PLANNING AND SCHEDULING OF COMPLEX, HIGH VALUE-ADDING SERVICE OPERATIONS

Sheneeta Williams White

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
In
Business; Management Science

Ralph D. Badinelli, Chairman

Cliff T. Ragsdale
Roberta S. Russell
Deborah F. Cook
Christopher W. Zobel

July 27, 2009
Blacksburg, Virginia

Keywords: service operations, resource planning, client involvement

PLANNING AND SCHEDULING OF COMPLEX, HIGH VALUE-ADDING SERVICE OPERATIONS

Sheneeta Williams White

ABSTRACT

This research takes the initial steps of evaluating resource planning for service operations in which the client is a direct resource in the service system. First, this research examines the effects of client involvement on resource planning decisions when a service firm is faced with efficiency and quality considerations. We develop a non-linear, deterministic, single-stage planning model that allows for examination of trade-offs among client involvement, efficiency and quality. Policy recommendations give service firms better insights into setting workforce, client intensity, and service generation levels.

Second, we examine the sensitivity of estimates of technology functions to data analysis and make policy recommendations to service providers on how to allocate resources when there are technology function uncertainties and uncontrollable inputs. Results show that resources are allocated to compensate for technology function uncertainties.

Third, we gain insights as to how resource decisions are made for multiple stages and for multiple clients. We extrapolate theoretical findings from the single-stage planning study to determine resource allocations across multiple services and stages. Results show that when the dynamic program in the single-stage study is extended there is trade-off between the cost of capacity changes and profits across multiple stages.

DEDICATION

To Terrence

ACKNOWLEDGEMENTS

Writing this dissertation has been a journey for me. I thank my Lord and Savior, Jesus Christ for leading and guiding me throughout this journey.

Dr. Badinelli, thank you for your patience and guidance throughout this process. I will never forget all that you have taught me.

Dr. Ragsdale, thank you for giving me the opportunity to be a part of the BIT family.

Drs. Russell, Cook, and Zobel thank you so much for supporting me throughout this dissertation. Your doors were always open whenever I needed to talk and I truly appreciate your help.

Gary, Jia, and Mauro you guys are the best. Thanks for all of the laughs. I wish each of you much success with all of your future endeavors.

I have to thank my husband, Terrence. You have supported me every step of the way. You believed in me when I couldn't believe in myself. Thank you for keeping my belly full of red beans and rice, greens, black-eyed peas, and enough fruit to feed a small country. I love you dearly.

Table of Contents

List of Figures

Chapter 4

Chapter 5

List of Tables

CHAPTER 1

RESEARCH PURPOSE, OBJECTIVES, AND MOTIVATION

1. RESEARCH PURPOSE

The purpose of this research is to develop a family of models that will advance resource planning for service operations. A resource plan is a medium-range capacity plan over a 6-18 month planning horizon. A resource plan specifies an optimal combination of production levels, workforce levels, backlog, and inventory over the planning horizon. This research is directed at developing models for resource planning for service firms in which the client is a direct resource in the service process.

In developing these models, we are concerned with the effects of both inaccurate estimates and mis-specifications of production functions on resource planning decisions. A production function is the mathematical function that maps inputs to outputs. In service organizations, production functions are difficult to estimate because the conversion process is unclear. Lack of clarity results in inaccurate production function estimates that can greatly affect resource planning decisions.

Conventionally, the function used to convert inputs to outputs is called the production function. Rather than using conventional terminology, this study will call the function used to convert inputs to outputs the technology function. Rousseau (1979) defines a technology "as the application of knowledge to perform work". In professional services, for example, labor is a primary input to the conversion process. Professional employees are required to use knowledge, experience, and judgment in order to meet the needs of clients. The family of models is developed to assist decision makers, in professional service organizations, in resource planning decisions. Therefore, the term technology function is appropriate. This break from convention is meant to acknowledge the presence of the knowledge worker in professional service organizations.

2. RESEARCH OBJECTIVES

This research has three objectives. The first objective is to measure sensitivity of resource planning models to efficiency and quality, which are functions of client intensity. Efficiency and quality affect effective capacity and quality performance.

The second objective is to measure sensitivity of resource planning models to mis-estimation and mis-specification of technology functions and uncontrollable for co-generated service operations. This research acknowledges that the functional form between inputs and outputs for service processes is often unknown or unclear (Nachum, 1999). We will also recommend policies that take into consideration mis-estimation and mis-specification of technology functions. Managerial policies will be based on nonlinear decision models. On the parameter side, this research recognizes mis-estimation and mis-specification of technology function form. The uncertainty in the parameters causes uncertainty in the performance measures.

The third objective is to measure sensitivity of resource planning models to multiple services and multiple stages.

This research consists of developing a family of three resource planning models for complex service operations for the purposes of achieving the aforementioned objectives.

The first model is a deterministic resource-planning model which serves as motivation for the research agenda. The objective of this paper is to examine the effects of client involvement on resource planning decisions when a service firm is faced with efficiency and quality considerations. This model is a single-stage resource planning model that incorporates efficiency and quality of the service process. Efficiency and quality are functions of both client involvement and the skill levels of clients and service providers. Nonlinear programming is used to find model-optimizing resource allocation plans.

The second model is a stochastic resource planning model. The objectives of this research are to examine the sensitivity of estimates of technology functions to data analysis and to make policy recommendations to service providers on how to allocate resources when there are technology function uncertainties and uncontrollable inputs.

The third model is a multi-stage, multi-service resource planning model for service supply chains. The objective of this study is to gain insights as to how resource decisions are made for multiple service processes over multiple services. We develop a non-linear, deterministic, multi-stage planning model that allows for examination of trade-offs among client intensity, efficiency and quality.

3. RESEARCH MOTIVATION – APPLICATION

The service sector is rapidly increasing and has grown to 70% of the US economy (Radding, 2006). This growth is evidence of the need for proper management of service operations. There have been numerous articles published in top journals focusing on service operations. According to a survey performed by Machuca et al. (2007), research in the area of planning, scheduling and control of service operations only accounts for only 5.1% of all service operations research published from 1997 - 2002. During this same period, only 4.5% of the published papers were theoretical. These authors also identified the primitive level of development of service operations theory for complex services, noting that only 5.8% of all published articles focus on professional and/or B2B service industries.

There are many definitions of service. For example, a service has been defined in the literature as:

- Sasser et al. (1978): "Intangible and perishable… created and used simultaneously".
- Lovelock (1983): A service is "characterized by its nature (type of action and recipient), relationship with customer (type of delivery and relationship), decisions (customization and judgment), economics (demand and capacity), mode of delivery (customer location and nature of physical or virtual space)".
- Chase and Aquilano (1992): A service business is the "management of organizations whose primary business requires interaction with the customer to produce the service."
- Fitzsimmons (2001): "A time-perishable, intangible experience performed for a customer acting in the role of co-producer."
- Spohrer et al. (2007): "….service is a kind of action, performance, or promise that's exchanged for value between provider and client."

Conventional definitions of services are applicable to recreation and leisure, retail, and transportation industries. These services are typically very routine and employ workers with low technical skills (Fitzsimmons and Fitzsimmons, 2004). I label these types of services as "simple" services. This research is directed at what I call "complex" services. Complex services are customized to fit a specific customer, have high customer contact, and employ knowledge workers. Examples of complex services are consulting, auditing, I/T development, and legal services (i.e., professional services). Schmenner (1986) designed a service matrix to be an initial step towards classification of services. See Figure 1. Although this service matrix has drawn criticism over the years, it serves the purpose of defining the scope of this research. For the purposes of this research, I define a service as the transformation of inputs into outputs such that value for the customer is created through a process that utilizes capabilities and capacities of both the customer and the provider.

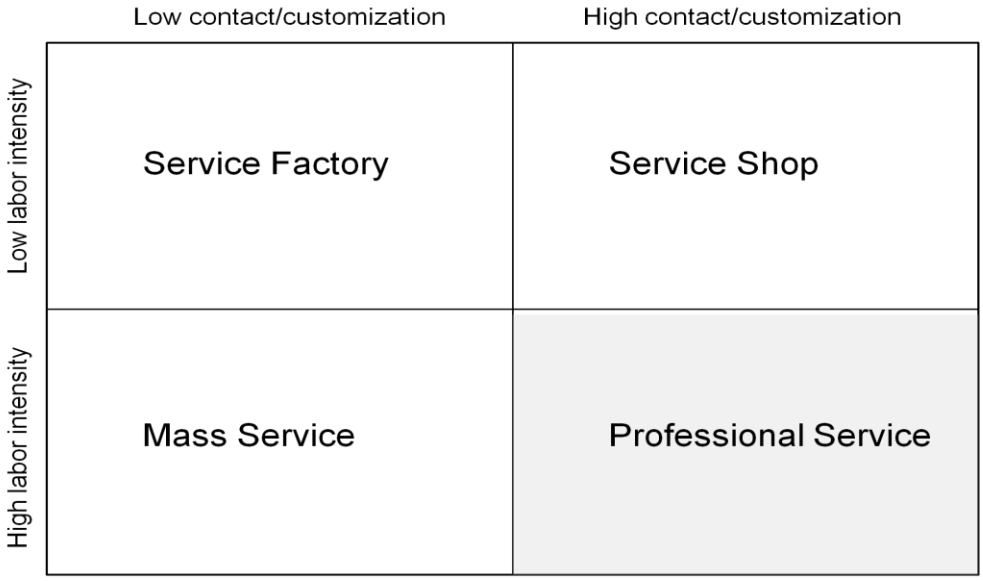

Figure 1: R.W. Schmenner's service matrix. Schmenner, R. W. (1986). "How can service businesses survive and prosper." Sloan Management Review **27**(3): 21-32. Permission to use granted by: Patti Newman, MIT Sloan Management Review (June 25, 2009)

It is important to define key terms that are used throughout this dissertation. *Efficiency* is defined as the ratio of output to input where outputs and inputs are measured as volumes of resources. *Quality* is a multi-dimensional scale that measures the degree of satisfaction of client

expectations, needs, wants and delights. *Productivity* is defined as a performance measure which combines the quality and efficiency of a process. The models presented in this dissertation are productivity-based resource planning models.

3.1 Characteristics of Services

Researchers have distinguished services from manufacturing by four characteristics often abbreviated IHIP. IHIP stands for intangibility, heterogeneity, inseparability (simultaneity), and perishability (Sasser et al., 1978). Although these characteristics are prevalent in literature, it is possible to argue that the IHIP characteristics do not adequately distinguish services from manufacturing operations. Sampson and Froehle's (2006) Unified Services Theory builds an impressive case against the claim that IHIP characteristics are only attributable to service processes. This research also questions the IHIP characteristics as only being attributable to services and suggests that there is a fifth essential and distinguishing characteristic – shared creation of value.

Intangibility is defined as an incapability of "being perceived by sense or touch" (Dictionary.com, 2008). A company that manufactures music CDs carries with it a host of intangible qualities. In professional services consultants are hired to provide guidance to a company and that guidance is indeed intangible, but that consultant can also provide that company with a document summarizing recommendations, which is tangible. Therefore, intangibility should not be seen as a characteristic to distinguish manufacturing from services (Laroche et al. 2001).

Heterogeneity is the characteristic that describes the variable, inconsistent, and nonstandard nature of services (Lovelock and Gummesson, 2004). In manufacturing mass customization has become a growing trend because of the ever increasing demand for variety. In services, for example, a medical procedure has to be customized to the patient because of individual customer needs. Contrary to conventional belief, heterogeneity should not be seen as a characteristic to distinguish manufacturing from services.

Perishability is the characteristic that output cannot be stored (Fitzsimmons and Fitzsimmons, 2004). If capacity does not allow output to meet demand in any time period, then a manufacturing firm can either hold inventory or backlog demand. In this dissertation we go

against conventional thinking and model inventory and backlog for services. We first quantify a service in terms of as the number of cycles of a process per unit of the service that is delivered to the client. A cycle is a single iteration of a process. For modeling purposes, we then define "inventory" as the number of completed cycles of a stage, which have been completed but for which a succeeding stage has not started. "Backlog" is measured by the number of units of services that are not completed by their due dates. Therefore, perishability should not be seen as a characteristic to distinguish manufacturing from services (Sampson and Froehle, 2006).

There are three types of resources – materials, capacities and capabilities. There are consumable resources such as raw materials. Labor (capacities) is also a type of resource. Consumable resources and labor are some of the resource types found in traditional manufacturing studies. This service-centric dissertation is focused on capacities as resources. As previously described, we capture services inventory by counting process cycles.

These process cycles are generated by service-provider and client labor. Lastly, there are renewable resources capabilities such as intellectual property, which can be enable a service process. Renewable resources have not been studied in great detail, but could be applied to services.Inseparability (simultaneity) is the simultaneous production and consumption of services. Manufacturing production, typically, happens in advance of demand and consumption. However, in services production and consumption happens simultaneously. This characteristic does distinguish services from manufacturing (Sampson and Froehle, 2006). The inseparability characteristic of services directly influences the models in this dissertation, because this characteristic allows for the presence of the customer in the service process. The proposed resource-planning models include the customer as a co-producer of the service.

This research is based on the idea that the shared creation of value between the service provider and the customer is a distinguishing characteristic of services. Shared creation of value is exhibited when both parties obtain value from the service. In manufacturing value is viewed as something that is added to products during the production process (Lusch et al. 2007). In services value is determined by the customer in the consumption process. Lusch et al. (2007) suggest that "there is no value until an offering is used – experience and perception are essential to value determination." The co-production relationship is the foundation for shared creation of value. The customer is a necessary part of the service process and is viewed as a co-producer. The

customer contributes information and his/her talents, knowledge, and experience to the service process.

3.2 Challenges to planning complex services

There are unmet challenges to resource and capacity planning for complex services which are the management of resources, the measurement of productivity and quality of the service, and the specification of technology-function form.

Managing resources is a challenge in service operations planning for two reasons. First, having the client as a resource introduces management challenges due to the variable nature of clients. Client variability is in the forms of knowledge, abilities, and motivation (Frei, 2006). Second, complex services must be flexible enough to deal with resource requirement changes. In complex services, there is a high-degree of customer contact and therefore requirement changes are common. These changes can range from the number of resources required, to the desired capability of each resource (Dietrich, 2006). Resource planning models must account for client variability and must be capable of handling resource requirement changes.

Measuring quality and productivity has long been viewed as a challenge in resource planning for complex services. Service quality is driven by customer perceptions, but how does a firm measure customer perceptions? This question may never be fully answered but service firms can ensure that requirements are clearly defined and continual communication with the client is maintained. Measuring productivity of service operations is challenging because of the variability in the performance of knowledge workers. This variation makes it is difficult to measure production output and often requires decision makers to measure attributes such as knowledge, skill level, and experience in order to determine potential output production (Nachum, 1999). Furthermore, there is no standard method for measuring the skills and capabilities of knowledge workers (Dietrich, 2006). In complex services, resources develop solutions to specific customer problems. Therefore, workers need to be creative and to think independently. This dynamic environment makes it difficult to determine, at any point in time, the knowledge that each worker truly possesses.

Specifying the form of the technology function is also a challenge for complex service firms. The transformation of inputs to outputs is usually not well understood in complex service

firms. This uncertainty is due to the unexplained variability in knowledge workers' performances over time.

4. RESEARCH MOTIVATION – THEORY

This research will develop productivity-based resource planning models for complex service operations. The technology functions that support any resource planning model has to be consistent with Data Envelopment Analysis (DEA) theory. DEA is a framework within our technology must fit. DEA is a linear programming technique developed by Charnes et al. (1978) for the purposes of measuring a service unit's efficiency. Service unit's efficiency is expressed as a ratio of outputs to inputs. DEA is described in detail in Chapter 2. This terminology in this dissertation will differ from conventional DEA terminology and define productivity as the ratio of outputs to inputs. This break from convention is meant to acknowledge the fact that performance measures in service-based resource planning models should combine both the quality and the efficiency of a process.

The resource planning models presented in this dissertation allocate resources to service units. In a DEA study, a service unit is called a decision making unit (DMU). A DMU consumes varying amounts of inputs to produce varying amounts of outputs. Consulting firms, bank branches, and hospitals are all examples of decision making units. This research disaggregates a DMU, noting that a DMU is comprised of several DMUs all having an individual technology function. For example, a bank branch is a DMU in the classic DEA sense. A bank branch is made up of individual workers who are also DMUs. A worker consumes varying amounts of inputs (e.g., experience, training) to produce varying amounts of outputs (e.g., skills, knowledge). The workers' technology functions are taken into consideration and resources are allocated among the appropriate DMUs.

The technology function determines the input-output coordinates of a DMU. It should be noted that DEA does not suggest a functional form linking inputs to outputs. However, this research specifies the technology function form for DMUs and shows how inaccurate estimates of technology functions effect resource planning decisions.

A *population* is defined as the set of all DMUs in a particular service sector. Each individual DMU, in the population, has a feasible input-output coordinate. The technology possibility set is the set of all feasible input-output coordinates of DMUs in the population.

The individual DMU's technology function is considered when making resource planning decisions. Each individual DMU's technology function can change, within a neighborhood, keeping that DMU feasible within the possibility set. The technology function can change as a result of workers receiving training or a company acquiring new information technology, for example. In Figure 2, service unit D has a neighborhood in which its technology function can change and it will remain feasible within the possibility set.

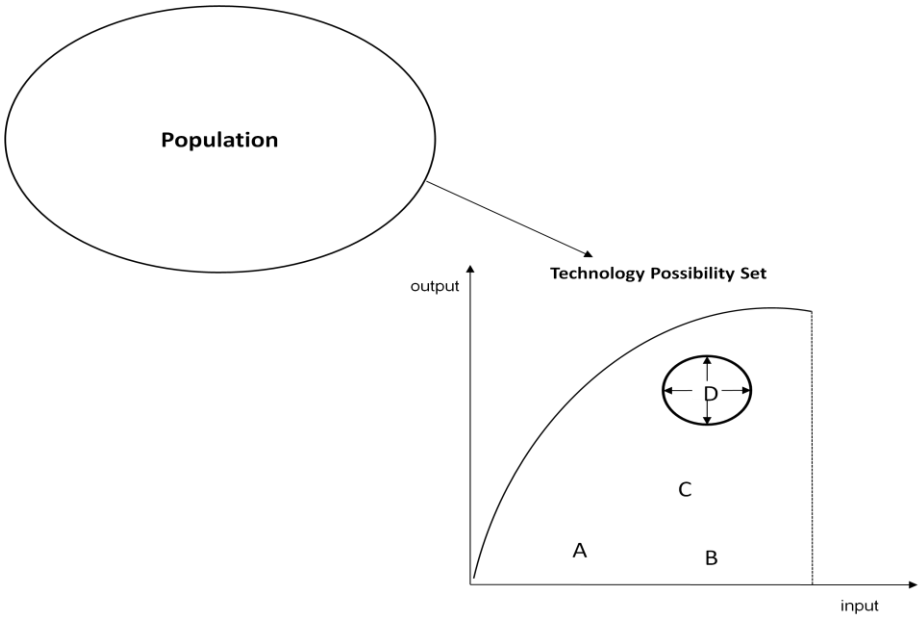

Figure 2: Technology possibility set (Created by author)

Previous approaches to resource planning for complex services make different assumptions about the extent of technology changes, which can take place over the planning horizon. In Gaimon (1997), technology changes occur when information technology costs decrease or the effectiveness of information technology increases, but these changes are allowable only at the organizational level not at the individual worker level. Some models assume that resources are fully productive (Anderson et al., 2006; Gaimon and Thompson, 1984). This means that there can be no improvements in the technology which is very unlikely in the real world. Anderson (2001) does allow unrestricted changes to the technology function at

the individual worker level. However, this model is not a multi-stage or multi-period service supply chain.

Previous approaches to defining explicit technology functions in resource allocation models have proposed deterministic forms of the function (Golany & Tamir, 1995; Golany et al., 2006). A shortcoming of resource allocation models for complex services is the lack of stochastic representations of technology functions. The nature of complex services is stochastic because of the variability in knowledge workers' performance and the inclusion of client in the production process. Variability is also present in the transformation process because of the high-degree of customization in complex service operations. Therefore, deterministic technology functions do not reflect true service operations planning.

Previous approaches to estimating implicit technology functions are based on assumptions that are too restrictive. Researchers have used DEA, as a technique to allocate resources in a service firm. DEA does not require an explicit technology function. Therefore, most DEA-based resource allocation models have implicit technology functions. Furthermore, limiting assumptions have been placed on technology changes in DEA-based resource allocation models. Some models allow only proportional changes to inputs and outputs (Golany et al., 1993; Thanassoulis, 1996; Korhonen and Syrjanen, 2004). Only allowing proportional changes takes a manufacturing-like approach to resource allocation. In manufacturing, if one machine produces ten widgets per hour, then by adding an additional machine twenty widgets per hour will be produced. This is not how complex service firms operate and decision makers cannot rely on such proportionality assumptions. Other models place limits on output production and thereby restrict changes to the underlying technology function (Beasley, 2003).

5. OUTLINE OF REMAINING CHAPTERS

The remainder of this dissertation is organized as follows. Chapter two contains a review of relevant literature. Chapters three, four, and five present the family of resource planning models, respectively. Chapter six summarizes research contributions and presents future research.

CHAPTER 2

REVIEW OF LITERATURE

The purpose of this chapter is twofold.

i. To review relevant literature in the area of service operations management
ii. To establish the contribution of this research

The remainder of the chapter is outlined as follows: Section 1 characterizes and classifies services; Section 2 describes planning and scheduling models for service operations; Section 3 is an introduction to Data Envelopment Analysis (DEA); Section 4 describes DEA-based resource allocation models.

1. SERVICE SCIENCE

The U.S. economy has moved from manufacturing-based to services-based with services now accounting for approximately 70% of jobs (Radding, 2006). Many researchers and top business executives agree that the world is now operating as a services-based economy. "The world is becoming a services system" said Jim Spohrer, the director of IBM Almaden Services Research. Levitt (1972) writes that "there are no such things as service industries. There are only industries whose service components are greater or less than those of other industries. Everybody is in service."

In recent years, the focus has shifted to services innovation. Services have now become a science. Services science is defined as "an emerging discipline that focuses on fundamental science, models, theories, and applications to drive innovation, competition, and quality of life through service" (Bitner et al., 2006). To build models for services there must be an understanding of the transformation of service inputs to outputs.

What are service inputs and outputs? An input is a factor of production. Service inputs can be divided into inputs from the service provider and inputs from the customer (Nachum, 1999; Sampson, 2007). Inputs from the service provider include labor, capital, and knowledge. The customer also provides inputs into the service transformation process in the form of

customer-provided information and customer-provided resources (Sampson and Froehle, 2006). An example of customer-provided information is a patient providing a doctor with their medical history. When a customer is used as a resource they are considered co-producers of the service.

Manufacturing and services have different types of output. An output is something that is produced. In manufacturing, an output is a good. The outputs of complex services, for example, are suggestions or advice about customer problems. The U.S. government recognized the difference in output between manufacturing and services and sought to classify "products" of services. In 1999 the North American Product Classification System (NAPCS) was developed to identify and define products (i.e., output) produced by industries in particular service sectors.[1] NAPCS was initiated to complement the North American Industry Classification System (NAICS). The NAICS classifies products produced from several service sectors such as: Arts, Entertainment and Recreation, Finance and Insurance, Educational, and Professional, Scientific and Technical services.

1.1 Definitions of Service

There have been many variations on the definition of service. Service definitions seem to focus on either the characteristics of service or the production process of service (Cook, Goh et al., 1999; Sampson and Froehle, 2006). Cook et al. (1999) gives an extensive review of service definitions spanning several decades. Many of those definitions are mentioned in this section.

Definitions focused on the characteristics of service (direct quotes):

- "Intangible and perishable… created and used simultaneously" (Sasser, Olsen et al., 1978)
- "intangibility of service output, the lack of inventories, the difficulty of portability, and complexity in definition and management" (Karmarkar and Pitbladdo, 1995)
- "services are deeds, processes, and performances" (Zeithaml and Bitner, 1996)
- "A time-perishable, intangible experience performed for a customer acting in the role of co-producer." (Fitzsimmons, 2001)

[1] The NAICS classification system was developed by statistical agencies from Canada, Mexico, and the United States. This industry classification scheme was supported by the North American Product Classification System (NAPCS). http://www.census.gov/epcd/www/naics.html, http://www.census.gov/eos/www/napcs/napcs.htm

Definitions focused on the production process of service:

- "Service is presumed to be performed by individuals for other individuals, generally on a one-to-one basis" (Levitt, 1972)
- "A service is a change in the condition of a person, or a good belonging to some economic entity brought about as the result of the activity of some other economic entity, with the approval of the first person or economic entity" (Hill, 1977).
- A service is defined by the extent of customer contact in the production process (Chase, 1978).
- "Services can be defined as economic activities that produce time, place, form, or psychological utilities." (Murdick, Render et al., 1990)
- A service business is the "management of organizations whose primary business requires interaction with the customer to produce the service." (Chase and Aquilano, 1992)
- "….service is a kind of action, performance, or promise that's exchanged for value between provider and client" (Spohrer, Maglio et al., 2007))

1.2 Classification of Services

A decision modeler can use information from service classifications to structure a decision model. A decision model uses data elements (e.g., decision variables, parameters, performance measures) and mathematical relationships to describe a decision problem. This research focuses on the decision problem of resource planning for complex services. Since there are not many resource planning models for complex services in current literature, literature that classifies services provides information that is helpful in the definition of data elements and the construction of the decision model.

There have been many approaches to classifying and categorizing services. Most classifications and characterizations of services stem from marketing literature. See Table 1.

Author	Categories/Groups
Judd, R.C. (1964)	• Rented Goods Services • Owned Goods Services • Non-goods Services
Shostack, G.L. (1977) Sasser, W.E. Jr., Olsen, R.P., Wyckoff, D.D. (1978)	Proportion of physical good and intangible services contained in each "product-service package"
Hill, T.P. (1977)	• Services affecting people vs. those affecting goods • Permanent vs. temporary effects of service • Physical vs. mental effects of service • Individual vs. collective services
Chase, R.B. (1978, 1981)	Degree of customer contact
Kotler, P. (1980)	• People vs. Equipment based • Extent of customers' presence • Public – Private vs. For-profit – Non-profit
Lovelock, C.H. (1983)	Five two-by-two classification matrices based on the following ideas: • Nature of service act • Relationship between service provider and customer • Customization • Demand and supply • Service delivery
Schmenner, R.W. (1986)	Service Process Matrix based on two dimensions: • Customer contact and customization • Labor intensity
Mersha, T. (1990)	Degree of customer contact. Definition of customer contact expanded to include active and passive contact
Kellogg & Nie (1995)	Service Product – Service Process Matrix

Table 1: Previous Service Classifications (Verma, R. and K. K. Boyer (2000). "Service classification and management challenges." Journal of Business Strategies **17**(1): 5.) Permission to use granted by: Dr. William Green, Editor, Journal of Business Strategies (June 29, 2009)

For example, Judd (1964) offered one of the first classifications of services. This classification divided services into three groups: rented goods services, owned good services, and non-goods services. Later, Hill (1977) developed five different ways in which to classify

services. Shostack (1977) and Sasser, Olsen et al. (1978) then offered the classification that focused on the proportion of physical goods and intangible parts in each service package.

Lovelock (1983) classifies services by asking the questions - who or what is the direct recipient of the service and what is the nature of the service act? Schmenner (1986) later presented a Service Process Matrix that builds off Chase (1981) and Lovelock (1983) and classifies services based on the degree of client contact/customization and the degree of labor intensity. The details of the Service Process Matrix are as follows:

- Service factories are low customer contact and low labor intensive service industries such as the transportation industry.
- Service shops increase in customer contact/customization. The health care industry is an example of service shop.
- Mass services have low degrees of customer contact/customization and a high degree of labor intensity. The banking and retail industries are examples of industries that produce mass services.
- Professional services have a high degree of customer contact/customization and high labor intensity. The legal and consulting industries are example of professional services.

Some classification schemes have been based on the degree of customer contact. Chase (1978) defined customer contact as "the physical presence of the customer in the system". Chase's Customer Contact Model (CCM) suggests separating those tasks that require the presence of the customer and those that can be performed in the back office. He identifies different worker skills needed for high-contact systems (HCS) versus low-contact systems (LCS). In HCS the worker needs public relations and interpersonal skills vs. in LCS the work needs analytical and technical skills. Kotler (1980) focused his classification approach on the extent of the customer presence and whether the service is for or not for profit. Mersha (1990) extends CCM by differentiating between active and passive customer contact. Active contact is defined as "direct contact between the customer and the service provider which involves direct customer-service system interaction". Passive contact is defined as "direct contact between the customer and the service system which does not involve customer-service system interaction". Kellogg and Nie (1995) introduced a service process/service package matrix where the service

process dimension focuses on the customer influence on service production and the service package dimension focuses on the degree of customization.

1.3 Characteristics of Services

Researchers have worked to develop a comprehensive list of service characteristics. The first such attempt came from Shostack (1977). She identifies intangibility, simultaneity, and co-production as characteristics that distinguish services from manufacturing. Since this time, in-depth reviews of service marketing literature have determined a different set of service characteristics. These four characteristics of services are intangibility, heterogeneity, inseparability, and perishability (Zeithaml, Parasuraman et al., 1985; Murdick, Render et al., 1990; Fisk, Brown et al., 1993). Today these characteristics are denoted, IHIP, and are the most notable characteristics of services.

Chapter 1 of this dissertation questions existing literature and explains how intangibility, heterogeneity, and perishability are not characteristics that distinguish services from manufacturing. It is shown that the distinguishable characteristics of services are inseparability and the shared creation of value between the customer and the service provider. Refer to Chapter 1 for definitions of each IHIP characteristic and further explanation of the indistinguishable nature of these characteristics.

Intangibility is defined as something that incapable of being seen or touched. This characteristic has been noted as the key distinction between manufacturing and services (Hill, 1977; Shostack, 1977). Distinctions have been made between physical intangibility (i.e., impalpable) and mental intangibility (i.e., cannot be grasped mentally) (Bateson, 1977).

Heterogeneity is that every service experience is different for every customer. The non-uniformity of service delivery, especially in high employee-customer contact and labor-intensive services, has implications on the way in which quality is measured and improved (Soteriou and Hadjinicola, 1999). Management can try to implement standards on the delivery of the service.

Inseparability is the simultaneous production and consumption of services. Inseparability is less of a problem in service factories than in professional services, because the customer is less involved in the service process (Schmenner, 2004). The inseparability characteristic also

introduces quality control problems, because the service cannot be inspected prior to consumption.

Perishability is the time-sensitive nature of services. Service capacity, like all capacity, is perishable. Since services cannot be inventoried, management is faced with the challenge of balancing capacity, utilization, and customer waiting times (Fitzsimmons and Fitzsimmons, 2004). Managers have a few options. They can try to smooth demand, adjust service capacity, and/or allow customers to wait. Of course these options vary depending on the type of service industry.

Researchers have begun to characterize services by the extent of customer involvement in the service system (Sampson and Froehle, 2006). Customer involvement in the service system can be in the customer provided information and customer co-production. This research extends the customer involvement characteristic and suggests that the co-creation of value is also a characteristic of services. Shared creation of value is exhibited when both parties obtain value from the service. Shared creation of value always exists for

1.4 Measuring Productivity and Quality of Services

There is an ongoing mission to improve productivity and quality measurements in service operations. Productivity and quality measures provide benchmarks for improving labor usage (McLaughlin and Coffey, 1990). Productivity measures provide information about the relationship between the output of the service and the input that is required for the service process. Service quality measures provide information about when and where to allocate resources in the service process. Since this research is focused on resource planning for service operations, it is essential that productivity and quality are considered.

The measure of productivity used in manufacturing is well understood. However, measuring productivity for services is not. Manufacturing productivity is the ratio of an organization's outputs to inputs. See Equation 2-1 for single input-output representation. The more output an organization produces per a given set of inputs, the more productive the organization.

$$\text{Productivity} = \frac{\text{output}}{\text{input}} \tag{2-1}$$

Customer involvement in the service process makes productivity difficult the measure. With this in mind, Chase (1981) developed the CCM for services. He determined that a service system's potential operating efficiency must account for the customer's contact time in the service system. See Equation 2-2.

$$\text{Potential Facility Efficiency} = f\left(1 - \frac{\text{customer contact time}}{\text{service creation time}}\right) \tag{2-2}$$

Further studies have offered another measure of service productivity. Gronroos and Ojasalo (2004) measure service productivity as a function of internal, external, and capacity efficiencies. See Equation 2-3. Internal efficiency is how well a firm utilizes its resources to produce outputs. External efficiency is how well a firm creates external interest in service output. Capacity efficiency is how well a service firm manages demand. The two efficiency measures, internal and external efficiency, were originally proposed by Ekholm (1984), but his measures ignored the element of quality and its interrelationship with productivity.

$$\text{Service Productivity} = f(\text{internal efficiency}, \text{external efficiency}, \text{capacity efficiency}) \tag{2-3}$$

Service productivity can also be measured using Data Envelopment Analysis (DEA). DEA is a technique for determining the productivity of service units with multiple inputs and outputs (Charnes et al., 1978). DEA is discussed in extensive detail in Section 2.3 of this chapter.

Measuring service quality happens during service delivery and is based on the perception of the customer. Parasurmanan et al. (1988) offer five dimensions of service quality:

- **Reliability**: ability to perform the promised service dependably and accurately
- **Tangibles:** physical facilities, equipment, and appearance of personnel
- **Responsiveness:** willingness to help customers and provide prompt service
- **Assurance:** knowledge and courtesy of employees and their ability to inspire trust and confidence
- **Empathy:** caring, individualized attention the firm provides its customers

These five dimensions are standard in most service operations or service marketing textbooks. Customers will differ on each dimension depending on their individual wants, needs, and expectations. Parasurmanan et al. (1988) developed SERVQUAL which is an instrument for measuring service quality based on these five dimensions and is the most commonly used instrument of its type.

Even with open lines of communication among service providers and customer, there can be gaps between customer expectations, perceptions, needs and wants. This gap is the measure of service quality (Fitzsimmons and Fitzsimmons, 2004). Improving market research, fostering better communication between management and employees, and reducing the number of management levels that distance the customer are all strategies for narrowing the gap. This dissertation does not seek to advance productivity and quality measurements of service, yet it acknowledges the importance of productivity and quality on resource planning.

2. PLANNING AND SCHEDULING MODELS FOR SERVICES

Production planning and control (PPC) is a key component in managing any productive system. PPC is the acquisition and allocation of resources in order to meet demand over a planning horizon (Sipper and Bulfin, 1997). A production plan determines how much of a product to make and when to make the product. Production planning is usually done at the aggregate level. Given a forecasted demand, an aggregate plan determines the production rates and workforce levels that minimize costs (e.g., hiring and firing costs, inventory costs, overtime costs). Aggregate production planning (APP) was introduced by (Holt, Modigliani et al., 1955). Hanssmann and Hess (1960) developed a linear programming approach to aggregate production planning. The linear APP approach is the foundation of the models presented in this dissertation.

A linear program is a mathematical model with a linear objective function and linear constraints. Mathematical models consist of decision variables, parameters, performance measures, constraints, and an objective function. A decision variable(s) describes the decision that is made by the model (e.g., the quantity of a product to produce). Parameters are values provided as inputs into the model. Performance measures describe how well a system is performing. Constraints are bounds or restrictions on the optimization problem (e.g. overtime

must be less than 10 hours per week). The objective function is a function of the decision variables that decision maker wants to minimize or maximize.

There are many service-based planning models in the literature. Most of these models are for planning and scheduling of simple services. This dissertation is centered on developing models for complex services, of which there are very few planning models. Table 2 shows differences in model formulation of service-based planning models. The papers are briefly described here. Abernathy et al. (1973): the purpose of this paper is to present a multi-period staffing and planning model for a nursing staff. The authors consider stochastic, stationary demand. Gaimon (1997): the purpose of this paper is to examine the effect information technology acquisitions have on knowledge-worker productivity. Soteriou and Hadjinicola (1999): the purpose of this paper is to develop a modeling framework to improve service quality in a multi-stage service system. The authors allocate a resource (i.e., a budget) in order to improve customers' perceptions of service quality. This nonlinear model minimizes the weighted loss in service quality perceptions per stage. Graves and Tomlin (2003): the purpose of this paper is to develop a multistage, multiproduct supply chain in order to minimize the shortfall of meeting stochastic demand. Napoleon and Gaimon (2004): the purpose of this paper to present two models for service operations. The models measure the impact worker knowledge

	Abernathy et al. (1973)	**Gaimon (1997)**	**Soteriou and Hadjinicola (1999)**	**Graves and Tomlin (2003)**	**Napoleon and Gaimon (2004)**	**Anderson and Morrice (2006)**	**Proposed Model (Model 1 –presented in Chapter 3)**
# of stages	Multiple	Single	Multiple	Multiple	Multiple	Multiple	Multiple
Objective	Costs	Profit	Minimize loss of service quality perceptions	Minimize shortfall in meeting demand	Profit	Costs	Profit
Resource capacity changes	Considered	Considered	Not considered	Considered	Considered	Considered	Considered
Resource training/learning	Considered	Considered	Not considered	Not considered	Considered	Not considered	Not Considered
Worker attrition	Considered	Considered	Not considered	Not considered	Not considered	Considered through costs of changing capacity	Not considered
Customer waiting	Not considered	Not considered	Considered through responsiveness factor	Not considered	Not considered	Considered through backlogging	Considered through backlogging
Inventory	Not considered	Not considered	Not considered	Not considered	Not considered	Not considered	Considered
Client Involvement	Not considered	Not considered	Not considered	Not considered	Not considered	Not considered	Considered

Demand	Stochastic	Deterministic	Not considered	Stochastic	Not considered	Stochastic	Deterministic
Methodology	Linear optimization	Optimal control theory	Non-linear optimization	Linear optimization	Optimal control theory	Optimal control theory	Non-linear programming

Table 2: Complex service-based planning models (Created by author)

has on performance gains from IT usage. Anderson et al. (2006): the purpose of this paper to evaluate coordination and information sharing in service supply chains.

2.1 Configuration of service supply chains

The goal of service supply chain management is similar to manufacturing supply chain management in that both are approaches to managing the delivery of a product to the end customer. Manufacturing and service supply chains are also similar because they have like components. Each type of supply chain receives inputs from a supplier. The inputs are then transformed into something of value to the customer and the service is delivered or the manufactured product is shipped or inventoried. Manufacturing supply chains have a unidirectional flow from supplier to customer, however, service supply chains have been determined to have a bidirectional flow (Sampson, 2000). Bidirectional flow means that production flows from the suppliers to customers and from customer to suppliers. For examples of service supply chain models, the reader is referred to Akkermans and Vos (2003) and Anderson et al. (2006).

3. INTRODUCTION TO DATA ENVELOPMENT ANALYSIS

Data Envelopment Analysis originated in the late 1970's in order to evaluate efficiency in nonprofit and service organizations. The original methodology of DEA was presented by Charnes, Cooper, and Rhodes (1978) in the seminal paper "Measuring Efficiency of Decision Making Units". The initial DEA model built on the previous work of Farrell (1957). Since this time hundreds of papers and books have been published on the subject of DEA.

The purpose of DEA is to analyze the performance (i.e., efficiency) of a sample of service units within an organization. The sample of service units are drawn from a population of service units. Each service unit, in the sample, has a measure of performance that is a ratio of its weighted outputs to weighted inputs. The weights are a measure of the decrease in efficiency with each unit reduction of output and a measure of the increase in efficiency with each unit reduction of input (Fitzsimmons and Fitzsimmons, 2004). The solution to a DEA analysis will determine the most favorable weights for each service unit. Multiple inputs and outputs are

aggregated to achieve an overall performance rating. This overall performance rating reflects which service units are efficient and which service units should be able to improve their inefficiency. The measurements can be used in performance evaluation and benchmarking (Cooper, Seiford et al., 2004).

In a DEA study, a service unit is called a decision making unit (DMU). A DMU is the unit whose efficiency is to be measured relative to that of other units of its kind (Charnes, Cooper et al., 1978). A DMU consumes varying amounts of *m* inputs to produce varying amounts of *s* outputs, see Figure 1. Mathematically stated, DMU_j consumes amounts $X_j = \{x_{ij} \mid i = 1,2,...,m\}$ and produces $Y_j = \{y_{rj} \mid r = 1,2,...,s\}$. DMUs within a DEA study must be homogenous units, meaning they use the same set of inputs to produce the same set of outputs (Charnes, Cooper et al., 1978). An input is a factor of production. Examples of inputs are labor, raw materials, and land. The set of inputs should include all resources which impact the outputs. An output is something that is produced, for example: sales, profit, transactions. The outputs should reflect all useful outcomes on which we wish to assess the DMU.

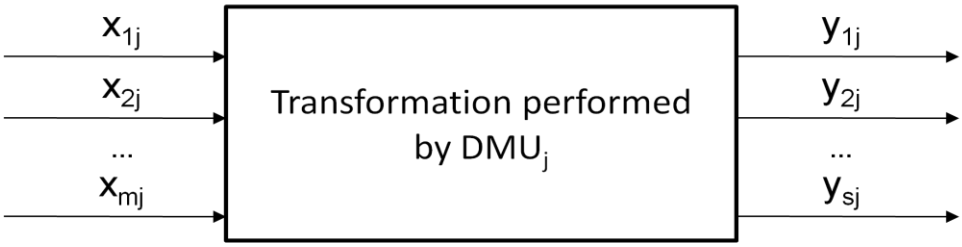

Figure 1: DMU representation (Created by author)

3.1 Introduction to Utility Theory

It is important, at this point, to define key terms that are used throughout this dissertation. *Efficiency* is defined as the ratio of outputs to inputs where outputs and inputs are measured as volumes of resources. *Quality* is a multi-dimensional scale which measures the degree of satisfaction of client expectations, needs, wants and delights. *Productivity* is defined as a performance measure which combines the quality and efficiency of a process. *Utility* is defined as benefits decision makers obtain from goods and services they consume (Maurice and Thomas, 1995). Efficiency, quality, and productivity can all be measures of a decision maker's utility. The

models presented in this dissertation are productivity-based resource planning models, because service resource planning should capture efficiency and quality as performance measures.

Utility theory assumes that every decision maker has preferences towards risks and return and that the decision maker will choose the alternative that maximizes his/her utility. It is assumed that every decision maker has a utility function by which preferences are made (Winston, 1991). Utilities are expressed on a scale from 0 to 1. A utility function is denoted mathematically as $U = (x)$. The "x" term can represent a single or multiple consequences or outcomes. For example, suppose that two people want to purchase a new car and both have different preferences that influence that decision. One person's preference, or utility function, is based on price $U = (\text{price})$, while the other person's preference is based on performance $U = (\text{performanc e})$. The measure of DEA efficiency is grounded in utility theory. The value of the input – output coordinate of each DMU_j is measured by its utility – $U = (X_j, Y_j)$ where X_j is a vector of input measures and Y_j is a vector of output measures.

3.1.1 Production Possibility Set and Returns-to-Scale

The Production Possibility Set (PPS) consists of all convex combinations of input-output coordinates. The following properties characterize the production possibility set (Banker, Charnes et al., 1984):

- *Convexity:* All points lying on the line segment joining two feasible input-output coordinates are also feasible (i.e., an input-output coordinate lying within the PPS).
- *Inefficient Production or "Free Disposability':* Inefficient production means that feasible input-output coordinates may use more inputs than required for output levels, or alternatively, may produce fewer output levels than feasible foe input levels.
- *Constant returns-to-scale:* if any feasible input-output coordinate is scaled up or down by some positive value, then a new feasible input-output coordinate is obtained. A PPS can reflect constant, decreasing, or increasing returns-to-scale.
- *Minimum Extrapolation:* the PPS is the smallest set containing all input-output coordinates.

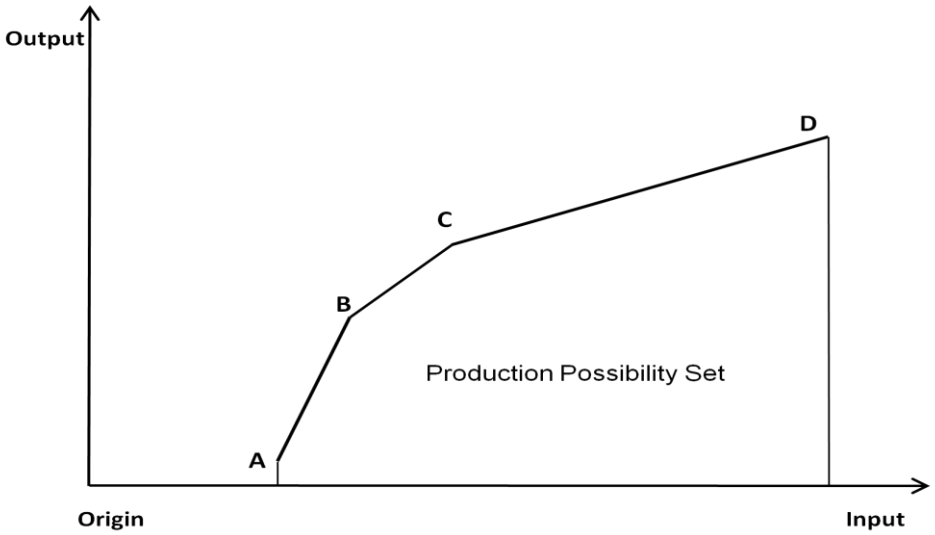

Figure 2: Production Possibility Set (Service Productivity Management: Improving service performance using Data Envelopment Analysis (DEA). New York, Springer Science+Business Media.) Permission to use granted by: Estella Jap A Joe, Springer/Kluwer Academic Publishers (July 16, 2009)

The impact of scale size on efficiency is measured by scale efficiency. Scale efficiency is the potential efficiency gains from a firm achieving its optimal size. Scale efficiency measures the difference between the efficiency rating of a DMU under constant returns-to-scale (CRS) and variable returns-to-scale (VRS) (Sherman and Zhu, 2006). Constant returns-to-scale is the increase in input levels that leads to a proportionate increase in output levels. Variable returns-to-scale is if an increase in input levels leads to a more than proportionate increase in output levels or an increase in input levels leads to a less than proportionate increase in output levels. The larger the deviation between CRS and VRS efficiency ratings, the smaller the value of scale efficiency will be.

Models can assess performance on constant, increasing, or decreasing returns-to-scale. An input-output coordinate is said to exhibit constant returns to scale when if its input levels are scaled up or down, there is another feasible proportionate input-output coordinate. Mathematically stated, let (x,y) represent an input-output coordinate. If a PPS reflects constant

returns-to-scale, then $(\alpha x, \alpha y)$ is a feasible input-output coordinate, given $\alpha > 0$ (Banker, Charnes et al., 1984).

When a line passes through an efficient input-output coordinate, the y-intercept term, say μ, reveals the presence of increasing, constant, or decreasing returns to scale.

If $\mu < 0$, increasing returns-to-scale (IRS),

If $\mu = 0$, constant returns-to- scale,

If $\mu > 0$, decreasing returns-to-scale

Figure 3 is a modified version of Sherman and Zhu's (2006) returns-to-scale representation and shows the regions for CRS, IRS and DRS. It should be mentioned that the Charnes, Cooper, Rhodes DEA model is known as the CRS model and the Banker et al. model is known as the VRS model.

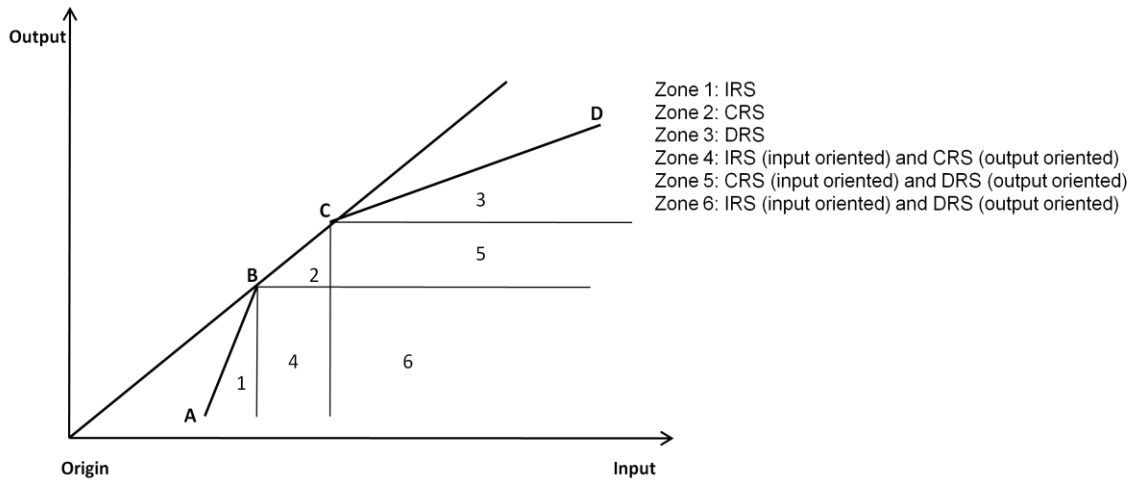

Figure 3: Returns-to-Scale Diagram (Service Productivity Management: Improving service performance using Data Envelopment Analysis (DEA). New York, Springer Science+Business Media.) Permission to use granted by: Estella Jap A Joe, Springer/Kluwer Academic Publishers (July 16, 2009)

3.1.2 Dominance, efficiency, Pareto efficiency, distance

Dominance occurs when one result is deemed better than another result. In DEA, higher output levels and fewer input levels are preferred. For example, if there exists a DMU A that has 1 bank teller (i.e., input) who can produce 5 bank transactions (i.e., output) in one hour and there is a DMU B that has 1 bank teller 1 bank teller (i.e., input) who can produce 10 bank transactions (i.e., output) in one hour, then DMU A is dominated by DMU B.

The efficient frontier is the set of all non-dominated solutions. This means that there is no feasible point which can offer higher output production without requiring higher input levels or can reduce input levels without reducing output production. The DMUs on the efficienct frontier are referred to as the best-practice units. In Figure 2.2, the efficient frontier is a linear combination of the input-output coordinates of DMUs A, B, C, and D. Each DMU not on the efficient frontier is considered inefficient. The level of inefficiency is determined by comparing the inefficient DMU to a referent DMU on the frontier that utilizes the same level of inputs to produce the same or greater level of outputs.

All points on the efficient frontier are Pareto-efficient. That is no point on this frontier is dominated by any other point in the PPS. Pareto-efficiency can be defined in terms of output and input orientation. When describing Pareto-efficiency in terms of output orientation, a DMU is Pareto-efficient if it is not possible to increase any one of its output levels without increasing at least one of its input levels (i.e. worsening its input level). When describing Pareto-efficiency in terms of input orientation, a DMU is Pareto-efficient if it not possible to lower anyone of its input levels without lowering one of its output levels.

If the result of a DEA analysis identifies a DMU as “efficient”, this means that the particular DMU is more efficient than another DMU or “relatively” efficient. A DMU that has been deemed efficient can be technically and/or scale efficient. Technical efficiency is the minimum inputs used to produce outputs. Scale efficiency is the potential efficiency gains from a firm achieving its optimal size. A DMU can also be technical and/or scale inefficient.

The purpose of distance is to measure the relative utility of any point in the PPS in order to determine the point’s technical efficiency or inefficiency. Distance is measured by the scaled ratio of the utility of a point on the efficient frontier to the utility of another point in the PPS.

A *direction* must be chosen when determining the point on the efficient frontier to use for comparison in the distance measure. In conventional DEA, there are two directions – vertical and horizontal. By moving vertically, we are trying to find a point in the PPS where output levels are greater for a given input level. By moving horizontally, we are trying to find a point in the PPS where input levels are less for a given output level.

The technical inefficiency of a DMU can be measured by Shephard's (1970) distance function. In Figure 2.4, $L(Y)$ is the set of feasible input values for a given value of Y such that each (X,Y) coordinate is contained within the PPS. The minimum amount of input used to produce a certain output level is determined by $h_x(X,Y) = \min\{h_x : h_x X \in L(Y), h_x \geq 0\}$. The point $h_x(X,Y)$ will lie on the efficient frontier. The efficiency of a point on the efficient frontier (point B in Figure 2.4) is measured by $g(X,Y) = 1/h(X,Y)$.

Shephard's distance function can also be used to measure distance vertically. In Figure 2.4, $P(X)$ is the set of feasible output values for a given value of X such that the (X,Y) coordinate is contained within the PPS. The maximum amount of output produced given a certain input level is determined by $h_y(X,Y) = \max\{h_y : h_y Y \in P(X), h_y \geq 0\}$. The efficiency of a point on the efficient frontier is measured by $g(X,Y) = 1/h(X,Y)$.

Suppose we want to find the relative inefficiency of point E, in Figure 4. We would first determine the set of feasible X values for the given Y value of Point E. Next, we would move horizontally to find any point(s) that used fewer inputs relative to Point E. We would determine that Point B used fewer inputs. This means that Point E is less efficient relative to Point B. In this example, if the efficiency of Point E equaled one-third then it can be determined, by using Shephard's distance function, that Point B is 3 times as efficient as Point E.

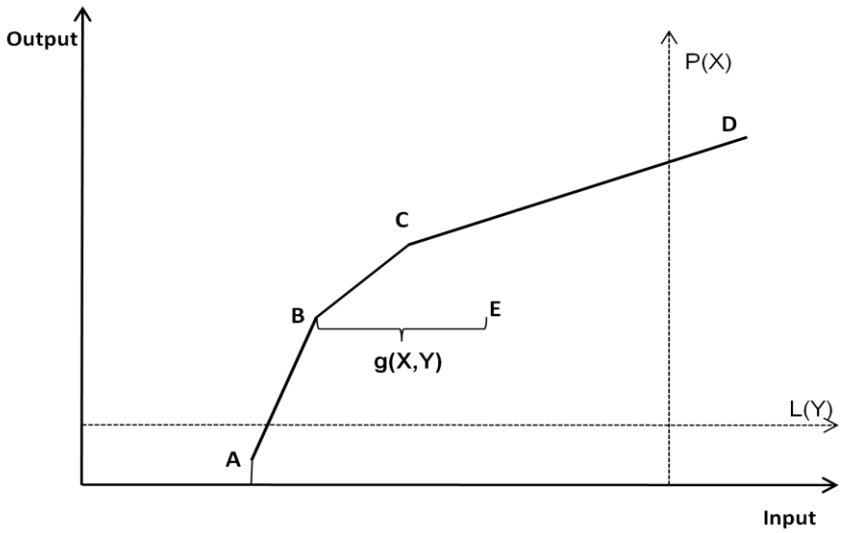

Figure 4: Measuring Distance (Created by author)

3.2 DEA Explained

DEA estimates the production possibility set of a population from a small sample of DMUs from that population. After this sample is drawn, the production possibility set is the smallest set of input-output coordinates. See Figure 3. DEA seeks to determine an efficient frontier that envelopes the PPS.

The following notation is used throughout the remainder of this section.

System Elements

$k =$ number of DMUs in the study; $j = 1,\ldots,k$

$DMU_j =$ decision making unit j

$S_{DMU} =$ set of DMU's in the study

$m =$ number of inputs; $i = 1,\ldots,m$

$s =$ number of outputs; $r = 1,\ldots,s$

$x_{ij} =$ input i of DMU_j

$y_{rj} =$ output r of DMU_j

Parameter

ε = lower bound on all weights

Decision variables

Each DMU is assigned a utility (efficiency) function by the DEA procedure for scaling DMU performance and estimating the efficient frontier. Hence, we define a efficiency function for each DMU in terms of the weights for the inputs and outputs.

v_i = weight assigned to input i

u_r = weight assigned to output r

λ_j = weight assigned to DMU_j

Performance measure

θ = efficiency rating

Data envelopment analysis is a method for measuring relative efficiency. DEA is comprised of a collection of models, all with different, yet associated, ways of evaluating performance. Table 2.3 highlights differences among the basic DEA models. Charnes, Cooper and Rhodes (CCR) introduced a ratio form of the DEA model in 1978 to measure efficiency. In the ratio model, the objective function maximizes the efficiency of the DMU under investigation (DMU_0) relative to the ratio of outputs to inputs of all other DMUs. There is a set of constraints that ensures the ratio of outputs to inputs for each DMU_j does not exceed 100%. The CCR ratio model can be mathematically represented as:

$$\max h_o(u,v) = \frac{\sum_{r=1}^{s} u_r y_{ro}}{\sum_{i=1}^{m} v_i x_{io}} \qquad (2\text{-}5)$$

subject to

$$\frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} \leq 1 \qquad \text{for } j = 1,\ldots,n$$

$$u_r, v_i \geq 0 \quad \text{for all } i \text{ and } r$$

DEA optimization problems have a primal (multiplier) and a dual (envelopment) formulation. The form of the model in which the decision variables are weights of inputs and outputs is the *multiplier* model and the form of the model where the decision variables are weights of individual units is the *envelopment* model. For example, if there are two bank branches being modeled, each having two inputs and two outputs, the multiplier model decision variables would be u_1, u_2, v_1, v_2 (one for each input and output). The envelopment model would have the following decision variables λ_1, λ_2 representing each branch. The name, envelopment model, is reference to the fact that the boundary of the PPS envelopes the input-output coordinates of the DMUs.

Because of the duality theorem in linear programming, the primal and the dual models result in the same optimal value of the objective function. The dual or envelopment model in DEA provides sustainable information about the relative efficiency of DMUs. The dual variables (one for each DMU) are the shadow prices related to the constraint sets that ensure the efficiency rating is no more than 1. As in linear programming, if the constraint is binding then the shadow price is positive and if the constraint is non-binding then the shadow price is zero. Therefore, when a DMU is deemed efficient the constraint is binding and the dual variable is positive.

Model	Originator(s)	Returns-to-scale	Efficient Frontier
CCR Input Model	"	Constant	Linear
CCR Output Model	"	Constant	Linear
BCC Input Model	Banker, Charnes, Cooper (1984)	Variable	Linear
BCC Output Model	"	Variable	Linear
Invariant Multiplicative Model	Charnes, Cooper, Seiford, Stutz (1982,1983)	Variable	Log-Linear
Variant Multiplicative Model	"	Constant	Log-Linear
Additive Model	Charnes, Cooper, Golany, Seiford, Stutz (1985,1987)	Variable	Linear

Table 3: Basic DEA Models (Created by author)

The remainder of this section gives the details of the models in Table 3, identifying unique characteristics of each. Each model has a primal and dual formulation, an input and an output orientation, and demonstrates either constant or variable returns-to-scale.

CCR – Linear Input Model

Primal (Multiplier Model)

$$\text{maximize} \sum_{r=1}^{s} u_r y_{ro} \tag{2-6}$$

subject to:

$$\sum_{r=1}^{s} u_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} \leq 0 \qquad \text{for all } j \in S_{DMU}$$

$$\sum_{i=1}^{m} v_i x_{io} = 1$$

$$u_r, v_i \geq \varepsilon > 0 \qquad \text{for } 1 \leq i \leq m, 1 \leq r \leq s$$

Charnes and Cooper developed an equivalent linear programming model of the ratio model. In the linear program, the denominator of the ratio is used as the normalizing constraint, by setting the constraint equal to 1, and the numerator of the ratio is the objective function (Charnes et al., 1978). The CCR – Linear Input model determines the optimal values of the decision variables u_r, v_i (i.e., weights). The CCR – Linear Input model optimizes the best possible weights for the DMU under investigation subject to the constraints. The objective function maximizes the weighted sum of the outputs of the DMU under investigation. This is a maximization problem since, ideally, DMUs would like to maximize outputs for a given set of inputs. Conversely, in the output model the inputs are minimized in the objective function. Constraint set 1 ensures that the ratio of weighted outputs to inputs of all of the DMUs will not exceed 1. Constraint set 2 is the normalizing constraint. Constraint set 3 is the non-negativity constraints for the decision variables. Epsilon is the lower bound for the multipliers.

Dual (Envelopment Model)

$$\underset{\bar{\lambda}, \theta, \bar{s}^-, \bar{s}^+}{\text{minimize}} \theta - \varepsilon \left(\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+ \right) \tag{2-7}$$

subject to:

$$\sum_{j=1}^{n} \lambda_j x_{ij} + s_i^- = \theta x_{io} \qquad i = 1,2,\ldots,m$$

$$\sum_{j=1}^{n} \lambda_j y_{rj} - s_r^+ = y_{ro} \qquad r = 1,2,\ldots,s$$

$\lambda_j \geq 0 \qquad\qquad j = 1,2,\ldots,n$

In the dual problem the Lagrange multipliers (i.e., decision variable) construct a convex combination of the points in the PPS which represent the input-output coordinates. The dual problem selects a point that allows for the maximal input reduction of the unit under investigation. The variable θ is the reduction applied to all inputs of the DMU under investigation to improve efficiency. Since theta is multiplied by the RHS of the input constraints, this gives decision makers information about how much the inefficient DMU should reduce its inputs by to improve its efficiency to that of the efficient unit(s) (i.e., can you produce the same amount with fewer inputs). For example, if DMU (3) is inefficient and theta = 0.90, then one way for DMU (3) to become more efficient could be to reduce its inputs by 90%. In effect we are moving the inefficient DMU to the left and projecting that DMU onto the frontier. The exact value to move the DMU, horizontally, is given by the slack variable s_i^- . The slack variables measure the distance of the point representing the input-output coordinate of the DMU under investigation from the point representing the input-output coordinate of the nearest DMU on the efficient frontier. The lower bound constant, ε , in the objective function is a relative weight on the maximizing of the slacks as compared to the minimizing of the efficiency, θ . The optimal solution to the dual problem will result in an efficiency score for the DMU under investigation (θ^*). The performance of a DMU is full (100%) efficient if and only if both (i) $\theta^* = 1$ and (ii) all slacks $s_i^- = s_r^+ = 0$. If any of the slack variables are not equal to zero, then the DMU is deemed inefficient. The decision variables are the Lagrange multipliers and θ in the dual problem. Constraint set 1 ensures that the weighted sum of the inputs (or the convex combination of input values) of all DMUs is equal to the inputs of the DMU under investigation multiplied by the efficiency score. Constraint set 2 ensures that the weighted sum of the outputs (or the convex combination of output values) of all DMUs is equal to the outputs of the DMU under investigation. Constraint set 3 is the non-negativity constraints for the decision variables.

The **CCR – Output Model** would modify the objective function from maximize to minimize to obtain the follow primal and dual problems, respectively:

CCR – Output Model

Primal (Multiplier Model)

$$\underset{\vec{u},\vec{v}}{\text{minimize}} \sum_{i=1}^{m} v_i x_{io} \tag{2-8}$$

subject to

$$\sum_{i=1}^{m} v_i x_{ik} - \sum_{r=1}^{s} u_r y_{rj} \geq 0 \quad \text{for all } j \in S_{DMU}$$

$$\sum_{r=1}^{s} u_r y_{ro} = 1$$

$$u_r, v_i \geq \varepsilon > 0 \qquad \text{for } 1 \leq i \leq m, 1 \leq r \leq s$$

Dual (Envelopment Model)

$$\underset{\varphi,\vec{\lambda},s^-,s^+}{\text{maximize}} \varphi + \varepsilon\left(\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+ \right) \tag{2-9}$$

subject to:

$$\sum_{j=1}^{n} \lambda_j x_{ij} + s_i^- = x_{io} \qquad i = 1,2,\ldots,m$$

$$\sum_{j=1}^{n} \lambda_j y_{rj} - s_r^+ = \varphi y_{ro} \qquad r = 1,2,\ldots,s$$

$$\lambda_j \geq 0 \qquad j = 1,2,\ldots,n$$

BCC – Linear Input Model

Primal (Multiplicative Model)

$$\text{maximize} \sum_{r=1}^{s} \mu_r y_{ro} - u_o \tag{2-10}$$

subject to:

$$\sum_{r=1}^{s} \mu_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} - u_o \leq 0 \qquad \text{for all } j \in S_{DMU}$$

$$\sum_{i=1}^{m} v_i x_{io} = 1$$

$$\mu_r, v_i \geq \varepsilon \qquad \text{for } 1 \leq i \leq m, 1 \leq r \leq s$$

$$u_0 \text{ free}$$

The BCC model differs only from the CCR model because of the addition of the variable u_o in the primal problem and the convexity constraint $(\vec{1}\lambda = 1)$ in the dual. The u_o variable is the y-intercept of the line (or hyperplane) passing through the point in the production possibility set representing the DMU under investigation, in the single (or multiple) input case. This term allows for the presence of variable returns-to-scale and measures scale efficiency. The u_o variable is free because the y-intercept can be positive, negative, or equal to zero. Refer to section 2.3.1.1 for a more detailed explanation of return-to scale.

Dual (Envelopment Model)

$$\underset{\vec{\lambda},\theta,\vec{s}^{-},\vec{s}^{+}}{\text{minimize}} \theta - \varepsilon \left(\sum_{i=1}^{m} s_i^{-} + \sum_{r=1}^{s} s_r^{+} \right) \tag{2-11}$$

subject to:

$$\sum_{j=1}^{n} \lambda_j x_{ij} + s_i^- = \theta x_{io} \qquad i = 1,2,\ldots,m$$

$$\sum_{j=1}^{n} \lambda_j y_{jj} - s_r^+ = y_{ro} \qquad r = 1,2,\ldots,s$$

$$\sum_{j=1}^{n} \lambda_j = 1$$

$$\lambda_j, s_i^-, s_r^+ \geq 0$$

The presence of the convexity constraint $(\vec{1}\lambda = 1)$ in the dual problem is a consequence of variable returns-to-scale. Ensuring convexity allows for the formulation of a piece-wise convex combinations of all DMUs in the possibility set. The convex combination of the efficient DMUs will uncover the "envelope". The envelope is the boundary of the PPS that contains of all non-dominated DMUs.

The **BCC – Output Model** follows the same transformation previously shown for the CCR – Output model. The **Invariant Multiplicative**, **Variant Multiplicative**, and **Additive** models are variations to the CCR and BCC models. The Invariant Multiplicative model is a contrast to the BCC model by the change from a linear to a log-linear efficient frontier. Following suit, the Variant Multiplicative model is a contrast to the CCR model by the change from a linear to a log-linear efficienct frontier. The Additive model is modeled as the BCC model without the normalizing constraint ($\sum_{i=1}^{m} v_i x_{io} = 1$).

4. EFFICENCY-BASED RESOURCE ALLOCATION MODELS

There have been previous approaches to efficiency-based resource planning in the literature. This section gives a brief synopsis of previous models. These models use DEA for resource allocation. DEA, by its design, was not intended for resource allocation but for measuring relative efficiency of service units. Table 4 highlights differences in technology function assumptions for efficiency-based resource planning models.

Golany, Phillips et al. (1993): The purpose of this paper is to use DEA to allocate a resource (i.e. a budget) in a way that is conducive to meeting overall organizational goals of effectiveness and efficiency. A central authority is assumed to allocate the budget and to set organizational goal targets. The authors develop three families of models, each under a different managerial situation. All models assume that an Additive DEA study was previously run and thus the relative efficiency ratings are known. There is an implicit production function that is ensured by the bounds on the allowable proportional changes in resources.

Golany and Tamir (1995): This paper focuses on using DEA to measure effectiveness, efficiency, and equity among service units. Effectiveness is defined as overall goals of the organization. Efficiency is defined as how well each service unit achieves set goals. Equity is defined as the proportionate allocation of resources to different service units. The purpose of the paper is to build a resource allocation model (DEA-RAM) to analyze trade-offs among effectiveness, efficiency, and equity. The authors use the Additive DEA model to maximize overall effectiveness of the organization. This is a multiple inputs - single output goal programming model. The authors use an empirical, multiplicative production function.

Athanassopoulos (1995): This paper focuses on integrating Goal programming and DEA (GoDEA) for multilevel planning problems. The purpose of the paper is to use GoDEA for services planning where there is a trade-off among organizational objectives. These objectives are efficiency (via resource allocation), effectiveness and equity. The author extends work by Thanassoulis and Dyson (1992) to handle global organizational input-output targets. The model allocates resources (i.e., grant monies) to service units via a centralized authority. The central authority sets global targets for controllable inputs and outputs while the remaining inputs and outputs are estimated based on the solution to the model. The authors note that these targets may not be achievable by GoDEA, so global targets used by GoDEA are "close to" the decision maker's preferences. There is an implicit technology that is ensured by global target setting.

Thanassoulis (1996): This paper focuses on the allocation of resources in education and health care services where the allocation is based on marginal resource levels. Marginal resource levels (MRLs) "are rates of resource entitlement per unit of each activity or output of an operating unit". In education this can represent the funding entitlements per pupil. The purpose of the paper is to develop a method for estimating MRLs which lead to DEA-efficient resource

allocation among service units. The method presented in this paper clusters the DMUs by the ratio its output levels are to one another. This is a constant returns-to-scale model. The authors do not assume an explicit production function. They use proportional scaling of inputs and outputs.

Fare et al. (1997): This paper focuses on the measuring of efficiency resulting from the reallocation of fixed resources. The purpose of the paper is to develop a model that measures efficiency by allowing for the allocations of a fixed, shared input (i.e., farmland). The authors develop a DEA – CCR output-oriented, multiplier model for measuring efficiency. The authors assume an empirical production function.

Athanassopoulos (1998): The purpose of this paper is to develop a model for resource allocation and target setting where efficiency, effectiveness, and equity are objectives. The author uses DEA and goal programming to achieve resource allocation. The author uses the principal-agent paradigm. The principal is the central decision maker and the agents are the service units. The resource allocation model is named the Target-Based Resource Allocation (TARBRA) model. The author uses DEA to estimate production function coefficients.

Kumar and Sinha (1999): The purpose of this paper is to use DEA to arrive at an efficiency-based production plan for situations where there are not requirements on data about costs of inputs or prices of outputs. The authors develop two production planning models which maximize the efficiency of each service unit. The first model is an input-oriented DEA-CCR model and the second model is an output-oriented DEA-CCR model. Each time period in the planning horizon is considered a service unit and units are linked by the inventory carried between them. The optimization is done over all service units simultaneously. This means the model is only run once, which differs from traditional DEA studies where the model is run for each service unit. The implicit production function is maintained by a set of constraints.

Beasley (2003): The purpose of this paper is to present general models for fixed cost allocation and resource allocation where the objective is to maximize average service unit efficiency. The author presents two models. The first linear model is for allocating fixed costs across multiple DMUs and the second linear model reallocates resources to achieve maximize

average service unit efficiency based on organizational goals. There is an implicit production function which is ensured by output targets and resource allocation limits.

Korhonen and Syrjanen (2004): The purpose of the paper was to develop a resource allocation model where there is a central decision maker that simultaneously controls all units. The authors perform a DEA study that is modeled as a multiple objective linear program (MOLP). In this model the decision maker seeks to maximize outputs while simultaneously minimizing the inputs. The current input – output values are used to construct the production possibility set (PPS) and units are only allowed to adjust their plans within that set. This paper extends Golany et al. (1993) and Golany and Tamir (1995) by making the problem an MOLP and by the addition of rules placed on the allowable movement of units within the PPS. The model was formulated output-oriented version of the models under constant and variable returns-to-scale. A Pareto Race is used to solve the optimization problem in the first case, where two objectives (targets) are set by the decision maker. Once the most preferred solution (the final solution) is found, then the decision maker has solved the allocation problem. There is an implicit production function. Since only proportional scaling of inputs and outputs are allowed, it can be assumed that the production function is held fixed.

Lozano and Villa (2004): The purpose of this paper is to develop general models for allocating resources where there is a centralized decision maker. The models seek to maximize efficiency of all service units simultaneously. Two BCC models are developed that project all service onto the efficient frontier. These models are differing from other DEA-based resource allocation models because they consider aggregate input reduction and aggregate output production. Major differences from conventional DEA: (1) instead of solving an independent LP model projecting each DMU – all DMUs are simultaneously projected, and (2) instead of reducing the inputs of any one DMU – aim to reduce the total input consumption of all DMUs. All inputs & outputs are assumed to take on quantitative values in a cardinal scale and are assumed to be discretionary. This paper assumes that the "parametric function connecting inputs to outputs is unknown", it is assumed that any of the existing units can be projected onto any feasible operation point belonging to the efficient frontier of the possibility set corresponding to the assumed technology. Lozano and Villa (2005): This paper is a minor modification to the

2004 paper with the allowance of a DMU to be closed. Downsizing (or closing) a DMU could be seen as a firing policy.

Asmild et al. (2006): The purpose of this paper is to develop a resource allocation model for an organization with a central controller. The central controller controls all service units. The paper extends Lozano & Villa (2004) and Lozano et al. (2005). The authors consider the case where only inefficient service units are allocated additional resources. The Lozano and Villa models can potentially relocate already efficient DMUs. This is a BCC model. In the "pre-processing stage" the DEA Additive model is run and the efficient and inefficient DMUs are identified. Then the Lozano and Villa (2004) model is run with only the inefficient DMUs. The authors offer extensions to their model by including non-transferable outputs and strictly non-discretionary variables. There is an unknown production function. Service units are only allowed to onto the efficient frontier of the PPS.

Golany et al. (2006): The purpose of this paper is to develop a DEA based model where the efficiency of the subsystems and the aggregate system is determined. The authors develop a deterministic linear acquisition model for the purposes of resource allocation among subsystems and a deterministic linear model for measuring efficiency of an aggregate system. The aggregate system is comprised of two subsystems (or stages) in tandem. Each subsystem uses capital and labor to produce a final product. Subsystem 1 produces an intermediate product which is an input into subsystem 2. The authors assume that this intermediate product can be sold for the same price on the open market that it charges subsystem 2. Resources can be exchanged between subsystems only when it is beneficial to both. Constant returns-to-scale are assumed throughout. This paper used a Cobb-Douglas production function to calculate the output for each subsystem.

Author(s)	Technology Identification	Technology form	Allowable Technology Changes	Allowable Scale Changes	Returns to Scale
Golany et al. (1993)	Implicit		No	Yes	VRS
Golany & Tamir (1995)	Explicit	Multiplicative	Yes	Yes	VRS
Athanassopoulos (1995)	Implicit		Yes	No	CRS
Thanassoulis (1996)	Implicit		No	No	CRS
Athanassopoulos (1998)	Explicit	Nonlinear	No	Yes	VRS
Fare et al. (1997)	Implicit		No	No	CRS (can be modified to VRS)
Kumar & Sinha (1999)	Implicit		Yes	No	CRS (can be modified to VRS)
Beasley (2003)	Implicit		No	No	CRS (can be modified to VRS)
Lozano & Villa (2004)	Implicit		No	Yes	VRS (can be modified to CRS)
Korhonen & Syrjanen (2004)	Implicit		No	No	CRS, VRS
Lozano & Villa (2005)	Implicit		No	Yes	VRS (can be modified to CRS)
Asmild et al. (2006)	Implicit		No	Yes	VRS
Golany et al. (2006)	Explicit	Cobb-Douglas	No	No	CRS

Table 4: Technology functions in DEA-based resource planning models (Created by author)

The models presented in this dissertation differ from the aforementioned models in several ways:

1. This research uses a stochastic technology functions for resource planning.
2. This research uses controllable and uncontrollable inputs in the models.
3. The resource planning models are of service supply chains (i.e., multi-stage and multi-service).

CHAPTER 3

THE EFFECTS OF EFFICIENCY AND QUALITY ON RESOURCE PLANNING FOR CO-GENERATED SERVICES

Abstract

In this paper we present a model for resource planning of the operations of a service enterprise in which the degree of client co-generation strongly influences the efficiency and quality of the enterprise. The purpose of our model is to develop service-operations plans that allow service firms to determine the optimal mix of client involvement, workforce size, and service output. Using this model we can determine the optimal level of client involvement and observe the effects of this participation on service efficiency and quality. It is shown that there are different cases for the form of the optimal policy.

Keywords: service operations, resource planning, client co-generation, service efficiency

1. Introduction

The objective of this paper is to examine the effects of client involvement on resource planning decisions when a service firm is faced with efficiency and quality considerations. We develop a non-linear, deterministic, single-period planning model that allows for examination of trade-offs among client involvement, efficiency and quality. Our model contributes to the literature as follows:

- To our knowledge, this is the first attempt to incorporate efficiency and quality performance measures, which are functions of client involvement, into a resource planning problem.
- We extend conventional resource planning approaches to include the extent of client involvement as a policy variable.
- We examine both theoretically and experimentally the effects on policy of varying levels of the workforce, efficiency and quality.

We position our model at the operational level of the business decision hierarchy. To date, most research into service management has been done on strategic decisions. Our research responds to the need for modeling of tactical and operational decisions (Machuca et al., 2007). Specifically, we focus on modeling operational problems which span the functional areas of generation of service and human-resource management.

Resource planning is a sequential decision process that strives to apply an organization's capacity most efficiently to meet demand (Holt et al., 1995). The plan covers a horizon of 6 to 12 months and aims to find optimal decisions concerning generation quantities, resource levels, inventory, and backorders. The typical context of the resource plan for a service provider, such as a software consulting firm, presents a planning horizon of 12 months, with monthly planning intervals. Demand patterns are typically non-stationary, showing seasonality and trend. In keeping with conventional practice in resource planning, we accommodate the uncertainty in demand through a deterministic forecast coupled with planned buffers and firming of the plan through rolling-horizon updates.

Traditionally, resource plans seek to minimize costs subject to constraints on capacity, workforce, and inventory. We have extended the conventional resource planning model to include the client as a source of direct labor. We have also added efficiency and quality measures which are functions of client involvement, client and worker skills, and the quality of the inputs to each service process. The efficiency function augments the customary capacity constraint to make our model realistic and relevant to business services. Traditional resource planning models have a linear objective function and constraints (Holt et al., 1995). Our model, however, introduces nonlinearity into the constraints with the inclusion of the efficiency and quality functions.

There are numerous mathematical models for the services sector. See Rust and Metter (1996) for a very comprehensive review of some of the most widely recognized models used by marketing researchers (Rust and Metters 1996). Models of services operations planning are primarily directed at the health care and transportation sectors and do not involve the aspect of client participation. In Table 1 we compare several service operations models that are relevant to our research and show how our model differentiates itself from the others.

	Ittig (1994)	**Gaimon (1997)**	**Soteriou and Hadjinicola (1999)**	**Napoleon and Gaimon (2004)**	**Anderson et al. (2006)**	**White and Badinelli**
# of stages	Single	Single	Multiple	Multiple	Multiple	Single-staged
Objective	Profit	Profit	Minimize loss of service quality perceptions	Profit	Costs	Profit
Resource capacity changes	Not considered	Considered	Not considered	Considered	Considered	Considered
Resource training/learning	Not considered	Considered	Not considered	Considered	Not considered	Not considered
Customer waiting	Considered	Not considered	Considered through responsiveness factor	Not considered	Considered through backlogging	Considered through backlogging
Inventory	Not considered	Not considered	Not considered	Not considered	Not considered	Considered
Client Involvement	Not considered	Not considered	Not considered	Not considered	Not considered	Considered
Demand	Stochastic	Deterministic	Not considered	Not considered	Stochastic	Deterministic
Methodology	Non-linear optimization	Optimal control theory	Non-linear optimization	Optimal control theory	Optimal control theory	Dynamic Programming
Service sector	Retail	None specified	Healthcare	Various examples	Oilfield service	Professional Services

Table 1: Comparison of Service Models (Created by author)

Client involvement greatly influences the service creation process. Clients can provide information to the service process and/or can be used as resources in the service process. One example of client-provided information is the medical history that a patient gives to her doctor. Without this information, the doctor will not have the preliminary knowledge needed to accurately diagnose the patient. Using the client as a resource is often referred to as client co-generation of a service. Co-generation requires the physical presence of the client in the service creation process; see Sampson and Froele (2006). Client co-generation is an essential element to this decision model.

We model efficiency and quality as a function of client involvement. Efficiency and quality increases as client involvement increases. It is assumed that the client brings certain skills, knowledge, and motivation to the service process. Capturing the effects of client involvement makes this model different from any others in literature.

Researchers are still debating the definition of a service. Over the past few decades there have been numerous attempts to define a service. The following is a sample of the most commonly used definitions of a service.

- Sasser, Olsen et al. (1978): "Intangible and perishable… created and used simultaneously".
- Lovelock (1983): A service is "characterized by its nature (type of action and recipient), relationship with customer (type of delivery and relationship), decisions (customization and judgment), economics (demand and capacity), mode of delivery (customer location and nature of physical or virtual space)".
- Murdick et al. (1990): Services can be defined as economic activities that produce time, place, form, or physiological utilities.
- Chase and Aquilano (1992): A service business is the "management of organizations whose primary business requires interaction with the customer to produce the service."
- Fitzsimmons (2001): "A time-perishable, intangible experience performed for a customer acting in the role of co-producer."

- Spohrer, Maglio et al. (2007): "….service is a kind of action, performance, or promise that's exchanged for value between provider and client."
- Vargo, Akaka (2009): "…function of service systems is to connect people, technology and information through value propositions with the aim of cocreating value for the service systems participating in the exchange of resources within and across systems."

We conclude that there are two primary elements needed to define a service, one is reliance on the client to produce the service and the other is the idea of shared creation of value. The reliance on the client is supported by Sampson and Froele's (2006) Unified Service Theory (UST). The UST distinguishes a service process from a non-service process by emphasizing that a service process relies on client inputs and clients as suppliers of information. Shared creation of value is the idea that a service can only be a service when there is a sharing of value between both the service provider and the client. The idea of Service-Dominant (SD) logic also embraces the idea of the co-creation of value; see Lusch and Vargo (http://www.sdlogic.net/). Acknowledging these essential elements and previous service definitions; we define a service as *the transformation of inputs into outputs such that value for the client is created through a process that utilizes capabilities and capacities of both the client and the provider.*

There are a variety of classification schemes for services. One of the earliest service typology is the Service Process Matrix (SPM) by R.W. Schmenner (1986). The SPM classifies services into four quadrants (service factory, service shop, mass service, professional service) by the level of client contact/customization and the level of labor intensity. Verma and Boyer (2000) studied management challenges among different service businesses in the four SPM quadrants. Their findings show that in reality there is not a clear line that distinguishes the types of services listed in the original Schmenner paper. Other classification schemes have focused on service output, the nature of the service, the role of the client, and the organizational structure. Cook et al. (1999) provide an outstanding summarization these classification schemes and they developed a unified representation of services based on commonalities they discovered among service typologies.

Another classification scheme is based on the amount of client involvement needed to produce the service. When there is a high-degree of client involvement, the term client-intensive

services is often used. This idea of client-intensive services or "high-contact" services was first introduced by Chase (1981) and Chase and Tansik (1983) in their discussion of the customer-contact model. This discussion showed how there are certain service businesses that rely heavily on client input and information. Kellogg and Nie (1995) developed a classification scheme that focused on customer influence in the service process. Some examples of client-intensive services are consulting, legal, software development services. The current paper focuses on service firms with a high degree of client involvement (e.g. professional services, business services).

2. The Resource Planning Decision

Trying to determine a resource plan can be a daunting task for managers. Consider a manager at a consulting service firm that has several clients being serviced across multiple industries. The manager's position requires her to determine which consultant to assign to each client based on client needs and the service firm's current workforce. She must also determine, based on the project plan, how much service to deliver to the client each period.

The decisions that the consulting firm manager is facing are all too familiar to operations managers. In operations literature these decisions are traditionally know as resource planning decisions or workforce management decisions. Managers typically develop a medium-range capacity plan over a 6-18 month planning horizon. This plan specifies an optimal combination of production levels, workforce levels, and inventory over the planning horizon. This research borrows concepts from manufacturing's approach to resource planning; however we are designing a model with service-specific elements.

This decision model determines the resource level sufficient to meet demand. It is assumed that the service firm has an initial staff of workers who are to be allocated to service jobs. Managers are allowed to hire more workers when needed, at a cost, and to lay-off workers when demand slows down. Due to limited fixed assets such as cubicles and/or computer workstations, there is an upper bound on the total workforce size. Similarly there is a lower bound on total workforce that represents the number of full-time employees needed to cover service agreements without exceeding a maximum backlog limit. In addition, this limit can represent any policies the company has on how low the workforce can go.

Managers must determine how much service to produce or which we call the service generation levels. The amount of service generation depends on the number of required service cycles. A service cycle is the required number of iterations of a service process. The cycle has a standard labor requirement and a standard lead time, which for the sake of simplicity of exposition, we set it to one. There is a maximum level of service generation that represents the level of generation needed to meet, but not exceed, demand. As well as a minimum level of service generation that represents the level of generation needed to meet demand without exceeding the maximum backlog. We assume that there is a maximum amount of allowable backlog that is set by the client or the service provider.

Since the clients are co-generators of the service, it is essential to include them as resources in the planning model. The client will supply the service firm with their workers as resources and information such as requirements and specifications. We specify the level of client involvement in a service process in terms of client intensity, defined as the ratio of the number of hours of client involvement in the process to the number of hours spent by the service provider in the service process. There is a limit to the amount of client intensity given by the client to the service provider in order to fulfill a service agreement. It is assumed that the client's resources have other obligations and therefore will only be able to provide a limited number of hours. As well as limit to the minimum level of client intensity required to deliver the service. This limit is determined by the minimum level of quality acceptable by the service provider, the client, or both. This boundary is needed by the service firm for benchmarking itself against competition and meeting basic client needs.

Decisions involve a cause-effect relationship between variables and performance measures. Constraints on service-provider and client capacities involve performance measures in this model. These constraints ensure that the amount of workforce that is staffed is adequate to meet demand. Constraints on backlog are enforced because some clients are only willing to wait so long before the service is completed. This backlog constraint is necessary to maintain client satisfaction. Service quality affects client satisfaction. Therefore, there is a constraint on how low service quality is allowed to drop.

Much literature supports the inclusion of client involvement as an argument of the efficiency and quality functions (Chase 1978; Fitzsimmons, 1985; Bowen, 1986; Mills and

Morris, 1986; Lance et al., 2002). The efficiency and quality of services are represented as nonlinear functions of client intensity. Client intensity is a function of client involvement and service generation. Client involvement is defined as the number of hours that client personnel are assigned to a service. Efficiency allows managers to determine service generation relative to the hours of client time.

Managers must be apprised of the efficiency of their workers because efficiency influences constraints on capacity. The amount of capacity needed to fulfill demand increases as efficiency decreases. The shape of the efficiency function is determined by parameters of the service process. By choosing different parameter values we are representing different service encounters. Each time the model is solved, it is solved for a unique client and service encounter.

Quality and backlog (i.e., client satisfaction) are interrelated performance measures which are affected by values of the decision variables. Client satisfaction is essential to any business that wants to stay competitive. In this model we capture client satisfaction through backlog and quality. If the client is more involved in the service creation process, then their perception of service quality will increase; see Soteriou and Hadjincola (1999). As client involvement increases, the service provider will see improved efficiency because there will be less speculation, misinterpretation, and miscommunication in the service process; see Soteriou and Hadjincola (1999). Like the efficiency function, the shape quality function is determined by parameters of the service process. Each time the model is solved, it is solved for a unique client and service encounter.

Costs (labor costs, hiring, firing, client, and backlog) are also performance measures in this decision model. Total labor costs are determined by the size of the workforce. The costs of changes to the size of the workforce are reflected in the total hiring and firing costs. The level of client intensity determines the total client costs. Total backlog costs are determined by any service generated backlog.

Parameters influence the cause-effect relationship between decision variables and performance measures. In this decision model there are several cost parameters (e.g., worker wages, client involvement, hiring and firing, backlog) that are reflected in the objective function and thus the performance measures. Additional parameters in this model are the limits/bounds on

the workforce, client involvement, backlog, and quality. Demand is a parameter that influences the cause-effect relationship between variables and parameters. There are also parameters that identify the process cycle requirements such as the standard labor hours required per service cycle and the standard number of cycles per service.

3. Modeling issues

The model constructs presented in this section contribute to the literature by providing ways in which the complex nature of services can be represented in a mathematical model. We describe the modeling challenges below.

Challenge 1 –Identifying efficiency and quality function characteristics

The efficiency and quality functions each strictly increase to a maximum value of 1.0 as a function of client intensity in the model. Literature reveals additional insights into the effects of client intensity on efficiency and quality. Chase (1978) determined that at some level, further client involvement is either ineffective or detrimental. Hence, there exists a level of client intensity at which efficiency reaches its maximum or saturation level. We did not model the saturation level, but we have included it as an efficiency characteristic. The effect of client intensity on quality does not reach a saturation level; the more the client is involved the better the quality of the service. The skill of the client and the service provider are also characteristics of the efficiency and quality functions. In addition, the quality from previous processes in the service process effect efficiency and quality of the current process, therefore the previous process's quality is also a characteristic.

Below, we summarize the characteristics of the efficiency and quality functions that follow from reasonable assumptions about the nature of services.

Efficiency function characteristics

1. Efficiency increases and exhibits diminishing marginal improvement as client involvement increases.

2. Efficiency increases at a diminishing marginal rate as a function of worker skill.

3. Efficiency increases at a diminishing marginal rate as a function of client skill.

4. Efficiency increases as a function of the quality of the preceding process.

5. There is a saturation level for client involvement such that, once this level has been reached, there are no more efficiency improvements.

Quality function characteristics

1. Quality increases and exhibits diminishing marginal improvement as client involvement increases.

2. Quality increases at a diminishing marginal rate as a function of worker skill.

3. Quality increases at a diminishing marginal rate as a function of client skill.

4. Quality increases as a function of the quality of the preceding process.

These assumptions determine characteristics of the shape of the efficiency and quality functions.

See Figures #1 and #2 for examples.

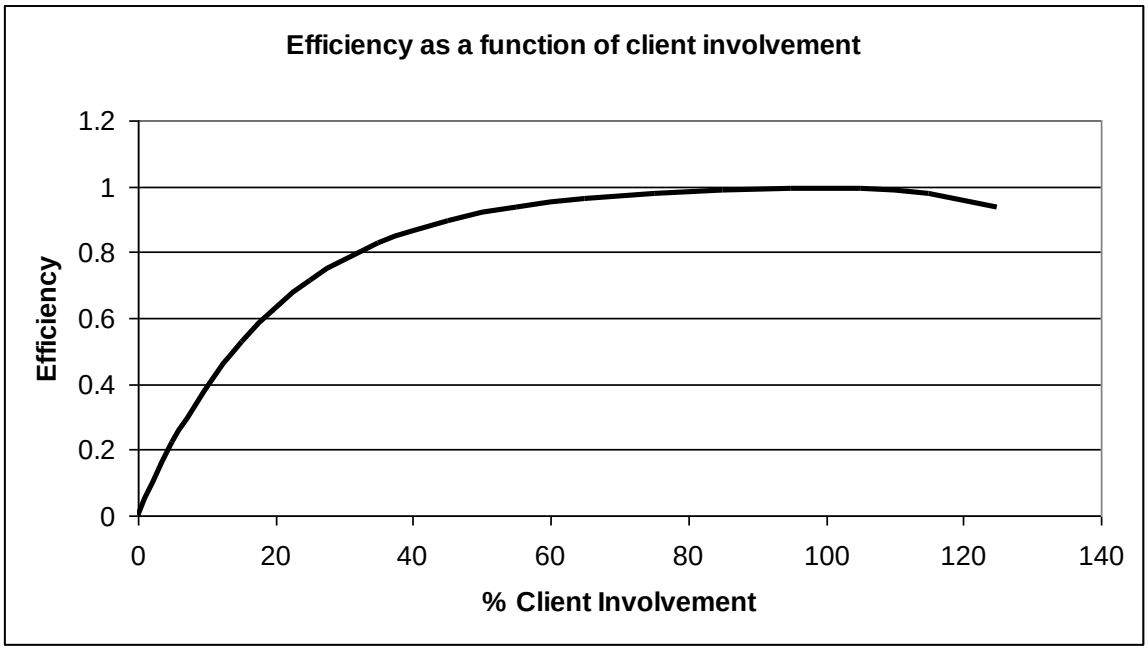

Figure 1: Example of an efficiency function (Created by author)

Figure 2: Example of a quality function (Created by author)

Challenge 2- How to take a service and quantify the volume of that service.

We define the deliverables of the service firm as "service types". We quantify the volume of a service as a *unit of service*. In the case of a software consulting firm, for example, service types can be database designs, webpage construction, code writing and testing, etc. A unit of service can be any code modules that make up the webpage, for example. We assume that the firm delivers each type of service through stages of processes. We assume that the precedence constraints among the processes that are required for a particular service type define a serial network of these processes. Hence, for each service type, there is a known serial chain of process stages. Each stage requires a certain number of "cycles" of a process per unit of the service that is delivered to the client. A cycle is the required number of iterations of a process.

Challenge 3- How are "inventory" and "backlog" represented?

There is no inventory in service processes, therefore we define "inventory" differently than conventional manufacturing processes. We define "inventory" as the number of cycles of a

stage, which have been completed but for which the succeeding stage has not started. Since this is a single-stage model, the parameter for the beginning inventory level represents the number of finished service cycles at the start of the period. The backlog variable measures the number of units of service that are not completed by their due dates.

Challenge 4- How are "inventory holding costs" and "backlog costs" measured?

The analogue of inventory holding cost is captured through the use of discounted cash flows in the objective function. If processes are completed prior to the times that they are needed, then the labor cost of generating these processes is recognized in the time periods of process generation and the revenue that is earned by the services for which these processes are generated is recognized at the due dates of the services. Hence, early generation will result is lower net present value. We will impose a cost on this backlog in order to capture the loss of goodwill penalties as well as the per-period cost of deferred revenues due to discounting of cash flows.

Despite our best efforts to make the model as realistic as possible we recognize and acknowledge its limitations. By borrowing concepts from manufacturing, we have opened the door to an over-simplification of the nature of services. Our services-focused planning model should be seen as a foundation for future service science research.

4. Descriptive Model

The intent of developing this model is to lay the foundation for future research; therefore we have assumed a single process stage, a single service job type, and a single time period. In future research we will show the relationship between this single-period model and a multi-period model and will try to illuminate the policies for the multi-period, multi-stage case based on the policies derived in this model.

The descriptive model is presented below.

4.1 Definitions and notation

Decision Variables

$g =$ generation of process cycles of the service (# of completed cycles)

$h =$ number of workers hired

$f =$ number of workers laid-off

$z =$ number of hours that client personnel are assigned the service

State Variables

$i =$ "inventory" of completed process cycles of the service at the end of a period (# of cycles)

$b =$ backlog of units of a service at the end of a period (# of units)

Performance measures

$w =$ size of workforce employed (# of workers)

$e =$ efficiency of the service as a function of client involvement

$q =$ quality of the service as a function client involvement

Parameters

Revenue & Cost Rates:

$c_h =$ cost of hiring a full-time regular employee

$c_f =$ cost of firing a full-time regular employee

$c_w =$ cost of wages per worker

$c_b =$ penalty cost of backlog of a service ($ per unit)

$c_z =$ cost of client involvement ($/labor-hour)

$v =$ revenue per unit of a service

The client intensity decision variable (y) is the ratio of client time/involvement (z) to the total worker time spent on the service ($r^h g$).

$$y = \frac{z}{r^h g} = \text{level of client intensity} \quad (1)$$

Other:

$d =$ forecasted demand of the service

$r =$ required number of cycles of the process per unit of service type

$r^h =$ number of standard labor hours required per cycle of the process of the service

$a^w =$ number of hours of worker availability per worker

$a^c =$ available client capacity for the service (hours)

$\bar{b} =$ maximum allowed backlog of the service (# units)

$\underline{y} =$ minimum required participation of client of a service as a fraction of standard process time spent on the service

$\bar{y} =$ maximum allowable participation of client of the service as a fraction of standard process time spent on the service

$\underline{q} =$ minimum required quality level

$$\underset{w,f,h,b,y,g}{\text{Maximize}} \quad dv - c_h h - c_f f - c_w w - c_b b - c_z r^h yg$$

Subject to:

$$i_0 + g - rd + rb = 0 \quad (0)$$
$$w - w_0 - h + f = 0 \quad (1)$$
$$\bar{w} - w \geq 0 \quad (2)$$
$$\bar{b} - b \geq 0 \quad (3)$$
$$a^w we(y) - r^h g \geq 0 \quad (4)$$
$$a^c - r^h yg \geq 0 \quad (5)$$
$$y - \underline{y} \geq 0 \quad (6)$$

$$\bar{y} - y \geq 0 \tag{7}$$

$$q(y) - \underline{q} \geq 0 \tag{8}$$

$$w, h, f, g, y \geq 0$$

The objective function maximizes profits and is expressed as revenue less the cost of hiring, firing, workforce level, backlog, and client involvement.

Constraint (0) is the balance constraint and expresses the inventory and backlog balance for the service. Through this constraint, any shortages of process completions of the service are recognized as backlog. Constraint (1) is a workforce balance constraint. Constraint (2) ensures that the maximum allowable workforce level is not exceeded. Constraint (3) ensures that the number of service jobs late (backordered) does not exceed the maximum allowable level.

Constraints (4) and (5) are the service provider and client capacity constraints, respectively. The efficiency term in the capacity Constraint (4) is needed to represent the effect of client involvement on the effective capacity of the workforce. If client involvement is negatively impacting efficiencies then the firm will not have enough capacity to meet forecasted demand.

In Constraints (6) and (7) we have set minimum and maximum amounts of client involvement, respectively. We assume that in business service such as consulting and IT development the client will always be part the service creation process, to a certain extent, and that the client cannot be the sole labor source in any process.

Typically organizations benchmark themselves against competitors in terms of quality and establish internal quality standards. Hence, constraint (8) imposes a minimum process quality level that must be achieved.

The inventory balance constraint, see Constraint (0) in the descriptive model above, can be written in terms of the backlog variable because the constraint is an equality constraint. We can remove the first constraint since it is an equality constraint. If we solve for b in terms of g, then we can replace b in the entire model with an expression in terms of g. The new expression

is $\frac{-i_0 - g + rd}{r} = b$ which represents the number of unfinished cycles at the end of the current period which is equal to the number of finished cycles at the start of the period less the number of generated cycles for the current period and the number of cycles required to produce the demanded number of service units. The new descriptive model is,

$$\underset{w,f,h,b,y,g}{\text{Maximize}} \quad dv - c_h h - c_f f - c_w w - \frac{c_b(-i_0 - g + rd)}{r} - c_z r^h yg$$

Subject to:

$$w - w_0 - h + f = 0 \tag{1}$$

$$\bar{w} - w \geq 0 \tag{2}$$

$$\bar{b} + \frac{i_0 + g - rd}{r} \geq 0 \tag{3}$$

$$a^w we(y) - r^h g \geq 0 \tag{4}$$

$$a^c - r^h yg \geq 0 \tag{5}$$

$$y - \underline{y} \geq 0 \tag{6}$$

$$\bar{y} - y \geq 0 \tag{7}$$

$$q(y) - \underline{q} \geq 0 \tag{8}$$

$$\frac{rd - i_0 - g}{r} \geq 0 \tag{9}$$

$$w, h, f, g, y \geq 0$$

4.2 Model Assumptions

Assumption 1: $e(y), q(y)$ are strictly increasing.

Assume that $q(y)$ is strictly increasing in y and approaches 1 perhaps asymptotically. Consequently, there is a unique value of y for which $q(y) = \underline{q}$. We shall denote this value y_q. Hence, Constraint #8 can be re-written as $y \geq y_q$. If $q(0) > \underline{q}$, then we set $y_q = 0$. A plot of this constraint's boundary in $y - g$ space is simply a vertical line at $y = y_q$.

Assumption 2: The parameters of the quality function and the maximum level of client involvement are set so that $y_q \leq \bar{y}$.

We need to ensure that the highest level of client involvement does not fail to provide the minimal level of quality, thereby ensuring a feasible solution to the resource allocation problem.

Assumption 3: a) The parameters of problem P1 must be set so that, if $y = y_{\max}$, then Constraint #4 places an upper limit on generation, $g_4(y_{\max})$ that is greater than $g_{\min}$. b) If $y = y_{\min}$, then Constraint #5 places a upper limit on generation, $g_5(y_{\min})$ that is greater than $g_{\min}$.

We must set the upper limit on generation greater than $g_{\min}$ so that we will not exceed the backlog limit enforced by Constraint #3.

Assumption 4: The parameters of Problem P1 must be set so that $g_{45} \geq g_{\min}$

Based on Assumption 3 and the fact that g_{45} is an upper bound on generation and that upper bound must be greater than $g_{\min}$.

Assumption 5: $\dfrac{c_b}{r} > c_z r^h \bar{y}$

By assuming the maximum cost of client involvement (i.e., client costs per labor hr x amount of client involvement x number of standard labor hours) to be less than the per cycle cost of backlog, then we prevent the automatic exchange of backorders for client involvement.

Assumption 6: $w \geq \dfrac{r^h g_{\min}}{a^w e(y_{\max})}$

$e(y_{\max})$ is the highest level of efficiency obtainable and $g_{\min}$ is the smallest number of cycles that can be generated. When the values of these two functions are true the workforce level can be reduced to a lower bound without violating the provider capacity constraint (Constraint #4). Assumption #6 mathematically forces this lower bound on the workforce level w .

5. Dynamic Program

We separate the hiring/firing/workforce decision variables from the rest of the decision variables by decomposing this problem into a two-stage dynamic program. Stage 1, the inner stage, is the optimization of the production plan for a given level of the workforce. The outer stage is the optimization of the combined workforce and production plans. The state variable that connects the outer stage to the inner stage is the workforce level, w .

Define,

$$z_1(w) = \frac{-c_b(-i_0 - g + rd)}{r} - c_z r^h yg \tag{2}$$

$$z_2 = z_1^*(w) - c_h h - c_f f - c_w w + revenue \tag{3}$$

First, we derive the solution of Problem P1 for the optimal production plan, given a workforce level, w . Once this solution is obtained, we can solve problem P2 to find the combination of the optimal workforce level and its associated production plan.

5.1 Optimality Conditions for Problem P1

The optimality conditions for Problem P1 are specified differently for cases and sub-cases. Service firms will find themselves in a specific case and sub-case based on parameter values and the value of service generation and client intensity at optimality. The cases and sub-cases stem from the shape of the feasible region. The shape of the feasible region is defined by the model constraints.

The constraints of Problem P1define a feasible region or a "box" when plotted in the $y-g$ plane. See Figure 3 for a graphical representation of the "box". Constraint #9 is the upper limit on generation ($g_{\max}$). Constraint #3 is the lower limit on generation ($g_{\min}$). Constraint #7 is the upper limit on client involvement ($y_{\max}$). Constraints #6 and #8 are the lower limit on

client involvement ($y_{\min}$). These boundaries will be vital in the determination of optimal policies.

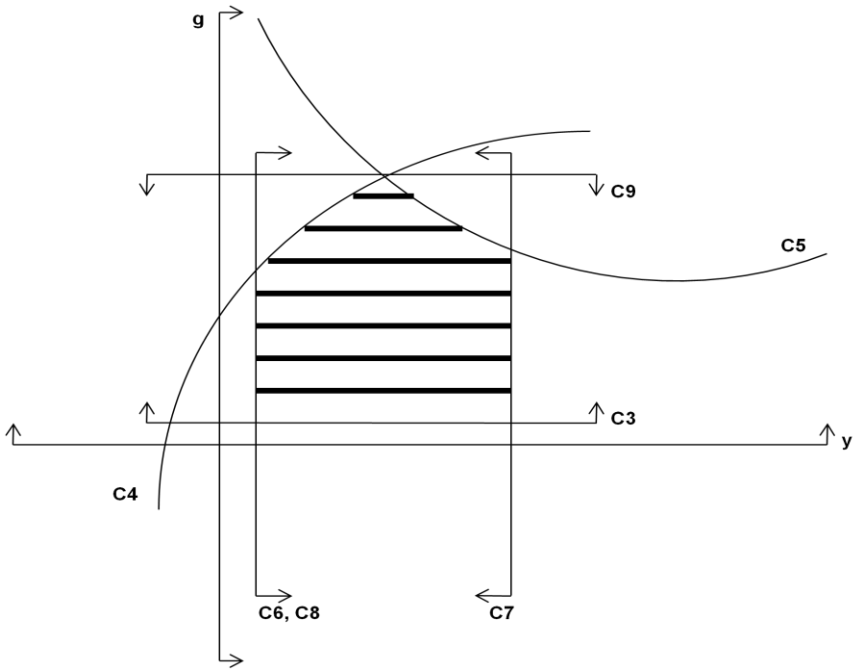

Figure 3: The feasible region for P1 (Created by author)

Constraints #4 and #5 are superimposed on the "box". Constraint #4 is the service provider's capacity constraint when binding this constraint can shift to the left (right) along the horizontal axis as the workforce level increases (decreases). This constraint is a strictly increasing function of y . Constraint #5 represents a constraint on client capacity and is not a function of the workforce level. Constraint # 5 is strictly decreasing in y . The intersection of Constraints #4 and #5 (y_{45}, g_{45}) plays an important role in determining the optimal solution. This intersection is defined by, y_{45} is the value of, *y*, that satisfies $ye(y)$. Where $ye(y) = \frac{a^c}{a^w w}$ and $g_{45} = \frac{a^w w}{r^h} e(y)$. Since *y* is strictly increases from 0 without bound and is positive in y and $e(y)$ is strictly increasing and positive in y , $ye(y)$ is strictly increasing in y .

Proposition #1 proves that at the intersection of these Constraint #4 and #5, with respect to g, is highest at g_{45}. Proposition #1also proves that at this upper limit on generation, client intensity is highest at y_{45}. The proof of Proposition #1 is in Appendix A.2.

Proposition 1: *The upper bound on generation imposed by Constraints #4 and #5 over all values of the level of client involvement is highest at* y_{45}. *That is,*

$$g \le \frac{a^w w}{r^h} e(y), g \le \frac{a^c}{r^h y} \Rightarrow g \le g_{45} \tag{4}$$

By determining which constraints are binding at optimality, we can better understand optimal policies. The proof of Proposition #2 is in Appendix A.3.

Proposition 2: *At optimality, one or more of Constraints #4, #5 or #9 is binding.*

The next three propositions give us insights into the behavior of the objective function. The proofs of Proposition #3 - #5 are in Appendix A.4 – A.6 respectively.

Proposition 3: *Along the arc formed by the boundary of Constraint #5, the objective function is decreasing in* y

Proposition 4: *Along the arc formed by the boundary of either Constraint #3 or Constraint #9, the objective function is decreasing in* y.

Proposition 5: *Along the arc formed by the boundary of Constraint #4, the objective function is concave in* y.

Lemma #1 is a stepping stone in establishing which constraint is binding at optimality. The proof of Lemma #1 is in Appendix A.1.

Lemma 1: *If* $y_{45} > y_{\min}$ *and* $y_4(g_{\max}) > y_{\min}$, *then the optimal solution must lie on Constraint #4.*

By Lemma #1, we have proven that the optimal solution to Problem P1 must lie on Constraint #4 and we are now motivated to look closer at Constraint #4. We set an upper bound

on y^* for the capacity constraint by defining $\bar{y}_4 = \min(y_4(g_{\max}), y_{\max}, y_{45})$. We also set a lower bound on y^* for the capacity constraint by defining $\underline{y}_4 = \max(y_4(g_{\min}), y_{\min})$. Hence, the optimal solution to Problem P1 prescribes the smallest or the largest possible value of y permitted by the service provider capacity constraint and the "box". By Proposition #6, $\bar{y}_4$ and $\underline{y}_4$ are decreasing in w. See Appendix A.7 for the proof of Proposition #6.

Proposition 6: $\bar{y}_4$, $\underline{y}_4$ *are decreasing in* w.

Propositions 7 and 8 determine how $y^*(w), g^*(w), z_1^*(w)$ move in relation to the workforce level. The proofs of these propositions are in Appendix A.8 and A.9 respectively.

Proposition 7: y^* is decreasing in w. g^* is increasing in w.

Proposition 8: $z_1^*(w)$ is increasing in w

If Constraint #4 is redundant, then the optimal solution must be on Constraint #5 or #9. See Appendix A.10 for the proof of Theorem #1.

Theorem 1: *If* $\bar{y}_4 < y_{\min}$, *then* $y^* = y_{\min}$, $g^* = \min(g_{\max}, g_5(y_{\min}))$

If Constraint #4 is not redundant, then there are three cases for the form of the optimal policy, which are defined in terms of the value of $\hat{y}$ relative to $y_{\min}$ and $y_{\max}$. Consequently, we should undertake a study of the dependency of $\hat{y}$ on model parameters. $\hat{y}$ is the point on Constraint #4 where the objective function is at its maximum amount. The condition that defines $\hat{y}$ is,

$$e(\hat{y}) = \left(\frac{c_b}{rr^h c_z} - \hat{y} \right) \left. \frac{\partial e}{\partial y} \right|_{y=\hat{y}} \tag{5}$$

As $\hat{y}$ is determined by the parameters, c_b, r, r^h, c_z and the shape of the efficiency function, $e(y)$, we can think of $\hat{y}$ as a derived parameter of each problem configuration. That is, for any service enterprise that is modeled, there is a fixed value of $\hat{y}$.

Theorem 2: *If* $\bar{y}_4 \geq y_{\min}$, *then*

$(i) y^* = \underline{y}_4, g^* = g_4(y^*), \hat{y} < \underline{y}_4$

$(ii) y^* = \hat{y},\ g^* = g_4(y^*), \underline{y}_4 \leq \hat{y} \leq \bar{y}_4$

$(iii) y^* = \bar{y}_4, g^* = g_4(y^*), \hat{y} > \bar{y}_4$

$$z_1(w) = -c_b\left(\frac{rd - i_0}{r} - \frac{g^*}{r}\right) - c_z r^h y^* g^*$$

Define, $\bar{y}_4 = \min(y_4(g_{\max}), y_{\max}, y_{45})$

$\underline{y}_4 = \max(y_4(g_{\min}), y_{\min})$

$\hat{y}$ is a point on Constraint #4 (see Proposition #5)

The proof of Theorem #2 is shown in Appendix A.11.

Theorem #2 is very important in the analysis of the optimal solution. By Lemma #1, we know that the optimal solution lies on Constraint #4. Depending on where $\hat{y}$ lies on Constraint #4, there are three cases to evaluate. If $\hat{y} < \underline{y}_4$, then $\hat{y}$ is not a feasible point, therefore $y^* = \underline{y}_4$ (Case 1). Likewise if $\hat{y} > \bar{y}_4$ then $\hat{y}$ is not a feasible point, therefore $y^* = \bar{y}_4$ (Case 3). In Case 2, $\bar{y}_4 \leq \hat{y} \leq \underline{y}_4$ therefore $\hat{y}$ is a feasible point. The optimal level of service generation will also lie on Constraint #4, $g^* = g_4(y^*)$.

Theorem #2 of the Optimality Analysis, when applied to each of these cases allows us to define sub-cases for different intervals of the state variable, w (see Figure 4). The solutions for Problem P1 over different intervals of the state variable, w, bind Constraint #4 with either the

generation, g, or the client intensity, y, fixed. Therefore, we are interested in the dependency of the optimal value of the variable that moves with respect to w.

The three cases and sub-cases are explained below.

Case 1: $\hat{y} \leq y_{\min}$

By Theorem 2, we can identify a specific form for the optimal solution to Problem P1, $(y^*(w), g^*(w))$ over each of two intervals of the variable, w.

Define w_1 = the value of the workforce level at which the capacity constraint is binding at the point $(y_{\min}, g_{\min})$. That is, $w_1 = \dfrac{r^h g_{\min}}{a^w e(y_{\min})}$

Sub-case 1.1: $\underline{w} \leq w < w_1$

In this sub-case, the optimal form of the solution to Problem P1 is,
$(y^*(w), g^*(w)) = (y_4(g_{\min}), g_{\min})$

Sub-case 1.2: $w_1 \leq w < \overline{w}$

In this sub-case, the optimal form of the solution to Problem P1 is,
$(y^*(w), g^*(w)) = (y_{\min}, g_4(y_{\min}))$

Case 2: $y_{\min} < \hat{y} \leq y_{\max}$

By Theorem 2, we can identify a specific form for the optimal solution to Problem P1, $(y^*(w), g^*(w))$ over each of three intervals of the variable, w.

Define w_1 = the value of the workforce level at which the capacity constraint is binding at the point $(\hat{y}, g_{\min})$. That is, $w_1 = \dfrac{r^h g_{\min}}{a^w e(\hat{y})}$.

Define $w_2 =$ the value of the workforce level at which the capacity constraint is binding at the point $(\hat{y}, g_{\max})$. That is, $w_2 = \dfrac{r^h g_{\max}}{a^w e(\hat{y})}$.

Sub-case 2.1: $\underline{w} \le w < w_1$

In this sub-case, the optimal form of the solution to Problem P1 is,

$(y^*(w), g^*(w)) = (y_4(g_{\min}), g_{\min})$

Sub-case 2.2: $w_1 \le w < w_2$

In this sub-case, the optimal form of the solution to Problem P1 is,

$(y^*(w), g^*(w)) = (\hat{y}, g_4(\hat{y}))$

Sub-case 2.3: $w_2 \le w < \overline{w}$

In this sub-case, the optimal form of the solution to Problem P1 is,

$(y^*(w), g^*(w)) = (y_4(g_{\max}), g_{\max})$

Case 3: $y_{\max} < \hat{y}$

By Theorem 2, we can identify a specific form for the optimal solution to Problem P1, $(y^*(w), g^*(w))$ over each of two intervals of the variable, w.

Define $w_1 =$ the value of the workforce level at which the capacity constraint is binding at the point $(y_{\max}, g_{\max})$. That is, $w_1 = \dfrac{r^h g_{\max}}{a^w e(y_{\max})}$.

Sub-case 3.1: $\underline{w} \le w < w_1$

In this sub-case, the optimal form of the solution to Problem P1 is,

$(y^*(w), g^*(w)) = (y_{\max}, g_4(y_{\max}))$

Sub-case 3.2: $w_1 \le w < \overline{w}$

In this sub-case, the optimal form of the solution to Problem P1 is,

$(y^*(w), g^*(w)) = (y_4(g_{\max}), g_{\max})$

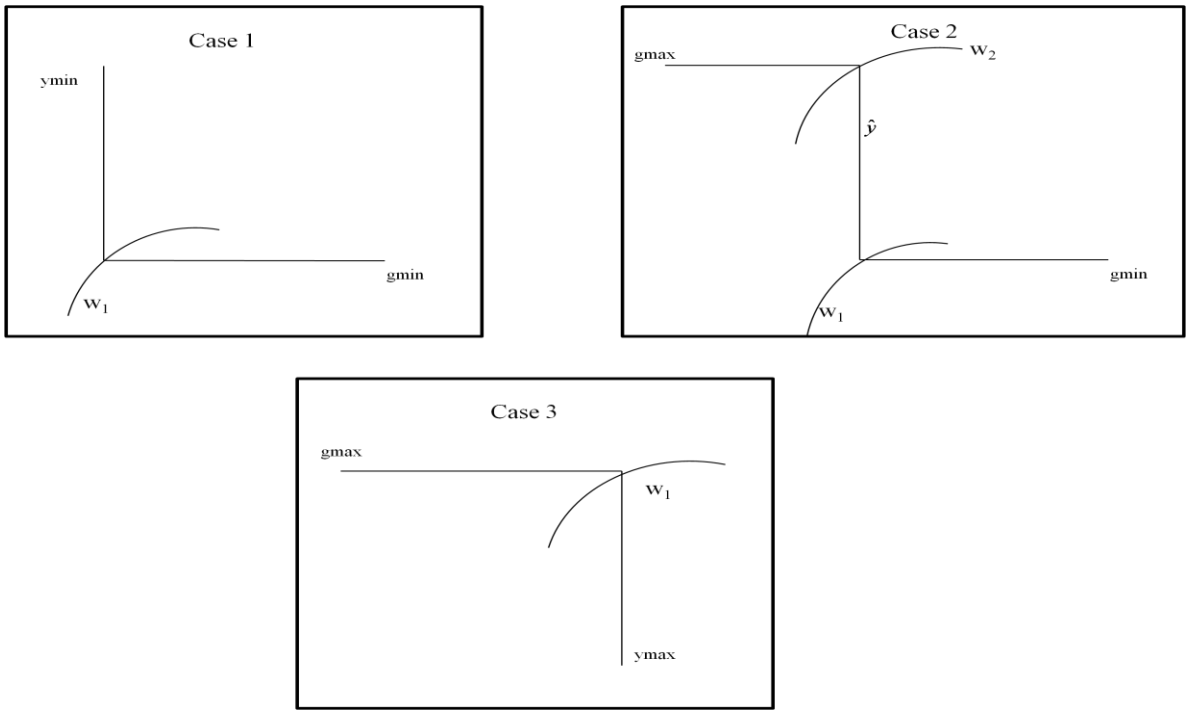

Figure 4: Cases for Problem P1 (Created by author)

Due to the complex nature of Theorem #2, an algorithm was developed to show the steps in obtaining the optimal solution.

Algorithm:

1. Determine $y_{\min}, y_{\max}, g_{\min}, g_{\max}$
2. Determine $\hat{y}_4, \underline{y}_4, \bar{y}_4$
3. If $\bar{y}_4 < y_{\min}$, then $y^* = y_{\min}$, $g^* = \min(g_{\max}, g_5(y_{\min}))$, else

$$(i) y^* = \underline{y}_4, g^* = g_4(y^*), \hat{y} < \underline{y}_4$$
$$(ii) y^* = \hat{y},\ g^* = g_4(y^*), \underline{y}_4 \leq \hat{y} \leq \bar{y}_4$$
$$(iii) y^* = \bar{y}_4, g^* = g_4(y^*), \hat{y} > \bar{y}_4$$

5.2 Optimality Conditions for Problem P2

Problem P2:

$$\max_{w,f,h} \text{imize } z_2 = -c_w w - c_h h - c_f f + z_1^*(w)$$

subject to:

$$w - w_0 - h + f = 0$$

$$\bar{w} - w \geq 0$$

$$w, f, h \geq 0$$

Using the equality constraint to eliminate one variable, we obtain the following optimization problem.

Problem P2'

$$\max_{w,f} \text{imize } z_2 = -(c_w + c_h) w - (c_h + c_f) f + z_1(w) + c_h w_0$$

$$\bar{w} - w \geq 0 \quad (1)$$

$$w - w_{\min} \geq 0 \quad (2)$$

$$w + f - w_0 \geq 0 \quad (3)$$

$$f \geq 0 \quad (4)$$

The KKT conditions for Problem P2 are as follows. See Winston (1990 pg. 651 – 652) for the theorem for the necessary condition for a maximization problem.

$$z_1^*(w) = dv - c_b\left(\frac{\text{rd} - \text{i}_0}{\text{r}} - \frac{g}{r}\right) - c_z y r^h g = cons\tan t + \frac{c_b}{r} g^* - c_z r^h y^* g^*$$

$$\frac{\partial L}{\partial w} = 0 \Rightarrow \left.\frac{dz_1^*}{dw}\right|_{w=w^*} - (c_w + c_h) = \lambda_1 - \lambda_2 - \lambda_3 \quad (1)$$

$$\frac{\partial L}{\partial f} = 0 \Rightarrow \lambda_3 + \lambda_4 = c_h + c_f \quad (2)$$

We place all possible solutions to the KKT conditions into five policy forms. These five policy forms are derived from possible values of the Lagrangian multipliers for each constraint in Problem P2. The Lagrangian multiplier is either equal to zero if the constraint is non-binding or is some positive value if the constraint is binding. After eliminating conflicting solutions, for example, we cannot hire up to the maximum workforce level, $\lambda_1 > 0$, and fire down to the minimum workforce level, $\lambda_2 > 0$, at optimality, we are left with five policy forms.

Policy Type 1: Hire up to the maximum workforce level

If the optimal policy calls for hiring up to the maximum workforce level, then $\lambda_1 > 0, \lambda_2 = 0, \lambda_3 = 0, \lambda_4 = c_h + c_f$ and KKT condition (1) implies,

$$\left.\frac{dz_1^*}{dw}\right|_{w=w^*} = (c_w + c_h) + \lambda_1 > c_w + c_h$$

Policy Type 2: Hire, but not up to the maximum workforce level

If the optimal policy calls for hiring, but not up to the maximum workforce level, then $\lambda_1 = 0, \lambda_2 = 0, \lambda_3 = 0, \lambda_4 = c_h + c_f$ and KKT condition (1) implies,

$$\left.\frac{dz_1^*}{dw}\right|_{w=w^*} = (c_w + c_h)$$

Policy Type 3: Neither hire nor fire.

If the optimal policy calls for maintaining the current workforce level, then $\lambda_1 = 0, \lambda_2 = 0, \lambda_3 + \lambda_4 = c_h + c_f$ and KKT condition (1) implies,

$$\left.\frac{dz_1^*}{dw}\right|_{w=w^*} = (c_w + c_h) - \lambda_3 \Rightarrow c_h - c_f < \left.\frac{dz_1^*}{dw}\right|_{w=w^*} < c_w + c_h$$

Policy Type 4: Fire, but not down to the minimum workforce level

If the optimal policy calls for firing, but not down to the minimum workforce level, then $\lambda_1 = 0, \lambda_2 = 0, \lambda_3 = c_h + c_f, \lambda_4 = 0$ and KKT condition (1) implies,

$$\left.\frac{dz_1^*}{dw}\right|_{w=w^*} = (c_w + c_h) - (c_h + c_f) = c_w - c_f$$

Policy Type 5: Fire down to the minimum workforce level

If the optimal policy calls for firing down to the minimum workforce level, then $\lambda_1 = 0, \lambda_2 > 0, \lambda_3 = c_h + c_f, \lambda_4 = 0$ and KKT condition (1) implies,

$$\left.\frac{dz_1^*}{dw}\right|_{w=w^*} = (c_w + c_h) - (c_h + c_f) - \lambda_2 = c_w - c_f - \lambda_2 < c_w - c_f$$

By Proposition 8, we know that the $\frac{dz_1^*}{dw}$ is increasing, but not strictly increasing, in w^*. Therefore, there can be more than one solution to the condition of setting $\frac{dz_1^*}{dw}$ equal to a constant. For this reason we need to allow an interval of values of the workforce level at which $\frac{dz_1^*}{dw} = c_w + c_h$ and at which $\frac{dz_1^*}{dw} = c_w - c_f$.

$w_{h1} =$ the smallest value of w at which $\frac{dz_1^*}{dw} = c_w + c_h$. If $\left.\frac{dz_1^*}{dw}\right|_{\overline{w}} > c_w + c_h$, then $w_{h1} = \overline{w}$. If $\left.\frac{dz_1^*}{dw}\right|_{\underline{w}} < c_w + c_h$, then $w_{h1} = \underline{w}$.

w_{h2} = the largest value of w at which $\frac{dz_1^*}{dw} = c_w + c_h$. If $\left.\frac{dz_1^*}{dw}\right|_{\overline{w}} > c_w + c_h$, then $w_{h2} = \overline{w}$. If $\left.\frac{dz_1^*}{dw}\right|_{\underline{w}} < c_w + c_h$, then $w_{h2} = \underline{w}$.

w_{f1} = the smallest value of w at which $\frac{dz_1^*}{dw} = c_w - c_f$. If $\left.\frac{dz_1^*}{dw}\right|_{\overline{w}} > c_w - c_f$, then $w_{f1} = \overline{w}$. If $\left.\frac{dz_1^*}{dw}\right|_{\underline{w}} < c_w - c_f$, then $w_{f1} = \underline{w}$.

w_{f2} = the largest value of w at which $\frac{dz_1^*}{dw} = c_w - c_f$. If $\left.\frac{dz_1^*}{dw}\right|_{\overline{w}} > c_w - c_f$, then $w_{f1} = \overline{w}$. If $\left.\frac{dz_1^*}{dw}\right|_{\underline{w}} < c_w - c_f$, then $w_{f1} = \underline{w}$.

The monotonicity of $\frac{dz_1^*}{dw}$ implies $w_{h2} < w_{f1}$.

Theorem #3 is motivated by the implications of the necessary KKT conditions. If there is a solution that we think is an optimal point, then it must satisfy the necessary KKT conditions. Theorem 3, below, establishes the form of the optimal solution for any given problem parameterization. As the Theorem states, each of the five policy types is applicable over certain intervals of values for the initial workforce level, w_0.

Theorem 3:

- *If $w_0 < w_{h1}$ then the optimal policy permits any solution that hires up to a workforce level in the interval $[w_{h1}, w_{h2}]$ (Policy Type 1, 2)*
- *If $w_{h1} \le w_0 \le w_{h2}$ then the optimal policy permits any solution that either maintains the current workforce level or hires up to a workforce level in the interval $(w_{0,} w_{h2}]$ (Policy Type 1, 2, 3)*
- *If $w_{h2} < w_0 < w_{f1}$ then the optimal policy maintains the current workforce (Policy Type 3)*

- *If* $w_{f1} \le w_0 \le w_{f2}$ *then the optimal policy either maintains the current workforce or lays off to a level within* $[w_{f1}, w_0)$ *(Policy Type 3, 4, 5)*

- *If* $w_{f2} < w_0$ *then the optimal policy lays off down to a level within* $\lfloor w_{f1}, w_{f2} \rfloor$ *(Policy Type 4, 5)*

Proof:

Follows from the requirements of the necessary KKT conditions ||

We will determine the values of w_{h1}, w_{h2} by setting $\frac{dz_1^*}{dw}$ equal to $c_w + c_h$ and w_{f1}, w_{f2} by setting $\frac{dz_1^*}{dw}$ equal to $c_w - c_f$. In each of the three cases for Problem P1, this derivative is evaluated either for the case of fixed generation (horizontal policy line) or for the case of fixed client intensity (vertical policy line). Hence, we are faced with two kinds of conditions that must be solved for w.

Sub-case Type 1: $g^*(w)$ is fixed for $w \in [w_1, w_2]$, then

$$\frac{dy^*}{dw} = -\frac{r^h g^*}{a^w w^2 e'(y^*(w))}$$

$$\frac{dz_1^*}{dw} = \frac{c_z (r^h g^*)^2}{a^w w^{*2} e'(y^*)}$$

Using the fact that the provider capacity constraint must be binding at optimality, we can substitute $a^w w^* e(y^*)$ for $r^h g^*$ to obtain,

$$\frac{dz_1^*}{dw} = \frac{c_z a^w e(y^*)^2}{e'(y^*)}$$

$$w_{h1} = w_{h2} = \frac{r^h g_{\min}}{a^w e(y)} \text{ where } \frac{e(y)^2}{e'(y)} = \frac{c_w + c_h}{c_z a^w}$$

$$w_{f1} = w_{f2} = \frac{r^h g_{\min}}{a^w e(y)} \text{ where } \frac{e(y)^2}{e'(y)} = \frac{c_w - c_f}{c_z a^w}$$

Since, $\frac{e(y(w))^2}{e'(y(w))}$ is decreasing in w there is at most one solution to each condition.

Interpretation: In this sub-case type, $\frac{\partial z_1^*}{\partial w}$ is the marginal effect on client cost of a unit change in workforce.

Sub-case Type 2: $y^*(w)$ is fixed for $w \in [w_1, w_2]$, then

$$\frac{dg^*}{dw} = \frac{a^w e(y^*)}{r^h}$$

$$\frac{dz_1^*}{dw} = (Y - y) c_z a^w e(y)$$

Since the left-hand side of this condition is fixed this condition is satisfied for all workforce levels for which $a^w we(y) = r^h g$ for $g \in [g_{min}, g_{max}]$. In other words, over an entire vertical segment of the policy line, this condition can be met.

$$w_{h1} = \frac{r^h g_{\min}}{a^w e(y)} \text{ where } (Y - y) e(y) = \frac{c_w + c_h}{c_z a^w}$$

$$w_{h2} = \frac{r^h g_{\max}}{a^w e(y)} \text{ where } (Y - y) e(y) = \frac{c_w + c_h}{c_z a^w}$$

$$w_{f1} = \frac{r^h g_{\min}}{a^w e(y)} \text{ where } (Y - y) e(y) = \frac{c_w - c_f}{c_z a^w}$$

$$w_{f2} = \frac{r^h g_{\max}}{a^w e(y)} \text{ where } (Y - y) e(y) = \frac{c_w - c_f}{c_z a^w}$$

Theorem 3 identifies the optimal values of the variables, w, f based on w_h, w_f. However, these critical policy parameters are different for the three cases defined earlier. Each sub-case defined earlier presents a specific functional representation for $(y^*(w), g^*(w))$. The solutions for w_h, w_f must be derived by evaluating $\frac{dz_1^*}{dw}$ for each sub-case. We examine the dependence of $\frac{dz_1^*}{dw}$ on w for these sub-cases below in Table 2.

Sub-Case	w_{h1} **(if not limited by $\underline{w}, \overline{w}$)**	w_{h2} **(if not limited by $\underline{w}, \overline{w}$)**	w_{f1} **(if not limited by $\underline{w}, \overline{w}$)**	w_{f2} **(if not limited by $\underline{w}, \overline{w}$)**
1.1	$w_{h1} = \frac{r^h g_{\min}}{a^w e(y)}$ where $\frac{e(y)^2}{e'(y)} = \frac{c_w + c_h}{c_z a^w}$	$w_{h2} = \frac{r^h g_{\min}}{a^w e(y)}$ where $\frac{e(y)^2}{e'(y)} = \frac{c_w + c_h}{c_z a^w}$	$w_{f1} = \frac{r^h g_{\min}}{a^w e(y)}$ where $\frac{e(y)^2}{e'(y)} = \frac{c_w - c_f}{c_z a^w}$	$w_{f2} = \frac{r^h g_{\min}}{a^w e(y)}$ where $\frac{e(y)^2}{e'(y)} = \frac{c_w - c_f}{c_z a^w}$
1.2	$w_{h1} = \frac{r^h g_{\min}}{a^w e(y_{\min})}$ where $(Y - y_{\min}) e(y_{\min}) = \frac{c_w + c_h}{c_z a^w}$	$w_{h2} = \frac{r^h g_{\max}}{a^w e(y_{\min})}$ where $(Y - y_{\min}) e(y_{\min}) = \frac{c_w + c_h}{c_z a^w}$	$w_{f1} = \frac{r^h g_{\min}}{a^w e(y_{\min})}$ where $(Y - y_{\min}) e(y_{\min}) = \frac{c_w - c_f}{c_z a^w}$	$w_{f2} = \frac{r^h g_{\max}}{a^w e(y_{\min})}$ where $(Y - y_{\min}) e(y_{\min}) = \frac{c_w - c_f}{c_z a^w}$
2.1	$w_{h1} = \frac{r^h g_{\min}}{a^w e(y)}$ where $\frac{e(y)^2}{e'(y)} = \frac{c_w + c_h}{c_z a^w}$	$w_{h2} = \frac{r^h g_{\min}}{a^w e(y)}$ where $\frac{e(y)^2}{e'(y)} = \frac{c_w + c_h}{c_z a^w}$	$w_{f1} = \frac{r^h g_{\min}}{a^w e(y)}$ where $\frac{e(y)^2}{e'(y)} = \frac{c_w - c_f}{c_z a^w}$	$w_{f2} = \frac{r^h g_{\min}}{a^w e(y)}$ where $\frac{e(y)^2}{e'(y)} = \frac{c_w - c_f}{c_z a^w}$
2.2	$w_{h1} = \frac{r^h g_{\min}}{a^w e(\hat{y})}$ where $(Y - \hat{y}) e(\hat{y}) = \frac{c_w + c_h}{c_z a^w}$	$w_{h2} = \frac{r^h g_{\max}}{a^w e(\hat{y})}$ where $(Y - \hat{y}) e(\hat{y}) = \frac{c_w + c_h}{c_z a^w}$	$w_{f1} = \frac{r^h g_{\min}}{a^w e(\hat{y})}$ where $(Y - \hat{y}) e(\hat{y}) = \frac{c_w - c_f}{c_z a^w}$	$w_{f2} = \frac{r^h g_{\max}}{a^w e(\hat{y})}$ where $(Y - \hat{y}) e(\hat{y}) = \frac{c_w - c_f}{c_z a^w}$

Sub-case	w_{h1} **(if not limited by $\underline{w}, \overline{w}$)**	w_{h2} **(if not limited by $\underline{w}, \overline{w}$)**	w_{f1} **(if not limited by $\underline{w}, \overline{w}$)**	w_{f2} **(if not limited by $\underline{w}, \overline{w}$)**
2.3	$w_{h1} = \frac{r^h g_{\max}}{a^w e(y)}$ where $\frac{e(y)^2}{e'(y)} = \frac{c_w + c_h}{c_z a^w}$	$w_{h2} = \frac{r^h g_{\max}}{a^w e(y)}$ where $\frac{e(y)^2}{e'(y)} = \frac{c_w + c_h}{c_z a^w}$	$w_{f1} = \frac{r^h g_{\max}}{a^w e(y)}$ where $\frac{e(y)^2}{e'(y)} = \frac{c_w - c_f}{c_z a^w}$	$w_{f2} = \frac{r^h g_{\max}}{a^w e(y)}$ where $\frac{e(y)^2}{e'(y)} = \frac{c_w - c_f}{c_z a^w}$
3.1	$w_{h1} = \frac{r^h g_{\min}}{a^w e(y_{\max})}$ where $(Y - y_{\max}) e(y_{\max}) = \frac{c_w + c_h}{c_z a^w}$	$w_{h2} = \frac{r^h g_{\max}}{a^w e(y_{\max})}$ where $(Y - y_{\max}) e(y_{\max}) = \frac{c_w + c_h}{c_z a^w}$	$w_{f1} = \frac{r^h g_{\min}}{a^w e(y_{\max})}$ where $(Y - y_{\max}) e(y_{\max}) = \frac{c_w - c_f}{c_z a^w}$	$w_{f2} = \frac{r^h g_{\max}}{a^w e(y_{\max})}$ where $(Y - y_{\max}) e(y_{\max}) = \frac{c_w - c_f}{c_z a^w}$
3.2	$w_{h1} = \frac{r^h g_{\max}}{a^w e(y)}$ where $\frac{e(y)^2}{e'(y)} = \frac{c_w + c_h}{c_z a^w}$	$w_{h2} = \frac{r^h g_{\max}}{a^w e(y)}$ where $\frac{e(y)^2}{e'(y)} = \frac{c_w + c_h}{c_z a^w}$	$w_{f1} = \frac{r^h g_{\max}}{a^w e(y)}$ where $\frac{e(y)^2}{e'(y)} = \frac{c_w - c_f}{c_z a^w}$	$w_{f2} = \frac{r^h g_{\max}}{a^w e(y)}$ where $\frac{e(y)^2}{e'(y)} = \frac{c_w - c_f}{c_z a^w}$

Table 2: Optimal Policies for Problem P2 (Created by author)

6. Managerial Interpretations

Service firms will find themselves in one of three cases. For any given workforce the service provider must determine client intensity and service generation levels. As the client intensity and service generation levels are exchanged, there is a trade-off between client costs and backlog costs. What distinguishes the three cases is the nature of the trade-off. The quantity that we have called $\hat{y}$ is the unique level of client intensity at which the trade-off is optimal. We define this point as the backlog-involvement balance point. Service providers would like to reach this point, but other constraints prohibit this from happening.

Consider Case 1. In this case, the backlog-involvement balance point is less than the minimum level of client intensity. The constraint that sets the minimum level of client intensity prohibits the service provider from reaching the backlog-involvement balance point. The client and service provider are motivated to reduce client intensity down to the minimum level. Of course reducing client involvement would reduce service generation resulting in higher backlog costs to the client. The trade-off favors higher backlog costs and lower quality over reduced client costs.

Consider Case 2. In this case, the backlog-involvement balance point is greater than the minimum level of client intensity and less than the maximum level of client intensity. The constraints that set the minimum and maximum levels of service generation can prohibit the service provider from reaching the backlog-involvement balance point. There are three intervals of the workforce level in this solution. If the given workforce level is low, then the client and service provider are motivated to reduce service generation down to the minimum level. Of course reducing client involvement would reduce service generation resulting in higher backlog costs to the client. The trade-off favors higher backlog costs and lower quality over reduced client costs. If the given workforce level is in the middle interval, then the client and service provider are motivated to set the level of client intensity to the backlog-involvement balance point. If the given workforce level is high, then the client and service provider are motivated to increase service generation up to the minimum level. Of course increasing client involvement would increase service generation resulting in lower backlog costs to the client. The trade-off favors lower backlog costs and higher quality over increased client costs.

Consider Case 3. In this case, the backlog-involvement balance point is greater than the maximum level of client intensity. The constraint that sets the maximum level of client intensity prohibits the service provider from reaching the backlog-involvement balance point. The client and service provider are motivated to increase client intensity up to the maximum level. Of course increasing client intensity would increase service generation resulting in lower backlog costs to the client. The trade-off favors lower backlog costs and higher quality over increased client costs.

6.1 Service Examples

Each of the cases described above can be applied to specific service industries. Case 1 can apply to higher education. The teacher is the service provider and the students are the clients. Imagine there is a particular student who has competing personal responsibilities which make the costs of being in attendance and turning in assignments on time extremely high. This student is willing to allow himself to work at the minimum allowable level (i.e., service generation). Therefore, he accepts penalty costs of late assignments and low results on exams (i.e., low quality) rather than the high costs of involvement.

Case 2 applies to realtor services. Once a realtor has been chosen as the service provider, the home seller or buyer can decide how much involvement he would like to contribute. For example, the realtor can tell the client that in order to sell the home in the timeframe desired several tasks such as, updating the kitchen, new carpeting and landscaping, should be completed. The client may chose to incur the cost of freshening up the landscaping only (i.e., setting client involvement between the maximum and minimum level). As result there will be some quality improvements and the home may not sell by the desired deadline (i.e., backlog costs).

Case 3 applies to some healthcare services. For severe illnesses patients are motivated to increase their involvement to the maximum level (e.g., follow-up visits, attending physical therapy sessions, taking medications as prescribed). The costs of remaining sick (i.e., backlog costs) are far too high for the patient to accept.

6.2 Numerical Results

To illustrate the model we develop a set of experiments using hypothetical data. In our experiments we chose an exponential form for the efficiency and quality functions. In the trade-

off experiment, we investigate the nature of a non-optimal, yet feasible, resource plan with respect to client intensity. In the what-if experiments, we investigate the nature of the optimal resource plan and the sensitivity of the resource plan with respect to various parameters of the service process. All experiments have an efficiency and quality scaling factor of 1.0. The experiments were run using Microsoft Excel Solver 2007. Table 3 shows the base case. The parameter values for the base case were chosen based on reasonable assumptions. For example, the service provider's available capacity per worker, a_{pt}^{w}, is set to 160 hours (40 hrs per wk x 4 wks per period). We acknowledge that an empirical study or case study needs to be performed in order to have more accurate estimates of the parameter values.

Parameter	Value	Parameter	Value	Parameter	Value
d_{xt}	40	c_z	50	$\bar{b}_x$	4
a_{pt}^{w}	160	c_h	10000	$\underline{y}_{sx}$	0.20
a_{sxt}^{c}	1000	c_f	5000	$\bar{y}_{sx}$	0.8
v_x	6000	c_w	8000	r_{sx}^{h}	80
r_{sx}	1	c_{bp}	5000	$\underline{q}_p$	0.70

Table 3: Base Case Parameters (Created by author)

Trade-off Experiment

Experiment #1

This experiment observes a trade-off between client costs and backlog costs when the level of client intensity is increased. We held fixed all parameters and increased the level of client intensity from 25% to 37% in increments of 2%. The level of client intensity used in this experiment was started at 25% because of the maximum backlog level had been reached and it was stopped at 37% because the maximum level of generation had been reached. Results show that as client intensity increases total client costs increase linearly and backlog costs decrease

(see Figure 5). This experiment shows the service provider that the backlog costs are nonlinear with respect to client intensity. Therefore, the decision of whether or not to incur backlog costs is not an easy decision to make.

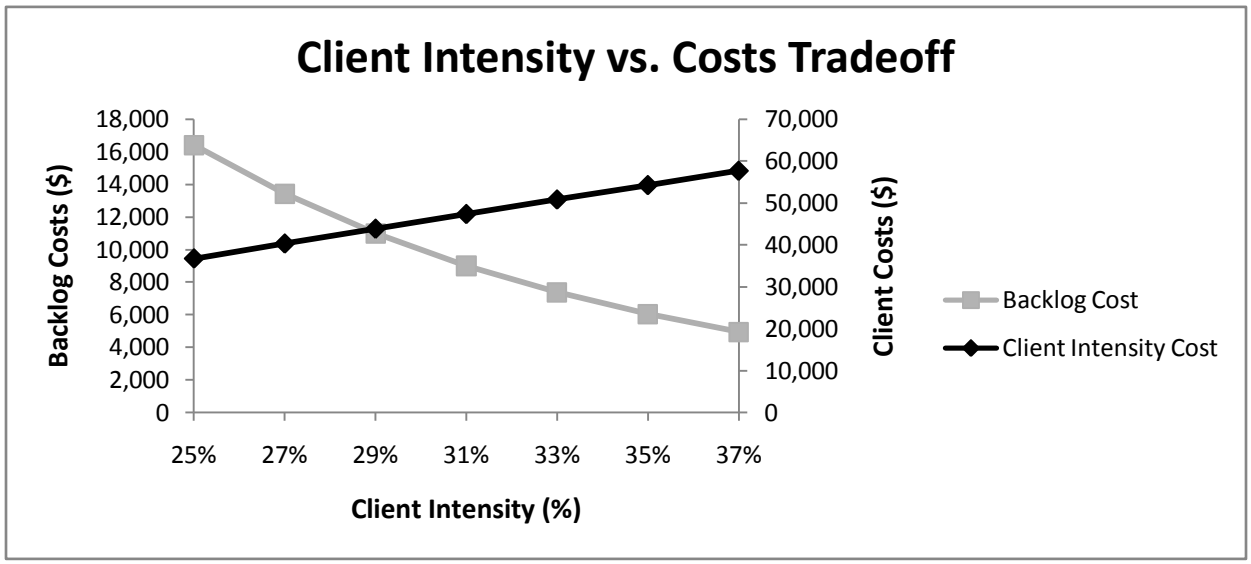

Figure 5: Client Intensity vs. Costs Tradeoff (Created by author)

From the results of Experiment #1, we see that there is a level of client intensity (29%) at which the sum of the client costs and the backlog costs is at its minimum. This point is denoted $\hat{y}$. Since this point has been reached, a service firm is in Case 2. In Case 2, the level of client intensity is constant at 29% over the workforce level interval from 17 to 19. If the workforce level is increased beyond 19, then the level of service generation will increase up to its maximum level (see Figure 6). If the workforce level is decreased below 17, then the level of service generation will decrease down to its minimum level.

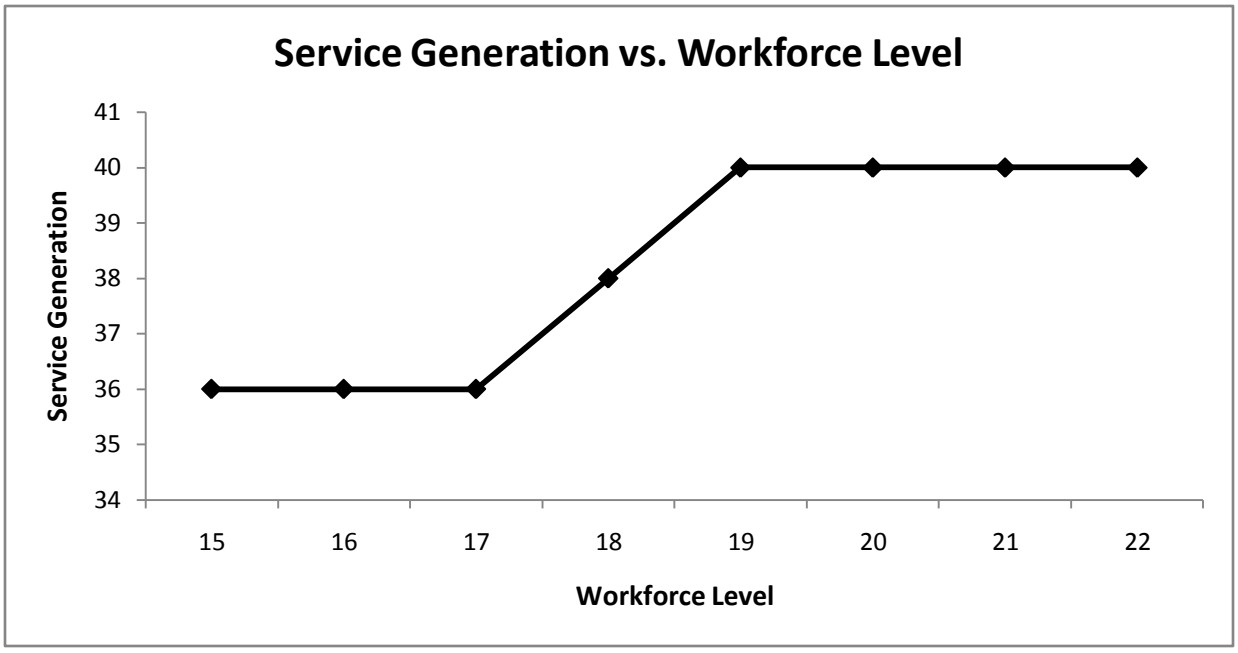

Figure 6: Service Generation vs. Workforce Level (Created by author)

What-If Experiments

The following experiments demonstrate what-if scenarios used to examine optimal solutions versus changes to the model parameters. All experiments assume a given workforce level. We use the term leverage rate often in the remainder of this section, so it is proper that we clarify the term at this time. The efficiency and quality leverage rates are coefficients used to convey the effectiveness of each unit of client intensity (% participation). By varying the rate parameter we change the concavity of the exponential function (see Figure 7). In the efficiency and quality functions below, the efficiency and quality scaling factors are denoted e_1 and q_1, respectively and the efficiency and quality leverage parameters are denoted λ_e and λ_q, respectively.

$$e(y) = e_1 (1 - \exp[-\lambda_e y])$$

$$q(y) = q_1 (1 - \exp[-\lambda_q y])$$

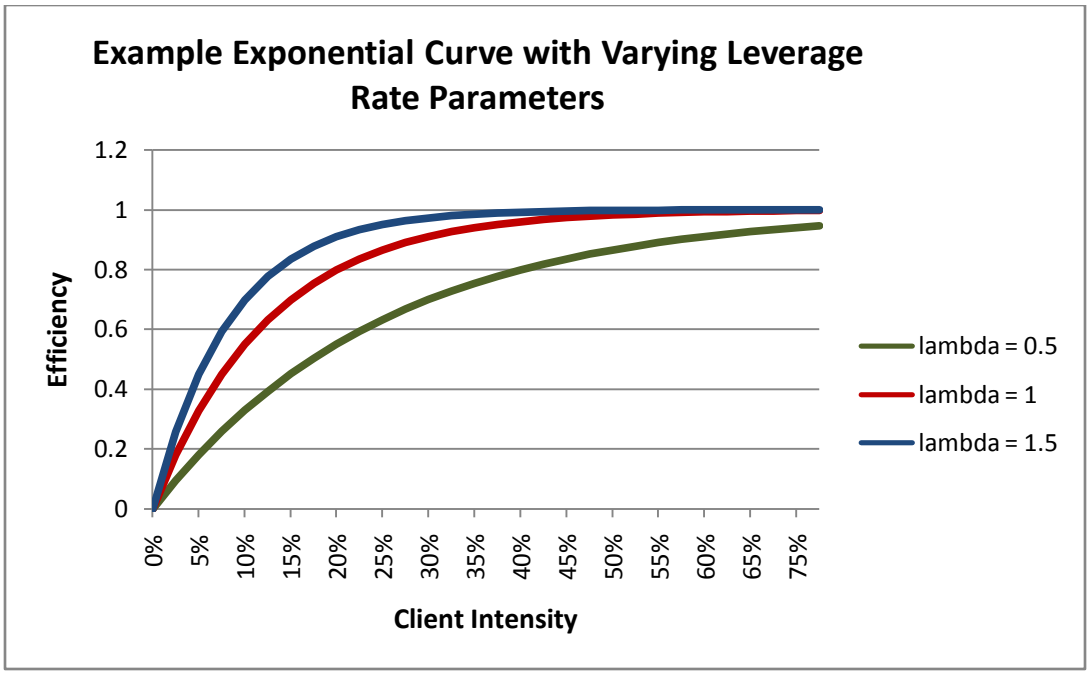

Figure 7: Example Efficiency Curve with Varying Leverage Parameter (Created by author)

Experiment #2

This experiment was designed to observe the changes in the optimal level of client participation when efficiency is improved in the service-creation process. Efficiency improvements can be modeled by increasing the leverage parameter for the function. We held fixed all parameters and increased the efficiency leverage parameter from 3.0 to 8.0 in increments of one. Results show that as improvements are made toward bettering efficiency of the service process, less client involvement is needed (see Figure 8).

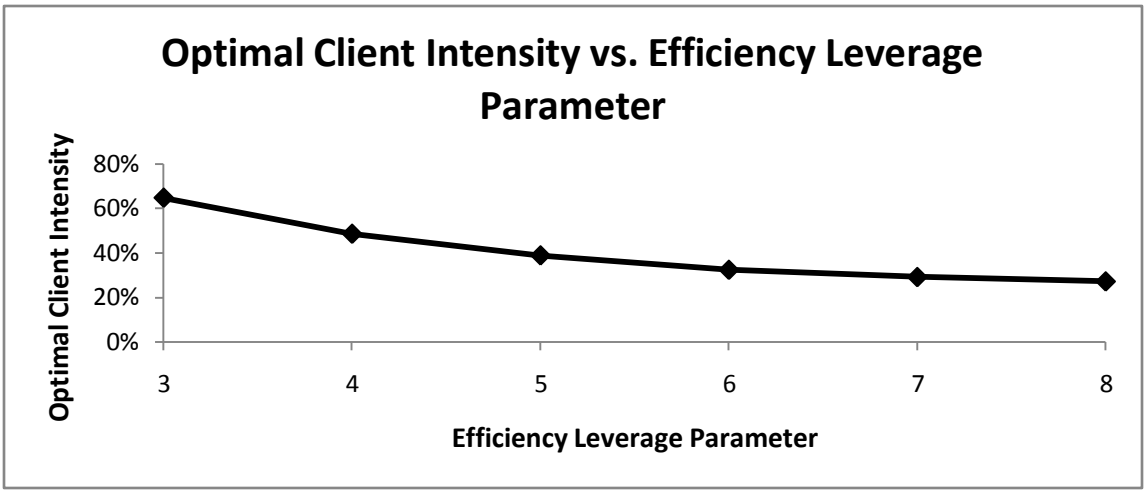

Figure 8: Optimal Client Intensity vs. Efficiency Leverage (Created by author)

Experiment #3

This experiment was designed to observe the changes in the optimal level of client participation when quality is improved in the service-creation process. We held fixed the value of the efficiency leverage parameter at 10.0 and increased the quality leverage parameter from 4.4 to 5.6 in increments of 0.2. The quality leverage parameter was started at 4.4, because quality had was at its lowest level allowed by the quality constraint and was stopped at 5.6 because the level of client intensity was no longer decreasing. Results show that as improvements are made toward bettering quality of the service process, less client involvement is needed (see Figure 9).

This experiment gives the service provider and client very important information because it shows that there is a level of client intensity at which quality improvements are no longer effective in lowering client costs. As client intensity decreases, client costs decrease. Since at optimality there is a level client intensity which, despite best efforts to improve quality, will not decrease any further there are no further reductions in client costs. This experiment also shows the transition from Case 2 to Case 1. In Case 2, the backlog-involvement balance point is obtainable. In Figure 7, the backlog-involvement balance point equals 25%. Reading the graph from right to left, we are decreasing the quality leverage parameter. By decreasing the quality leverage parameter, we increase the lower bound on client intensity beyond backlog-involvement balance point and thus transition to Case 1 where $\hat{y} < y_{\min}$.

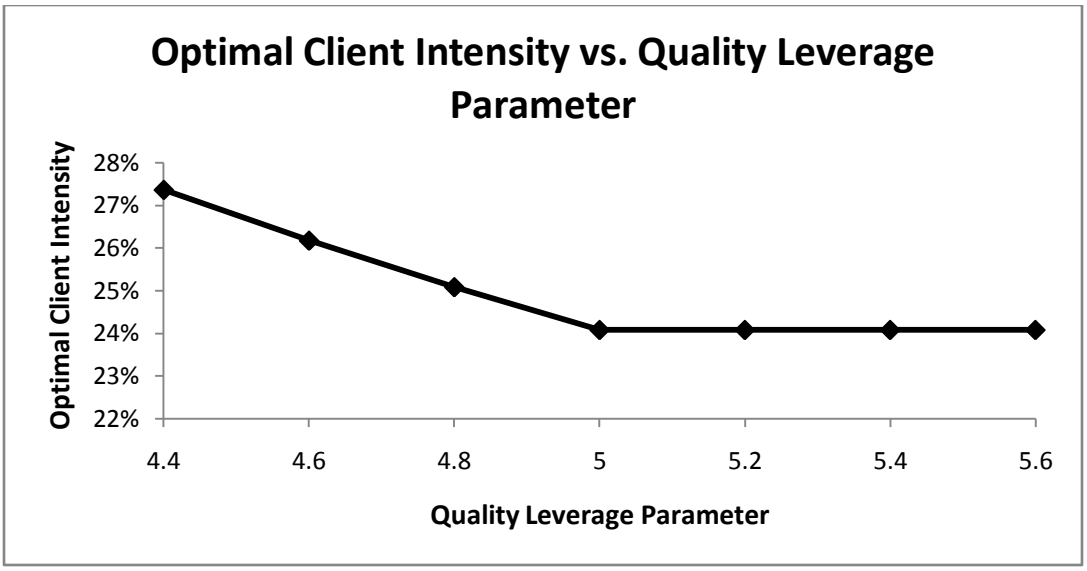

Figure 9: Optimal Client Intensity vs. Quality Leverage (Created by author)

Experiment #4

This experiment was designed to observe the changes in the optimal level of client participation when the maximum level of client intensity is increased. By increasing the maximum level of client intensity the client can become becomes more involved in the service creation process. We held fixed all parameters and increased the minimum level of client intensity parameter from 19% to 26% in increments of one percent. The maximum client intensity parameter was started at 19% because this was the smallest value allowed that satisfied the maximum backlog constraint and was stopped at 26% because the optimal level of client intensity no longer increased beyond this maximum level. Results show that when the client is allowed to become more involved in the service process, the optimal level of client involvement is increases up to a point (see Figure 10).

This experiment shows the transition from Case 2 to Case 3. In Case 2, the backlog-involvement balance point is obtainable. In Figure 10, the backlog-involvement balance point equals 22%. Reading the graph from right to left, we are decreasing the maximum client intensity level. By decreasing the maximum client intensity parameter we decrease the upper bound on client intensity lower than the backlog-involvement balance point and thus transition to Case 3 where $\hat{y} > y_{\max}$.

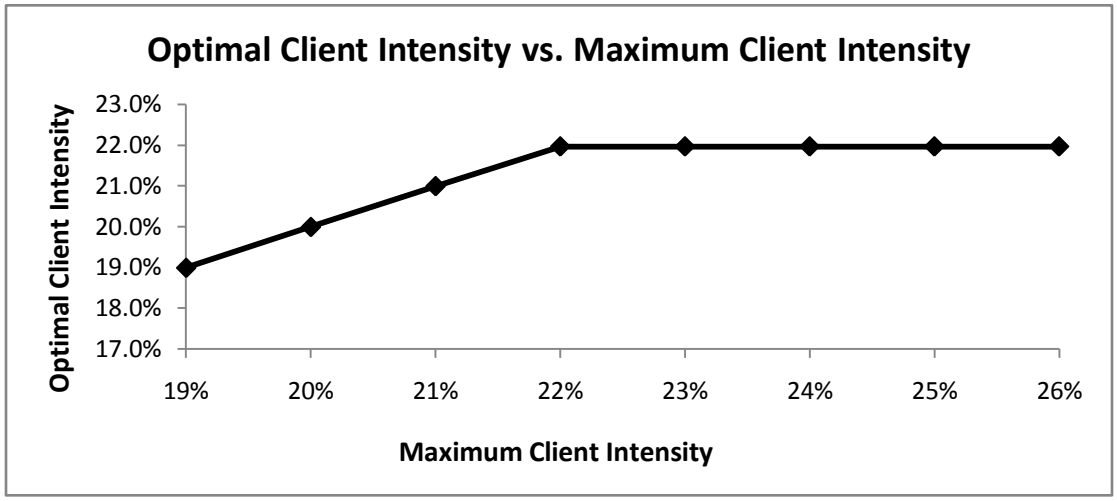

Figure 10: Optimal Client Intensity vs. Maximum Client Intensity (Created by author)

Experiment #5

This experiment was designed to observe the changes in the optimal level of client intensity and the optimal level of service generation when the workforce level is increased. By increasing the size of the workforce, we have more available service provider capacity and therefore optimal client intensity decreases and optimal service generation increases (see Figure 11). The percentage of client involvement levels out at 15% due to the minimum client involvement constraint we have in the model. Likewise, the level of service generation levels out at 40 because demand has been met.

This experiment demonstrates Proposition #7. Proposition #7 proves that optimal client intensity is decreasing in the workforce level and optimal service generation is increasing in the workforce level.

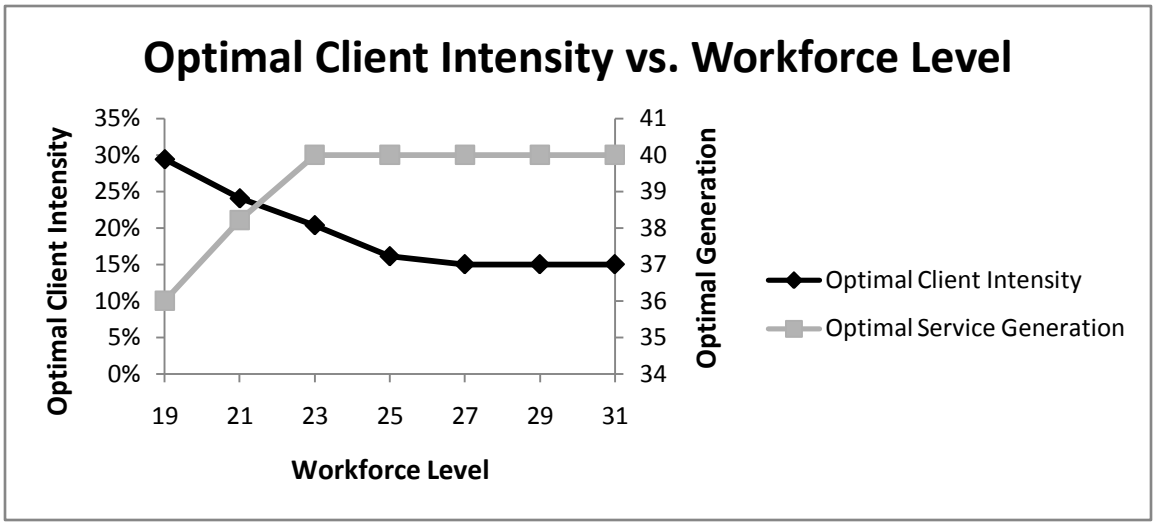

Figure 11: Optimal Client Intensity and Service Generation vs. Workforce
(Created by author)

7. Conclusion and Future Research

In this paper, we have developed a resource-planning model that will serve as a foundation for future analytic service science research. Our model is unique by the incorporation of the client as a resource in the planning model. We also capture the effects of client intensity on efficiency and quality in our model. Policy recommendations will give service enterprises valuable information regarding their hiring and firing policies, level of client intensity, and service generation.

Theoretical results show that service firms will find themselves in one of three cases based on parameter values. When the service provider determines client intensity and service generation levels there is a trade-off between client costs and backlog costs. What distinguishes the three cases is the nature of the trade-off. The backlog-involvement balance point is the unique level of client intensity at which the trade-off is optimal. An experiment was performed to show this trade-off.

What-if experiments were performed to demonstrate theoretical policy recommendations. The results of these experiments show that as improvements are made toward bettering efficiency and quality of the service process, less client involvement is needed. We also found

that optimal client intensity is decreasing in the workforce level and optimal service generation is increasing in the workforce level.

We contribute to the literature by offering our definition of services and modeling constructs. Our definition of services is not a departure from existing literature, but it does emphasize client co-generation and the shared creation of value which are two necessary elements in defining service. The manner in which we define "inventory" and "backlog" in this paper is unique. We also identify one method for quantifying the volume of a service.

Like other analytic models, our model is based on certain assumptions. We construct a single-period, single-process, single-stage model which makes for a simplistic view of resource planning for services. A natural extension of our model is to model the multi-period, multi-stage case. By observing the legacy effects of resource planning decisions we can better understand the true nature of services. Service science is a burgeoning field and should offer a wealth of research opportunities in the future.

References

Anderson, E. G., D. J. Morrice, et al. (2006). "Stochastic Optimal Control for Staffing and Backlog Policies in a Two-Stage Customized Service Supply Chain." Production and Operations Management **15**(2): 262.

Chase, R. (1978). "Where Does the Customer Fit in a Service Operation." Harvard Business Review: 137-142.

Chase, R. B. (1981). "The Customer Contact Approach to Services: Theoretical Bases and Practical Extensions." Operations Research **29**(4): 698.

Chase, R. B. and N. J. Aquilano (1992). Production & Operations Management. Boston, Irwin.

Cook, D. P., C.-H. Goh, et al. (1999). "Service typologies: A state of the art survey." Production and Operations Management **8**(3): 318.

Fitzsimmons, J. A. (2001). Service Management. Boston, McGraw-Hill.

Gaimon, C. (1997). "Planning Information Technology-Knowledge Worker Systems." Management Science **43**(9).

Holt, C. C., F. Modigliani, et al. (1955). "A Linear Decision Rule for Production and Employment Scheduling." Management Science (pre-1986) **2**(1): 1.

Ittig, P. T. (1994). "Planning service capacity when demand is sensitive to delay." Decision Sciences **25**(4): 541.

Kellogg, D. L. and W. Nie (1995). "A framework for strategic service management." Journal of Operations Management **13**(4): 323-337.

Lovelock, C. H. (1983). "Classifying Services to Gain Strategic Marketing Insights." Journal of Marketing **47**(3): 9-20.

Machuca, J. A. D., M. d. M. González-Zamora, et al. (2007). "Service Operations Management research." Journal of Operations Management **25**(3): 585.

Murdick, R. D., R. Render, Russell, R.S. (1990). Service operations management. Boston, Allyn and Bacon.

Nachum, L. (1999). "The Productivity of Intangible Factors of Production: Some Measurement Issues Applied to Swedish Management Consulting Firms " Journal of Service Research **2**(2): 123-137.

Napoleon, K. and C. Gaimon (2004). "The Creation of Output and Quality in Services: A Framewotk to Analyze Information Technology-Worker Systems." Production and Operations Management **13**(3).

Sampson, S. E. and C. M. Froehle (2006). "Foundations and Implications of a Proposed Unified Services Theory." Production and Operations Management **15**(2): 329.

Sasser, W. E., R. P. Olsen, et al. (1978). Management of service operations: text, cases, and readings. Boston, Allyn and Bacon.

Schmenner, R. W. (1986). "How can service businesses survive and prosper." Sloan Management Review **27**(3): 21-32.

Soteriou, A. C. and G. C. Hadjinicola (1999). "Resource allocation to improve service quality perceptions in multistage service systems." Production and Operations Management **8**(3).

Spohrer, J., P. Maglio, et al. (2007). Steps Toward a Science of Service Systems. Computer**:** 71-77.

Vargo, S. L. and A. Archupru (2009). "Service-Dominant Logic as a Foundation for Service Science: Clarifications." Service Science **1**(1): 32-41.

Verma, R. and K. K. Boyer (2000). "Service classification and management challenges." Journal of Business Strategies **17**(1): 5.

Winston, W. L. (1991). Operations research: applications and algorithms. Boston, PWS-Kent Publishing Company.

Appendix of Proofs

(A. 1) Proof of Lemma #1

Based on the model constructs, the feasible region is bounded by Constraints #3, #6 - #9. We know that Constraint 4 is strictly increasing in y and that it intersects the region bounded by Constraints #3, #6 - #9. By Proposition #2, the optimal solution must be on the boundary of Constraint #4, Constraint #5 or Constraint #9. Suppose the optimal solution lies on Constraint #5 at any point where the value of y is not the smallest, feasible value of y. By Proposition #3, the objective function is decreasing in y along Constraint #5 at this point. This point cannot be the optimal solution because the objective function can be improved. The objective function can be improved by going to another point on Constraint #5 that has a smaller value of y. Suppose the optimal solution lies on Constraint #9 at any point where the value of y is not the smallest, feasible value of y. By Proposition #4, the objective function is decreasing in y along Constraint #5 at this point. This point cannot be the optimal solution because the objective function can be improved. The objective function can be improved by going to another point on Constraint #9 that has a smaller value of y. Therefore, the optimal solution must lie on Constraint #4. ||

(A.2) Proof of Proposition #1

$\frac{\partial}{\partial g} = \frac{r^h}{e} > 0$ Constraint #4 is strictly increasing

$\frac{\partial}{\partial g} = -r^h y < 0$ Constraint #5 is strictly decreasing ||

(A.3) Proof of Proposition #2

$$\frac{-c_b(-i_0 + rd)}{r} + g\left(\frac{c_b}{r} - c_z r^h y\right) \quad (1)$$

$$\frac{\partial z_1}{\partial g} = \frac{c_b}{r} - c_z r^h y > 0 \quad (2)$$

By lines1 and 2 above and Assumption #5, the objective function is strictly increasing in g for any point in the interior of the feasible region. ||

(A.4) Proof of Proposition #3

The boundary of Constraint #5 is defined by the following relation:
$\frac{a^c}{r^h} - yg = 0$ which defines a strictly decreasing function, $g(y)$. A tangent vector to this constraint boundary can be written as follows:

Let, $f = a^c - r^h yg$

$\nabla f^T t = 0 \Rightarrow -gt_1 - yt_2 = 0 \Rightarrow \frac{t_2}{t_1} = -\frac{g}{y} = -\frac{a^c}{r^h y^2}$. We arbitrarily set $t_1 = 1$

$$t = \left(1, \frac{dg}{dy}\right) = \left(1, -\frac{a^c}{r^h y^2}\right)$$

The rate of change of the objective function along this feasible arc is then,

$$\nabla z_1^T t = -c_z r^h g + \left(\frac{c_b}{r} - c_z r^h y\right)\left(\frac{-a_c}{r^h y^2}\right)$$

$\nabla z_1^T t$ is less than zero.

If $\frac{c_b}{r} > c_z r^h \bar{y}$, then $(\frac{c_b}{r} - c_z r^h y) > 0$ and $[(\frac{c_b}{r} - c_z r^h y)(\frac{-a_c}{r^h y^2})] < 0$.

Also, $-c_z r^h g < 0$. Therefore, $\nabla z_1^T t$ is less than zero. ||

(A.5) Proof of Proposition #4

The boundary of Constraint #3 and Constraint #9 is defined by the relations, $g(y) = g_{\min}$, $g(y) = g_{\max}$, respectively. Hence, movement along this boundary can be parameterized in terms of y. A tangent vector to this constraint boundary can be written,

$t = (1, \frac{dg}{dy}) = (1,0)$

The rate of change of the objective function along this feasible arc is then,

$\nabla z_1^T t = -c_z r^h g$

Since $-c_z r^h g < 0$ then $\nabla z_1^T t$ is always less than zero. ||

(A.6) Proof of Proposition #5

The boundary of Constraint #4 is defined by the following relation: $g - \frac{a^w w}{r^h} e(y) = 0$. Hence, movement along this boundary can be parameterized in terms of y. A tangent vector to this constraint boundary can be written,

$t^T = \left(1, \frac{dg}{dy}\right) = \left(1, \frac{a^w w}{r^h} \frac{\partial e}{\partial y}\right)$

The Hessian matrix for z_1 is,

	y	g
y	0	$-c_z r^h$
g	$-c_z r^h$	0

$$t^T H(z_1) = \left(-c_z r^h \left(\frac{a^w w}{r^h} \frac{\partial e}{\partial y} \right), \; -c_z r^h \right)$$

$$t^T H(z_1) t = 2 \left[\left(c_z r^h \left(\frac{a^w w}{r^h} \frac{\partial e}{\partial y} \right) \right) \right]$$

The above expression is negative. Hence, any movement along the constraint boundary will produce concave behavior in z_1.

||

The optimality condition for z_1 on the boundary of Constraint #4 is,

$\nabla z_1 t = 0$

$$\nabla z_1 t = -c_z r^h g + \left(\frac{c_b}{r} - c_z r^h y \right) \left(\frac{a^w w}{r^h} \frac{\partial e}{\partial y} \right) = 0$$

(A.7) Proof of Proposition #6

$y_4(g_{\max})$ is decreasing in w. $y_{\max}$ is not a function of the workforce level. y_{45} is the value of y that satisfies $ye(y) = \frac{a^c}{a^w w}$. y_{45} is decreasing in w.

$y_4(g_{\min})$ is decreasing in w . $y_{\min}$ is not a function of the workforce level. ||

(A.8) Proof of Proposition #7

Values of $y*$ are either decreasing in w or are constant in w and y^* is continuous, proves that y^* is decreasing in w.

y^* will take on one of the following values $\hat{y}_4, \underline{y}_4, \bar{y}_4, y_{\min}$

(a) By the expression given in Proposition #5, $\hat{y}$ is not a function of the workforce level, w.

(b) Proposition #6 proves that $\bar{y}_4$, $\underline{y}_4$ are decreasing in w.

(c) $y_{\min}$ is not a function of the workforce level

Continuity

The solution for y* takes on different forms for different values of w.

Region (i) = $y^* = \underline{y}_4$

Region (ii) = $y^* = \hat{y}$

Region (iii) = $y^* = \bar{y}_4$

Region (iv) = $y^* = y_{\min}$

Definitions:

- w_1 is the value of the workforce level where $\underline{y}_4(w) = \hat{y}(w)$
- w_2 is the value of the workforce level where $\hat{y}(w) = \bar{y}_4$
- w_3 is the value of the workforce level where $\bar{y}_4 = y_{\min}$
- The values of w are ordered as follows: $w_1 > w_2 > w_3$

g^* is continuous in w and all possible values of g^* are either increasing in w or not a function of w. Therefore, g^* is increasing in w.

- $g^* = g_4(y^*)$
- y^* is continuous in w
- g_4 is continuous in y*

||

(A.9) Proof of Proposition #8

$$z_1(w) = -c_b\left(\frac{\mathrm{rd} - \mathrm{i}_0}{\mathrm{r}} - \frac{g^*}{r}\right) - c_z r^h y^* g^*$$

In general, $\dfrac{\partial z_1^*(w)}{\partial w} = \left(\dfrac{c_b}{r}\dfrac{\partial g^*}{\partial w}\right) - c_z r^h y^* \dfrac{\partial g^*}{\partial w} - c_z r^h g^* \dfrac{\partial y^*}{\partial w}$

Proposition #7 proves that the transitions in $y^*(w)$ and $g^*(w)$ from one form of the solution to another as w increases are continuous. Therefore, the transitions in $z_1^*(w)$ from one form of the solution to another as w increases are continuous.

In conclusion, it has been proven that $z_1^*(w)$ is increasing in w in each region and that it is continuous. Therefore, $z_1^*(w)$ is increasing in w. ||

(A.10) Proof of Theorem 1

Theorem #1 proves that Constraint #4 is redundant. Then by Proposition #2, the optimal solution will be on Constraint #5 or #9. The optimal solution is the smallest, feasible value of y on Constraint 5 or 9. By Proposition #3, the objective function is decreasing in y along Constraint #5. This means that we will move along Constraint #5 until we hit a boundary. That boundary is $y_{\min}$ .Therefore, $y^* = y_{\min}$. By Proposition #4, the objective function is decreasing in y along Constraint #9. This means that we will move along Constraint #9 until we hit a boundary. That boundary is $y_{\min}$. Therefore, $y^* = y_{\min}$.

$g_5(y_{\min}) > g_{\max}$ means that a portion of Constraint #5 lies above Constraint #9. Therefore , $g^* = g_{\max}$.

||

(A.11) Proof of Theorem 2

Define, $\bar{y}_4 = \min(y_4(g_{\max}), y_{\max}, y_{45})$

$\underline{y}_4 = \max(y_4(g_{\min}), y_{\min})$

$\hat{y}$ is a point on Constraint #4 (see Proposition #5)

The conditions of Lemma #1 are met by the conditions of this theorem. Therefore, the optimal solution is on Constraint #4. Therefore, $g^* = g_4(y^*)$. ||

CHAPTER 4

SERVICES RESOURCE PLANNING WITH TECHNOLOGY FUNCTION UNCERTAINTIES

Abstract
The way in which services transform inputs into outputs is typically uncertain or unknown. Consequently decision makers, at best, can only make estimates of the underlying technology function. The objectives of this research are to examine the sensitivity of estimates of technology functions to data analysis and to make policy recommendations to service providers on how to allocate resources when there are technology function uncertainties and uncontrollable inputs.

Keywords: service operations, stochastic, technology function, production function

1. Introduction

The goal of this research is to offer policy recommendations to service providers regarding how to best allocate resources when the mathematical function that maps inputs to outputs is unclear. We measure sensitivity of a resource planning model to uncontrollable inputs, mis-estimation of function parameters, and mis-specification of the function form that defines the transformation process. A resource plan identifies capacity levels needed to meet demand. The "engine" of the resource plan is the function that maps inputs to outputs. When a service provider is unsure of the structure of the function (mis-specification) or unsure of the parameter values of the function (mis-estimation) or the uncontrollable inputs the resource plan may be inefficient.

This paper begins with a discussion of what defines a service. A service has been defined as "a kind of action, performance, or promise that's exchanged for value between provider and client"; see Spohrer, Maglio et al. (2007). It can also be seen as "a time-perishable, intangible experience performed for a customer acting in the role of co-producer"; see Fitzsimmons (2001). Vargo and Akaka (2009) state that the "…function of service systems is to connect people, technology and information through value propositions with the aim of co-creating value for the service systems participating in the exchange of resources within and across systems." We define a service process as the transformation of inputs into outputs such that value for the client is

created through a process that utilizes capabilities and capacities of both the client and the provider.

This study focuses on resource planning for a particular service type within a particular service firm. A service type is a well-defined, value-adding experience that is offered by the provider to the client, such as processing a mortgage loan application, admitting a patient to a hospital, or teaching a class.

A set of different process types are needed to deliver a service type. For example, the service type of processing a mortgage loan application includes the process types of interviewing the applicant, gathering required documents, and submitting the application for underwriting. Each process type uses multiple inputs to generate multiple outputs. See Figure 1.

An input is a component of production. Examples of inputs are labor, raw materials, skills, and knowledge. An output is something that is produced or generated, such as sales, profit, transactions. The set of inputs should include all resources which impact the outputs. The outputs should reflect all useful outcomes on which we wish to assess the service engagement.

A service engagement represents an instance of a service type. For example, a service engagement is a particular person completing a mortgage loan application that needs to be processed or a particular person requiring admission to a hospital.

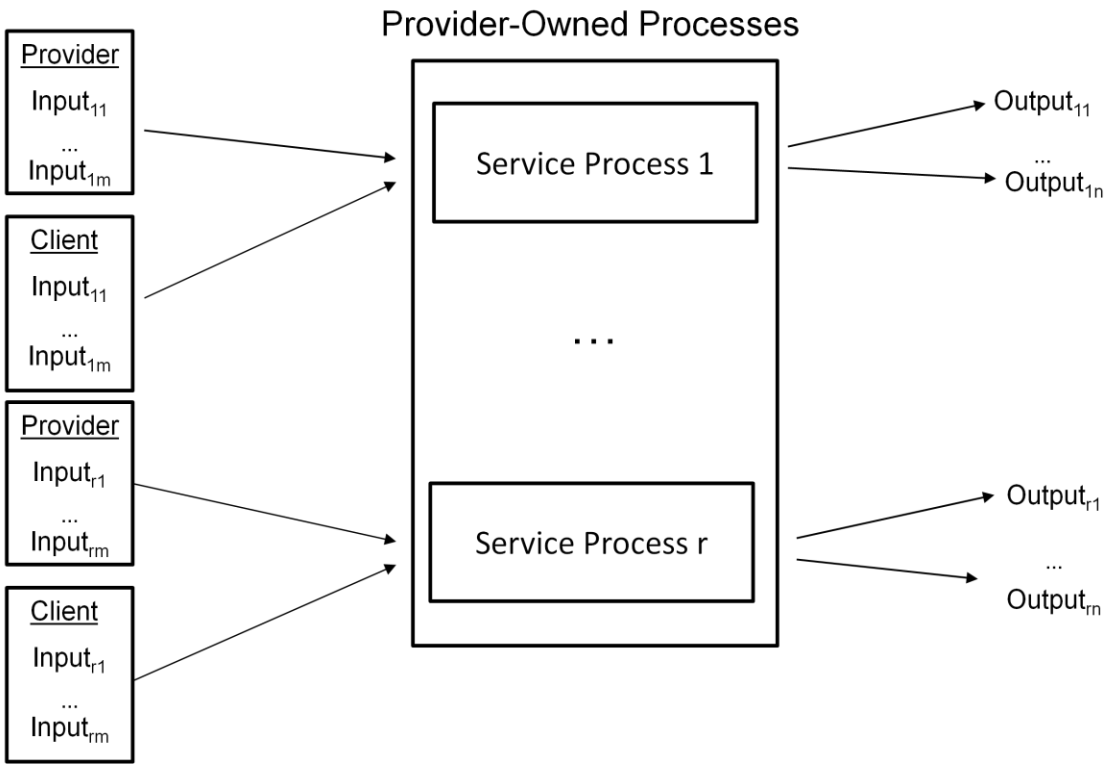

Figure 1: Model of a Service Type (Created by author)

Services should be classified along a continuum for the purpose of motivating model types in resource planning. Services along the continuum require different modeling constructs. The dimensions of the continuum are standardization vs. customization and client co-generation. Classifying services along a continuum acknowledges the fact that many different types of services share similar characteristics, but at differing degrees.

One end of the continuum identifies services that are standardized and have limited client co-generation. Services with a more standardized service process may vary some from client to client, but typically employees execute a routine process, see (Fitzsimmons and Fitzsimmons, 2004). These services, at certain levels of the business organization, have limited client co-generation. Although there is an abundance of client contact, the client may not be essential to all stages of the process. Some back office services, recreational and leisure services are examples.

On the other end of the continuum are services that have a high degree of client co-generation in the service creation process. These services are less standardized and more

customized. They typically employ knowledge workers, see Nachum (1999). This research is directed at resource planning decisions for service firms with a high degree of client co-generation or what we are terming high value-adding service operations. We are considering high value-adding services to be services such as consulting services, software development/IT services, and some hospital services. Resource planning for high-value adding services is difficult due to the lack of clarity regarding how inputs are converted into outputs.

In this model, the degree of client involvement is discretionary. Client involvement becomes discretionary when the service provider has the ability to limit or increase the amount of client contact in some way. For example, consultants can decide how many times of day they want to meet with a particular client. In mathematical modeling, when the degree of client involvement is discretionary its identifier changes from a parameter to a decision variable.

Manufacturing and service operations differ in the process used to transform inputs into outputs, see Mills and Moberg (1982). In manufacturing, machines and the flow of jobs through the facility typically identifies the transformation process. Often there is limited to no human interaction. Conversely, high-value adding service workers use their knowledge, experience, skills (i.e., inputs) to deliver a service (i.e., output) to the client. For high-value adding services, this transformation could be considered a technology. Rousseau (1979) identifies a technology as a process of transforming inputs into outputs and defines technology as "the application of knowledge to perform work". Hence, this research adopts the term technology function when describing the input to output conversion process.

A technology function is the mathematical function that maps inputs to outputs. In service organizations, technology functions are difficult to estimate because the conversion process is unclear. Lack of clarity results in inaccurate technology function estimates that can greatly affect resource planning decisions. In manufacturing, the model of the conversion of inputs to outputs is typically known. The production function must be well defined in manufacturing capacity planning decisions. In contrast, there exist many sources of uncertainty in the transformation process for services, see Dietrich (2006). Inherit variability in the transformation process is a source of uncertainty. As well as mis-estimation and mis-specification of the technology function.

There have been previous approaches to resource planning for services in the literature. Most papers use Data Envelopment Analysis (DEA) as the methodology. DEA, by its design, was not intended for resource allocation but for measuring relative efficiency of service units. In DEA no assumptions are made about the underlying transformation process. See Table 1 for a brief description of a few DEA-based resource allocation models and their representations of the transformation function.

The contributions of this research are as follows.

- This research uses a stochastic technology function for resource planning. Some service-based models use Chance-Constrained DEA, see Land et al. (1993), to incorporate stochastic considerations in inputs and output measures, but these models are not used for resource allocation.
- We have a foundation for modeling resource allocation decisions for high-value adding service systems when technology functions are mis-estimated and/or mis-specified. The field of service science is still in its infancy, therefore any policy recommendations that can be generated regarding resource-planning for service systems is greatly needed.
- We identify three sources of uncertainty in the technology function.
    - Uncontrollable inputs
    - Parameter mis-estimation
    - The structure of the transformation function is unknown or mis-specified
- This model that behaves intuitively. This is a well-constructed model that accurately captures the effect of resource allocations on service output levels. We investigate the trade-offs between inefficiencies, risk, and costs versus resource allocations.

Author(s)	Technology Identification	Technology form	Allowable Technology Changes	Allowable Scale Changes	Returns to Scale
Golany et al. (1993)	Implicit		No	Yes	VRS
Golany & Tamir (1995)	Explicit	Multiplicative	Yes	Yes	VRS
Athanassopoulos (1995)	Implicit		Yes	No	CRS
Thanassoulis (1996)	Implicit		No	No	CRS
Athanassopoulos (1998)	Explicit	Nonlinear	No	Yes	VRS
Fare et al. (1997)	Implicit		No	No	CRS (can be modified to VRS)
Kumar & Sinha (1999)	Implicit		Yes	No	CRS (can be modified to VRS)
Beasley (2003)	Implicit		No	No	CRS (can be modified to VRS)
Lozano & Villa (2004)	Implicit		No	Yes	VRS (can be modified to CRS)
Korhonen & Syrjanen (2004)	Implicit		No	No	CRS, VRS
Lozano & Villa (2005)	Implicit		No	Yes	VRS (can be modified to CRS)
Asmild et al. (2006)	Implicit		No	Yes	VRS
Golany et al. (2006)	Explicit	Cobb-Douglas	No	No	CRS

Table 1: Resource Planning Models (Created by author)

2. The Resource-Planning Decision

The service process has inputs from both the service provider and the client. The service provider and the client must determine the values of the decision variables of this resource-planning problem which are the quantities of inputs to allocate to each service process. Managers want to allocate resources in order to improve their day-to-day operations and ultimately to better position themselves in the market place.

Decision variables influence the performance measures in a decision model. The performance measures in this study are the total costs, the loss functions that represent expected underachievement of individual output targets, and the slack capacities.

Utility theory assumes that every decision maker has preferences towards risks and return and that the decision maker will choose the alternative that maximizes his/her utility. Based on the decision makers' preferences, there are different weights placed on the loss functions and the weights reflect the relative value of each loss function.

Parameters influence the performance measures. Parameters include the targets placed on the quantity of each input and output for each process in the service system. These targets come from benchmarking or historical data. There are usage rate parameters for each input and yield rate parameters for each output of every service process. Next, there are capacity parameters for each input from the service provider and from the client to be used for the service processes. Lastly, there are cost parameters for over allocation of service provider and client resources per process and cost parameters for under production of output per process.

The challenge for the service provider and the client is that there is uncertainty in the technology function. For any resource planning model, the technology function is needed to make resource planning decisions. When that function is unclear, resources may not be allocated properly. Service providers and clients want to ensure that they are allocating resources in the most efficient manner based on what they do know about the technology function in an effort to minimize the costs of underproduction and over allocation.

3. The Model

Two optimization models are developed in order to determine the optimal resource allocation plan. We take a two-tiered approach. The results from the first model are passed to the second model.

The first optimization model is a deterministic optimization model. The results from the first model are benchmark levels of input resources from the service provider and client. The benchmark technology function, which is comprised of given usage and yield rates, is what the service provider believes to be the most efficient function mapping inputs to outputs.

The second optimization model is a stochastic, resource-planning model. This model takes the benchmark input levels from the first model and develops the resource allocation plan for the service engagement. The stochastic element of this model is represented by a probability density function that captures the deviation from the benchmark output levels reflecting both inefficiencies and uncertainties.

3.1 Model Assumptions

We measure the output level of each process independently. We assume each process has its own technology function. There are no setup costs associated with re-allocating resources to a service engagement. It is assumed that the provider and the client each have fixed capacities per input to be allocated to all service processes. There is a linear cost for each input.

The benchmark usage and yield rates are estimated using historical data. Target output levels for each process are provided by the client to each optimization model.

We assume there is a density function of each output level. The input levels are parameters of this density function. As resource allocations change the shape of this density function will change. We assume that the density function represents a service engagement's deviation from the benchmark output level. This deviation reflects both inefficiency and uncertainty in the technology function form and parameters.

We assume a form of the technology function that follows Athanassopoulos (1998).

3.2 Descriptive Models

Indices (general to both models)

p	process type index
i	input index
j	output index
r	number of process types; $p = 1,...,r$
m_w	number of inputs from the service provider; $i = 1,...,m_w$
m_c	number of inputs from the client; $i = 1,...,m_c$
n	number of outputs; $j = 1,...,n$

Optimization Model #1

The first optimization model applies a benchmark technology function to obtain benchmark input levels for given target output quantities. The model minimizes the input levels from the service provider and the client across all processes. Constraints (M1.1) ensure that the number process cycles that can be generated by each service provider resource is greater than or equal to the number of completed process cycles. Constraints (M1.2) ensure that the number process cycles that can be generated by each client resource is greater than or equal to the number of completed process cycles. Constraints (M1.3) ensure that the number process cycles needed to generate a particular given output target is less than or equal to the number of completed process cycles. Constraints (M1.1 – M1.3) balance the resources received by each service process with that process' number of generated cycles, (i.e., inflow = outflow). Constraint (M1.4) and (M1.5) are the service provider and client capacity constraints, respectively.

Decision Variables

$\hat{x}_{ip}^{w}$	benchmark quantity of resource i allocated to process p by the service provider
$\hat{x}_{ip}^{c}$	benchmark quantity of resource i allocated to process p by the client
v_p	the number of benchmark process cycles of process p which are completed

Parameters

$\hat{y}_{jp}$	target quantity of output j by process p
a_i^w	available service provider capacity of resource i
a_i^c	available client capacity of resource i
β_{pi}	benchmark generation rate of process p that is supported by resource i (cycles/unit of resource)
α_{pj}	benchmark generation rate of process p that is required by output j (cycles/unit of resource)
$\mu_{pi} = \frac{1}{\beta_{pi}}$	benchmark usage rate of resource i per cycle of process p (units of resource/cycle)
$\gamma_{pj} = \frac{1}{\alpha_{pj}}$	benchmark generation rate of output j per cycle of process p (units of resource/cycle)

Minimize $\sum_{i,p} \hat{x}_{ip}^{w} + \sum_{i,p} \hat{x}_{ip}^{c}$

Subject to

$$\hat{x}_{ip}^{w} / \mu_{pi} \geq v_p \quad \text{for all } i = 1,\ldots,m_w,\ p = 1,\ldots,r \tag{M1.1}$$

$$\hat{x}_{ip}^{c} / \mu_{pi} \geq v_p \quad \text{for all } i = 1,\ldots,m_c,\ p = 1,\ldots,r \tag{M1.2}$$

$$\hat{y}_{jp} / \gamma_{pj} \leq v_p \quad \text{for all } j = 1,\ldots,n,\ p = 1,\ldots,r \tag{M1.3}$$

$$a_i^{w} - \sum_{p} \hat{x}_{ip}^{w} \geq 0 \quad \text{for all } i = 1,\ldots,m_w,\ p = 1,\ldots,r \tag{M1.4}$$

$$a_i^{c} - \sum_{p} \hat{x}_{ip}^{c} \geq 0 \quad \text{for all } i = 1,\ldots,m_c,\ p = 1,\ldots,r \tag{M1.5}$$

all variables are nonnegative

Technology Matrix

For the technology functions applied in the research, the resources must be procured according to usage rates of a process cycle and outputs are generated according to yield rates of a process cycle. The way in which we represent a technology function is a special case of the linear constant returns-to-scale (CRS) technology function used by Athanassopoulos (1998). A function exhibits constant returns-to-scale if, for any change in an input variable, each output variable is changed proportionately. We use the term "recipe" to describe technology functions in this research. The following example shows how this label is appropriate for the type of technology functions we assume for this research. If a recipe requires 5 cups of flour, ½ cup of water, ½ teaspoon of salt, and 3 packages of yeast to make one loaf of bread and we double each input we will then be able to make two loaves of bread. Therefore, the "recipe" for inputs and outputs per process cycle forces all outputs and inputs to be in fixed proportions with respect to one another.

The linear CRS technology function can be written,

$$T\,\vec{x}_p = \vec{y}_p \tag{1}$$

where, $T = \lfloor t_{jip} \rfloor_{nxm}$

We define the component of the technology matrix for row j and column i as, $t_{jip} = \frac{\partial y_{jp}}{\partial x_{ip}}$

$$t_{jip} = \frac{y_{jp}}{x_{ip}} = \frac{\beta_{pi}}{\alpha_{pj}} = \frac{\gamma_{pj}}{\mu_{pi}} = \gamma_{pj}\beta_{pi} \tag{2}$$

We note that $\frac{\partial y_{jp}}{\partial x_{ip}} = \frac{y_{jp}}{x_{ip}}$ due to the linear, constant returns-to-scale form of the technology function.

Fixed proportions of inputs and outputs are inherent in these formulas for the elements of the technology function matrix. For example,

$$\frac{x_{pi}}{x_{pk}} = \frac{t_{1\,pk}}{t_{1\,pi}} = \frac{t_{2\,pk}}{t_{2\,pi}} = \ldots = \frac{t_{npk}}{t_{npi}} = \frac{\beta_{pk}}{\beta_{pi}} \tag{3}$$

$$\frac{y_{pj}}{y_{pk}} = \frac{t_{jp\,1}}{t_{kp\,1}} = \frac{t_{jp\,2}}{t_{kp\,2}} = \ldots = \frac{t_{kpn}}{t_{jpn}} = \frac{\alpha_{pk}}{\alpha_{pj}} \tag{4}$$

Optimization Model #2

Optimization Model #2 is a stochastic, resource-planning model. The first part of the objective function, the integral, is similar to the stock-out loss function in a newsboy model. This integral captures the effect of a service engagement not generating output at the target output levels. There are penalties/weights placed on underproduction. The weights on underproduction and the target output levels are determined by the service provider and the client. The second part of the objective function captures the costs of allocating more than the benchmark input quantities of service provider and client resources to a service engagement. The constraints are service provider and client capacity constraints, respectively. Optimization can be done via standard search routines.

The density function of actual output is a function of inputs. As resource allocations change, the shape of the density function of y_{jp} will change. This is because the resource levels are incorporated into the parameters of the distribution; see equations (10) and (11) below.

Decision Variables

x_{ip}^{w} quantity of input i allocated to process p provided by the service provider

x_{ip}^{c} quantity of input i allocated to process p provided by the client

Performance Measures

y_{jp} actual quantity of output j achieved by process p

Parameters

$\hat{y}_{jp}$ target quantity of output j by process p

$f_{y_{jp}}(y_{jp};x_{ip})$ the probability density function of output j for process p

$\hat{x}_{ip}^{w}$ benchmark quantity of input i allocated to process p provided by the service provider; this quantity is obtained from the solution of Optimization Model #1

$\hat{x}_{ip}^{c}$ quantity of input i allocated to process p provided by the client; this quantity is obtained from the solution of Optimization Model #1

c_{jp}^{u} weight applied to under-production of output j from process p

c_{ip}^{o} cost of over allocation of service provider input i for process p

c_{ip}^{o} cost of over allocation of client input i for process p

a_{i}^{w} available service provider capacity of input i

a_{i}^{c} available client capacity of input i

$$\underset{x_{ip}^{w},\,x_{ip}^{c}}{\text{Minimize}} \sum_{p=1}^{r}\sum_{j=1}^{n} c_{jp}^{u}\int_{0}^{\hat{y}_{jp}} (\hat{y}_{jp}-y_{jp})\, f_{y_{jp}}(y_{jp};x_{ip})dy_{jp} + \sum_{i\in S_w} c_{ip}^{o}(x_{ip}^{w}-\hat{x}_{ip}^{w}) + \sum_{i\in S_c} c_{ip}^{o}(x_{ip}^{c}-\hat{x}_{ip}^{c})$$

Subject to

$$a_{i}^{w} - \sum_{p} x_{ip}^{w} \geq 0 \quad \text{for all } i,p \qquad \text{(M2.1)}$$

$$a_{i}^{c} - \sum_{p} x_{ip}^{c} \geq 0 \quad \text{for all } i,p \qquad \text{(M2.2)}$$

all variables are nonnegative

There exists a unique solution to Optimization Model #2; see Badinelli (2008).

Consider random variation in the elements of the matrix $\vec{\gamma}_p \vec{\beta}_p^T$.

Define,

$$\vec{g}_p = \vec{\gamma}_p - \vec{z}_{gp} \quad (5)$$

$$\vec{b}_p = \vec{\beta}_p - \vec{z}_{bp} \quad (6)$$

where $\vec{z}_{gp}, \vec{z}_{bp}$ are non-negative random variables.

Approximation:

$$\left(\vec{z}_{gp}\vec{\beta}_p^T + \vec{\gamma}_p\vec{z}_{bp}^T + \vec{z}_{gp}\vec{z}_{bp}^T\right) \approx \left[\tau_{jip} + \varepsilon_{jip}\right] \quad (7)$$

Where τ_{jip} = constant, $\varepsilon_{jip} \sim N(0, \sigma_{ji})$ and $\frac{\tau_{jip}}{\sigma_{jip}} > 3$

$\left[\tau_{jip} + \varepsilon_{jip}\right]$ is an nxm matrix of normal random variates with positive mean values and negligible probabilities of negatives values. The constant τ_{jip} represents the overall level of inefficiency of the process type. We considered deviations from the benchmark recipe as inefficiencies.

Therefore,

$$\vec{y}_p = \frac{\vec{\gamma}_p\vec{\beta}_p^T}{m}\vec{x}_p - \left[\tau_{jip} + \epsilon_{jip}\right]\vec{x}_p \quad (8)$$

or

$$\vec{y}_p = T\vec{x}_p - \left[\tau_{jip} + \epsilon_{jip}\right]\vec{x}_p \quad (9)$$

Which implies, $y_{jp} \sim N(\mu_{yjp}, \sigma_{yjp})$

Where,

$$\mu_{yjp} = \sum_{i=1}^{m}(T_{jip} - \tau_{jip})x_{ip} \quad (10)$$

$$\sigma_{y_{jp}}^2 = \sum_{i=1}^{m} x_{ip}^2\sigma_{jip}^2 \quad (11)$$

The Integral

$$\int_0^{\hat{y}_{jp}} (\hat{y}_{jp} - y)\, f_{y_{jp}}(y; x_{ip})\, dy$$

$$= \int_0^{\hat{y}_{jp}} \hat{y}_{jp}\, f_{y_{jp}}(y; x_{ip})\, dy - \int_0^{\hat{y}_{jp}} y\, f_{y_{jp}}(y; x_{ip})\, dy$$

$$= \hat{y}_{jp} F_{y_{jp}}(\hat{y}_{jp}) - y F_{y_{jp}}(y)\Big|_{y=0}^{\hat{y}_{jp}} + \int_0^{\hat{y}_{jp}} F_{y_{jp}}(y)\, dy$$

$$= \hat{y}_{jp} F_{y_{jp}}(\hat{y}_{jp}) - \hat{y}_{jp} F_{y_{jp}}(\hat{y}_{jp}) + \int_0^{\hat{y}_{jp}} F_{y_{jp}}(y)\, dy$$

$$= \int_0^{\hat{y}_{jp}} F_{y_{jp}}(y)\, dy$$

4. Numerical Results

A series of cases were run via Microsoft Excel Solver®. The experiments show plots of optimal solutions for different parameter settings, such as capacities, benchmark output levels, risk levels, inefficiency levels, and penalty costs. For illustrative purposes we chose eight input types (four from the service provider and four from the client), two output types, and three process types. The base case is specified in Table 2 and the benchmark provider and client resource quantities are specified in Table 3. The output targets are specified in Table 4. The parameter values for the base case were chosen based on reasonable assumptions. We acknowledge that an empirical study or case study needs to be performed in order to have more accurate estimates of the parameter values.

Each case/experiment was designed to illuminate model behavior under specific conditions. We want to highlight the elements of the model that distinguish this study from others in literature and gain managerial insights. For example, we designed cases to show the behavior of the loss function under certain conditions. We also designed cases highlighting technology function uncertainties. Other cases show the effects of the two-tiered modeling approach by examining how parameters of the first optimization model effect resource

allocations in the second optimization (e.g., changes in benchmark output rates). Note: we averaged the service provider and client resource quantities when displaying the results.

a_i^w & a_i^c	c_{ip}^o	c_{jp}^u	β_{pi}	α_{pj}	γ_{pj}	τ_{jip}	σ_{jip}
150	20	50	10	1	1	3	1

Table 2: Base Case Parameters (Created by author)

$\hat{x}_{ip}^w$	Process 1	Process 2	Process 3
Resource 1	40	50	40
Resource 2	40	35	35
Resource 3	80	75	70
Resource 4	10	10	10
$\hat{x}_{ip}^c$	Process 1	Process 2	Process 3
Resource 1	40	40	30
Resource 2	10	20	12
Resource 3	30	30	20
Resource 4	50	25	53

Table 3: Benchmark Resource Quantities (Created by author)

$\hat{y}_{jp}$	Process 1	Process 2	Process 3
Output Target 1	2000	2000	2000
Output Target 2	2000	2000	2000

Table 4: Benchmark Output Targets (Created by author)

Case 1: Resource Allocation vs. Output Target

In this experiment we varied the output target of Output 1, while keeping fixed the output target of Output 2 at the base case value.

The results of this experiment show as the output target increased for Output 1, the average allocation of service provider and client resources increased in order to meet the output target; see Figure 2. The resource quantities leveled off once all capacities had been used. Since there is no hiring included in this model, managers should understand that they will incur penalty costs if the workforce level is not sufficient to meet changes to output targets.

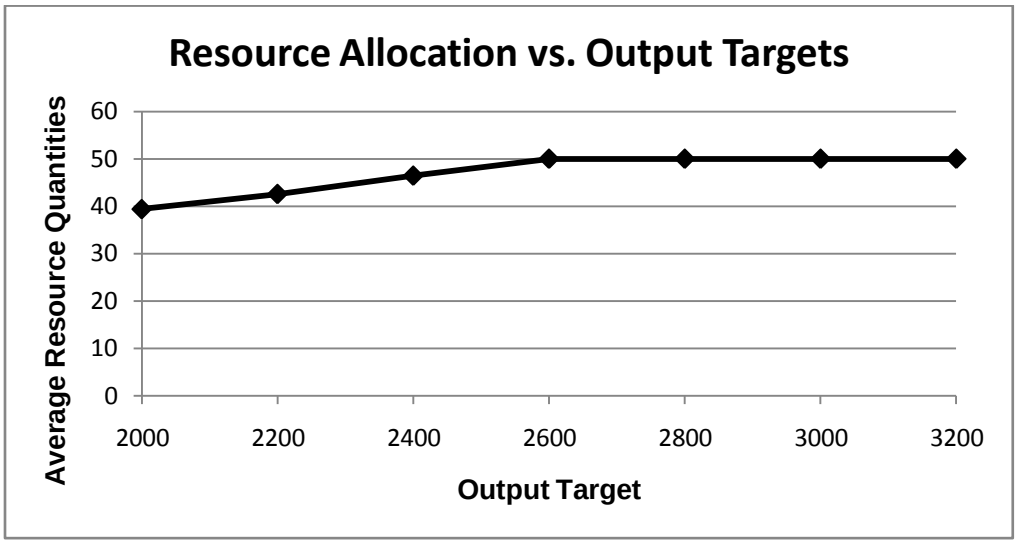

Figure 2: Resource Allocation vs. Output Target (Created by author)

There are other interesting results of this experiment. If we compare the two loss functions, we see that, as the output target level increases, the penalty of the loss function for Output 1 also increases; see Figure 3. The loss function for Output 1 is increasing despite the increase in resources. The loss function representing Output 2 decreases because resource quantities are increasing and it is easier to meet targets with more resources. We only show output targets up to 2600 in Figure 3, so that we can highlight the difference between the two loss functions.

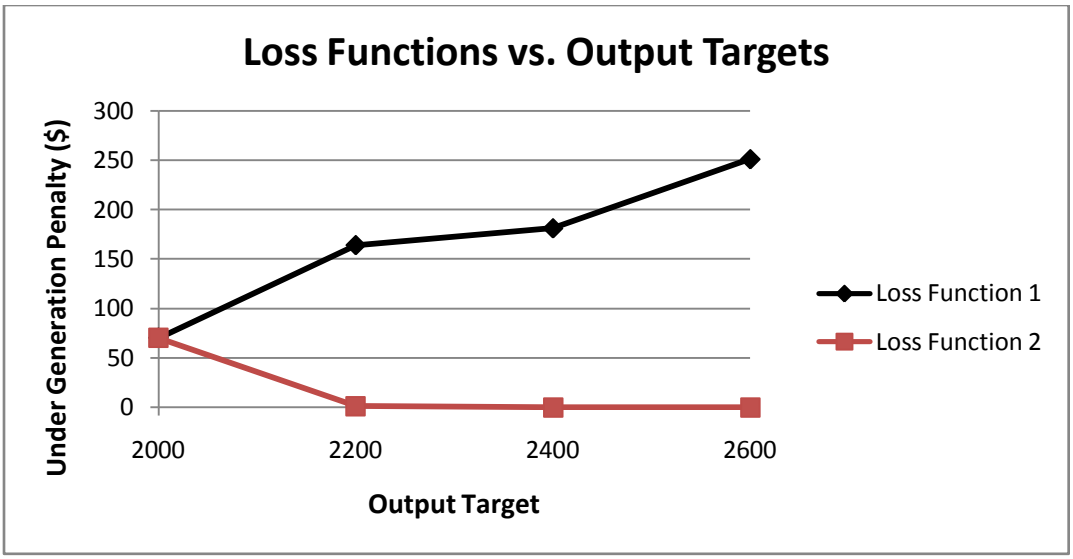

Figure 3: Loss Functions vs. Output Targets (Created by author)

Case 2: Resource Allocation vs. Worker Capacities

In this experiment we varied the service provider capacities across all inputs and held fixed the client capacities at the base-case value.

The results of this experiment show that when service provider capacities are low there is a need to use more client resources in order to meet output target levels; see Figure 4. As service provider capacities are increased, fewer client resources were needed. This behavior levels off after output target levels are achieved.

Service capacity constraints are not like those in manufacturing. When service provider capacity constraints are binding, client resources are allocated in order to keep the costs of missing the output targets low. As service provider capacities are increased, there is no longer a need to allocate more client resources because the cost of adding more resources outweighs the costs of missing the output targets. This is a behavior that is not typical in manufacturing. There is more flexibility in services. The technology function matrix allows for the exchange between service provider and client resources. There are variations in the "recipe" – more than one way to achieve the output.

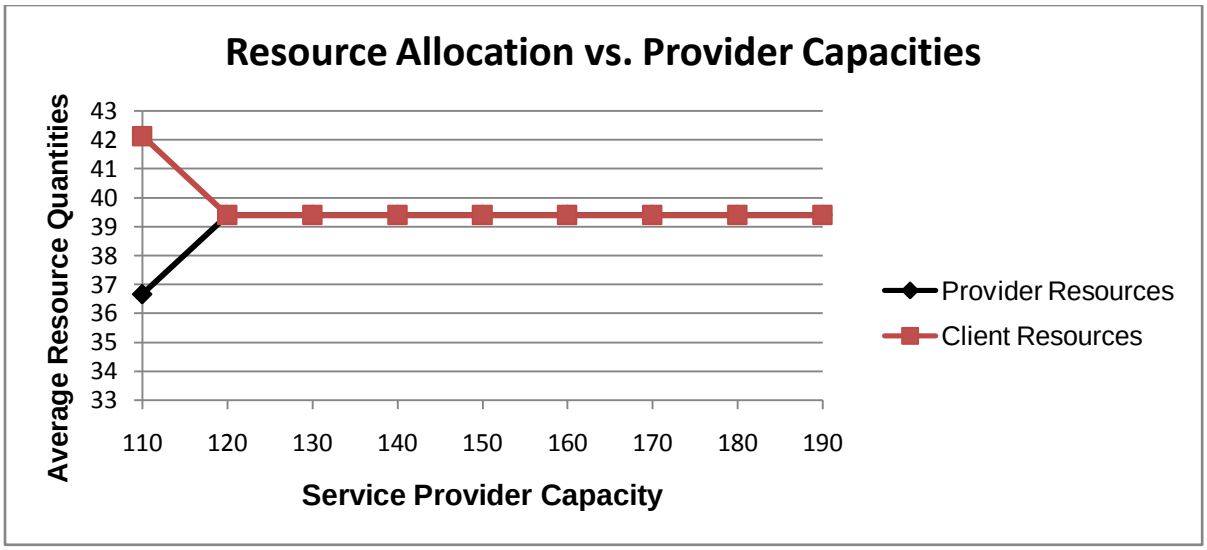

Figure 4: Resource Allocation vs. Service Provider Capacities (Created by author)

In Figure 5, the values of loss function for both output types are equal. When the service provider's capacity level increases from 110 to 120 there in an increase in provider resources and as a result there is a decrease in both loss function penalties. It is easier to meet output targets with more resources. The penalties of missing output targets level off as provider and client resource allocations level off.

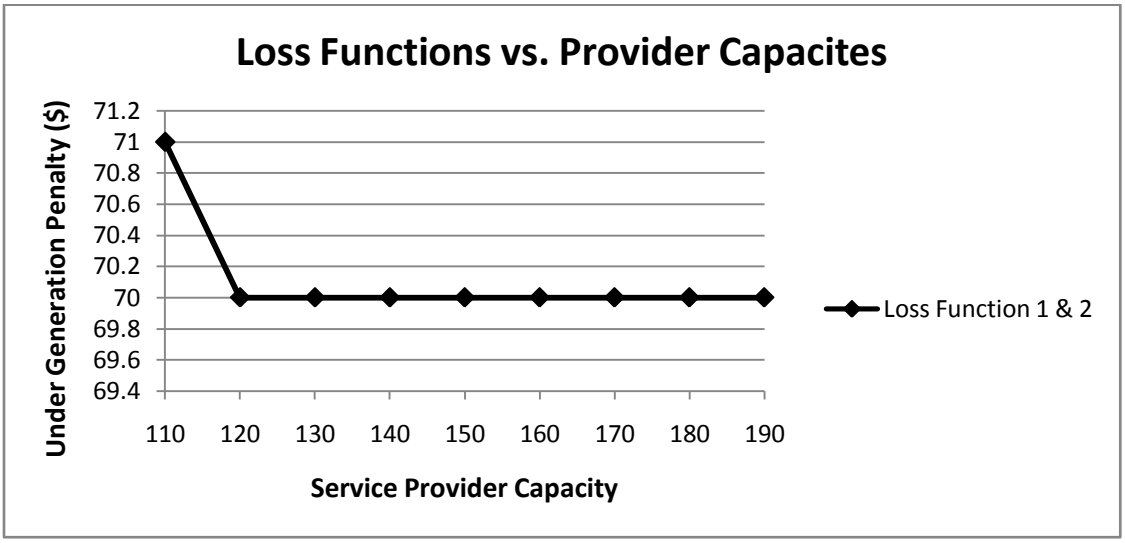

Figure 5: Loss Functions vs. Service Provider Capacities (Created by author)

Case 3: Resource Allocation vs. Underproduction Penalty

In this experiment we varied the weight on the loss function for Output 1, while keeping fixed the weight on the loss function of Output 2 fixed at the base case value.

The results of this experiment show that as the weight on the first loss functions is increased, more service provider and client resources are allocated in order to meet the target output levels and keep costs low.

The concavity seen in Figure 6 is due to the nonlinearity in the objective function. When the plot of the average resource quantities is concave, it exhibits diminishing marginal investment as the loss function weights increase. The resource costs and the loss make up the objective function. At optimality, there is a trade-off between resource costs and loss. The decision model finds an optimal solution that is a compromise between the increasing resource costs versus the increase in loss.

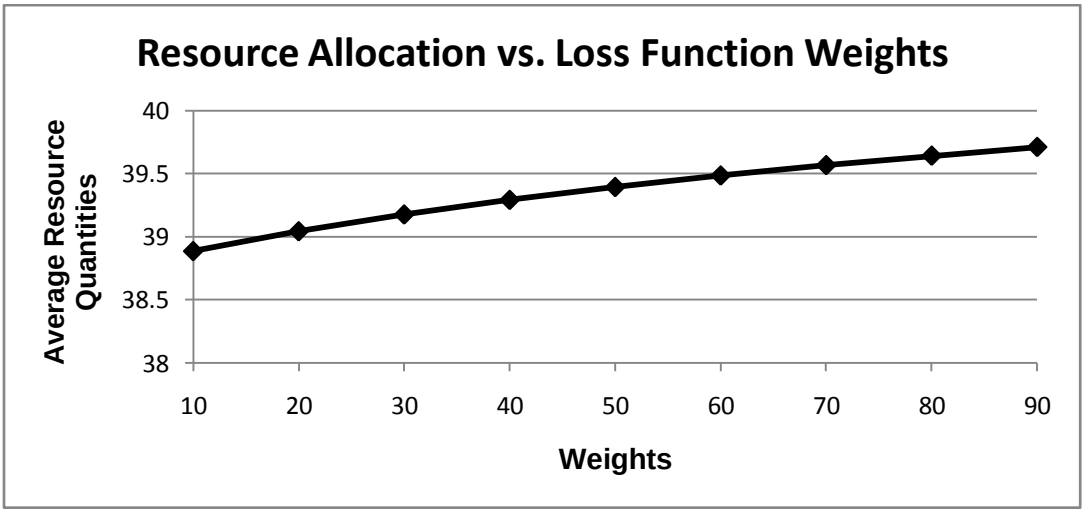

Figure 6: Resource Allocation vs. Loss Function Weights (Created by author)

A plot of the two loss functions for this experiment is shown in Figure 7. As we increased the weights on Output 1for missing output targets, the costs for Output1 of the missing output targets also increased. Due to the increase in resource allocations as seen in Figure 6, output targets for Output 2 are easily met. Hence the decrease in the costs of loss function 2.

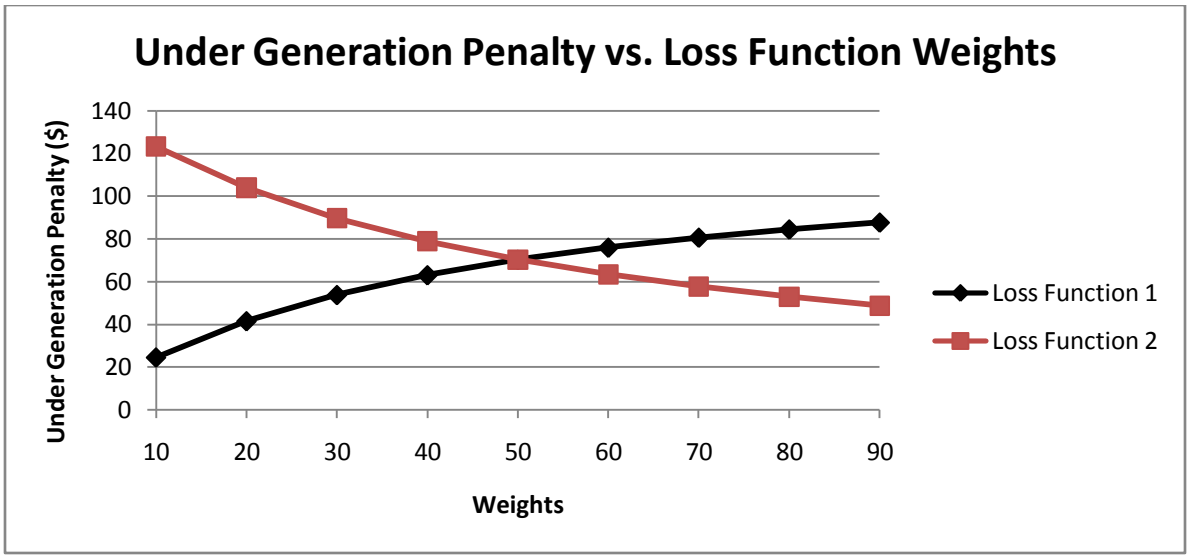

Figure 7: Under Generation Costs vs. Loss Function Weights (Created by author)

Case 4: Resource Allocation vs. Benchmark Generation rate

In this experiment we varied the benchmark generation rate for Process 1, while keeping fixed the benchmark generation rate of Process 2 and Process 3. By increasing the benchmark generation rate for Process 1, we made the process more efficient. The process is more efficient because the number of process cycles generated per unit of service provider and client resource increased.

The results of this experiment show that as the number of process cycles per unit input increases the service firm received more "bang for the buck", because fewer resources are needed to meet target output levels. Figure 8 also shows when the benchmark generation rate = 6, we are stealing from Process 2 & Process 3 because we have hit our provider and client capacity constraints.

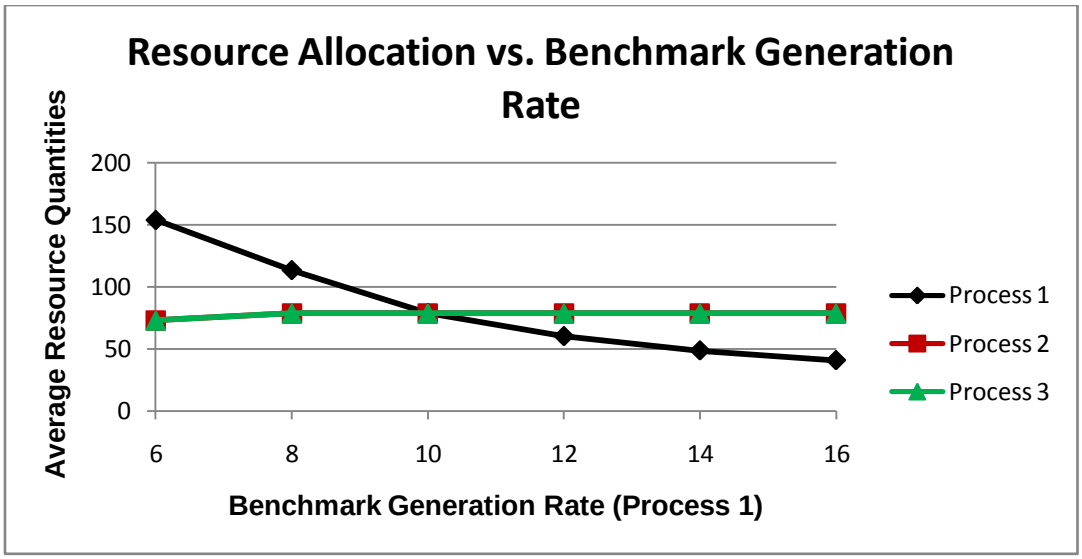

Figure 8: Resource Allocation vs. Benchmark Generation rate (Created by author)

Case 5a: Resource Allocation vs. Inefficiency Level

In this experiment we increased the inefficiency level, τ_{jip}, of Process 2 while the inefficiency level of other processes was held fixed at the base case value.

The results of this experiment show that as Process 2 becomes more inefficient, more resources are allocated to that process; see Figure 9. When the inefficiency level ≤ 3, Process 2 is more efficient than Process 1 and Process 3 and when the inefficiency level > 3 Process 2 is less efficient. As Process 2 becomes more inefficient, more resources are allocated to that process in order to minimize penalty costs.

In Figure 10 we see that as Process 2 becomes more inefficient the penalty cost of not meeting output targets increases because it is harder for Process 2 to meet its target output level. Process 1 and 3 penalty costs of not meeting output targets are equal and constant.

The convexity seen in Figure 9 and 10 is due to the nonlinearity in the objective function. There is an increasing marginal investment in resource costs and loss as processes become more inefficient. The significance of an increasing marginal investment of average resource quantities and loss is that the service provider and the client should ensure processes are as efficient as possible in order to keep resource costs and loss low.

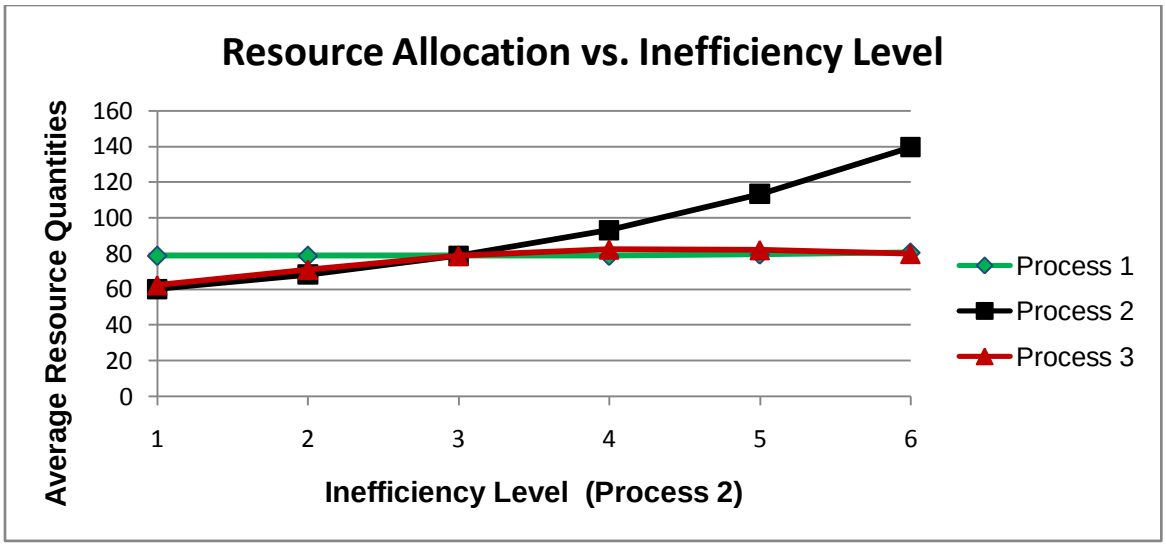

Figure 9: Resource Allocation vs. Inefficiency Level (Created by author)

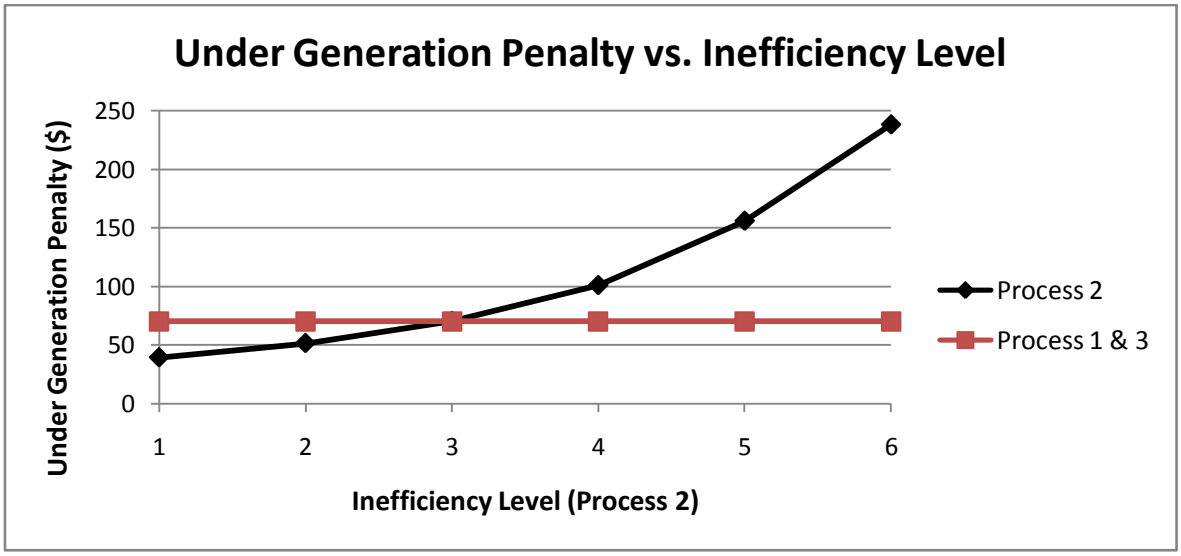

Figure 10: Under Generation Penalty vs. Inefficiency Level (Created by author)

Case 6a: Resource Allocation vs. Risk Level

In this experiment we increased the randomness (risk) of Process1, while keeping fixed the randomness of Process 2 and Process 3.

The results of this experiment show that as the risk level of Process 1 is increased, more service provider and client resources were allocated to that process in order to keep penalty costs low. The resource allocations for Processes 2 and 3 were evenly allocated.

In Figure 12 we see that as the risk (randomness) of Process 1 is increased, the penalty cost of not meeting output targets increases because it is harder for Process 1 to meet its target output level. Process 2 and 3 penalty costs of not meeting output targets are equal and constant.

The convexity seen in Figure 11 and 12 is due to the nonlinearity in the objective function. There is an increasing marginal investment in average resource quantities and loss as process risk increases. The significance of this experiment is that the service provider and the client now have more insights into the effects of increased process risk on resource costs and loss.

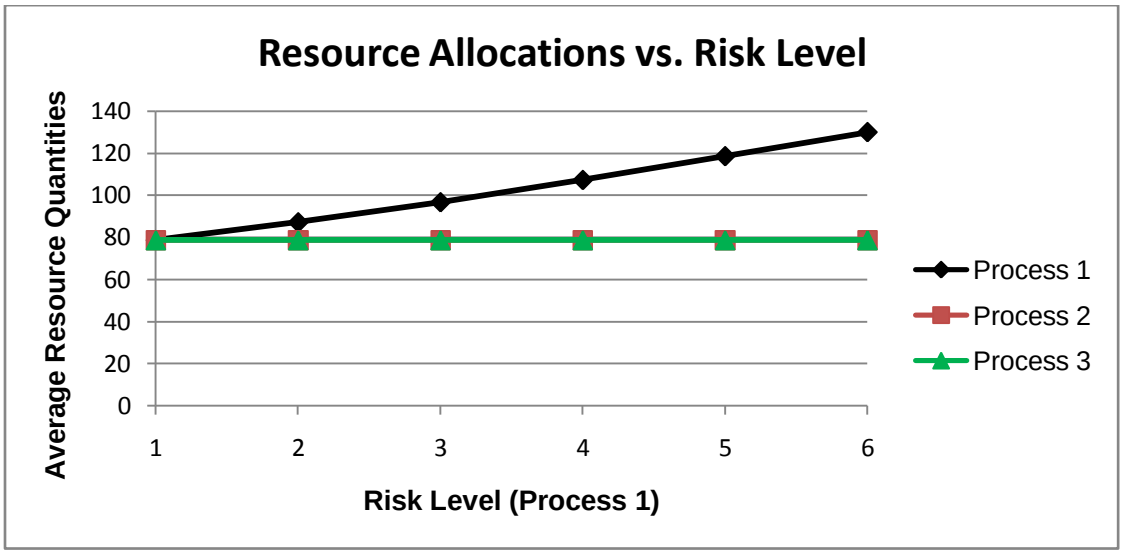

Figure 11: Resource Allocation vs. Risk Level (Created by author)

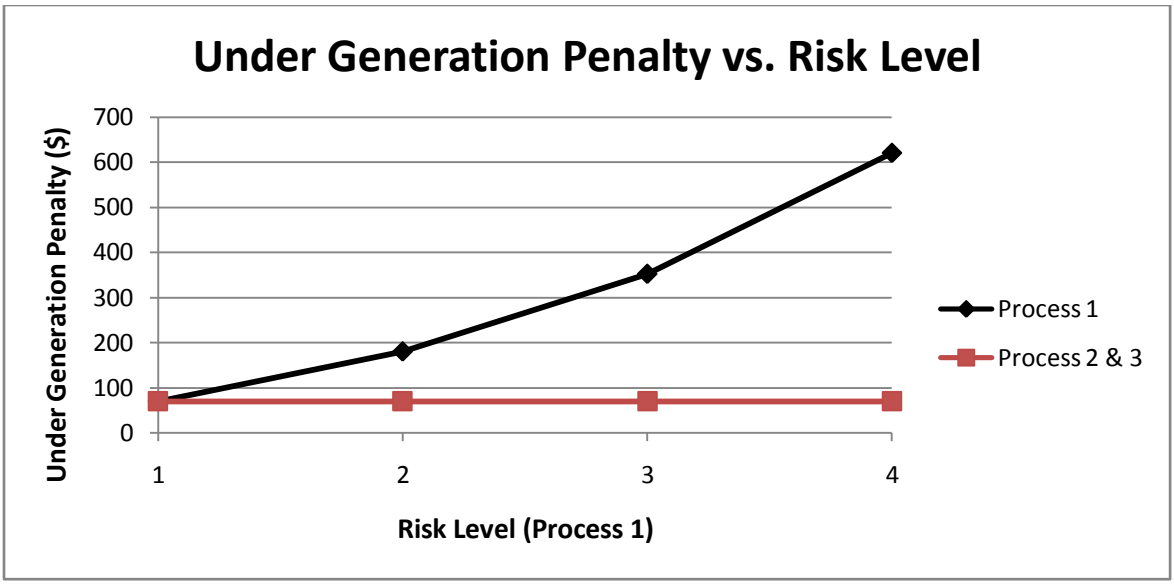

Figure 12: Under Generation Penalty vs. Risk Level (Created by author)

Case 6b: Resource Allocation vs. Risk Level

In this experiment we increased the risk level of all processes simultaneously. The vertical axis is total resource allocations summed across all input types. The resource quantities were the same for the service provider and the client in this experiment.

The results of this experiment show that as the risk level is increased, more service provider and client resources are allocated; see Figure 13. Resource quantities leveled out at 150 because of the capacity constraint.

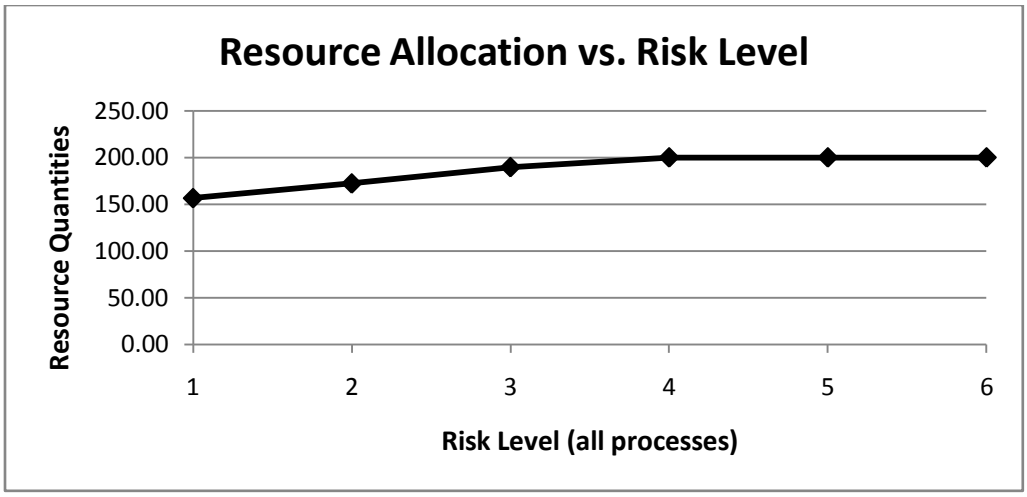

Figure 13: Resource Allocation vs. Risk Level (all processes) (Created by author)

5. Conclusion and Future Research

This paper has examined the effects of uncertainties of the structure of the technology function (mis-specification), uncertainties of the parameter values of the technology function (mis-estimation), and the uncontrollable inputs on resource plans for a service engagement.

Two resource-planning models were developed. The first model is a deterministic, resource-planning model that applies a benchmark technology function to obtain benchmark input levels for given target output quantities. We apply linear technology functions in which the inputs must be procured according to usage rates of a process cycle and outputs are generated according to yield rates of a process cycle.

The second model is a stochastic, resource-planning model. This model receives benchmark input levels and target output levels as parameters from the first model and allocates resources at a minimal cost. In the objective function of this model, we incorporate a probability density function that captures the deviation from the benchmark output levels reflecting both inefficiencies and uncertainties.

We prove that in the presence of technology function uncertainties, service firms will compensate by allocating resources so that penalty costs are minimized. The service provider and the client should put forth every effort to minimize uncertainties. Although there are

uncontrollable inputs, efforts should be made to improve technology function parameter inaccuracies and technology function form specification.

In the future, we will examine the effects of technology function uncertainties on a multi-period service supply chain. We will also extend this research by using various distributions for representing inefficiency randomness. Additionally, it would be beneficial to the field of service research to perform a case study to add further validity our model.

References

Abramowitz, M. and I. A. Stegun (1965). Handbook of Mathematical Functions: with Formula, Graphs, and Mathematical Tables, Dover Publications.

Asmild, M., J. C. Paradi, et al. (2006). "Centralized resource allocation BCC models." Omega **In Press, Corrected Proof**.

Athanassopoulos, A. D. (1995). "Goal programming & data envelopment analysis (GoDEA) for target-based multi-level planning: Allocating central grants to the Greek local authorities." European Journal of Operational Research **87**(3): 535-550.

Athanassopoulos, A. D. (1998). "Decision support for target-based resource allocation of public services in multiunit and multilevel systems." Management Science **44**(2): 173.

Badinelli, R. (2008). Resource Allocation for Service Supply Chains. INFORMS Annual Meeting. Washington, DC.

Banker, R. D., H. Chang, et al. (2003). "The public accounting industry production function." Journal of Accounting and Economics **35**(2): 255-281.

Charnes, A., W. W. Cooper, et al. (1978). "Measuring the efficiency of decision making units." European Journal of Operational Research **2**: 429-444.

Chase, R. (1978). "Where Does the Customer Fit in a Service Operation." Harvard Business Review: 137-142.

Chase, R. B. (1981). "The Customer Contact Approach to Services: Theoretical Bases and Practical Extensions." Operations Research **29**(4): 698.

Fare, R., R. Grabowski, et al. (1997). "Efficiency of a fixed but allocatable input: A non-parametric approach." Economics Letters **56**(2): 187-193.

Fitzsimmons, J. A. (2001). Service Management. Boston, McGraw-Hill.

Fitzsimmons, J. A. and M. J. Fitzsimmons (2004). Service Management. New York, McGraw-Hill.

Golany, B., S. T. Hackman, et al. (2006). "An efficiency measurement framework for multi-stage production systems." Annals of Operations Research **145**(1): 51.

Golany, B., F. Y. Phillips, et al. (1993). "Models for improved effectiveness based on DEA efficiency results." IIE Transactions **25**(6): 2.

Golany, B. and E. Tamir (1995). "Evaluating efficiency-effectiveness-equality trade-offs: A data envelopment analysis approach." Management Science **41**(7): 1172.

Koopmans, T. C. (1951). Analysis of Production as an efficient combination of activities. New York, Wiley.

Korhonen, P. and M. Syrjänen (2004). "Resource Allocation Based on Efficiency Analysis." Management Science **50**(8): 1134.

Kumar, C. K. and B. K. Sinha (1999). "Efficiency based production planning and control models." European Journal of Operational Research **117**(3): 450-469.

Land, K., C. A. Knox Lovell, et al. (1993). "Chance-Constrained Data Envelopment Analysis." Mangerial and Decision Economics **14**(6): 541-554.

Lozano, S. and G. Villa (2004). "Centralized Resource Allocation Using Data Envelopment Analysis." Journal of Productivity Analysis **22**(1/2): 143.

Lozano, S. and G. Villa (2005). "Centralized DEA models with the possibility of downsizing." The Journal of the Operational Research Society **56**(4): 357.

Mills, P. K. and D. J. Moberg (1982). "Perspectives on the Technology of Service Operations." The Academy of Management Review **7**(3): 467-478.

Nachum, L. (1999). "The Productivity of Intangible Factors of Production: Some Measurement Issues Applied to Swedish Management Consulting Firms " Journal of Service Research **2**(2): 123-137.

Rousseau, D. M. (1979). "Assessment of technology in organizations: Closed versus open systems approaches." Academy of Management. The Academy of Management Review (pre-1986) **4**(000004): 51.

Spohrer, J., P. Maglio, et al. (2007). Steps Toward a Science of Service Systems. Computer**:** 71-77.

Thanassoulis, E. (1996). "A data envelopment analysis approach to clustering operating units for resource allocation purposes." Omega **24**(4): 463-476.

Vargo, S. L. and A. Archupru (2009). "Service-Dominant Logic as a Foundation for Service Science: Clarifications." Service Science **1**(1): 32-41.

CHAPTER 5

MULTI-STAGE, MULTI-SERVICE RESOURCE PLANNING

Abstract

Planning and allocating resources for service operations in which the client is a co-generator of the service is an important problem facing decision makers. The problem becomes more daunting when the resources must be allocated to multiple service processes over multiple time periods. Therefore, the objective of this study is to gain insights as to how resource decisions are made for multiple stages and for over multiple clients. The model in this study allows decision makers to determine the optimal workforce level, the level of client involvement, and the level of service generation.

Keywords: service operations, resource planning, client involvement, multi-stage

1. Introduction

The objective of this study is to gain insights as to how resource decisions are made for multiple service processes over multiple time periods. We develop a non-linear, deterministic, multi-stage planning model that allows for examination of trade-offs among client intensity, efficiency and quality. Our model contributes to the literature as follows:

- We derive guidelines for resource planning policies that are specific to client co-generated services.
- We examine the sensitivity of policies to the effectiveness of client involvement on efficiency and quality.
- We examine the effects on policy of multi-stage, multi-service decisions.

Inherent in services, such as consulting and business services, is a close client-provider relationship, which is a determinant of the successful delivery of the service. This relationship is

known a client co-generation. This kind of service process typically starts with a contract or agreement between the client and the service provider, which describes the acceptable lead time of the job, the structure of the job and payment, and the responsibilities of all parties involved (Dietrich, 2006). The provider gathers pertinent information from the client regarding the requirements, specifications, and delivery date of the job. Throughout this process the client is an active participant. Our model captures the involvement of the client throughout the service process when the client is seen as a co-generator of the service. We assume that, for each client, there is an established service agreement and delivery date for the work and that the processes needed to complete the job have been identified.

We define a service as a transformation of inputs into outputs such that value for the client is created through a process that utilizes capabilities and capacities of both the client and the provider. Like our definition of services, other definitions highlight the fact that the client and the provider are involved in the service-creation process and that there is a shared creation of value (Spohrer et al., 2007). Sampson (2007) offers a similar definition and what he classifies as a paradigm - "The customer-supplier service paradigm holds that customer inputs are a necessary and sufficient condition for a service process to be a service process, and the lack of customer inputs characterizes all non-service processes".

We position our model at the operational level of the business decision hierarchy. To date, most research into service management has been done on strategic decisions. Our research responds to the need for modeling of tactical and operational decisions (Machuca et al., 2007). Specifically, we focus on modeling operational problems which span the functional areas of production of service and human-resource management.

Resource planning is a sequential decision process that strives to apply organization's capacity most efficiently to meet demand (Holt et al., 1995). The plan is typically for a horizon of 6 to 12 months and aims to find optimal decisions concerning production quantities, resource levels, inventory, and backorders.

Traditionally, resource plans seek to minimize costs subject to constraints on capacity, workforce, and inventory. We have extended the conventional resource planning model to

include the client as a source of direct labor. Researchers have noted that having the client involved in the service creation process improves productivity (Chase, 1981; Fitzsimmons, 1985; Bowen, 1986; Mills and Morris, 1986). We have also added efficiency and quality measures which are functions of client intensity, client and worker skills, and the quality of the inputs to each service process. The efficiency function augments the customary capacity constraint to make our model realistic and relevant to co-generated services. Traditional resource planning models have a linear objective function and constraints (Holt et al., 1995). Our model, however, introduces nonlinearity into the constraints with the inclusion of the efficiency and quality functions.

2. Literature review

Researchers continue to debate an official definition of services. Below is a sampling of definitions from the literature.

- Sasser et al (1978): "Intangible and perishable… created and used simultaneously".
- Lovelock (1983): A service is "characterized by its nature (type of action and recipient), relationship with customer (type of delivery and relationship), decisions (customization and judgment), economics (demand and capacity), mode of delivery (customer location and nature of physical or virtual space)".
- Chase and Aquilano (1992): A service business is the "management of organizations whose primary business requires interaction with the customer to produce the service."
- Fitzsimmons (2001): "A time-perishable, intangible experience performed for a customer acting in the role of co-producer."
- Spohrer, Maglio, Bailey, Gruhl. (2007): "….service is a kind of action, performance, or promise that's exchanged for value between provider and client."
- Vargo, Akaka (2009): "…function of service systems is to connect people, technology and information through value propositions with the aim of co-creating value for the service systems participating in the exchange of resources within and across systems."

One trend to notice is that some of the more recent definitions of service highlight the role and/or participation of the client. Our definition is no different in this regard; we define services as the transformation of inputs into outputs such that value for the customer is created through a process that utilizes capabilities and capacities of both the customer and the provider. Several mathematical models explicitly define functions for service quality and efficiency. See Gaimon (1997); Carrillo and Gaimon (2004); Napoleon and Gaimon (2004). Our model focuses on the effects client intensity on service quality and productivity. By contrast the aforementioned papers focus on the effects of information technology on services.

Service management has been studied for many years and is attracting the interests of more researchers. The attention is growing as the economy is largely dominated by workers employed in the services sector (Fitzsimmons and Fitzsimmons, 2004). There are far too many articles on service management to review in this paper, so we will highlight the ones specific to client participation in the service process and those focusing on service operations. Chase and Levitt (1972) and Chase and Garvin (1989) are among the pioneers and have laid the foundation for research in this field.

There are numerous pieces of literature that focus on customer participation in service operations and act as empirical support and motivation for our model. Bowen (1986) states that service productivity can be increased by having the customer perform certain service operation tasks. Lovelock and Young (1979) also suggests that by having the customer involved in the service system productivity can be increased. Service performance can also be improved by acknowledging the customer as a partial employee (Bowen, 1986; Mills and Morris 1986). Fitzsimmons (1985) explained that productivity can be improved by directly substituting customer labor for provider labor. Mills and Morris (1986) focus on how firms can affect costs saving by having customer involvement in service creation. Bitner et al. (1997) developed two frameworks for assisting decision makers and to direct future research related to customer participation. Martin et al. (2000) argue that a measurement of productivity that does not capture the client side of the encounter is inadequate in the case of a business service such as consulting. The current paper presents a mathematical resource-planning model that explicitly represents customer intensity in the service creation process using explicit efficiency and quality functions.

There are numerous mathematical models for the services sector. See Rust and Metter (1996) for a very comprehensive review of some of the most widely recognized models used by marketing researchers. Models of services operations planning are primarily directed at the health care and transportation sectors and do not involve the aspect of customer participation. In Table 1 we compare several service operations models that we believe are relevant to our research and show how our model differentiates itself from the others.

	Abernathy, Baloff et al. (1973)	**Ittig (1994)**	**Gaimon (1997)**	**Soteriou and Hadjinicola (1999)**	**Napoleon and Gaimon (2004)**	**Anderson et al. (2006)**	**White and Badinelli**
# of stages	Multiple	Single	Single	Multiple	Multiple	Multiple	Multiple
Objective	Costs	Profit	Profit	Minimize loss of service quality perceptions	Profit	Costs	Profit
Resource capacity changes	Considered	Not considered	Considered	Not considered	Considered	Considered	Considered
Resource training/learning	Considered	Not considered	Considered	Not considered	Considered	Not considered	Not considered
Customer waiting	Not considered	Considered	Not considered	Considered through responsiveness factor	Not considered	Considered through backlogging	Considered through backlogging
Inventory	Not considered	Not considered	Not considered	Not considered	Not considered	Not considered	Considered
Client Involvement	Not considered	Not considered	Not considered	Not considered	Not considered	Not considered	Considered
Demand	Stochastic	Stochastic	Deterministic	Not considered	Not considered	Stochastic	Deterministic
Methodology	Linear optimization	Non-linear optimization	Optimal control theory	Non-linear optimization	Optimal control theory	Optimal control theory	Non-linear Optimization
Service sector	Healthcare	Retail	None specified	Healthcare	Various examples given	Oilfield service firm	Professional services

Table 1: Comparison of Service Models (Created by author)

3. Descriptive model

In this section we present a model of a multistage service-operations plan in terms of the number of process cycles generated, capacity changes, and the amount of client intensity. We define the deliverables of the service firm as *service types*. In the case of a software consulting firm, for example, service types could be database designs, web-page construction, code writing and testing, etc. We assume that the firm delivers each type of service through stages of processes. We assume that the precedence constraints among the processes that are required for a particular service type define a serial network of these processes. Hence, for each service type, there is a known serial chain of process stages. Each stage requires a certain number of *cycles* of a process per unit of the service that is delivered to the client. A cycle is a single iteration of a process. The cycle of each stage has a standard labor requirement and a standard lead time, which for the sake of simplicity of exposition, we set it to one. If a process stage requires more than one time period, then we split the stage into multiple stages so that each stage can be accomplished in one time period. Processes are grouped into categories called process types. All processes within a process type require similar labor expertise.

3.1 Definitions and notation

Indices

t = time period index

x = service job index

p = process index

s = stage index

Index Sets:

T = number of planning periods (months)

S_x = set of service types

S_p = set of process types

Decision Variables

g_{sxt} = generation of process cycles of stage s of service x in period t (# of completed cycles)

h_{pt} = number of workers hired for process p in period t

f_{pt} = number of workers laid-off from process p in period t

z_{sxt} = number of hours that client personnel are assigned to stage s of service x in period t

State Variables

i_{sxt} = "inventory" of completed process cycles of stage s of service x at the end of period t (# of cycles)

b_{xt} = backlog of units of service x at the end of period t (# of units)

We define "inventory" as the number of completed cycles of a stage, which have been completed but for which the succeeding stage has not started. The analogue of inventory holding cost is captured through the use of discounted cash flows. If processes are completed prior to the times that they are needed, then the labor cost of generating these processes is recognized in the time periods of process generation and the revenue that is earned by the services for which these processes are generated is recognized at the due dates of the services. Hence, early generation will result is lower net present value.

The backlog variable measures the number of units of services that are not completed by their due dates. We will impose a cost on this backlog in order to capture the loss of goodwill penalties as well as the per-period cost of deferred revenues due to discounting of cash flows.

Performance measures

w_{pt} = size of workforce employed for process type p at time t (# of workers)

e_{sxt} = efficiency of the process of stage s of service x in period t as a function of client intensity, worker skill, client skill, and the quality of the input from the previous process.

q_{sxt} = quality of the process of stage s of service x in period t as a function client intensity, worker skill, client skill, and the quality of the input from the previous process.

Parameters

Revenue & Cost Rates:

c_h = cost of hiring a full-time regular employee

c_f = cost of firing a full-time regular employee

c_w = cost of wages per worker-period

c_b = penalty cost of backlog of service x ($ per unit-period)

c_z = cost of client intensity ($/labor-hour)

v_x = revenue per unit of service x

Other:

d_{xt} = forecasted demand of service x in period t

l_x = total standard lead time for service x (# periods).

p_{sx} = the process of stage s of product x

r_{sx} = required number of cycles of the process of stage s per unit of service job x as

$s \in \{1,2,\ldots l_x\}$

r_{sx}^h = number of standard labor hours required per cycle of the process of stage s of service job x

a_{pt}^{w} = number of hours of worker availability for process type p per worker-period

a_{sxt}^{c} = available client capacity for stage s of service x in period t (hours)

$\bar{b}_x$ = maximum allowed backlog of service job x (# units)

$\underline{y}_{sx}$ = minimum required participation of client in stage s of service x as a fraction of standard process time spent on the stage

$\bar{y}_{sx}$ = maximum allowable participation of client in stage s of service x as a fraction of standard process time spent on the stage

$\underline{q}_p$ = minimum required quality level of process type p

u = financial discount rate per period

Key Performance Indicator

N = discounted net profit over the planning horizon

3.2 Efficiency and Quality functions

The dependencies of the efficiency and quality functions on client intensity and skill levels, which we capture in our model of these functions, are motivated and supported by empirical literature. Much literature supports the inclusion of client intensity as an argument of the efficiency and quality functions (Chase 1978; Fitzsimmons, 1985; Bowen, 1986; Mills and Morris, 1986; Lance et al., 2002). Another argument of the function is service provider skill. To support this inclusion we refer to Bettencourt et al. (2002), which states that clients and employees must be trained and motivated for the co-generation relationship to work. Bowen (1986) supports the dependency on client skill, stating that customers can be better managed in the service creation process when they have the skills for the required tasks. We note that

provider and client skills were not explicitly modeled in this study. The last argument of both functions is the quality of the preceding stage's output, reflecting the fact that the quality of work that is passed from one process stage greatly affects the efficiency and quality of the succeeding process stage.

Efficiency function characteristics

We assume that the efficiency function increases to a maximum value of 1.0 as a function of client intensity. Below, we summarize the characteristics of the efficiency and quality functions that follow from reasonable assumptions about the nature of services.

We define the level of client intensity for a given stage of a given service type in terms of the ratio of client time to the total worker time spent on that stage.

$y_{sxt} = \dfrac{z_{sxt}}{r_{sx}^{h} g_{sxt}}$ = level of client intensity in period t for stage s of service x

We express efficiency and quality as functions of their arguments as,

$$e_{sxt}\left(y_{sxt}, k^{c}_{p_{sx}t}, k^{w}_{p_{sx}t}, q_{s\text{-}1,x,t\text{-}1}\right)$$

$$q_{sxt}\left(y_{sxt}, k^{c}_{p_{sx}t}, k^{w}_{p_{sx}t}, q_{s\text{-}1,x,t\text{-}1}\right)$$

For simplicity of presenting the fundamental characteristics of these functions below, we suppress the subscripts wherever this will not cause confusion.

1. Efficiency increases at a diminishing marginal rate as client involvement increases. $\partial e(y,k^{c},k^{w},q)/\partial y \geq 0\,;\ \partial^{2} e(y,k^{c},k^{w},q)/\partial y^{2} \leq 0$

2. Efficiency increases at a diminishing marginal rate as a function of worker skill. $\partial e(y,k^{c},k^{w},q)/\partial k^{w} \geq 0\ \ \partial^{2} e(y,k^{c},k^{w},q)/\partial k^{w\,2} \leq 0$

3. Efficiency increases at a diminishing marginal rate as a function of client skill. $\partial e(y,k^c,k^w,q)/\partial k^c \geq 0,\ \partial^2 e(y,k^c,k^w,q)/\partial k^{c\,2} \leq 0$

4. Efficiency increases as a function of the quality of the preceding process. $\partial e(y,k^c,k^w,q)/\partial q \geq 0$

Quality function characteristics

1. Quality increases at a diminishing marginal rate as client involvement increases. $\partial q_s(y,k^c,k^w,q_{s-1})/\partial y \geq 0;\ \partial^2 q_s(y,k^c,k^w,q_{s-1})/\partial y^2 \leq 0$

2. Quality increases at a diminishing marginal rate as a function of worker skill. $\partial q_s(y,k^c,k^w,q_{s-1})/\partial k^w \geq 0\ \ \partial^2 q_s(y,k^c,k^w,q_{s-1})/\partial k^{w\,2} \leq 0$

3. Quality increases at a diminishing marginal rate as a function of client skill. $\partial q_s(y,k^c,k^w,q_{s-1})/\partial k^c \geq 0,\ \partial^2 q_s(y,k^c,k^w,q_{s-1})/\partial k^{c\,2} \leq 0$

4. Quality increases as a function of the quality of the preceding process. $\partial q_s(y,k^c,k^w,q_{s-1})/\partial q_{s-1} \geq 0$

These assumptions determine characteristics of the shape of the efficiency and quality functions.

See Figures 1 and 2 for examples.

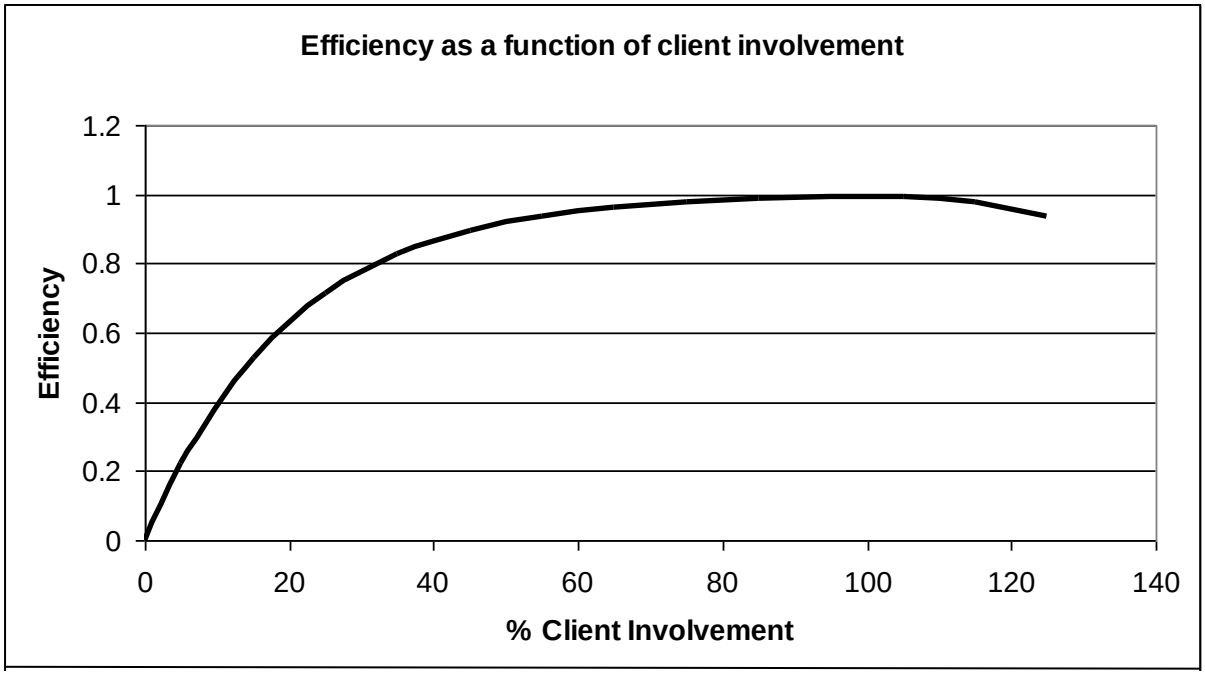

Figure 1: Example of an efficiency function (Created by author)

Figure 2: Example of a quality function (Created by author)

3.3 Model formulation

$$\text{Maximize} \sum_{t=1}^{T} e^{-ut} \left\{ \sum_{x \in S_x} \left(q_{xt} v_x - c_{bx} b_{xt} \right) - \sum_{p \in S_p} \left(c_h h_{pt} + c_f f_{pt} + c_w w_{pt} \right) - \sum_{x \in S_x} \sum_{s=1}^{l_x} c_z y_{sxt} r_{sx}^h g_{sxt} \right\}$$

Subject to:

$$i_{sxt-1} + g_{sxt} - \frac{r_{sx}}{r_{s+1x}} g_{s+1,x,t+1} - i_{sxt} = 0 \qquad \text{, for all } x \in S_x, p < l_x \qquad (1)$$

$$i_{l_x xt-1} + g_{l_x xt} - r_{l_x x} d_{xt} - i_{l_x xt} + r_{l_x x} b_{xt} = 0 \qquad \text{, for all } x \in S_x \qquad (2)$$

$$w_{pt} - w_{pt-1} - h_{pt} + f_{pt} = 0 \qquad \text{, for all } p \in S_p \qquad (3)$$

$$\bar{b}_x - b_{xt} \geq 0 \qquad \text{, for } x \in S_x \qquad (4)$$

$$a_{pt}^w w_{pt} - \sum_{p_{sx}=p} \frac{g_{sxt} r_{sx}^h}{e_{sxt}} \geq 0 \qquad \text{, for all } p \in S_p \qquad (5)$$

$$a_{sxt}^c - y_{sxt} r_{sx}^h g_{sxt} \geq 0 \qquad \text{, for all } x \in S_x, \ s \in \{1,2,\ldots l_x\} \qquad (6)$$

$$y_{sxt} - \underline{y}_{sx} \geq 0 \qquad \text{, for all } x \in S_x, \ s \in \{1,2,\ldots l_x\} \qquad (7)$$

$$\bar{y}_{sx} - y_{sxt} \geq 0 \qquad \text{, for all } x \in S_x, s \in \{1,2,\ldots l_x\} \qquad (8)$$

$$q_{sxt} - \underline{q}_{p_{sx}} \geq 0 \qquad \text{, for all } x \in S_x, \ s \in \{1,2,\ldots l_x\} \qquad (9)$$

Constraints (1) and (2) are the inventory balance constraints. The first constraint ensures that any work performed at any stage of a service is supported by the requisite process completions at the previous stage. The second constraint expresses the inventory and backlog balance for the final stage of a service. Through this constraint, any shortages of process completions of the final stage of a service are recognized as backlog. Constraint (3) is a

traditional workforce balance constraint. Constraint (4) ensures that the number of service jobs late (backordered) does not exceed the maximum allowable amount.

Constraints (5) and (6) are capacity constraints. The efficiency term in the capacity constraint (5) is needed to represent the effect of client involvement, skill levels and input quality on the effective capacity of the workforce. If client involvement, skills, or quality is negatively impacting efficiencies, then the firm will not have enough capacity to meet forecasted demand.

In constraints (7) and (8) we have set minimum and maximum amounts of client intensity, respectively. We assume that in co-generated service such as consulting and IT development the client will always be part the service creation process, to a certain extent, and that the client cannot be the sole labor source in any process.

Typically organizations benchmark themselves against competitors in terms of quality and establish internal quality standards. Hence, constraint (9) imposes a minimum process quality level that must be achieved.

4. Analytic Results: Parallels to the single-stage model

In the paper titled, "The Effects of Efficiency and Quality on Resource Planning for Co-Generated Services" (Chapter 3), a single-stage version of the model in this study was formulated and solved. We extrapolate theoretical findings from the single-stage model to discover insights as to how resources are allocated in a multi-stage, multi-service scenario.

In the single-period study, all possible solutions to the KKT conditions of Problem P2 were placed into five policy types. Policy Type 1 called for hiring up to the maximum workforce level. Policy Type 2 called for hiring, but not up to the maximum workforce level. Policy Type 3 called for maintaining the current workforce level. Policy Type 4 called for firing workers, but not down to the minimum level. Policy Type 5 called for firing workers down to the minimum level.

Theorem 3, of the single-stage study, established the forms of the optimal solution for any give problem parameterization. As the theorem states, each of the five policy types is applicable over certain intervals of the values of the initial workforce level, w_0. The workforce level of one stage is a state variable connecting the single-stage model and the multi-stage model.

It is assumed that each stage has a different required number of cycles and a different number of standard labor hours required per cycle of the process of stage. Therefore, the workforce will not merely be balanced between the stages and services.

4.1Multi-stage effects

Consider a single-service, single-process, two-stage problem. Assume a single worker type is used across all stages. We make this restrictive assumption in setting up the problem so that we can easily draw parallels to the findings of the single-stage study. This assumption allows us to focus on the tradeoffs of coordinating a resource plan across stages. In the objective functions listed below, the first subscript denotes the problems in the dynamic program (P1, P2 from the single-stage problem, and P3 for the multi-stage problem). The second subscript denotes the stage.

The objective function for stage 2 (exclusive of inventory and backlog cost), given an initial workforce of w_1 is denoted,

$$z_{22}(w_2, f_2, z_{12}(y_2, g_2; w_2); w_1, i_1) \tag{10}$$

$$z_{32}(w_1, i_1) = z_{22}(w_2, f_2, z_{12}(y_2, g_2; w_2); w_1, i_1) \tag{11}$$

$$z_{31}(w_0, i_0) = z_{21}(w_1, f_1, z_{11}(y_1, g_1; w_1); w_0, i_0) + h_1 i_1 + z^*_{32}(w_1, i_1) \tag{12}$$

Let's examine the behavior of the optimal workforce level of stage 2, given the workforce level of stage 2, $w^*_2(w_1)$. By Theorem 3, in the single-stage study, we know that

- *If* $w_0 < w_{h1}$ *then the optimal policy permits any solution that hires up to a workforce level in the interval* $[w_{h1}, w_{h2}]$.
- *If* $w_{h2} < w_0 < w_{f1}$ *then the optimal policy maintains the current workforce.*
- *If* $w_{f2} < w_0$ *then the optimal policy lays off down to a level within* $\lfloor w_{f1}, w_{f2} \rfloor$.

We know that the performance measures motivate adjustments in the workforce level from one stage to the next. When allocating resources across stages there is a trade-off between the costs of capacity changes versus profits.

Suppose stage 2 is given a workforce level from stage 1, $w_1 < w_{h1}$. Then the optimal policy permits any solution that hires workers up to a workforce level in the interval $[w_{h1}, w_{h2}]$. For every value of w_1 within this interval the service provider incurs hiring costs for stage 2. If the value of w_1 increases, then the cost of hiring decreases (i.e., if stage 2 starts with more workers then less workers need to be hired). Of course, the service provider would like to keep stage 2 hiring costs a low as possible. The only way to achieve this goal is to have stage 1 give stage 2 a larger workforce, thus decreasing stage 1's profits. Stage 1's costs increase from their optimal value. This dilemma presents a trade-off between the increased hiring costs in stage 2 versus lower profits in stage 1.

Suppose stage 2 is given a workforce level from stage 1, $w_{h2} < w_1 < w_{f1}$, then the optimal policy maintains this give workforce through stage 2. For every value of w_1 within this interval the service provider does not incur any hiring or firing costs in stage 2. By Proposition 8, in the single-stage study, we know that $z_{12}^{*}(w)$ is increasing and concave in w. Therefore, as w_1 increases within this interval, profits increase for stage 2. This dilemma also presents a trade-off between the increased hiring costs in stage 2 versus lower profits in stage 1."

Suppose stage 2 is given a workforce level from stage 1, $w_{f2} < w_1$, then the optimal policy lays off workers down to a level within $\lfloor w_{f1}, w_{f2} \rfloor$. For every value of w_1 within this interval the service provider incurs firing costs for stage 2. If the value of w_1 increases, then the cost of firing increases (i.e., if stage 2 starts with fewer workers then less workers need to be

fired). Of course, the service provider would like to keep stage 2 firing costs a low as possible. The only way to achieve this goal is to have stage 1 give stage 2 a smaller workforce, thus decreasing stage 1's profits. Stage 1's profits decrease because of higher backlog costs (i.e., fewer workers decrease service generation). This dilemma presents a trade-off between the increased firing costs in stage 2 versus lower profits in stage 1.

4.2 Multi-service effects

Consider a multi-service, single-process type, single-stage problem. Assume a single worker type is used across all processes. We make this restrictive assumption so that we can easily draw parallels to the findings of the single-stage study. This assumption allows us to focus on the trade-offs of apportioning resources across services. In the objective functions listed below, the first subscript denotes the inner stage of the dynamic program from the single-stage study (P1) and the second subscript denotes the service.

$z_{1a}(y_a, g_a; w_a) =$ the objective function for service a (exclusive of labor/hiring/firing cost), given an allocation of the workforce of w_a.

$z_{1b}(y_b, g_b; w_b) =$ the objective function for service a (exclusive of labor/hiring/firing cost), given an initial allocation of the workforce of w_b.

$z_{1a}(y_a, g_a; w_a) + z_{1b}(y_b, g_b; w_b) + c_w(w_a + w_b) + c_h h + c_f f =$ objective function where,

$$w - h + f = w_0$$

$$w = w_a + w_b$$

Theorem 2, in the single-stage study, proved that there are three cases for the optimal solution of the inner problem of the dynamic program. By applying Theorem 2 to the multi-service case, we can determine the behavior of optimal workforce, client intensity, and generation levels for each service.

We refer the reader to Chapter 3, Section 5.1 of this dissertation for a detailed explanation of the cases/subcases derived from Theorem 2. These cases/subcases identify specific forms for the optimal solution to Problem P1 of the dynamic program. Each service a or b can fall into any one of these seven cases/subcases, making forty-nine possible combinations. It should not be assumed that service a and b will be in the same case/subcase. We summarize these cases/subcases below.

For each case we define the minimum and maximum possible workforce levels as $\overline{w}, \underline{w}$.

$\overline{w}$ = the value of the workforce level at which the capacity constraint is binding at the point $(y_{\min}, g_{\max})$.

$\underline{w}$ = the value of the workforce level at which the capacity constraint is binding at the point $(y_{\max}, g_{\min})$.

Furthermore, for every case there are values of the workforce level at which the form of the policy changes from fixed generation to fixed client intensity relative to the workforce level. These values are denoted w_1, w_2. See Chapter 3 for clarification.

Case 1: $\hat{y} \leq y_{\min}$, $w_1 = \dfrac{r^h g_{\min}}{a^w e(y_{\min})}$

Sub-case 1.1: $\underline{w} \leq w < w_1$, $(y^*(w), g^*(w)) = (y_4(g_{\min}), g_{\min})$

Sub-case 1.2: $w_1 \leq w < \overline{w}$, $(y^*(w), g^*(w)) = (y_{\min}, g_4(y_{\min}))$

Case 2: $y_{\min} < \hat{y} \leq y_{\max}$, $w_1 = \dfrac{r^h g_{\min}}{a^w e(\hat{y})}$, $w_2 = \dfrac{r^h g_{\max}}{a^w e(\hat{y})}$

Sub-case 2.1: $\underline{w} \leq w < w_1$, $(y^*(w), g^*(w)) = (y_4(g_{\min}), g_{\min})$

Sub-case 2.2: $w_1 \leq w < w_2$, $(y^*(w), g^*(w)) = (\hat{y}, g_4(\hat{y}))$

Sub-case 2.3: $w_2 \leq w < \overline{w}$, $(y^*(w), g^*(w)) = (y_4(g_{\max}), g_{\max})$

Case 3: $y_{\max} < \hat{y}$, $w_1 = \dfrac{r^h g_{\max}}{a^w e(y_{\max})}$

Sub-case 3.1: $\underline{w} \le w < w_1$, $(y^*(w), g^*(w)) = (y_{\max}, g_4(y_{\max}))$

Sub-case 3.2: $w_1 \le w < \overline{w}$, $(y^*(w), g^*(w)) = (y_4(g_{\max}), g_{\max})$

For a given total workforce, w, the optimal allocation to w_a, w_b is determined by the condition,

$$\frac{\partial z_{1a}}{\partial w_a} = \frac{\partial z_{1b}}{\partial w_b} \tag{13}$$

Recall, that by Proposition 8, in the single-stage study, we know that $z_1^*(w)$ is increasing and concave in w. Let's examine how these derivatives can achieve equality when service a and b have different parameter values.

If the slope of z_{1a} is shallow for large values of w_a and the slope of z_{1b} is shallow for smaller values of w_b, then these derivatives will achieve equality when the parameter settings of service b are such that a smaller workforce is needed from service b. A service requires a smaller workforce when any of the following parameter settings are true: 1) the number of standard labor hours required per cycle is small; 2) the number of required cycles of stage is small; 3) the cost of client intensity is lower than the cost of provider wages (i.e., at optimality client resources will be used instead of provider resources); 4) demand is low; 5) higher backlog limit; 6) higher quality limit. If these parameters are set in the manner described above, the results of Chapter 3 tell us that the feasible regions for the two services are such that $\overline{w}_a > \overline{w}_b$ and $\underline{w}_a > \underline{w}_b$. Furthermore, the parameters will affect how condition 13 is satisfied. Condition 13 is satisfied at a solution which favors higher client intensity and lower workforce level for service b compared to service a. The derivatives in condition 13 will favor solutions for service b in sub-cases 1.1, 2.1, and 3.1 and for service a in sub-cases 2.3, 3.2, and 3.1.

5. Numerical Results

To illustrate the model we develop a set of cases/experiments using hypothetical data. In our experiments we chose an exponential form for the efficiency and quality functions. The cases are designed to investigate the nature of the optimal resource plan and the sensitivity of the resource plan with respect to various parameters of the service process. Experiments are run to show how resources are not merely balanced across multiple stages. Recall we assumed that each stage has a different required number of cycles and a different number of standard labor hours required per cycle of the process of stage.

Our test cases were based on a two-stage service supply chain. All cases have an efficiency and quality scaling factor of 1.0. The efficiency and quality leverage rate is varied among the different cases. We use the term leverage rate often in the remaining sections of the paper, so it is proper that we clarify the term at this time. The efficiency and quality leverage rate are coefficients used to convey the effectiveness of each unit of client intensity. By varying the rate parameter we change the concavity of the exponential function; see Figure 3.

In the efficiency and quality functions below, the efficiency and quality scaling factors are denoted e_1 and q_1, respectively and the efficiency and quality leverage parameters are denoted λ_e and λ_q, respectively.

$$e(y) = e_1(1 - exp[-\lambda_e y])$$

$$q(y) = q_1\left(1 - exp\left[-\lambda_q y\right]\right)$$

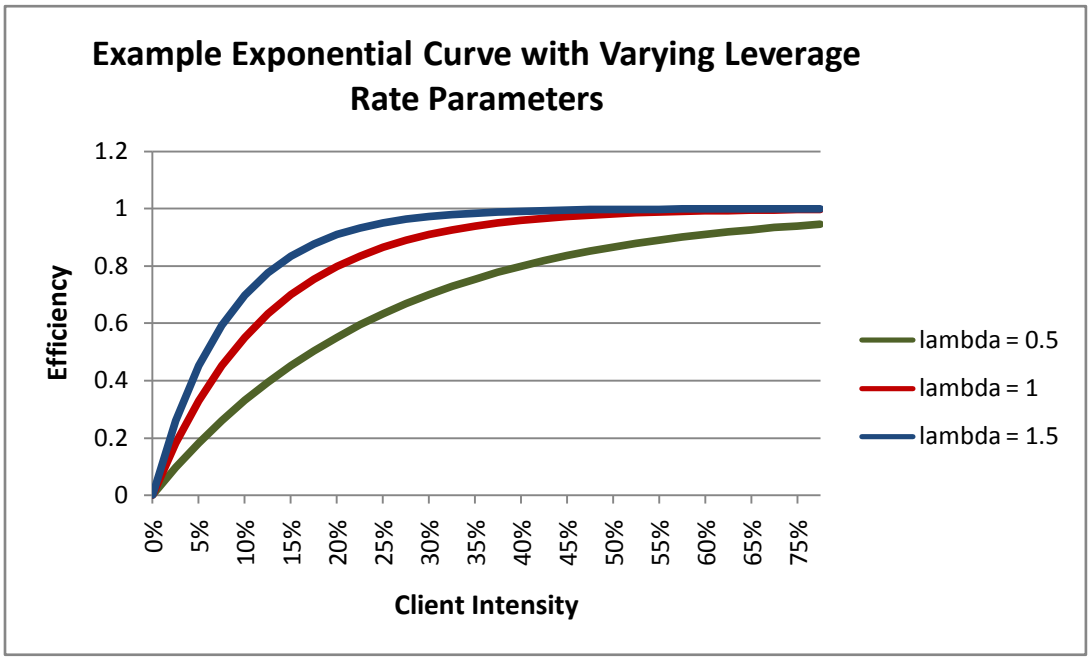

Figure 3: Example Efficiency Curve with Varying Leverage Parameter (Created by author)

The model was coded in the AIMMS 3.7 modeling software and solved with the KNITRO 5.1 optimization solver (Byrd, Nocedal et al., 2006). Table 2 shows the base case. The parameter values for the base case were chosen based on reasonable assumptions. For example, the service provider's available capacity per worker, a_{pt}^{w}, is set to 160 hours (40 hrs per wk x 4 wks per period). We acknowledge that an empirical study or case study needs to be performed in order to have more accurate estimates of the parameter values.

Parameter	Value	Parameter	Value	Parameter	Value
d_{xt}	40	c_z	50	$\bar{b}_x$	4
a^w_{pt}	160	c_h	10000	$\underline{y}_{sx}$	0.20
a^c_{sxt}	1000	c_f	5000	$\bar{y}_{sx}$	0.8
v_x	6000	c_w	8000	r^h_{sx}	80
r_{sx}	1	c_{bp}	5000	$\underline{q}_p$	0.70

Table 2: Base Case & Experimental Parameter Data (Created by author)

Case #1: Workforce Size vs. Required standard labor hours

In this case we increase the number of standard labor hours required per cycle of stage 2 while holding fixed the number of standard labor hours required per cycle of stage 1. The initial workforce size = 20 workers.

In Figure 4, workers are hired to support the requirements of stage 1 (workforce size increased from 20 to 22). Since stage 2's requirements are less than those of stage 1, the workforce size is reduced from 22 to 14. Fewer workers are fired for stage 2 as the number of standard labor hours required per cycle of stage 2 increases. Once the number of standard labor hours required per cycle of stage 2 is greater than those of stage 1, workforce increases up to a level that supports the requirements of stage 2.

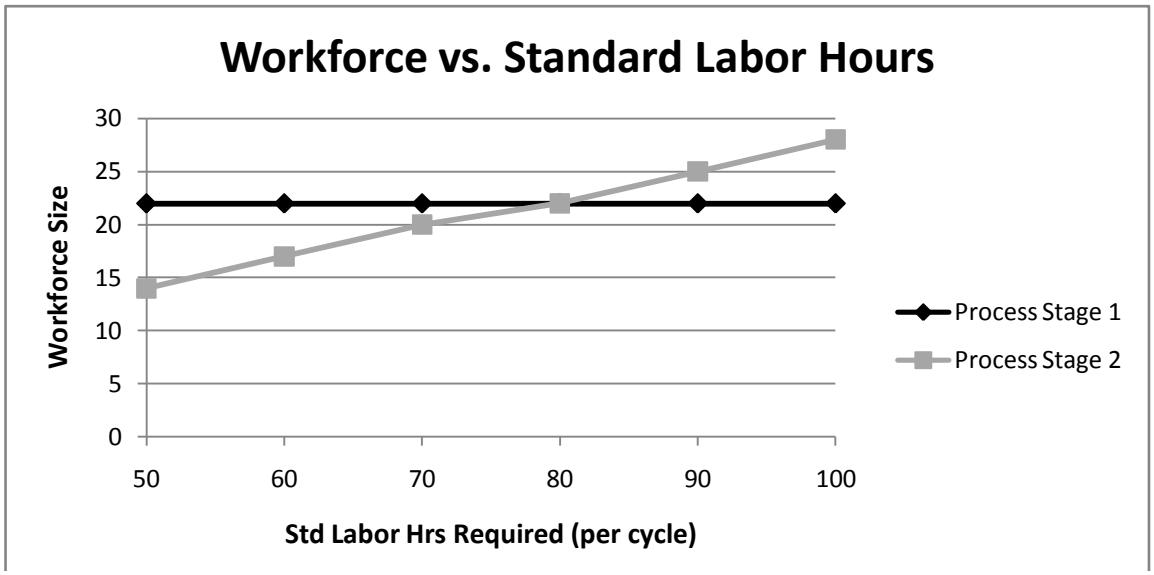

Figure 4: Workforce Size vs. Required Standard Labors (Created by author)

Case #2: Workforce Size vs. Required number of cycles

In this case we increased the number of required cycles of stage 1 and decreased the number of required cycles of stage 2. The number of required process cycles of a stage is equal to the demand for the service multiplied by the parameter, r_{sx}. The initial workforce size = 20 workers.

In Figure 5, workers are fired down to a level that supports the requirements of stage 1 (workforce size decreased from 20 to 2). Since stage 2 requires more cycles than those of stage 1, the workforce size is larger (although one worker was fired). As the required number of process cycles per stage increases (decreases), workers are hired (fired) accordingly.

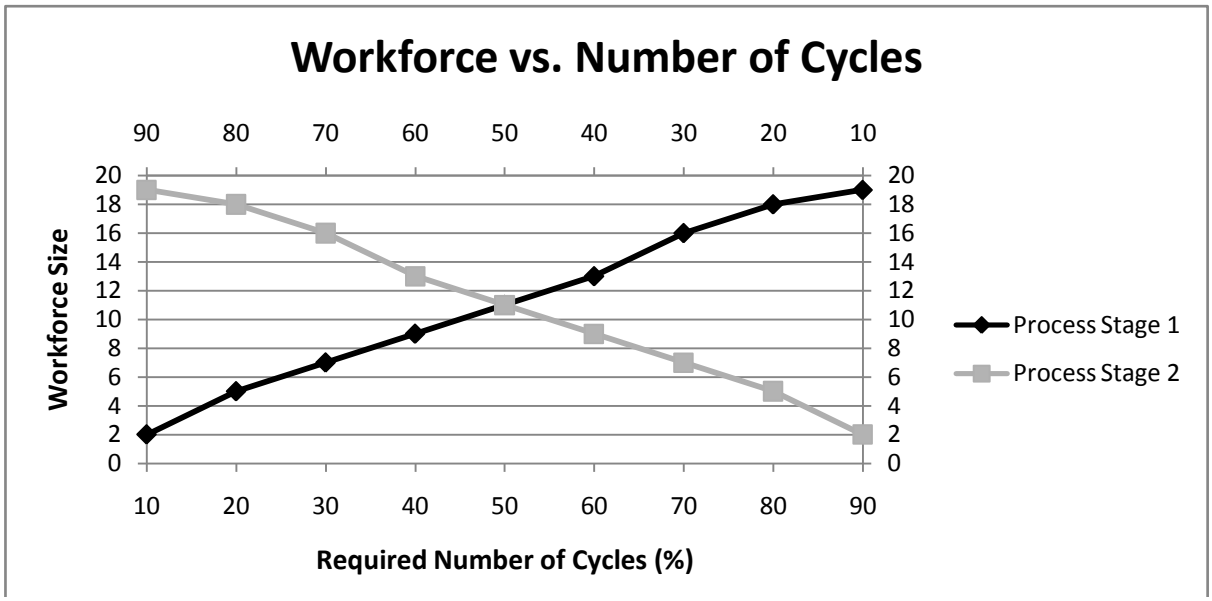

Figure 5: Workforce Size vs. Required Number of Service Cycles (Created by author)

Case #3: Client Intensity vs. Leverage rate parameters

In this case we investigate the optimal level of client involvement by observing the behavior between multiple processes when the efficiency and quality leverage parameters are increased. This case has also been divided into two similar experiments. In the first experiment, the efficiency leverage parameter is varied and the quality leverage parameter is held fixed. Increasing the leverage parameter of our exponential efficiency function has the effect of increasing the concavity of this function (see Figure 6). When the processes are evaluated individually an interesting trade-off is noticed.

We found that, due to the nonlinearity of the efficiency and quality functions, trade-offs between the amount of client involvement and the cost of labor per process are complex. The total cost of labor per process includes hiring and firing costs and wages. By increasing the leverage parameter of the efficiency function we increase the concavity of the function.

In Figure 6, the optimal level of client intensity shifts towards the downstream process as efficiency leverage increases. When efficiency leverage is low, client intensity must be kept high. When client intensity is kept high, the quality of the first (upstream) process is high (85%). When efficiency increases, the service provider is motivated to reduce total client intensity. When the service provider reduces total client intensity, quality of process 1 deteriorates (70%). In order to compensate for the effect of low quality of process1 on the efficiency of process 2, the service provider shifts client intensity to process 2. The shift happens when quality reaches 70% due to a minimum quality constraint in the model.

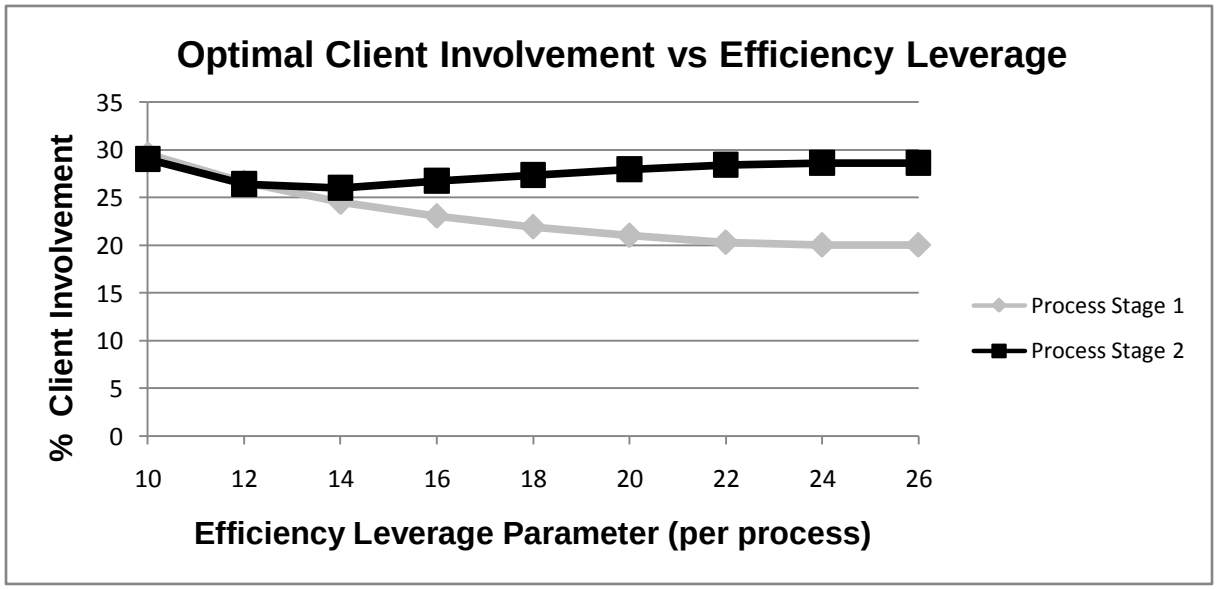

Figure 6: Optimal client intensity vs. efficiency leverage (per process) (Created by author)

In Figure 7, we see a similar dynamic. In this case we held fixed the efficiency leverage parameter at 10.0 and varied the quality leverage parameter. As we increased the quality leverage parameter the optimal level of client intensity shifts towards the downstream process. When quality leverage is low, client intensity must be kept high. When client intensity is kept high, the

efficiency of the first (upstream) process is high (93.5%). When quality increases, the service provider is motivated to reduce total client intensity. When total client intensity is reduced, efficiency of process 1 deteriorates (92.3%).

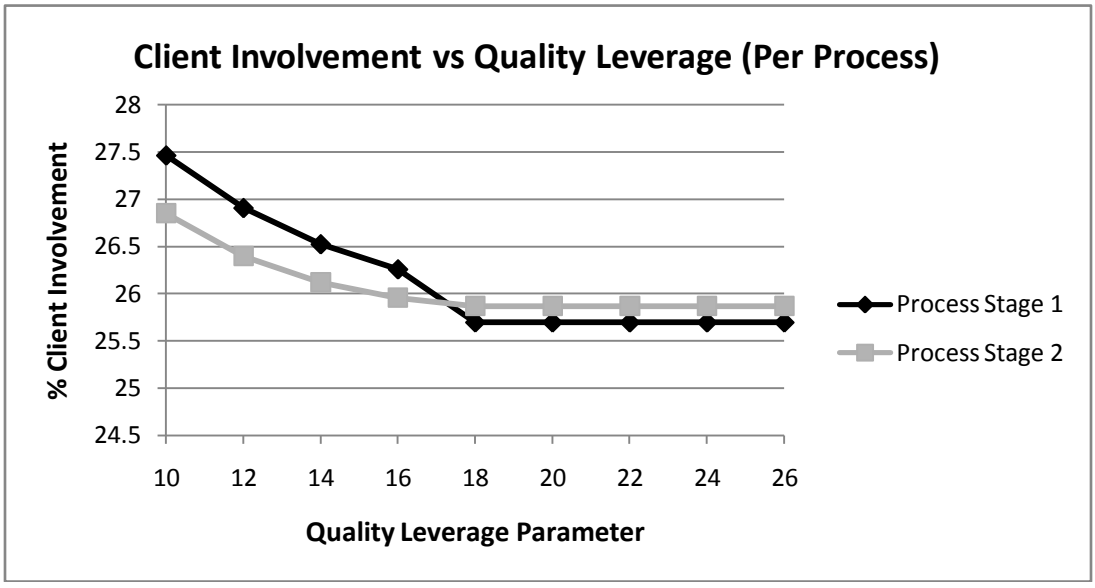

Figure 7: Percent client intensity vs. quality leverage parameter (per process) (Created by author)

6. Conclusions

In this paper we have developed a model for services such as consulting firms to assist decision makers in determining the optimal workforce and client intensity levels in the service-creation process when there are multiple processes and multiple time periods. We approached this problem by taking the conventional resource planning model, typically used for manufactured goods, and significantly modifying it for application to services. Contributions to the descriptive modeling of resource planning have been made by the addition of client intensity as a decision variable to the classical resource planning model, the addition of efficiency and quality functions as performance measures, and the representation of efficiency as a function of client intensity and quality. In addition, traditional resource planning models have a linear

objective and constraints. Our model, however, introduces nonlinearity into the constraints with the inclusion of the efficiency and quality functions.

Results of the model give rise to several managerial policy recommendations:

1. In firms where there are multiple services the optimal allocation to w_a, w_b is determined by the condition $\frac{\partial z_{1a}}{\partial w_a} = \frac{\partial z_{1b}}{\partial w_b}$, given total workforce, w.

2. In firms in which jobs must go through multiple processes and where the quality of the downstream process depends upon the quality of the upstream process, it can be advantageous to allocate more client intensity to the downstream process.

3. It can be beneficial to set a higher quality standard on downstream processes rather than upstream processes. This recommendation is counterintuitive since traditional manufacturing practices suggest that higher quality standards should be placed on upstream processes.

References

Abernathy, W. J., N. Baloff, et al. (1973). "A Three-Stage Manpower Planning and Scheduling Model - A Service-Sector Example." Operations Research **21**(3): 693.

Anderson, E. G. (2001). "The nonstationary staff-planning problem with business cycle and learning effects." Management Science **47**(6): 817.

Bettencourt, L. A., A. L. Ostrom, et al. (2002). "Client Co-Production in Knowledge-Intensive Business Services." California Management Review **44**(4): 100-128.

Bordoloi, S. and H. Matsuo (2001). "Human resource planning in knowledge-intensive operations: A model for learning with stochastic turn-over." European Journal of Operational Research **130**: 169-189.

Bowen, D. E. (1986). "Managing Customers as Human Resources in Service Organizations." Human Resource Management (1986-1998) **25**(3): 371.

Byrd, R. H., J. Nocedal, et al. (2006). KNITRO: An Integrated Package for Nonlinear Optimization. Large-Scale Nonlinear Optimization. D. Pillo, Gianni, Roma and Massimo, Springer-Verlag. **83:** 35-59.

Carrillo, J. E. and C. Gaimon (2004). "Managing Knowledge-Based Resource Capabilites Under Uncertainty." Management Science **50**(11).

Chase, R. (1978). "Where Does the Customer Fit in a Service Operation." Harvard Business Review: 137-142.

Chase, R. B. (1985). "The 10 Commandments of Service System Management." Interfaces **15**(3): 68.

Chase, R. B. and D. A. Garvin (1989). "The Service Factory." Harvard Business Review: 61-66.

Dietrich, B. (2006). "Resource Planning for Business Services." Communication of the ACM **49**(7): 62-64.

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

Fitzsimmons, J. A. and M. J. Fitzsimmons (2004). Service Management. New York, McGraw-Hill.

Gaimon, C. (1997). "Planning Information Technology-Knowledge Worker Systems." Management Science **43**(9).

Holt, C. C., F. Modigliani, et al. (1955). "A Linear Decision Rule for Production and Employment Scheduling." Management Science (pre-1986) **2**(1): 1.

Ittig, P. T. (1994). "Planning service capacity when demand is sensitive to delay." Decision Sciences **25**(4): 541.

Lance, A. B., L. O. Amy, et al. (2002). "Client co-production in knowledge-intensive business services." California Management Review **44**(4): 100.

Levitt, T. (1972). "Production-Line Approach to Service." Harvard Business Review.

Machuca, J. A. D., M. d. M. González-Zamora, et al. (2007). "Service Operations Management research." Journal of Operations Management **25**(3): 585.

Napoleon, K. and C. Gaimon (2004). "The Creation of Output and Quality in Services: A Framewotk to Analyze Information Technology-Worker Systems." Production and Operations Management **13**(3).

Rust, R. T. and R. Metters (1996). "Mathematical models of service." European Journal of Operational Research **91**: 427-439.

Sampson, S. E. (2007). A Customer-Supplier Paradigm for Service Science. 2007 DSI Services Science Miniconference, Pittsburgh.

Soteriou, A. C. and G. C. Hadjinicola (1999). "Resource allocation to improve service quality perceptions in multistage service systems." Production and Operations Management **8**(3).

Spohrer, J., P. Maglio, et al. (2007). Steps Toward a Science of Service Systems. Computer**:** 71-77.

Vargo, S. L. and A. Archupru (2009). "Service-Dominant Logic as a Foundation for Service Science: Clarifications." Service Science **1**(1): 32-41.

CHAPTER 6

SUMMARY, CONCLUSIONS AND FUTURE RESEARCH

The function of a service system is to transform inputs into outputs such that value for the service provider and the client is created through a process that utilizes capabilities and capacities of both the client and the provider. Services are the exchange of resources and information for the purposes of co-creating value for both the service provider and the client. Vargo and Akaka (2009) explain that the "function of a service system is connect people, technology and information through value propositions with the aim of co-creating value for the service systems participating in the exchange of resources within and across systems". Co-creation or shared creation of value is at the heart of services.

Services are a growing sector of the U.S. economy. This growth is evidence of the need for proper management of services. The core of service management is the planning and scheduling of resources. Just as there are well-established methods and models for planning and scheduling in manufacturing, there is now a need for these types of models for services.

This chapter is a summary of the conclusions drawn as a result of this research. Section 1 revisits the research motivation. Section 2 summarizes the findings of the single-stage resource planning model (Chapter 3). Section 3 summarizes the findings of the stochastic resource planning model (Chapter 4). Section 4 summarizes the findings of the multi-stage, multi-service resource planning model (Chapter 5). Section 5 suggests future research initiatives that can be pursued.

1. Research Motivation

The impetus of this research was to introduce a family of models for the advancement of resource planning for service operations. By developing these models, we were able to describe the behavior of optimal resource plans for service systems. We derived guidelines for resource planning policies that are specific to client co-generated services. We examined the sensitivity of policies to: 1) the effectiveness of client involvement on efficiency and quality, 2) uncontrollable inputs, the mis-estimation and mis-specification of technology functions, and 3) multi-process and multi-stage service systems.

This research responds to the need for analytic solutions applied to high value-adding service operations. We identified high value-adding services as services in which the client is a direct resource in the service process and the provider and the client have shared value in the outcome of the service. Therefore, we introduced the client as a co-generator of the service in all three resource-planning models.

This research responds to the need for modeling of tactical and operational decisions. To date, most research into service management has been done on strategic decisions. We position this research at the operational level of the business decision hierarchy and apply analytic methodologies to resource planning.

2. Single-Stage Resource Planning

We developed a non-linear, deterministic, single-stage resource planning model for co-generated service operations. In this study we examined both theoretically and experimentally the effects on policy of varying levels of client intensity. We incorporated efficiency and quality performance measures as a function of client intensity into the resource planning problem. Results show that as improvements are made toward improving efficiency and quality of the service process, less client involvement is needed.

Dynamic programming was used to solve this resource-planning problem. Problem P1is the optimization of the production plan for a given workforce level. Problem P2 is the optimization of the combined workforce and production plans. The state variable that connects the P2 to P1 is the workforce level.

The optimal policies for Problem Pl show that service firms will find themselves in one of three cases based on parameter values. If the backlog-involvement balance point is less than the minimum client intensity and quality levels, then the service firm is in Case 1. If the backlog-involvement balance point is greater than the l minimum client intensity and quality levels but less than the maximum client intensity level, then the service firm is in Case 2. If the backlog-involvement balance point is greater than the maximum client intensity level, then the service firm is in Case 3. The backlog-involvement balance point is the level of client intensity at which the trade-off between client costs and backlog costs are optimal. In all three cases the service

provider and the client are motivated to either decrease or increase client intensity to reach this point.

For Problem P2, we found five policy forms for all possible solutions to the KKT conditions. We derived the policy forms from possible values of the Lagrange multipliers for each constraint. Policy type 1 and 2 are the hiring policies. Policy type 3 is the policy in which the workforce is unchanged. Policy type 4 and 5 are the firing policies. Policy recommendations give service firms better insights into setting workforce, client intensity, and generation levels.

3. Stochastic Resource Planning

We developed a non-linear, stochastic, single-stage resource planning model with technology function uncertainties. The first objective of this study was to measure sensitivity of resource planning models to uncontrollable inputs, mis-estimation and mis-specification of technology functions for service operations. The second objective was to recommend policies that take into consideration uncontrollable inputs, mis-estimation and mis-specification of technology functions.

We take a two-tiered approach by developing two optimization models to achieve the objectives. The first optimization model is a deterministic model. In this model we allocate resources given an output target level and benchmark usage and yield rates. The results of this model are benchmark levels of input resources from the service provider and the client. The benchmark usage and yield rates make up the benchmark technology function. These benchmark levels of input resources and the benchmark technology function are passed to the second resource-planning model.

The second model is a stochastic, resource-planning model. A resource plan is generated based on benchmark input levels and target output levels. Deviations from the benchmark output levels are captured through a probability density function in the objective function. The deviations from the benchmark output levels reflect both inefficiencies and uncertainties.

The results of this study show that resources are allocated to compensate for technology function uncertainties. In numerical experiments, we tested the effects of technology function parameter changes on resource quantities. We found that as parameter values (e.g., randomness,

loss function weights, risk) were modified resource allocations from the service provider and the client changed in order meet output targets.

We also found that service capacity constraints are not like those in manufacturing. Service provider resources and client resources are exchanged when capacity limits are reached. This is a behavior that is not typical in manufacturing. There is more flexibility in services. The structure of the technology function matrix allows for the exchange between service provider and client resources, because there is more than one way to achieve output targets.

A trade-off exists between resource costs and loss because these two performance measures are in the objective function. When resource capacities were increased to meet demand resource costs increase and loss decreases. This trade-off presented itself in many of the experimental results.

4. Multi-Stage, Multi-Service Resource Planning

In this study we gained insights about how resources are allocated across multiple stages and multiple services. We develop a non-linear, deterministic, multi-stage planning model that allows for examination of trade-offs among client intensity, efficiency and quality. We borrowed resource planning model constructs, typically used for manufactured goods, and significantly modified them for application to services. We introduced client intensity as a decision variable to the descriptive modeling of resource planning and we added efficiency and quality as functions of client intensity as performance measures. We extrapolate theoretical findings from the single-stage planning study to determine resource allocations across multiple services and stages.

Results show that when the dynamic program in the single-stage study is extended there is trade-off between the cost of capacity changes and profits across multiple stages. If the workforce level of stage 1 given to stage 2 requires stage 2 hire workers, then there are increased hiring costs for stage 2. Stage 2 would like to be given a larger workforce from stage 1, but this will increase inventory holding costs and reduce profits for stage 1. If the workforce level of stage 1 given to stage 2 requires stage 2 fire workers, then there are increased firing costs for stage 2. Stage 2 would like to be given a smaller workforce from stage 1, but this will increase inventory costs and reduce profits for stage 1.

In the multi-service scenario, we found that the optimal allocation of resources across services happens when the slope of the objective function of one service equals the slope of the objective function of the other service. The slopes of the objective functions change as the parameter values change.

Results also show that in firms in which jobs must go through multiple stages and the quality of the downstream process depends upon the quality of the upstream process, it can be advantageous to allocate more client intensity to the downstream process. In addition, it can be beneficial to set a higher quality standard on downstream processes rather than upstream processes. This recommendation is counterintuitive since traditional manufacturing practices suggest that higher quality standards should be placed on upstream processes.

5. Future Research

The family of models presented in this research is merely the tip of the iceberg of what can be pursued in terms of analytic solutions to service operation challenges. This research can be extended in the following ways:

1. Assume different technology function forms. In the stochastic planning we assumed a linear, constant returns-to-scale form of the technology function. An extension to this research would be to assume a functional form such as the Cobb-Douglas function, the translog function, or the multiplicative function. Another extension to this research would be to assume variable returns-to-scale. The relationship of an output to inputs changes when different technology function forms are chosen.

2. Consider utility functions that are more complex and more accurately represent co-generated value than the loss function. We assumed very simplistic utility functions in the stochastic resource planning model. Service providers' or clients' utility functions represent preferences towards an outcome.

3. Perform case studies to further validate models. The parameter values in this research were chosen for illustrative purposes only. Although reasonable values were chosen, it would be beneficial to have actual cost, revenue, and demand values for example. In addition, the models have been constructed based on reasonable assumptions, but there

may be some "real-life" elements that are missing from the model. A modeler should make every attempt to build a model as realistically as possible.

4. Use various functions for representing efficiency and quality. In this research an exponential form of efficiency and quality were used for illustrative purposes. The reasoning for this extension is the same as #3 in this list. Cases studies should be performed in order to gather the most realistic data possible.

5. Obtain the optimal policy form of the solution to the multi-stage model. The policy guidelines for the multi-stage study were extrapolated from the optimal policies in the single-stage study. The KKT conditions should be derived to determine a unique solution to the multi-stage, multi-service problem.

6. Model resource planning for service value networks. In this research we have developed resource plans for a service firm that is not part of a network. If a firm is part of a network, then there is a sharing of the same suppliers among its competitors. This type of network structure is well known by supply chain management scholars. Lush, Vargo et. al (2000) describe a value network as "a spontaneously sensing and responding spatial and temporal structure of largely loosely coupled value proposing social and economic actors interacting through institutions and technology" of which a supply chain is a sub-part.

REFERENCES

Abernathy, W. J., N. Baloff, et al. (1973). "A Three-Stage Manpower Planning and Scheduling Model - A Service-Sector Example." Operations Research **21**(3): 693.

Akkermans, H. and B. Vos (2003). "Amplification in Service Supply Chains: An Exploratory Case Study from the Telecom Industry." Production and Operations Management **12**(2): 204.

Anderson, E. G. (2001). "The nonstationary staff-planning problem with business cycle and learning effects." Management Science **47**(6): 817.

Anderson, E. G., D. J. Morrice, et al. (2006). "Stochastic Optimal Control for Staffing and Backlog Policies in a Two-Stage Customized Service Supply Chain." Production and Operations Management **15**(2): 262.

Asmild, M., J. C. Paradi, et al. (2006). "Centralized resource allocation BCC models." Omega **In Press, Corrected Proof**.

Athanassopoulos, A. D. (1995). "Goal programming & data envelopment analysis (GoDEA) for target-based multi-level planning: Allocating central grants to the Greek local authorities." European Journal of Operational Research **87**(3): 535-550.

Athanassopoulos, A. D. (1998). "Decision support for target-based resource allocation of public services in multiunit and multilevel systems." Management Science **44**(2): 173.

Baloff, N., J. Hershey, et al. (1973). "A Three-Stage Manpower Planning and Scheduling Model--A Service-Sector Example." Operations Research **21**(3): 693.

Banker, R. D., A. Charnes, et al. (1984). "Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis." Management Science (pre-1986) **30**(9): 1078.

Bateson, J. E. (1977). Do We Need Service Marketing? Marketing Consumer Services: New Insights, Marketing Science Institute.

Beasley, J. E. (2003). "Allocating fixed costs and resources via data envelopment analysis." European Journal of Operational Research **147**(1): 198-216.

Bettencourt, L. A., A. L. Ostrom, et al. (2002). "Client Co-Production in Knowledge-Intensive Business Services." California Management Review **44**(4): 100-128.

Bitner, M. J., W. T. Faranda, et al. (1997). "Customer contributions and roles in service delivery." International Journal of Service Industry Management **8**(3): 193.

Bitran, G. and M. Lojo (1993). "A framework for analyzing the quality of the customer interface." European Management Journal **11**(4): 385-396.

Bordoloi, S. and H. Matsuo (2001). "Human resource planning in knowledge-intensive operations: A model for learning with stochastic turn-over." European Journal of Operational Research **130**: 169-189.

Bowen, D. E. (1986). "Managing Customers as Human Resources in Service Organizations." Human Resource Management (1986-1998) **25**(3): 371.

Bowers, M. R., C. L. Martin, et al. (1990). "Trading Places: Employees As Customers, Customers As Employ." The Journal of Services Marketing **4**(2): 55.

Brady, M. K. and J. J. Cronin (2001). "Customer Orientation: Effects on Customer Service Perceptions and Outcome Behaviors." Journal of Service Research **3**(3).

Bretthauser, K. M. (204). "Service Management." Decision Sciences **35**(3).

Byrd, R. H., J. Nocedal, et al. (2006). KNITRO: An Integrated Package for Nonlinear Optimization. Large-Scale Nonlinear Optimization. D. Pillo, Gianni, Roma and Massimo, Springer-Verlag. **83:** 35-59.

Carrillo, J. E. and C. Gaimon (2000). "Improving Manufacturing Performance through Process Change and Knowledge Creation." Management Science **46**(2).

Carrillo, J. E. and C. Gaimon (2002). "A Framework for Process Change." IEEE Transactions on Engineering Management **49**(4).

Carrillo, J. E. and C. Gaimon (2004). "Managing Knowledge-Based Resource Capabilities Under Uncertainty." Management Science **50**(11).

Charnes, A., W. W. Cooper, et al. (1994). Data Envelopment Analysis: Theory, Methodology, and Applications. Boston, Kluwer Academic Publisher.

Charnes, A., W. W. Cooper, et al. (1978). "Measuring the efficiency of decision making units." European Journal of Operational Research **2**: 429-444.

Chase, R. (1978). "Where Does the Customer Fit in a Service Operation." Harvard Business Review: 137-142.

Chase, R. and D. Tansik (1983). "The Customer Contact Model for Organization Design." Management Science **29**(9): 1037-1050.

Chase, R. B. (1981). "The Customer Contact Approach to Services: Theoretical Bases and Practical Extensions." Operations Research **29**(4): 698.

Chase, R. B. and N. J. Aquilano (1992). Production & Operations Management. Boston, Irwin.

Chase, R. B. and D. A. Garvin (1989). "The Service Factory." Harvard Business Review: 61-66.

Cook, D. P., C.-H. Goh, et al. (1999). "Service typologies: A state of the art survey." Production and Operations Management **8**(3): 318.

Cook, W. D., D. Chai, et al. (1998). "Hierarchies and Groups in DEA." Journal of Productivity Analysis **10**(2): 177-198.

Cooper, W., L. Seiford, et al. (2004). Handbook on Data Envelopment Analysis: History, Models and Interpretations (Chp.1). Boston, Kluwer Academic Publisher.

Dietrich, B. (2006). "Resource Planning for Business Services." Communication of the ACM **49**(7): 62-64.

Fare, R., R. Grabowski, et al. (1997). "Efficiency of a fixed but allocatable input: A non-parametric approach." Economics Letters **56**(2): 187-193.

Fare, R. and S. Grosskopf (2000). "Network DEA." Socio-Economic Planning Sciences **34**(1): 35-49.

Färe, R. and C. A. Knox Lovell (1978). "Measuring the technical efficiency of production." Journal of Economic Theory **19**(1): 150-162.

Fisk, R. P., S. W. Brown, et al. (1993). "Tracking the evolution of the services marketing literature." Journal of Retailing **69**(1): 61-103.

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

Fitzsimmons, J. A. and M. J. Fitzsimmons (2004). Service Management. New York, McGraw-Hill.

Forsund, F. R. and N. Sarafoglou (2002). "On the Origins of Data Envelopment Analysis." Journal of Productivity Analysis **17**(1/2): 23.

Frei, F. X. (2006). "Customer-Introduced Variability in Service Operations." Harvard Business Review.

Gaimon, C. (1997). "Planning Information Technology-Knowledge Worker Systems." Management Science **43**(9).

Gaimon, C. and G. L. Thompson (1984). "A distributed parameter cohort personnel planning model that uses cross-sectional data." Management Science **30**(6).

Gill, P. E., W. Murray, et al. (2006). User's Guide for SNOPT Version 7: Software for Large-Scale Nonlinear Programming.

Golany, B., S. T. Hackman, et al. (2006). "An efficiency measurement framework for multi-stage production systems." Annals of Operations Research **145**(1): 51.

Golany, B., F. Y. Phillips, et al. (1993). "Models for improved effectiveness based on DEA efficiency results." IIE Transactions **25**(6): 2.

Golany, B. and E. Tamir (1995). "Evaluating efficiency-effectiveness-equality trade-offs: A data envelopment analysis approach." Management Science **41**(7): 1172.

Graves, S. C. and B. T. Tomlin (2003). "Process flexibility in supply chains." Management Science **49**(7): 907.

Grönroos, C. and K. Ojasalo (2004). "Service productivity: Towards a conceptualization of the transformation of inputs into economic results in services." Journal of Business Research **57**(4): 414-423.

Harvey, J. (1998). "Service quality: a tutorial." Journal of Operations Management **16**(5): 583-597.

Hill, T. P. (1977). "On Goods and Services." Review of Income and Wealth **23**(4): 315-338.

Holt, C. C., F. Modigliani, et al. (1955). "A Linear Decision Rule for Production and Employment Scheduling." Management Science (pre-1986) **2**(1): 1.

Holt, C. C., F. Modigliani, et al. (1956). "Derivation of a Linear Decision Rule for Production and Employment." Management Science (pre-1986) **2**(2): 159.

Ittig, P. T. (1994). "Planning service capacity when demand is sensitive to delay." Decision Sciences **25**(4): 541.

Judd, R. C. (1964). "The Case for Redefining Services." Journal of Marketing **28**(1): 58-59.

Karmarkar, U. S. and R. Pitbladdo (1995). "Service markets and competition." Journal of Operations Management **12**(3-4): 397-411.

Karmarkar, U. S. (2007). The Global Information Economy and Service Industrialization: The UCLA BIT Project, UCLA Anderson School of Management**:** 1-5.

Kaulio, M. A. (1998). "Customer, consumer and user involvement in product development: a framework and a review of selected methods." Total Quality Management **9**(1): 141-149.

Kellogg, D. L. and R. B. Chase (1995). "Constructing an empirically derived measure for customer contact." Management Science **41**(11): 1734.

Kellogg, D. L. and W. Nie (1995). "A framework for strategic service management." Journal of Operations Management **13**(4): 323-337.

Koopmans, T. C. (1951). Analysis of Production as an efficient combination of activities. New York, Wiley.

Korhonen, P. and M. Syrjänen (2004). "Resource Allocation Based on Efficiency Analysis." Management Science **50**(8): 1134.

Kumar, C. K. and B. K. Sinha (1999). "Efficiency based production planning and control models." European Journal of Operational Research **117**(3): 450-469.

Lance, A. B., L. O. Amy, et al. (2002). "Client co-production in knowledge-intensive business services." California Management Review **44**(4): 100.

Laroche, M., J. Bergeron, et al. (2001). "A three-dimensional scale of intangibility." Journal of Service Research : JSR **4**(1): 26.

Levitt, T. (1972). "Production-Line Approach to Service." Harvard Business Review.

Lewis, H. F. and T. R. Sexton (2004). "Network DEA: efficiency analysis of organizations with complex internal structure." Computers & Operations Research **31**(9): 1365-1410.

Liang, L., Y. Feng, et al. (2006). "DEA models for supply chain efficiency evaluation." Annals of Operations Research **145**(1): 35.

Lovelock, C. and E. Gummesson (2004). "Whither Services Marketing? In Search of a New Paradigm and Fresh Perspectives." Journal of Service Research : JSR **7**(1): 20.

Lovelock, C. H. (1983). "Classifying Services to Gain Strategic Marketing Insights." Journal of Marketing **47**(3): 9-20.

Lozano, S. and G. Villa (2004). "Centralized Resource Allocation Using Data Envelopment Analysis." Journal of Productivity Analysis **22**(1/2): 143.

Lozano, S. and G. Villa (2005). "Centralized DEA models with the possibility of downsizing." The Journal of the Operational Research Society **56**(4): 357.

Lusch, R. F., S. L. Vargo, et al. (2007). "Competing through service: Insights from service-dominant logic." Journal of Retailing **83**(1): 5-18.

Machuca, J. A. D., M. d. M. González-Zamora, et al. (2007). "Service Operations Management research." Journal of Operations Management **25**(3): 585.

Martin, C. R. J., D. A. Horne, et al. (2001). "A perspective on client productivity in business-to-business consulting services." International Journal of Service Industry Management **12**(2): 137.

McLaughlin, C. P. and S. Coffey (1990). "Measuring Productivity in Services." International Journal of Service Industry Management **1**(1).

Mersha, T. (1990). "Enhancing the customer contact model." Journal of Operations Management **9**(3): 391-405.

Metters, R. D., F. X. Frei, et al. (1999). "Measurement of Multiple Sites in Service Firms with Data Envelopment Analysis." Production and Operations Management **8**(3).

Mills, P. K., R. B. Chase, et al. (1983). "Motivating the Client/Employee System as a Service Production Strategy." Academy of Management. The Academy of Management Review **8**(2): 301.

Mills, P. K. and D. J. Moberg (1982). "Perspectives on the Technology of Service Operations." The Academy of Management Review **7**(3): 467-478.

Mills, P. K. and J. H. Morris (1986). "Clients as "Partial" Employees of Service Organizations: Role Development in Client Participation." The Academy of Management Review **11**(4): 726.

Murdick, R. D., R. Render, et al. (1990). Service operations management. Boston, Allyn and Bacon.

Nachum, L. (1999). "Measurement of productivity of professional services: An illustration on Swedish management consulting firms." International Journal of Operations & Production Management **19**(9): 922.

Nachum, L. (1999). "The Productivity of Intangible Factors of Production: Some Measurement Issues Applied to Swedish Management Consulting Firms " Journal of Service Research **2**(2): 123-137.

Parasuraman, A., V. A. Zeithaml, et al. (1988). "Servqual: A Multiple-Item Scale for Measuring Consumer Perc." Journal of Retailing **64**(1): 12.

Parasuraman, A., V. A. Zeithaml, et al. (1994). "Reassessment of Expectations as a Comparison Standard in Measuring Service Quality: Implications for Further Research." Journal of Marketing **58**(1): 111-124.

Radding, A. (2006). How IBM is Applying Science to the World of Service. Consulting Magazine.

Rousseau, D. M. (1979). "Assessment of technology in organizations: Closed versus open systems approaches." Academy of Management. The Academy of Management Review (pre-1986) **4**(000004): 51.

Ruggiero, J. (1998). "Non-discretionary inputs in data envelopment analysis." European Journal of Operational Research **111**(3): 461-469.

Rust, R. T. and R. Metters (1996). "Mathematical models of service." European Journal of Operational Research **91**: 427-439.

Sampson, S. E. (2000). "Customer-supplier duality and bidirectional supply chains in service organizations." International Journal of Service Industry Management **11**(4): 348.

Sampson, S. E. (2007). A Customer-Supplier Paradigm for Service Science. 2007 DSI Services Science Miniconference, Pittsburgh.

Sampson, S. E. and C. M. Froehle (2006). "Foundations and Implications of a Proposed Unified Services Theory." Production and Operations Management **15**(2): 329.

Sasser, W. E., R. P. Olsen, et al. (1978). Management of service operations: text, cases, and readings. Boston, Allyn and Bacon.

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.

Schmidt, P. (1985). "Frontier production functions." Econometric Reviews **4**(2): 289 - 328.

Schmidt, P. and C. A. K. Lovell (1980). "Estimating stochastic production and cost frontiers when technical and allocative inefficiency are correlated." Journal of Econometrics **13**(1): 83-100.

Shephard, R. W. (1970). Theory of Cost and Production Functions. Princeton, NJ, Princeton University Press.

Sherman, H. D. and J. Zhu (2006). Service Productivity Management: Improving service performance using Data Envelopment Analysis (DEA). New York, Springer Science+Business Media.

Shostack, G. L. (1977). "Breaking Free from Product Marketing." Journal of Marketing **41**(2): 73-80.

Sipper, D. and R. L. Bulfin (1997). Produciton Planning, Control, and Integration. Boston, Mc-Graw Hill.

Soteriou, A. C. and G. C. Hadjinicola (1999). "Resource allocation to improve service quality perceptions in multistage service systems." Production and Operations Management **8**(3).

Soteriou, A. C. and Y. Stavrinides (1997). "An internal customer service quality data envelopment analysis model for bank branches." International Journal of Operations & Production Management **17**(8): 780.

Spohrer, J., P. Maglio, et al. (2007). Steps Toward a Science of Service Systems. Computer**:** 71-77.

Stigler, G. J. (1950). "The Development of Utility Theory I." The Journal of Political Economy **58**(4): 307-327.

Stigler, G. J. (1950). "The Development of Utility Theory II." The Journal of Political Economy **58**(5): 373-396.

Tang, C. S. (1990). "The Impact of Uncertainty on a Production Line." Management Science **36**(12): 1518.

Thanassoulis, E. (1996). "A data envelopment analysis approach to clustering operating units for resource allocation purposes." Omega **24**(4): 463-476.

Thanassoulis, E. (2001). Introduction to the Theory and Application of Data Envelopment Analysis: A foundation text with integrated software. Boston, Kluwer Academic Publishers.

Thanassoulis, E. and R. G. Dyson (1992). "Estimating preferred target input-output levels using data envelopment analysis." European Journal of Operational Research **56**(1): 80-97.

Verma, R. and K. K. Boyer (2000). "Service classification and management challenges." Journal of Business Strategies **17**(1): 5.

Verma, R. and G. M. Thompson (1999). "Managing service operations based on customer preferences." International Journal of Operations & Production Management **19**(9): 891.

Winston, W. L. (1991). Operations research: applications and algorithms. Boston, PWS-Kent Publishing Company.

Youngdahl, W. E. and D. L. Kellogg (1997). "The relationship between service customers' quality assurance behaviors, satisfaction, and effort: A cost of quality perspective." Journal of Operations Management **15**(1): 19-32.

Zeithaml, V. A. and M. J. Bitner (1996). Services Marketing. New York, McGraw Hill.

Zeithaml, V. A., A. Parasuraman, et al. (1985). "Problems and Strategies in Services Marketing." Journal of Marketing **49**(2): 33-46.