014) Pricing and operational performance in discretionary services. *Production Oper. Management* 23(4):689–703.

Topkis DM (1998) *Supermodularity and Complementarity* (Princeton University Press, Princeton, NJ).

Vargo SL, Lusch RF (2004) Evolving to a new dominant logic for marketing. *J. Marketing* 68(1):1–17.

von Hippel E (1994) "Sticky information" and the locus of problem solving: Implications for innovation. *Management Sci.* 40(4):429–439.

Wang Y, Yin R (2014) Product value enhancement under target pricing: Effort complementarity and commitment sequence. Working paper, Arizona State University, Tempe.

Wemmerlöv U (1990) A taxonomy for service processes and its implications for service design. *Internat. J. Service Indust. Management* 1(1):20–40.

Whitaker GP (1980) Coproduction: Citizen participation in service delivery. *Public Administration Rev.* 240–246.

White SW, Badinelli RD (2012) A model of efficiency-based resource integration in services. *Eur. J. Oper. Res.* 217(2):439–447.

Xue M, Field JM (2008) Service coproduction with information stickiness and incomplete contracts: Implications for consulting services design. *Production Oper. Management* 17(3):357–372.

Xue M, Hitt LM, Harker PT (2007) Customer efficiency, channel usage, and firm performance in retail banking. *Manufacturing Service Oper. Management* 9(4):535–558.