Wharton

Financial
Institutions
Center

Measuring the Efficiency of Service Delivery Processes: With Application to Retail Banking

by
Frances X. Frei
Patrick T. Harker

96-31-B

The Wharton School
University of Pennsylvania

THE WHARTON FINANCIAL INSTITUTIONS CENTER

The Wharton Financial Institutions Center provides a multi-disciplinary research approach to the problems and opportunities facing the financial services industry in its search for competitive excellence. The Center's research focuses on the issues related to managing risk at the firm level as well as ways to improve productivity and performance.

The Center fosters the development of a community of faculty, visiting scholars and Ph.D. candidates whose research interests complement and support the mission of the Center. The Center works closely with industry executives and practitioners to ensure that its research is informed by the operating realities and competitive demands facing industry participants as they pursue competitive excellence.

Copies of the working papers summarized here are available from the Center. If you would like to learn more about the Center or become a member of our research community, please let us know of your interest.

Anthony M. Santomero
Director

The Working Paper Series is made possible by a generous grant from the Alfred P. Sloan Foundation

Measuring the Efficiency of Service Delivery Processes: With Application to Retail Banking [1]

October 1996

Abstract: This paper presents a methodology that determines the role of design in calculating the efficiency of service delivery processes. The efficiency of these processes is determined by using a variation of frontier estimation (DEA-like) techniques. The methodology is then applied to a particular service delivery process in retail banking. The methodology allows one to address the question of how much inefficiency in a business process is due to the wrong process design, and how much is due to the right design, poorly executed. Consistent with expectations, the results show that no single process design dominates. However, for a particular institution, the methodology demonstrates the tradeoffs, and, offers specific recommendations for either improving an existing process or radically changing to a different design.

Keywords: Data Envelopment Analysis, Productivity, Process Management, Retail Banking, Services

Frances X. Frei, Simon School of Business, University of Rochester, Rochester, NY 14627, frei@mail.ssb.rochester.edu

Patrick T. Harker, Department of Operations and Information Management, University of Pennsylvania, Philadelphia, PA 19104-6366, harker@opim.wharton.upenn.edu

1. Introduction

The design and implementation of service delivery processes plays a key role in the overall competitiveness of modern organizations. For example, Roth and Jackson (1995) provide clear evidence that process capability and execution are major drivers of performance due to their impact on customer satisfaction and service quality. Thus, any study of the efficiency of service organizations *must* focus on the role of process design and performance.

Traditional efficiency studies measure the performance of a firm by its ability to transform inputs to outputs. However, the actual way in which these inputs are transformed to outputs is often overlooked. That is, each firm's operation is conceptualized as a black box: inputs go in, outputs come out, and little analytical attention is paid to the inner workings of the transformation process. This paper examines this "black box" and argues that the actual design of the transformation process is a critical component in the performance of a firm. Further, this paper submits that the design of the transformation mechanism, or the *process design*, must be fully studied and integrated into performance analysis in order to provide useful managerial recommendations.

This paper presents and illustrates a methodology that determines the role of process design in calculating process efficiency. Using a variation on frontier estimation (DEA-like) techniques, this methodology permits one to address the question of how much of process inefficiency is due to the wrong process design for the desired output maximization, and how much inefficiency is due to poor execution of the correct design. The analysis of a service delivery process illustrates the efficacy of this new approach.

This study concentrates on one aspect of organizational performance; the role of process design. By focusing on the process as the unit of analysis, the impact of information technology (IT), human resources, and, most importantly, the interaction between the two, on performance is analyzed. Why focus on processes? The traditional approach to process management and control develops an optimal schema for work and then encodes this into the organizational culture and resulting information systems. For example, the standard industrial engineering (IE)/operations research approach to process improvement studies the process, simulates it, and then proposes a new process design. The design is then implemented by the creation of new job descriptions, incentives, and information systems. In this sense, the new process becomes encoded by the process definition and supporting IT and incentive structures. *Process reengineering* has reinvented this traditional IE approach, but in the context of white-collar work (Davenport and Short 1990). The important contribution of this reengineering phenomenon has shifted the focus onto *processes* as the central unit of management.

In the analysis of customer service delivery processes, in particular, the focus should not be solely on traditional measures of productivity or financial performance such as transactions per full-time equivalent employee (FTE), return on assets (ROA), or return on equity (ROE). Rather, comparisons based on intermediary measures that evaluate the drivers of performance from the perspective of all participants in the service delivery process must also be considered. The empirical example in this paper demonstrates the use of such intermediary measures. That is, the outputs of the process described in Section 4 do not explicitly represent the amount of money produced by a process, but rather, they indicate the performance of that particular process. It is the opinion of the authors that performance on a set of service delivery processes will coincide with firm-level performance (for evidence of this result in the retail banking industry see Frei, Kalakota, and Marx, 1997). However, this does not imply that there is not much to be learned from analysis of a single process. Single process analysis will allow the determination of how much inefficiency in a business process is due to the wrong process design, and how much is due to the right design, poorly executed. Equipped with this knowledge, managers will be better able to determine when large reengineering projects are necessary versus minor adjustments to existing business processes.

The structure of the remainder of this paper is as follows. The next section reviews the related literature on efficiency, processes, and frontier estimation. Section 3 introduces process technology and specifically addresses the role of process design in productivity and performance analysis. Section 4 presents an empirical application of this methodology in the context of retail banking service-delivery processes, and Section 5 describes this paper's contribution as well as avenues of further research.

2. Literature Review

There is a rich history of literature demonstrating the importance of processes in analyzing firm performance (Chase, 1981 and 1983; Levitt, 1972; Roth and van der Velde, 1991; Roth and Jackson, 1995, Shostack, 1987). In addition, there has been a stream of literature on strategic frameworks to help conceptualize performance in retail banking specifically (Chase, Northcraft, and Wolf, 1984; Huete and Roth, 1987; Sherman and Gold, 1985, Haag and Jaaska, 1995, Roth and van der Velde, 1989). The framework presented in Roth and Jackson (1995) describes how process capabilities and people impact business performance. Their work provides the framework from which we are able to talk about how much inefficiency in process performance is due to the wrong design and how much is due to poor performance.

While it is beyond the scope of this paper to reference the entire literature on performance in retail banking, we refer the reader to an excellent review of branch and bank level studies of performance by Berger and Humphrey (1997). Their review paper summarizes the results of x papers in this area. Our work adds to this literature by including process level data in our analysis.

When estimating the performance of processes, the first consequence to note is that there are usually multiple outputs. These multiple outputs preclude the use of standard statistical regressions involving a single dependent variable. The estimation methods described in this paper deal with these multiple outputs by using deterministic frontier estimation. Specifically, Data Envelopment Analysis (DEA) is used to determine relative performance amidst multiple inputs and outputs. Charnes, Cooper, and Rhodes (1978) introduced DEA as a new way to measure efficiency of decision-making units (DMUs). Since then, there have been over 400 articles that have used variations of DEA in analyzing performance (see Seiford 1994).

The original DEA method determines the relative efficiency measure for a DMU by maximizing the ratio of weighted outputs to inputs subject to the condition that similar ratios for every DMU are not larger than one. The solution of this problem results in a set of efficiency scores less than or equal to one, as well as a reference set of efficient DMUs. This method has come to be known as the 'input-oriented method' as its efficiency score is determined by holding outputs constant and assessing to what extent inputs would have to be improved (decreased) in order for a DMU to be considered efficient. The 'output-oriented method' is similar to the input-oriented method although, in this case, the ratio of weighted inputs to outputs is minimized in order to determine the amount that each DMU's outputs can be improved (increased) while holding the inputs constant. In either case, an efficient DMU has no potential improvement, whereas inefficient DMUs have efficiency scores reflecting the potential improvement based on the performance of other DMUs. In order to determine the relative efficiency scores, a linear program must be run for each DMU. By using a linear objective function, the implicit assumption is that the efficient frontier is piecewise linear.

An extension was made over the methods described above when Frei and Harker (1995) allowed the efficiency calculation to incorporate simultaneous changes in the input and output space. That is, in opposition to previous methods that calculated efficiency along either the input or the output space, Frei and Harker determined the efficiency based on the least-norm projection of an inefficient DMU to the efficient frontier. This improvement allows the directions for improvement to cover the range of inputs and outputs simultaneously and thus, yields less restrictive managerial recommendations. It should be noted that the recommendations from the least-norm projection method, which are used in this paper, are significantly different than that of the oriented method as demonstrated in Frei and Harker (see Appendix B for a summary of this methodology). In addition, our work is based on the Baker, Charnes, and Cooper (1984) (BCC) model of DEA that extended the original constant returns to scale model to allow for variable returns to scale. For an excellent review of DEA models, see Seiford and Thrall (1990).

3. Process Technology and Efficiency

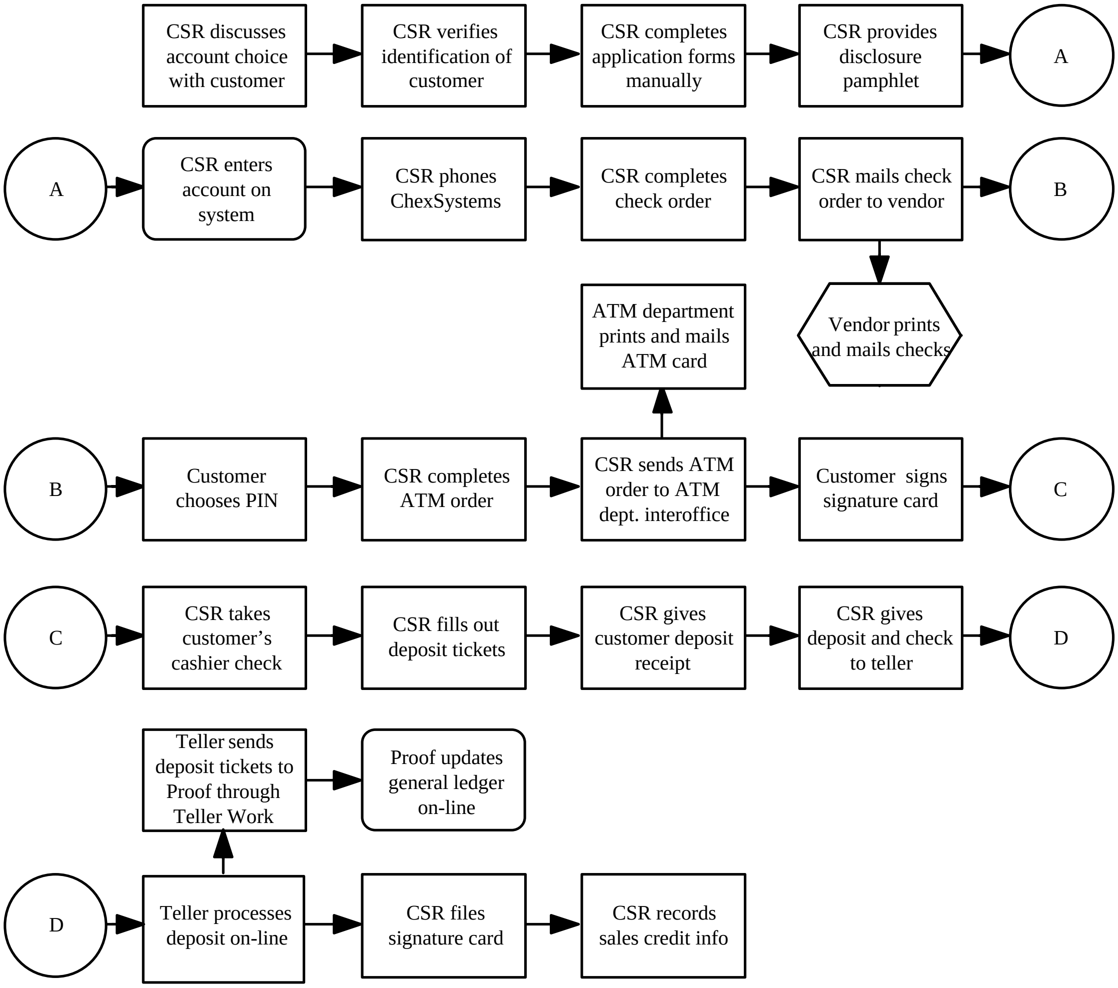
This section describes the proposed technique for evaluating the performance of service-delivery processes. Creating a process map is the standard first step in the evaluation of processes (Shostack, 1987; Kingman-Brundage, 1992). However, even after a careful study of process maps, it is still difficult to determine how efficient a single process is or, from a group of processes, which are *better*. To illustrate this, consider the two processes shown in Figures 1 and 2. Each process entails opening a checking account, which will be described in detail in Section 4, however the processes differ from one another in terms of number of steps, order of work, use of technology, and level of service provided. Upon inspection, it is difficult to evaluate and compare these two processes, let alone to compare a collection of processes across a large number of firms. Thus, after a careful understanding of the process, what is the next step? The technique that is described herein builds on existing frontier estimation methods to provide a way to evaluate processes with multiple inputs and outputs, at least some of which have non-market values[1]. The first step is to determine the relative efficiency of a given process, the other firms that might be used for benchmarking this process, and the managerial implications of the choices involved.

Frei and Harker (1995) introduced a method for determining relative efficiency scores by projecting to the closest point on the efficient frontier. This section adds the effect of process technology to the previously described methodology. That is, the performance of DMUs is already being assessed over multiple inputs and outputs, with capital and labor presumably as inputs, and some notion of performance scores as outputs. The question addressed in this section is what role process design has in this evaluation.

When using the process as the unit of analysis, the mechanism by which a DMU converts its inputs to outputs is addressed. That is, the process, or its design, is either an additional input or output like labor and capital, or else it exhibits other characteristics. It is difficult to think of design characteristics as either inputs or outputs, as they are neither created nor consumed as other inputs and outputs. Rather, design characteristics provide the structure for the creation and consumption of the existing inputs and outputs. The process, according to Morroni (1992), actually defines how capital and labor interact in order to produce outputs.[2] Thus, the process design defines the production technology for the organization.

[1] The non-market values prohibit the use of a profit function in an econometric evaluation.

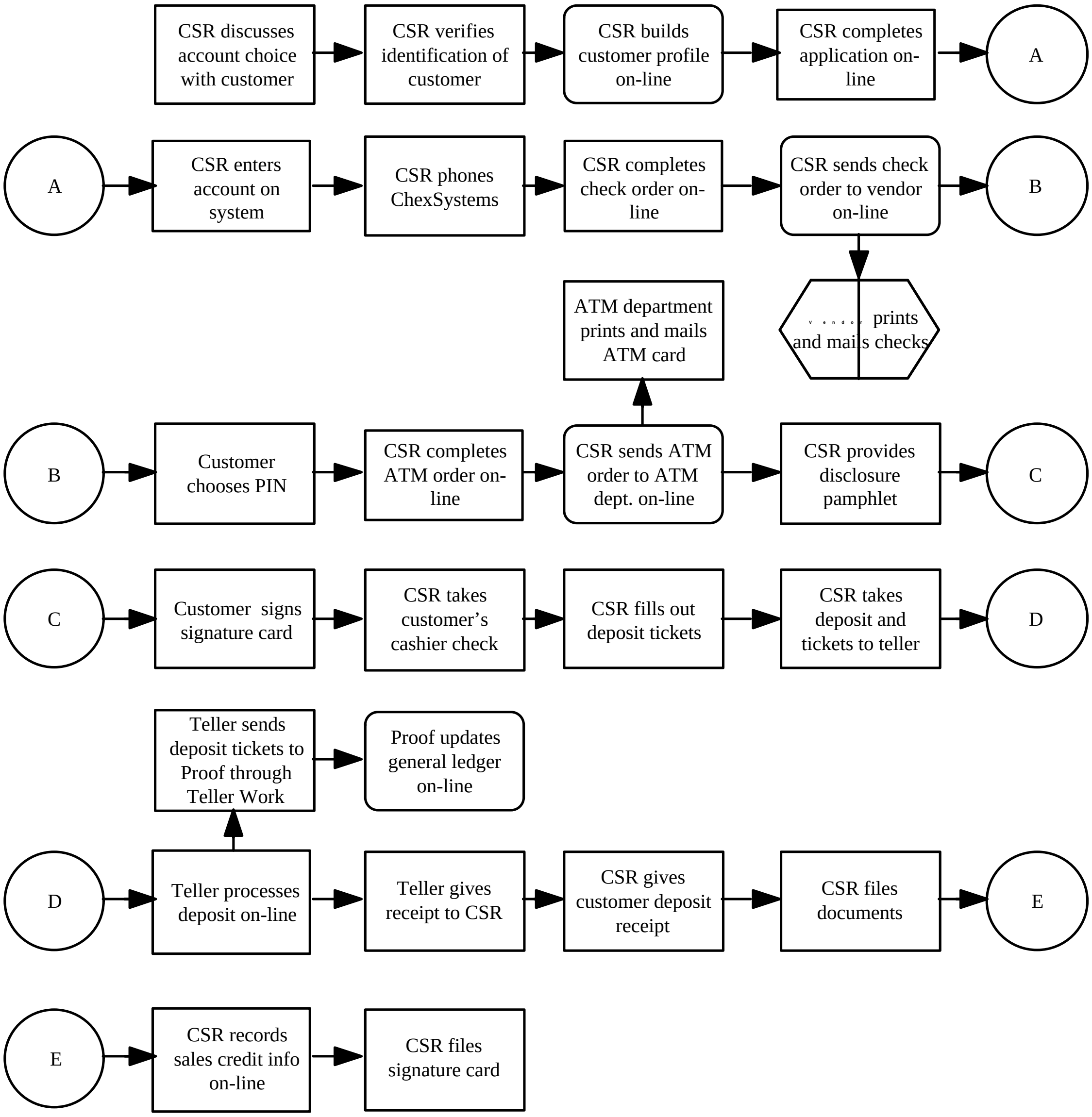
[2] An illustrative example used in Morroni's book describes ten ditch diggers with ten shovels. If an eleventh shovel is added, the process of ditch digging must be changed. That is, an additional shovel will not benefit the one-person, one-shovel process. The point here is that a process typically defines the relationship between capital and labor, and thus they are not immediately interchangeable.


CSR discusses account choice with customer
CSR verifies identification of customer
CSR completes application forms manually
CSR provides disclosure pamphlet
A
A
CSR enters account on system
CSR phones ChexSystems
CSR completes check order
CSR mails check order to vendor
B
ATM department prints and mails ATM card
Vendor prints and mails checks
B
Customer chooses PIN
CSR completes ATM order
CSR sends ATM order to ATM dept. interoffice
Customer signs signature card
C
C
CSR takes customer's cashier check
CSR fills out deposit tickets
CSR gives customer deposit receipt
CSR gives deposit and check to teller
D
Teller sends deposit tickets to Proof through Teller Work
Proof updates general ledger on-line
D
Teller processes deposit on-line
CSR files signature card
CSR records sales credit info

Legend:
CSR = customer service representative
ATM = automated teller machine
PIN = personal identification number
Square box = manual step
Rounded edge box = online step

Figure 1. An Example of the Open Checking Account Process

Given that process designs are not the same as inputs like labor, how can the method described in Frei and Harker (1995) be extended to account for process designs, and to analyze their relative efficiency? First, it is important to recognize that when determining the empirical production function or efficient frontier, one defines a production function F(x,y,z), where x is the set of inputs, y is the set of outputs, and z is the set of other characteristics that are not explicitly inputs or outputs. The focus of this paper is the subset of z representing process-design characteristics. Recognizing that there are always externalities contributing to production, the proposed methodology attempts to capture the specific effect of design characteristics. Obviously, these design characteristics are neither consumed nor produced, but they still play an important role in production. It is to this end that the additional step to treat the designs as defining alternative production functions is taken. In addition to looking at the empirical production function or efficient frontier of the entire data set, it is useful to look at the set of empirical production functions represented by each of a class of design groups. This step is essential as it determines how well a DMU is performing relative to its own design group, as well as how well a DMU is performing overall.


CSR discusses account choice with customer
CSR verifies identification of customer
CSR builds customer profile on-line
CSR completes application on-line
A
A
CSR enters account on system
CSR phones ChexSystems
CSR completes check order on-line
CSR sends check order to vendor on-line
B
ATM department prints and mails ATM card
Vendor prints and mails checks
B
Customer chooses PIN
CSR completes ATM order on-line
CSR sends ATM order to ATM dept. on-line
CSR provides disclosure pamphlet
C
C
Customer signs signature card
CSR takes customer’s cashier check
CSR fills out deposit tickets
CSR takes deposit and tickets to teller
D
Teller sends deposit tickets to Proof through Teller Work
Proof updates general ledger on-line
D
Teller processes deposit on-line
Teller gives receipt to CSR
CSR gives customer deposit receipt
CSR files documents
E
E
CSR records sales credit info on-line
CSR files signature card

Legend:
CSR = customer service representative
ATM = automated teller machine
PIN = personal identification number
Square box = manual step
Rounded edge box = online step

Figure 2. A Second Example of the Open Checking Account Process

The work of Brockett and Golany (1996) introduces the concept, in the context of Data Envelopment Analysis (DEA), of organizing DMUs into subgroups in order to determine if one subgroup outperforms another. This logic is easily transferred to processes, where it is recognized that while all processes require the same categories of inputs to produce similar categories of outputs, there are vastly different ways of organizing the way in which the work occurs. Brockett and Golany determine the efficient frontier for each subgroup in order to determine which input and output scenarios are dominant within each subgroup. Although it is easy to visually understand which process-design group is dominant in two dimensions, Brockett and Golany provide no means of determining this in higher dimensions. Thus, although Brockett and Golany's idea of comparing frontiers is useful in two dimensions, it is quite difficult in higher dimensions.

This paper takes Brockett and Golany's (1996) methodology and implements it differently. The standard first step of determining the efficient frontier of the entire set of DMUs along with the overall efficiency ratings for each DMU is utilized herein. Then, a test is performed to see if the overall efficiency ratings for the DMUs in one process-design group are statistically different from the overall efficiency ratings of the DMUs in another process-design group. After this statistical test, the efficient frontier for each process-design group is determined yielding design-group efficiency ratings for each DMU. The challenge is to determine which portion of the overall efficiency rating is due to poor execution, and which is due to belonging to the wrong process-design group. For each of these stages, rather than using the standard DEA efficiency scores, the least-norm projection algorithm described in Frei and Harker (1995) is utilized to allow efficiency scores to be based on a simultaneous improvements of inputs and outputs rather than one or the other.

Figure 3 illustrates an example of seven DMUs that belong to two process-design groups, represented by squares (Group I) and circles (Group II) respectively. It is clear from this figure that the overall frontier contains D1 though D4, with all other DMUs inefficient. The methodology presented herein addresses the question of how to make the inefficient DMUs (D5 through D7) efficient. In order to do this, it is necessary to determine, for each DMU, when it is beneficial to attempt to improve performance within the same process-design group, and when it is beneficial to change process-design groups. In Figure 3, the dotted lines show the process-design groups' individual frontiers. Each inefficient DMU now has two projections, one to the overall frontier, and one to the DMU's design-group frontier. The task is to determine all three frontiers along with all of the associated distance measurements, projections, and reference sets. The implications of this will be illustrated with a data set of 126 banks in the next section.

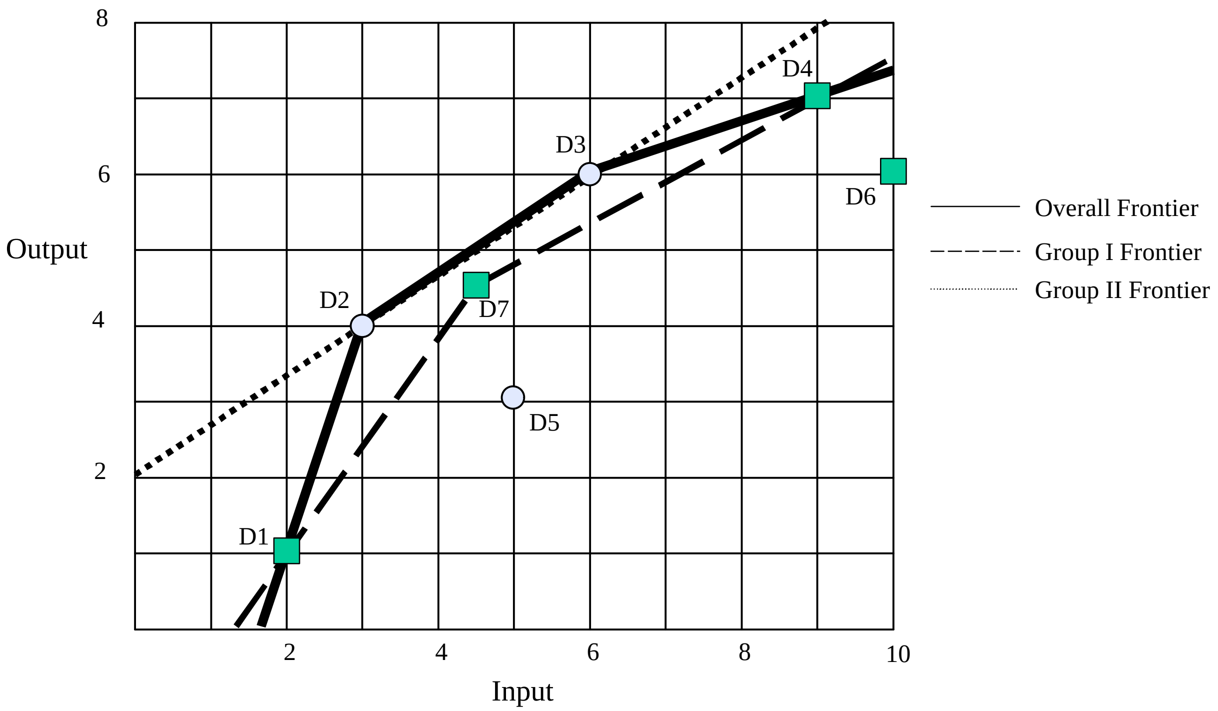


Figure 3. Overall Frontier for Two Process-Design Groups

There are three categories that a DMU might fall in when performing this analysis. The first category is comprised of DMUs that are efficient overall and therefore, efficient in *any* design group (D1 through D4). The second category is comprised of DMUs that are inefficient overall, but efficient in their design groups (D7). There are two possibilities for improvement in this category: one is for the DMU to stay within the same design group and to seek to improve this group's efficiency by expanding the production frontier. The second possibility is to seek improvement by changing design groups. The third category of DMUs is comprised of DMUs that are inefficient overall and inefficient in their own design groups (D5 and D6). In this case, the definition of the reference set is not clear. If the least-norm projection to the design group frontier lands on the overall frontier, it is likely that the inefficiency is due to poor performance rather than the wrong process design. However, if the least-norm projection to the design-group frontier does not land on the overall frontier, the inefficiency can be due to performance or design-group choice. This last scenario is the situation for which a numerical example is described in Section 4.

The usefulness of this method hinges upon selecting appropriate process design groupings. We have found there to be two ways to select these groupings. The first method is appropriate when there are obvious managerially actionable design characteristics available. In this instance, the characteristics need only be identified and separated into distinct categories such as high and low scores for each characteristic or, if the size of the data set allows, finer segmentation. For example, in Section 4 we present a situation where we use process design groupings of high and low technology and the level of extra services provided by the bank (one, two, or three).

Technology was chosen, as it is obviously a distinguishing factor in process design. Similarly, the level of service provided is also a distinguishing factor. For example, based on the customer base served, a bank may choose to offer the bare minimum of service or, conversely, routinely provide extra service even at the expense of taking up more of the consumer's (and employee's) time. Each of these decisions is managerially actionable in that they require strategic decisions on the part of the manager to determine their level.

The second method is to perform cluster analysis on a set of design characteristics. This method should be chosen when there are not obvious managerially actionable design characteristics or when these characteristics do not break along obvious lines.

The methodology for evaluating a process can be summarized as follows:

Step 1. Use the methodology developed in Frei and Harker (1995) to determine the efficiency score, reference set, and frontier projection for the entire data set.

Step 2. Separate the data into process-design groups using obvious break points in the managerially actionable design characteristics or, more formally, through cluster analysis.

Step 3. Determine if the set of overall efficiency scores calculated in Step 1 is statistically different for each of the process design groups defined in Step 2.

Step 4. As described in Step 1, determine the efficiency score, reference set, and frontier projection separately for each process-design group.

Step 5. Isolate the portion of overall inefficiency that is due to poor performance, and the portion that is due to the wrong process design. The overall and design-group efficiency scores have been determined in Steps 1 and 3, respectively. If the DMU is inefficient in both cases, then the portion of the overall inefficiency that is due to poor execution is $\frac{Efficiency_{DesignGroup}}{Efficiency_{Overall}}$ while the portion of inefficiency due to the wrong process design is 1- $\frac{Efficiency_{DesignGroup}}{Efficiency_{Overall}}$. If the DMU is inefficient overall but efficient in its design group, then the overall inefficiency is attributable to the wrong design.

Step 6. Determine the specific managerial recommendations for improvement, both within a design group, as well as for the entire set of banks. This step is obviously context dependent however we provide general guidelines for interpreting the results with an illustration of the retail banking industry provided in the next section.

4. Empirical Study of Retail Banking Efficiency

This paper is a result of the work undertaken by the retail banking study at the Wharton Financial Institutions Center. The retail banking study is an interdisciplinary research effort aimed at understanding the drivers of competitiveness in the industry, where competitiveness means not simply firm performance but the relationship between industry trends and the experiences of the retail banking labor force. The main phase of the study entailed a detailed survey of technology, work practices, organizational strategy, and performance in 135 U.S. retail banks. The survey focused on the largest banks in the country and was not intended as a random sample of all U.S. banks. In the end, the approach gained the participation of banks holding over 75% of the total assets in the industry in 1994. Participation in the study was confidential, but not anonymous, enabling the team to match survey data with data from publicly available sources. The scope and scale of this survey make it the most comprehensive survey to date on the retail banking industry.

Using the conceptual framework for financial services described in Frei, Harker, and Hunter (1996), data has been collected on a variety of service delivery processes, which represent the bulk of the work that occurs at a typical retail bank branch. The type of information collected on each process included the number of steps, number of automated steps, number of potentially automated steps, number of hand-offs, number of departments involved, number of approvals required, number of different people involved, customer's time required in the process, activity time, and cycle time. The process data was collected both on-site and through a mailed survey. See Frei (1996) for a complete description of the data collection process.

To illustrate the methodology proposed herein, consider one of the retail banking customer service delivery processes: the process of opening a checking account, which is depicted in Figures 1 and 2. The scenario for the process of opening a checking account involves a customer who has no existing relationship with the bank. This customer walks into a branch with a $500 cashier's check drawn on a different bank, and requests to open a consumer checking account with an ATM card. The data categories used for this analysis are described in Table 1 (see Appendix A for the complete data set).

When using any frontier estimation technique, the subsequent analysis is only as good as the initial decision of inputs and outputs. These variables are to be selected such that the inputs represent the resources consumed by the decision-making units and the outputs represent the performance of the decision-making units. For the open checking account process, we identified the labor and capital required in opening a checking account. The capital required was in the form of information technology. We initially attempted to use the cost of the technology, but found that there was no correlation between the money spent on this technology and its capabilities (Frei 1996). Thus, we used the capability of the technology as an input in order to separate the investment process performance with the performance of the service delivery process. The actual functionality of the technology was determined by surveying customer service representatives in six different branches for each retail bank in our study.

Along similar lines, we used activity time as an input as opposed to the cost of the labor in order to separate variation in pay with variation in performance. We determined the total time required by branch personnel to complete each step in the process. We calculated activity time using both the specific numbers for each bank as well as a standard number for each step in the process. There was little statistical difference in using either measure. The results reported in this paper are those found by using the standard times for each step. For a complete description of the surveys, see Frei (1996).

In terms of the output of the process, we found that the most appropriate measures for performance of the process were the amount of customer's time required in the process, and the amount of time until the customer received their ATM card and checks. As all three of these output measures are not of the "more is better" form required by all DEA techniques, we have used the affine transformation proposed by Ali and Seiford (1990). These measures were decided upon after extensive conversations with retail bank personnel and customers. While we have confidence in the appropriateness of these measures, there is a glaring absence of a quality measure. Unfortunately, retail banks do not consistently measure the quality of individual service processes in the branch and thus our analysis does not include any quality measures. It is left for future research to work with banks in an attempt to gather these quality measures.

Table 1. Inputs and Outputs for Process Efficiency Measurement

Inputs		
1.	Activity Time	The amount of time (in minutes) bank personnel spent performing the process.
2.	Technology Functionality	The functionality of the technology in place. Note that this is not the technology cost as, oddly, there is not typically a strong correlation between Technology Cost and Technology Functionality (see Frei 1996). This is measured in units of functionality were a higher number indicates more functionality.
Outputs		
1.	Check Cycle Time	The elapsed time (in days) from when a customer enters the bank to open a checking account, until the time the customer receives his/her (non-starter) checks.
4.	ATM Cycle Time	The elapsed time (in days) from when a customer enters the bank to open a checking account, until the time the customer receives his/her ATM card.
3.	Customer Time	The elapsed time (in minutes) from when a customer enters the bank to open a checking account, until the customer leaves the bank. That is, the amount of the customer's time required in the process.

The inputs to the open checking account process are labor and capital, with labor represented by the amount of time bank personnel spend performing the process, and capital appearing in the form of functionality of computer stations. This functionality measures the capabilities of the platform computer stations in the branch. It is important to realize that the technology input is not the cost of the machines, but rather the capability of the machine. This distinction is necessary because in retail banking, there are often orders of magnitude differences in the amount spent for similar functionality. Functionality is used rather than cost in order to separate out the performance of the IT procurement process from the performance of the open checking account process (see Frei 1996). Thus, by using Technology Functionality as an input, it is asserted that banks that use less Technology Functionality for similar levels of outputs are more efficient. The specific outputs of the open checking account process are the amount of time that a customer is involved in the process, and the elapsed time until the customer receives his/her checks and ATM card.

The inputs and outputs for the 126 banks as well as their efficiency scores are shown in Appendix A. Several aspects of these results are important. First, each bank that has an efficiency score of zero is efficient, and is thus on the efficient frontier. In addition, the efficiency score for the non-efficient banks is actually the distance of that bank from the frontier. Thus, the further away from the frontier a bank is, the greater its inefficiency. It is important to recognize that unlike most of the methodologies of other published work on DEA, the distance calculation used in this paper actually calculates this distance to the frontier. Most other DEA applications are restricted to calculate efficiency solely along either the input space or the output space, however, the method used herein, developed in Frei and Harker (1995), allows for the distance to be calculated along both spaces simultaneously, and thus represents the true distance to the frontier**.**

4.1 The Efficient Banks

Five of the 126 banks are efficient for the process of opening a checking account. The processes of these five banks range from the simplest process that clearly concentrates on speed, to those that are more complicated but endeavor to provide other benefits. Table 2 shows the inputs and outputs of the five efficient banks, the efficient banks' average, and the 126-bank average.

Table 2. The Efficient Bank Set

	Technology	Activity	Customer	Check Cycle	ATM Cycle
ID	Score	Time	Time	Time	Time
58	3 (0.41)	52 (0.97)	35 (1.18)	2.5 (3.05)	1 (2.34)
70	4 (0.55)	61 (1.14)	30 (1.38)	10 (0.76)	10 (0.23)
86	3 (0.41)	41 (0.77)	41 (1.01)	8.5 (0.90)	8.5 (0.28)
115	8 (1.09)	27 (0.50)	24 (1.72)	10 (0.76)	1 (2.34)
118	3 (0.41)	55 (1.03)	33 (1.25)	10 (0.76)	10 (0.23)
Eff. Avg	4.2 (0.57)	47.2 (0.88)	32.6 (1.29)	8.2 (1.31)	6.1 (1.08)
126-Bk Avg	7.34 (1.00)	53.57 (1.00)	42.21 (1.00)	8.69 (1.00)	6.11 (1.00)

The associated, scaled values are provided in parenthesis next to each input and output. These scaled values were determined according to the method proposed by Haag and Jaska (1995) which divides each value by the average of all institutions. In addition, as the frontier estimation analysis requires inputs to be of the *less is better* variety, and outputs of the *more is better* type, an affine transformation was used, as suggested by Ali and Seiford (1990), to account for the fact that even though time is an output of the process, it is not the case that more time is better. That is, for the three outputs on the open checking account process the transformation y = M – x is used, where x and y are the original and transformed observations and M is the largest observation. The result of this scaling is that the average of each category will be 1.0. Thus, for inputs, a value less than one is better than average, whereas for outputs, a value less than one is worse than average. As expected, the efficient banks have better inputs and outputs, on average, than the overall set of banks. However, even though a bank is efficient, it is not necessarily performing well along each of these dimensions. For example, one of the efficient banks, Bank 115, has a technology score and check cycle time worse than the overall average.

4.2 The Role of Inputs and Outputs on Process Efficiency

The efficiency score represents the relative ability of the bank to transform the inputs to outputs for the process of opening a checking account. By analyzing the relationship of each of these inputs and outputs, it can be determined whether a single input or output dominates the efficiency score relative to the others. For the overall data set, Technology Functionality has a much stronger relationship with the efficiency score (R^2 of 62%, statistically significant positive slope) than any of the other inputs and outputs as shown in Table 3. The slope implies that for each increase of one unit of functionality, a bank is 0.1 units further from the frontier. The direction of this relationship is expected as Technology Functionality is an input to the process and more input should lead to being further from the frontier. Activity Time, which is also an input to the process, is not strongly correlated with the distance from the frontier (R^2 of 4%), which means that as Activity Time increases, the distance from the frontier does not necessarily increase. In the case of the outputs, it is expected that as the amount of time required in the process decreases, the distance from the frontier should also decrease. The amount of time that the customer is involved in the process has a moderately strong relationship with distance from the frontier (R^2 of 24%) and the direction of the relationship is as expected (positive slope). The amount of time it takes for the customer to receive their checks has no relationship to the distance from the frontier. In addition, the amount of time it takes for the customer to receive their ATM card has a moderate relationship with distance from the frontier (R^2 of 12%) in the direction opposite than expected (negative slope). The counter-intuitive direction of the ATM relationship could be due to the fact that ATM Cycle Time plays a relatively less prominent role in the efficiency analysis due to the high variance of these times as compared to the other outputs (see Table 3 for the coefficient of variation of each of the inputs and outputs).

Table 3. Regression of Process Efficiency Score with Process Inputs and Outputs

	R^2	p-statistic	slope	coefficient of variation
Technology Functionality	60%	<= 0.0001	0.10	0.36
Activity Time	4%	0.02	0.01	0.11
Customer Time	24%	<= 0.0001	0.03	0.14
Check Cycle Time	0%	0.48	0.00	0.34
ATM Cycle Time	12%	<=0.0001	-0.02	0.90

4.3 The Role of Process Design Characteristics on Process Efficiency

This section explores the relationship between process-design characteristics and efficiency scores by determining the effect of specific characteristics such as the number of steps involved in a process, or the number of extra services provided in a process. In the open checking account process, these extra services consist of the provision of a check starter kit and the mailing of a thank you card. The number of steps has a very small effect on the efficiency score (a simple linear regression of the efficiency score versus the number of steps results in an R^2 = 2% and a slope of 0.02 with a p-value of 0.06). That is, the variation in efficiency score cannot be explained by the variation in number of steps required to complete the process. Similarly, performing an ANOVA demonstrates that the group of banks that provide no extra services has statistically lower efficiency scores than either the group of bank that provides one extra service or the group of banks that provides two extra services. There is no statistical difference between the group of bank that provides one extra service and the group of banks that provides two extra services.

It is surprising that the effect on explaining efficiency is not greater in either the number of steps or in providing extra-services as each of these increases Activity Time and potentially increases Customer Time. The fact that there is virtually no effect suggests that banks that are not providing these services are not performing worse than banks that are providing them. In fact, by providing these extra services, such banks may be paying more attention to their process designs and thus, making up for the increase in inputs (Activity Time) in other ways.

In order to determine how well banks manage the customer's time, it is necessary to look at when in the process certain steps are undertaken. There are three steps that can potentially be done after a customer leaves, but that many banks perform while the customer is still present. The Activity Time (input) is the same under either scenario, but the Customer Time (output) can be significantly impacted if the order of steps is addressed. The three steps that can be done after the customer leaves are sending the check order to the vendor, sending the ATM order to the vendor, and recording the sales credit. All 126 banks send the check order to the vendor, but only 113 of the banks send the ATM order (the rest make the ATM card on the spot), and 118 of the banks record a sales credit when a checking account is opened.

There seems to be no discernible effect on efficiency score by any of the three steps that can be performed either before or after the customer leaves. The effect of each of these steps individually on efficiency is determined by creating two groups of banks based on when the step occurred. For example, in order to compare the efficiency scores from the set of banks that sent the check order before the customer left with the set of banks that waited until after the customer left, an ANOVA is run with the null hypothesis that the groups have the same mean. The ANOVA yields a p-value of 0.21, which means that the null hypothesis cannot be rejected and thus, it cannot be asserted that the two groups are different. Similar results are achieved when comparing the banks that sent the ATM card while the customer was still there, with banks that waited until after the customer left (p-value 0.14) and when comparing the banks that recorded the sales credit while the customer was still there, with banks that waited until after the customer left (p-value 0.84). However, one should not conclude that the placement of these steps in the process is not important. Rather, the conclusion is that, by itself, the placement of these steps is not demonstrably important. However, the placement of these steps is precisely the area within which individual banks can improve their efficiency. For certain design groups, this placement is more of an issue than for others.

4.4 Process Design Groups

The previous section determined the effect of process design characteristics and process inputs and outputs on efficiency. In order to locate the cause of inefficiency, banks must be separated into process-design groups. These groups will allow further explanation of the implications and tradeoffs between improving an existing process and performing a radical process redesign. The process design groups were separated according to the schema described in Table 4, which yields four process design groups. The separation is according to two dimensions: the amount of Technology Functionality available in the process, and the number of extras offered in the process. It is important to keep in mind that splitting the data into design groups is an important step and while it was relatively straightforward to determine the managerially actionable characteristics to use in this example, it is not always straight-forward and often times, cluster analysis is necessary (see Section 3).

Table 4. Process Design Groups

	Low *Extras* (0 or 1)	High *Extras* (2)	Total
Low Tech (<= 7)	21 (Group 1)	43 (Group 2)	64
High Tech (> 7)	21 (Group 3)	41 (Group 4)	62
	42	84	126

After calculating the efficiency score for each of the 126 banks (average = 0.69), the scores are separated by process design group, and determined the average score for each group (0.42, 0.54, 0.91, and 0.27 respectively for groups 1 through 4). In order to determine which groups have statistically different efficiency scores, a Mann-Whitney test is performed, where the null hypothesis that the means are equal, for each of the six two-group combinations. The null hypothesis could not be rejected in only two of the six cases, with Groups 1 and 2 and with

Groups 3 and 4. That is, Group 1 is statistically different from Groups 3 and 4, but is not statistically different from Group 2. Similarly, Group 2 is statistically different from Groups 3 and 4, and Group 3 is not statistically different from Group 4. Thus, the low-tech groups are different from the high-tech groups, but the low and high extras have much less of an effect.

The above result illustrates the fact that banks with great deals of Technology Functionality in their operations have a very difficult time exploiting this functionality in terms of increased efficiency. The low-tech banks tend to be more efficient overall. With a lower level of functionality, there are simply fewer process design choices, and hence less variability. With investments in technology, the choices, and the ability to make incorrect ones, increases.

From the above analysis it is clear that there are differences between at least some of the process design groups; namely, those that are high-tech versus low-tech. How does this difference contribute to analysis of processes? In order to generate a more complete understanding of an inefficient bank, it is useful to know how well that bank is performing within its own design group, and what the potential benefit is from switching design groups. The knowledge of which groups are different from one another will help in uncovering the potential gain from switching from one design to another. To illustrate this, an analysis of an individual bank in demonstrated in the next section.

4.5 Individual Bank Analysis

One of the 126 banks, Bank 6, is analyzed in detail in this section in order to illustrate the proposed methodology. In order to analyze Bank 6, information from the tables in Frei (1996) is duplicated below. It is important to keep in mind that all of this analysis is done on a relative basis with scaled data. That is, the data from the 126 banks are scaled according to the method proposed by Haag and Jaska (1995) prior to the implementation of the models. As a result of this scaling, the average value of any input or output is 1.0. In addition, because this analysis requires inputs to be of the *less is better* variety and the outputs to be *more is better*, each of the three time-measuring outputs has been replaced by an affine transformation before scaling.

4.5.1 Analysis of Bank 6

Table 5 characterizes Bank 6 as a bank with high Technology Functionality, average Activity Time, below average customer and Check Cycle Time, and above average ATM cycle. In addition, Bank 6 is in process design Group 4, which is characterized by high Technology Functionality and providing extra services. The projection onto the 126-bank frontier, as shown in Table 5, shows that the efficient firms most like Bank 6 require less Technology Functionality and Activity Time (28% and 48% less, respectively), significantly improved customer time (84% improvement), moderately improved Check Cycle Time (22% improvement), and an ATM Cycle Time that is actually 8% *worse*. The implication of the ATM Cycle Time component of the projection is that while making the other improvements, Bank 6 can actually become a bit worse on this dimension and still be efficient.

Bank 6, which is in Group 4, has a reference set for the observable projection which contains banks in Groups 1, 2, and 3. Thus, in order for Bank 6 to become efficient overall, it will likely need to change process-design groups as no other banks within Group 4 are efficient. Often, changing design groups implies a much more significant change than trying to improve within a design group. Thus, Bank 6 has an important decision to make. It can either substantially change what it is currently doing, or it can limit the scope of this change and do the best it can given its design. If the decision is to be guided by this projection, then Bank 6 will undergo a radical change that will involve changing process designs. On the other hand, if Bank 6 is not in a position to implement radical changes, and is constrained to stay within its process design group, then an analysis of Group 4, in isolation, would be helpful. Understanding the tradeoffs involved in terms of anticipated improvement, as well as magnitude of change required, can only be accomplished by looking both at the overall analysis and the design group analysis which is performed in the next section.

Table 5. Bank 6 Projection onto the 126-Bank "Observable Frontier"

	Technology	Activity	Customer	Check	ATM	Design
ID	Score	Time	Time	Cycle Time	Cycle Time	Group
6	1.3622	1.0827	0.8801	0.7627	2.3388	4
Projection onto 126-bank *observable* frontier:						
	0.98	0.56	1.62	0.93	2.15	
% change:						
	-28%	-48%	84%	22%	-8%	
Reference Set:						
58	0.41	0.97	1.18	3.05	2.34	2
86	0.4	0.77	1.01	0.90	0.28	1
115	1.09	0.50	1.72	0.76	2.34	3

4.5.2 Bank 6 Analysis Within Process Design Group 4

Through the above analysis of the entire set of banks, it was determined that the projection onto the efficient frontier can require a substantial change in practices. In this section, the effect of analyzing a bank solely within its process-design group will be shown. The intent here is twofold. First, the aim is to separate the portion of the 126-bank efficiency score that is due to poor performance, and the portion that is due to an inappropriate process design. Second, in the instance that Bank 6 is not in a position to make radical changes, reasonable recommendations are still necessary. Before looking at the specific projection within the design-group, it is helpful to look at the relationship of the efficiency score with the inputs and outputs for those banks in Group 4. Recall from Table 3 that Technology Functionality had the largest correlation with the 126-bank efficiency score. The implication of this correlation was that those banks with more Technology Functionality tended to be more inefficient. However, within Group 4 there is a low correlation with Technology Functionality, which means that Group 4 banks with high functionality are not necessarily less efficient (see Table 6). This result is not surprising as Group

4 is a high Technology Functionality design group and thus, there is less variation of Technology Functionality within this group than there is when analyzing all 126 banks.[3]

Table 6. Correlation of Process Score, Inputs, and Outputs in Group 4

	R^2	p-statistic	slope	coefficient of variation
Technology Functionality	5%	0.17	0.02	0.14
Activity Time	42%	<= 0.0001	0.02	0.10
Customer Time	25%	<= 0.0001	0.01	0.12
Check Cycle Time	12%	0.02	0.01	0.39
ATM Cycle Time	0%	0.82	0.00	0.92

Note that within Group 4 (41 of the original 126 banks are in Group 4), there is a new *relative* scaling of Bank 6 as can be seen in Table 7. The scaling now represents the relative performance of Bank 6 within process design Group 4. Although Bank 6 had higher than average Technology Functionality for the overall data set, within Group 4 it is of average Technology Functionality. However, other than Technology Functionality, the relative scaling of the inputs and outputs are quite similar for the entire set of banks and Group 4 banks. Although the relative scaling of Bank 6 is similar with each set of banks, the projections onto the frontier are significantly different. For example, Bank 6 is significantly further away from the 126-bank observable frontier than from the Group 4 observable frontier (1.02 versus 0.25). Thus, it is expected that the projection to the observable frontier within Group 4 would be significantly different from the projection to the observable frontier of the 126-bank data set. If this is true, then the inefficiency of Bank 6, while partially due to poor performance (as indicated by being inefficient in its own design group), is also substantially due to the wrong process design. That is, by implementing the recommendations for the Group 4 projection, Bank 6 will certainly improve. However, compared to the entire data set, this improvement is not substantial.

The projection to the design-group frontier indicates that there are efficient banks that are using less Activity Time (17% less) achieving better Customer time and Check Cycle Time (8% and 7% improvement, respectively), and similar ATM Cycle Time than Bank 6. It is important to recognize that within Group 4, the projection to the frontier indicates that if Bank 6 follows the other recommendations, it can actually have higher Technology Functionality and still be efficient. If Bank 6 stays within the high-tech, high-extras design group, it should concentrate its efforts on improving Activity Time, Customer Time, and Check Cycle Time in order to become efficient. The Activity Time can be improved by automating some existing steps, and determining if all of the steps currently performed are necessary. The customer time can be improved by either removing a step performed when the customer is involved in the process, or by delaying a step until after the customer leaves the process. The Check Cycle Time can be improved either by automating some of the existing steps or by negotiating faster response time with the supplier.

[3] A coefficient of variation of 0.14 for Group 4 banks versus 0.36 for the 126-bank data set.

Again, because this is the projection to the “observable frontier”, Bank 6 is assured that the current recommendations have been realized by other banks in its design group and thus, are not unrealistic.

In terms of isolating which portion of the efficiency score is due to the poor execution and which is due to the wrong process design, note that Bank 6 is inefficient both overall and within process design Group 4. Thus, the portion of the efficiency score that is due to poor execution is $\frac{Efficiency_{DesignGroup}}{Efficiency_{Overall}} = \frac{0.2534}{1.0159} = 25\%$, and the portion that is due to an inappropriate process-design group is 100% - 25% = 75%. Thus, Bank 6 can expect a limited improvement within its own design group, but would yield a much larger improvement if it followed the implication from the projection to the 126-bank frontier as indicated in Table 5. However, this increased improvement must be weighed against the difficulty of potentially switching from high Technology Functionality to low, and from high customer *extras* to levels lower. Thus, the technique quantifies the potential improvement and thus sets the stage for the ability to perform the trade-off analysis that is, of course, context specific.

Table 7. Bank 6 Projection onto the Group 4 “Observable Frontier”

126	Group 4	Technology	Activity	Customer	Check Cycle	ATM Cycle
ID	ID	Score	Time	Time	Time	Time
6	2	1.06	1.07	0.92	0.77	2.29
Projection to *observable* frontier:						
6	2	1.22	0.89	0.99	0.82	2.29
% change:						
6	2	15%	-17%	8%	7%	0%
Reference Set:						
7	3	1.28	0.86	1.00	0.55	2.29
8	4	1.28	0.94	1.03	1.28	2.29
54	19	1.06	0.77	1.23	0.90	0.27
83	26	0.85	1.05	0.94	2.55	2.29

5. Contribution & Future Research

This paper presents a methodology for measuring process efficiency. The core of this methodology is the notion that process design is critical, especially in service delivery systems. That is, it is not sufficient to look at the amount of inputs that a process transforms into outputs, but also to understand the way in which this occurs. To this end, we introduce process design groups that are a way to group together processes according to specific design characteristics. It is through analyzing a process relative to its design group, as well as relative to the entire set of organizations that we are able to determine which portion of inefficiency is due to the wrong process design and which portion is due to the right design, poorly executed.

The empirical analysis presented herein analyzed a single process across 126 retail banks. For each bank, the analysis produced a relative efficiency score and a guide to becoming efficient within the bank's own design group, as well as to become efficient overall. The result of this analysis shows that there are potentially vast differences in recommendations resulting from analyzing a bank in each of these two ways. The implication of these differences addresses the question of how much inefficiency is due to the wrong design versus poor execution of the right design.

A limitation of this work is the lack of quality measures for the performance of the service delivery process. We have found these measures difficult to obtain, as most banks do not collect quality measures of this sort. However, the measures are still necessary and we are working with many retail banks in order to determine feasible, adequate measures. Future research will be aimed at refining the methodology proposed herein to account for errors in the measurement of inputs and outputs. Service delivery *systems* rarely consist of a single *process*. Thus, in order to benchmark systems rather than processes, a new methodology is required; this is also left for future research.

Another limitation of the proposed methodology is the link between these efficiency measures and overall profitability. Future work will be devoted to the integration of Return on Quality (Rust, Zahorik, Keiningham 1995; Rust and Zahorik 1993) measures with process design characteristics.

References

Ali, A. and Seiford, L (1990), “Translation Invariance in Data Envelopment Analysis,” Operations Research Letters, 9, 403-405.

Banker, R., Charnes, A., and Cooper, W (1984), “Some Models for Estimating Technical and Scale Efficiencies in Data Envelopment Analysis,” Management Science, 30, 1078-1092.

Berger and Humphrey (1997), “Efficiency of Financial Institutions: International Survey and Directions for Future Research,” working paper, Federal Reserve Board, Washington DC.

Brockett, P. L. and Golany, B. (1996), “Using Rank Statistics for Determining Programmatic Efficiency Differences in Data Envelopment Analysis,” Management Science, 42, 466-472.

Charnes, A., Cooper, W. W., and Rhodes, E. (1978), “Measuring the Efficiency of Decision Making Units,” European Journal of Operational Research, 2, 429-444.

Chase, R. B. (1981), “The Customer Contact Approach to Services: Theoretical Bases and Practical Extensions,” Operations Research, 29, 698-706.

Chase, R. B. and Tansik, D. A. (1983), “The Customer Contact Approach to Organization Design,” Management Science, 29, 1037-1050.

Chase, R. B., Northcraft, G. B. and Wolf, G. (1984), “Designing High Contact Service Systems: Applications to Branches of a Savings and Loan,” Decision Sciences, 15, 542-556.

Davenport, T. H. and Short, J. E. (1990), “The New Industrial Engineering: Information Technology and Business Process Redesign,” Sloan Management Review, 31, 11-27.

Frei, F. X. (1996), “The Role of Process Designs in Efficiency Analysis: An Empirical Investigation of the Retail Banking,” Unpublished Dissertation, The Wharton School, University of Pennsylvania, Philadelphia, PA.

Frei, F. X. and Harker, P. T. (1995), “Projections Onto Efficient Frontiers: Theoretical and Computational Extensions to DEA,” Working Paper, Wharton Financial Institutions Center, The Wharton School, University of Pennsylvania, Philadelphia, PA.

Frei, F.X., Harker, P. T., and Hunter, L. W. (1995), “Performance in Consumer Financial Services Organizations: Framework and Results from the Pilot Study,” Working Paper, Wharton Financial Institutions Center, The Wharton School, University of Pennsylvania, Philadelphia, PA.

Frei, F. X. Kalakota, K., and Marx L. (1997), “Process Variation as a Determinant of Service Quality and Bank Performance: Evidence from the Retail Banking Study,” Working Paper, Operations Management Group, Simon School of Business, University of Rochester, Rochester, NY.

Haag, S. and Jaska, P. (1995), “Interpreting Inefficiency Ratings: An Application of Bank Branch

Operating Efficiencies," Managerial and Decision Economics, 16, 7-14.

Huber, G. P. and Power, D. J. (1985), "Retrospective Reports of Strategic-Level Managers: Guidelines for Increasing Their Accuracy," Strategic Management Journal, 6, 171-180.

Huete, A. and Roth, A. V. (1987), I'm looking this up.

Kingman-Brundage, J. (1992), "The ABCs of Service System Blueprinting," in C.H. Lovelock (ed.), Measuring Services, Second Edition, Englewood Cliffs: Prentice Hall.

Levitt, T. (1972), "Production Line Approach to Service," Harvard Business review, 50, 41-52.

Morroni, M. (1992). Production Processes and Technical Change, London: Cambridge University Press.

Ortega, J.M., and Rheinbaldt, W.C. (1970), "Iterative Solution of Nonlinear Equations in Several Variables." New York: Academic Press.

Roth, A. V. and Jackson, W. E. (1995), "Strategic Determinants of Service Quality and Performance: Evidence from the Banking Industry," Management Science, 41, 1720-1733.

Roth, A. V. and van der Velde, M. (1989), "Investing in Retail Delivery Systems Technology," Journal of Retail Banking, 11, 23-34.

Roth, A. V. and van der Velde, M. (1991), "Operations as Marketing: A Competitive Service Strategy," Journal of Operations Management, 10, 303-328.

R.T. Rust, A.J. Zahorik, and T.L. Keiningham (1995), "Return on quality (ROQ): making service quality financially accountable," Journal of Marketing 59, 58-70.

R.T. Rust and A.J. Zahorik (1993), "Customer satisfaction, customer retention, and market share," Journal of Retailing 69, 193-215.

Seiford, L. M. (1994), "A Bibliography of Data Envelopment Analysis," Technical Report, Department of IEOR, University of Massachusetts, Amherst, MA.

Seiford, L. M. and Thrall, R. (1990), "Recent Developments in Data Envelopment Analysis: The Mathematical Approach to Frontier Analysis," Journal of Econometrics.

Sherman, H. and Gold, F. (1985), "Bank Branch Operating Efficiency", Journal of Banking and Finance, 9, 297-315.

Shostack, G. L. (1987), "Service Positioning Through Structural Change," Journal of Marketing, 51, 34-43.

Appendix A: Productivity in Retail Banking: The Wharton/Sloan Study

Table A1. Inputs and Outputs for 126-Bank Data Set

ID	Tech Score	Act Time	Cust Time	Check Cycle	ATM Cycle	Eff. Score
1	6	57	44	8.5	8.5	0.4483
2	7	54	52	10	1	0.8279
3	6	53	44	4	4	0.2918
4	11	54	39	7	8.5	0.5252
5	6	57	43	5	1	0.3291
6	10	58	47	10	1	0.7542
7	12	47	43	14	1	0.9224
8	12	51	42	6	1	0.6988
9	4	54	40	8.5	8.5	0.2577
10	11	43	43	10	10	0.7177
11	7	63	44	10	17.5	0.6238
12	9	52	41	7.5	5	0.3899
13	3	62	45	8.5	8.5	0.4519
14	8	56	52	7	1	0.766
15	9	47	42	8.5	3.5	0.4533
16	4	52	37	8.5	11	0.213
17	9	61	49	10	14	0.7727
18	8	62	59	5	1	1.1249
19	9	49	44	8.5	1	0.5278
20	8	51	43	10	10	0.4868
21	9	54	41	8.5	7	0.4341
22	9	67	58	12	1	1.2426
23	10	48	39	6	12	0.4477
24	9	47	42	9	9	0.4669
25	7	59	49	8.5	1	0.6621
26	7	53	43	10.5	1	0.5294
27	6	54	44	8.5	1	0.4735
28	6	56	42	8.5	1	0.4287
29	7	56	44	10	3	0.5165
30	9	66	54	10	1	1.0038
31	10	45	31	5	5	0
32	8	55	41	10	10	0.4617
33	6	45	36	10	14	0.354
34	7	49	36	7	10	0.2103
35	7	51	37	10	10	0.3743
36	4	52	46	6	6	0.4098
37	3	55	40	10	10	0.2825
38	3	57	45	8.5	8.5	0.3739
39	11	50	43	8.5	8.5	0.6611
40	10	44	39	7.5	5	0.4446
41	5	62	50	8.5	1	0.7333
42	7	41	36	8.5	2	0.2344
43	4	50	48	8.5	8.5	0.4491
44	12	61	49	6	1	0.9145
45	9	59	45	8.5	8.5	0.5442
46	7	55	46	10	1	0.5892
47	8	48	41	8.5	4	0.3609
48	12	46	37	10	5	0.5314
49	6	57	45	7	6	0.4207

(cont’d)

ID	Tech Score	Act Time	Cust Time	Check Cycle	ATM Cycle	Eff. Score
50	10	49	35	4.5	4.5	0.23
51	8	57	50	6	3	0.6364
52	11	61	49	10	1	0.9054
53	10	51	45	5	10	0.5171
54	10	42	35	8.5	8.5	0.3339
55	7	63	44	3.5	1	0.3882
56	4	56	40	8.5	8.5	0.2883
57	6	53	38	10	5	0.3887
58	3	52	35	2.5	1	0
59	3	57	43	2.5	1	0.4001
60	9	57	36	7.5	10	0.3971
61	12	53	42	7	12	0.7354
62	7	50	40	3	7.5	0
63	11	53	53	7.5	10.5	0.9586
64	11	51	42	21	2	1.1732
65	9	50	37	8.5	8.5	0.3615
66	9	43	35	8.5	8.5	0.3083
67	6	52	44	10	10	0.4958
68	3	49	37	10	10	0.2034
69	6	57	45	8.5	1	0.5163
70	4	61	30	10	10	0
71	4	55	42	6	6	0.2917
72	4	53	46	8.5	8.5	0.3845
73	7	53	52	7.5	7.5	0.7311
74	5	60	49	7.5	1	0.6781
75	8	51	45	6.5	1	0.4292
76	6	58	47	8.5	1	0.5915
77	8	50	40	10	10	0.4248
78	4	54	44	3.5	10	0.4195
79	4	51	39	8.5	7	0.1999
80	9	70	55	8.5	1	1.0388
81	6	59	52	7	10	0.7537
82	11	38	30	10	7	0.3274
83	8	57	46	3	1	0.4356
84	10	59	45	5	2	0.564
85	11	48	44	7	3	0.6542
86	3	41	41	8.5	8.5	0
87	12	55	43	10	1	0.8207
88	10	56	37	14	14	0.7334
89	7	52	35	14	14	0.5409
90	7	63	47	10	1	0.6621
91	8	45	34	10.5	3	0.3048
92	8	53	43	10	1	0.5206
93	10	55	48	12	1	0.8503
94	4	49	36	8.5	10.5	0.1755
95	9	53	35	6	8.5	0.2756
96	4	56	31	8.5	7	0.0145
97	9	51	43	7.5	7.5	0.4232
98	4	52	35	10	10	0.1473
99	5	54	37	8.5	12	0.312

(cont’d)

ID	Tech Score	Act Time	Cust Time	Check Cycle	ATM Cycle	Eff. Score
100	5	58	46	10	1	0.5623
101	6	54	40	10	1	0.457
102	3	61	44	8.5	8.5	0.4084
103	10	54	37	10	10	0.5088
104	5	44	37	7	10	0.2092
105	3	53	47	7	2	0.4608
106	7	61	44	6	6	0.3588
107	7	49	41	8.5	4	0.3414
108	5	54	37	14	30	0.7113
109	10	54	40	6	6	0.4455
110	9	58	41	10	10	0.5322
111	8	57	39	4	4	0.2263
112	5	62	50	14	1	0.8071
113	8	48	44	10	5	0.5099
114	10	56	48	10	1	0.7788
115	8	27	24	10	1	0
116	6	62	46	8.5	1	0.5795
117	3	53	34	9	12	0
118	3	55	33	10	10	0
119	4	57	35	14	37.5	0.8609
120	8	51	43	8.5	8.5	0.4026
121	4	53	37	10	2	0.2943
122	6	58	39	8.5	1	0.3904
123	12	47	41	24	3	1.3214
124	10	59	50	8.5	1	0.8158
125	11	54	45	12	1	0.8502
126	8	58	41	14	6	0.6841
average	**7.34**	**53.57**	**42.21**	**8.69**	**6.11**	

Appendix B: Summary of Methodology

The intent of frontier estimation is to deduce empirically the production function in the form of an efficient frontier. That is, rather than knowing how to convert functionally inputs to outputs, these methods take the inputs and outputs as given, map out the best performers, and produce a relative notion of the efficiency of each. The problem with the existing methods is that they each measure efficiency in a conceptually suspect, albeit computationally effective, way. If the DMUs are plotted in their input/output space, then an efficient frontier that provides a tight envelope around all of the DMUs can be determined. The main function of this envelope is to get as close as possible to each DMU without passing by any others. A simple example of an efficient frontier (using variable returns to scale) is shown in Figure 6.

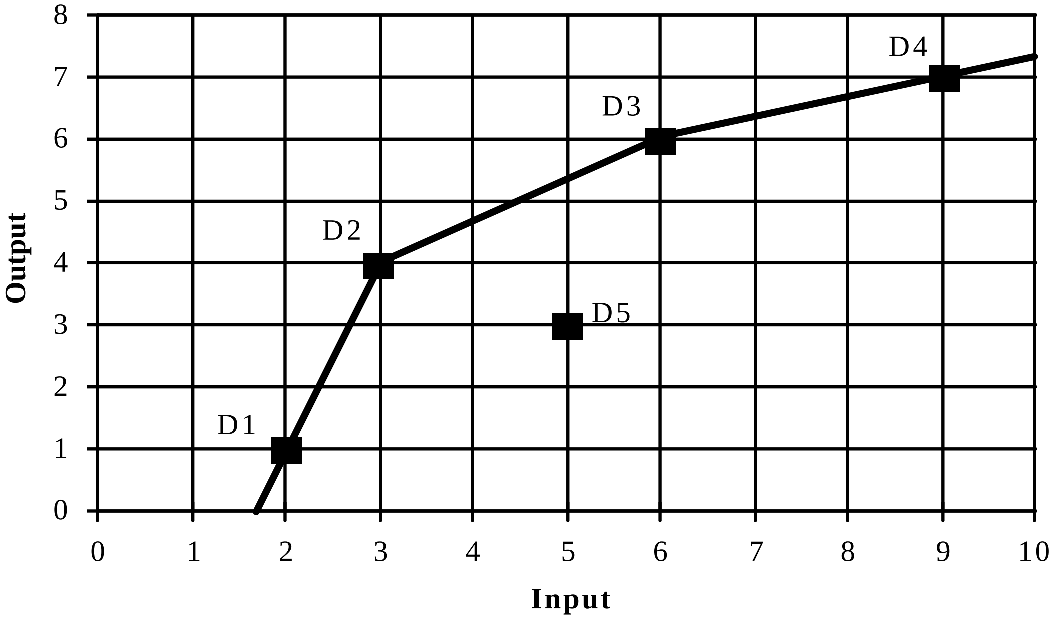


Figure 4. The Efficient Frontier

Each DMU along the frontier is considered efficient while those falling below the frontier, (e.g., D5) are considered inefficient. The method of determining the efficiency score for D5 varies according to the technique employed. Of the two classic methods, the input- or output-oriented methods, the efficiency score is determined, in effect, by determining the projection directly along the horizontal axis (holding outputs constant), or along the vertical axis (holding inputs constant). The method developed in this section determines the least-norm projection from an inefficient DMU to the frontier, in both the input and output space. In fact, one could develop a G-norm projection method (Ortega and Rheinbaldt 1970) wherein the input, output, and least two-norm projection are derived as special cases.

The oriented methods require a series of LPs to be solved, one for each DMU. The method described in this appendix also requires the solution of a series of liner programs – the series represented by the multiplier dual to the additive DEA model – and then algebraically determines the least-norm projection. See Frei and Harker (1995) for a full description of the algorithm.

This least two-norm projection is illustrated in Figure 7 along with each of the oriented methods. Intuitively, the least-norm projection measure seems more appropriate than the other measures in that a DMU's benchmarking or reference set should be those efficient DMUs that offer some resemblance to the DMU. It is easy to imagine either of the oriented measures projecting to points on the frontier much farther away than the least-norm projection, thus yielding a reference set with presumably less in common than that of the closest projection. As can be seen, the efficient set of DMUs does not change with any of these methods, but the efficiency score, or relative projection, will be different in each case. These efficiency scores are useful in that they determine the relative inefficiency of a DMU; however, the real impact of this method is in the determination of the precise coordinates of the DMU, or convex combination of DMUs, against which a DMU can benchmark. For example, via the least-norm projection measure, D5 will reference D2 and D3 as benchmarks for efficient performance. D1 and D2 should only be the benchmarks under the special circumstance of an inability to alter outputs. It is in this benchmarking or reference set that the dominance of this new method can be illustrated. Conceptually, it does not make sense to benchmark against firms that are potentially far away, except under special circumstances when either the inputs or outputs are restricted.

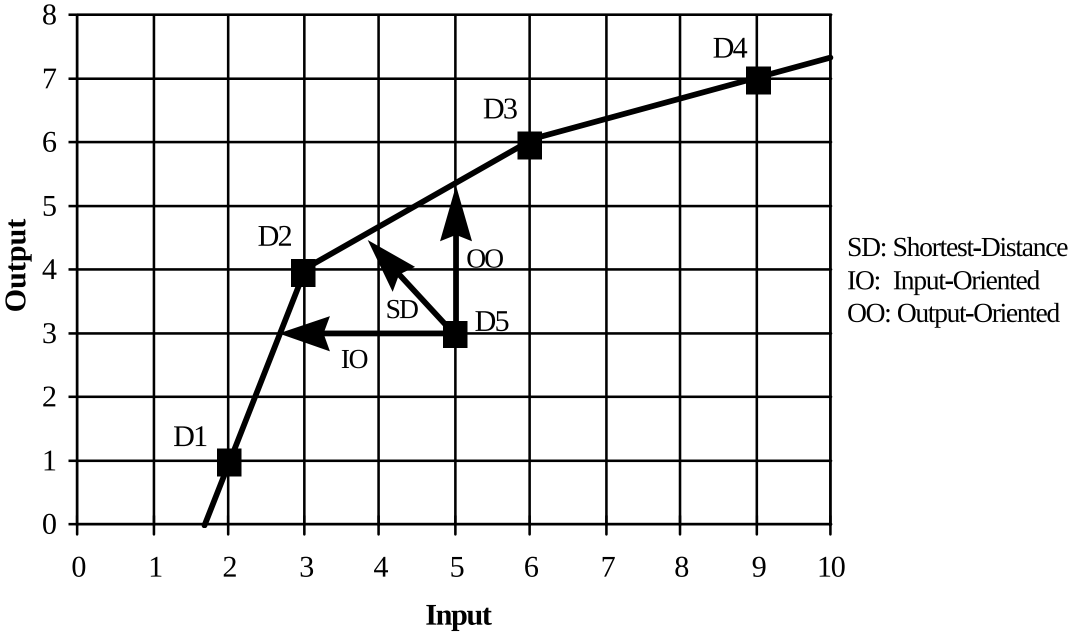


Figure 5. Single Input, Single Output Example

The mathematics of the method for determining this least-norm projection measure is described in detail in Frei and Harker (1995). However, it is important to note that this method requires the same number of LPs in its solution as each of the other methods. In addition, it is shown that the least-norm projection from an inefficient DMU is never at the intersection of two or more supporting hyperplanes of the efficient frontier. It is easy to see that in two dimensions, again referring to Figure 7, where no interior point will be closest to DMU 2.

Although the oriented methods, in effect, project onto the frontier, they do not construct the frontier in their solution. This is significant because in understanding which facet of the frontier an inefficient DMU projects to, it is possible to determine the returns-to-scale of the projection. For example, the input-oriented projection in Figure 7 lands on the frontier at an increasing returns-to-scale segment while the least-norm projection and output-oriented each project onto a decreasing returns-to-scale segment. The returns-to-scale of a particular segment is directly attributable to the sign of the intercept of the supporting hyperplane. That is, if the sign is negative, the returns-to-scale is increasing, and if the sign is positive, the returns-to-scale is decreasing. If the intercept is at zero, there is constant returns-to-scale. Using the projections that actually construct the facets of the efficient frontier in their solution to the set of LPs, it is immediately apparent where a DMU projection lies and the relative returns-to-scale of that facet.

The significance of understanding the returns-to-scale is that in determining if the cost associated with improving performance is worthwhile, it is important to know the benchmarking environment of a DMU. Thus, the input-oriented projection shows that within the production possibility set there are other firms that are not only producing the same with less, but they are also enjoying increasing returns-to-scale (that is, the ability to get a relatively bigger jump in outputs than is required from the inputs). The least-norm projection measure shows that the closest portion of the frontier, one that involves the ability to change inputs and outputs simultaneously, will leave the DMU in a decreasing returns-to-scale environment. The managerial implications are obviously significant.

In addition, by using the oriented methods the projection onto the frontier is determined by reducing (increasing) the inputs (outputs) by an identical percentage based on the efficiency rating. Thus, if there are three inputs, then using the input-oriented approach not only requires that outputs are held constant, but also mandates that the reduction in inputs along each of the input dimensions is identical (and equal to the efficiency score). Sherman and Gold (1985) improve upon this by determining a projection based on the reference set of the DMU and the corresponding weight of each DMU in the reference set (the dual price of each DMU constraint as shown in Section 2). That is, the reference set composite DMU is used as the focus of what an inefficient DMU would look like if efficient.

Least-Norm Projection Algorithm

The fundamental problem with computing the least-norm projection (least two-norm) projection onto the efficient frontier is that, although the production set is convex, the frontier is non-convex. Thus, one must solve a least two-norm projection onto a piecewise-linear, non-convex surface. To overcome the problems caused by the non-convexity, the following iterative procedure is used.

Step 1. Solve a linear program for each DMU in order to create the supporting hyperplanes $\mathbf{H} = \{H_i\}$ of the efficient frontier.

For each DMU:

Step 2. Calculate the least-norm projection d_i and the location of the projection algebraically for each hyperplane H_i.

Step 3. Compute $d_* = \min\{ d_i \}$, the least-norm projection to all hyperplanes **H**.

As in each of the previously described methods, the algorithm described herein requires the solution of an LP for each DMU in order to determine the equations of the supporting hyperplanes on the efficient frontier. The general form of these n linear programs is:

$$
\begin{aligned}
&\underset{m,n,u_0}{\max} \quad w_0 = m^T \mathbf{Y}_0 - n^T \mathbf{X}_0 + u_0 \\
&\text{subject to} \\
&m^T \mathbf{Y} - u^T \mathbf{X} - u_0 \vec{\mathbf{1}} \le 0 \\
&-m^T \le \cdot \vec{\mathbf{1}} \\
&-u^T \le \cdot \vec{\mathbf{1}} \\
&u_0 \text{ free}
\end{aligned}
\qquad \textbf{(1)}
$$

The solution to these n linear programs determines the efficient frontier. In order to find the closest point on the frontier to each DMU, the following nonlinear program is solved in order to compute the least two-norm projection of (**X**,**Y**) onto a hyperplane denoted by the equation.

$$
\begin{aligned}
&\underset{\overline{X},\overline{Y}}{\min} \quad \sqrt{\left\|\mathbf{Y} - \overline{\mathbf{Y}}\right\| + \left\|\mathbf{X} - \overline{\mathbf{X}}\right\|} \\
&\text{subject to} \\
&m^T \overline{\mathbf{Y}} - n^T \overline{\mathbf{X}} + u_o = 0
\end{aligned}
\qquad \textbf{(2)}
$$

This problem can be solved algebraically to produce the projection $\left(\overline{\mathbf{X}}, \overline{\mathbf{Y}}\right)$ and the least-norm projection D_{1HP}:

$$\overline{\mathbf{Y}} = \mathbf{Y} - \frac{m\left(m^T \mathbf{Y} - v^T \mathbf{X} - u_o\right)}{m^T m + n^T n}$$

$$\overline{\mathbf{X}} = \mathbf{X} + \frac{n\left(m^T \mathbf{Y} - v^T \mathbf{X} - u_o\right)}{m^T m + n^T n}$$

$$D_{1HP} = \sqrt{\frac{\left(m^T \mathbf{Y} - v^T \mathbf{X} - u_o\right)^2}{m^T m + n^T n}} = \frac{\left(m^T \mathbf{Y} - v^T \mathbf{X} - u_o\right)}{\sqrt{m^T m + n^T n}} \qquad \textbf{(3)}$$

Thus, the least-norm projection from each DMU to a given supporting hyperplane is known in closed form. By computing the least-norm projection for each hyperplane, one can then compute the minimum overall in Step 3 of the algorithm.

The Observable Frontier

In utilizing projections from an inefficient DMU onto the efficient frontier, it is important to understand the implications of where the projection lands. That is, there are issues about returns-to-scale depending on the segment of the frontier, as described above, as well as observability issues. The observability issue is one in which the efficient frontier can be split into two sections, one that represents either an observable or convex combination of observed input-output combinations, and another that represents extrapolations of observable scale. There are many instances in which it is not practical to have a benchmark that extends beyond the scale observed by any existing DMUs, and thus, independent projections onto the observable frontier are necessary. To illustrate this, Figure 3 duplicates the two-dimensional example from above with the observable portion of the frontier represented by a thicker line than the non-observable portion. In addition, another inefficient DMU has been added in order to illustrate that this DMUs shortest projection onto the entire frontier lands at a non-observable scale. That is, the projected output for D6 is 7.2 whereas the largest observed output, that of D4, is only 7. There may very well be instances in which the shortest projection onto the observable frontier, in this case directly onto D4, is the best solution. Thus, it is necessary to report both the overall efficiency score as well as the observable efficiency score.

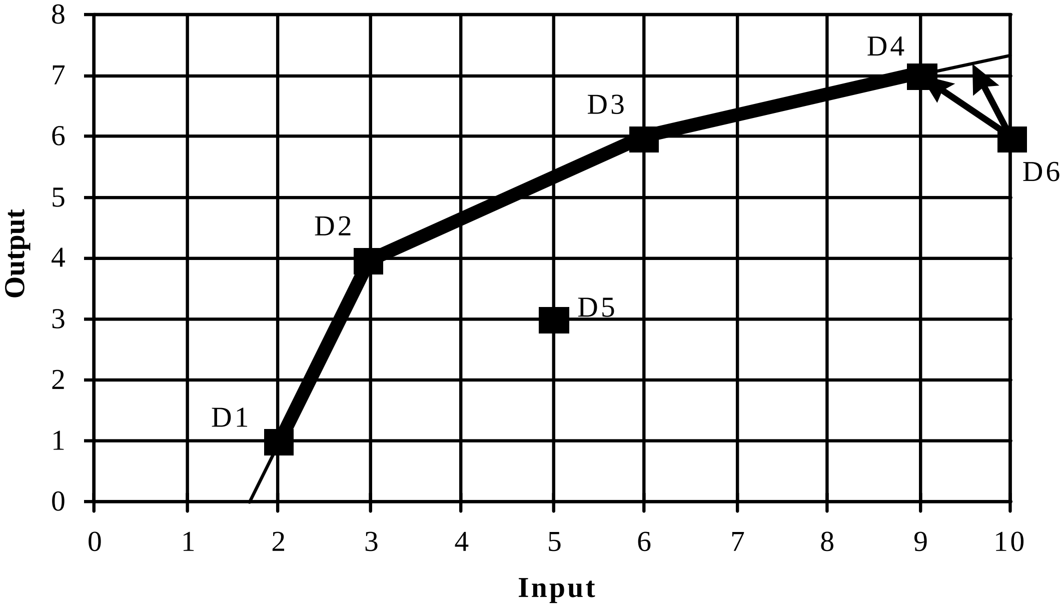


Figure 6. The Observable Frontier

Observable Frontier Algorithm

The following algorithm is used to compute the projection to the observable frontier:

Step 1. Repeat steps 2 - 4 of the shortest project algorithm for each unique hyperplane H_i that makes up a facet of the efficient frontier.

Step 2. Determine all DMUs that are on the supporting hyperplane H_i; these DMUs are the reference set for this hyperplane.

Step 3. Determine the convex hull of the reference set.

Step 4. Determine the least-norm projection from each inefficient DMU to the convex hull; call this distance D_i. In order to determine the least-norm projection from a DMU to the convex hull, one must solve the following non-linear program (10). In this problem, *I* represents the set of DMUs in the reference set of H_i:

$$\begin{gathered}
\min \quad \left\| \begin{pmatrix} y \\ x \end{pmatrix} - \begin{pmatrix} y_H \\ x_H \end{pmatrix} \right\|^2 \\
\text{subject to} \\
\begin{pmatrix} y_H \\ x_H \end{pmatrix} \in S \\
S = \left\{ \begin{pmatrix} y \\ x \end{pmatrix} : \begin{pmatrix} y \\ x \end{pmatrix} = \sum_{i \in I} \lambda_i \begin{pmatrix} y_i \\ x_i \end{pmatrix}, e^T \lambda = 1, \lambda \geq 0 \right\}
\end{gathered} \tag{4}$$

Step 5. The least-norm projection to the observable frontier is the smallest of the D_i for each inefficient DMU.