Operations in Financial Services—An Overview

Emmanuel D. (Manos) Hatzakis
Goldman Sachs Asset Management, Goldman, Sachs & Co., New York, New York 10282-2198, USA, manos.hatzakis@gs.com

Suresh K. Nair
Department of Operations and Information Management, School of Business, University of Connecticut, Storrs, Connecticut 06269-1041, USA, suresh.nair@business.uconn.edu

Michael L. Pinedo
Department of Information, Operations and Management Science, Leonard N. Stern School of Business, New York University, New York, New York 10012-1106, USA, mpinedo@stern.nyu.edu

We provide an overview of the state of the art in research on operations in financial services. We start by highlighting a number of specific operational features that differentiate financial services from other service industries, and discuss how these features affect the modeling of financial services. We then consider in more detail the various different research areas in financial services, namely systems design, performance analysis and productivity, forecasting, inventory and cash management, waiting line analysis for capacity planning, personnel scheduling, operational risk management, and pricing and revenue management. In the last section, we describe the most promising research directions for the near future.

Key words: financial services; banking; asset management; processes; operations

History: Received: November 2009; Accepted: July 2010 by Kalyan Singhal, after 2 revisions.

1. Introduction

Over the past two decades, research in service operations has gained a significant amount of attention. Special issues of Production and Operations Management have focused on services in general (Apte et al. 2008), and various researchers have presented unified theories (Sampson and Froehle 2006), research agendas (Roth and Menor 2003), literature surveys (Smith et al. 2007), strategy ideas (Voss et al. 2008), and have discussed the merits of studying service science as a new discipline (Spohrer and Maglio 2008). A few books and a special issue of Management Science have focused on the operational issues in financial services in particular (see Harker and Zenios 1999, 2000, Melnick et al. 2000). However, financial services have still been given scant attention in much of the literature relative to other service industries such as transportation, health care, entertainment, and hospitality. The dilution of focus, by concentrating on more general distinguishing features does not do justice to financial services where some of these characteristics are not central. (The more general features that are typically being considered include intangibility, heterogeneity, contemporaneous production and consumption, perishability of capacity, waiting lines (rather than inventories), and customer participation in the service delivery.)

In this overview, we mean by financial services primarily firms in retail banking, commercial lending, insurance (other than health), credit cards, mortgage banking, brokerage, investment advisory, and asset management (mutual funds, hedge funds, etc.).

1.1. Importance of Financial Services

Financial services firms are an important part of the service sector in an economy that has been growing rapidly over the past few decades. These firms primarily deal with originating or facilitating financial transactions. The transactions include creation, liquidation, transfer of ownership, and servicing or management of financial assets; they could involve raising funds by taking deposits or issuing securities, making loans, keeping assets in custody or trust, or managing them to generate return, pooling of risk by underwriting insurance and annuities, or providing specialized services to facilitate these transactions.

Services is a large category that encompasses firms as diverse as retail establishments, transportation firms, educational institutions, consulting, information, legal, taxation, and other professional, real estate, and healthcare. Even within financial services, there is a wide variety of firms which are characterized by unique production processes and specialized skills. The processes and skills required for banking are quite distinct from solicitations for credit cards, acquisitions of new insurance accounts, or the handling of equity dividends and proxy voting, for example.
Table 1  Employment within Financial Services

<table>
<thead>
<tr>
<th>NAICS code</th>
<th>Title within financial services</th>
<th>Total employees</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>5221</td>
<td>Depository credit intermediation (banks, credit unions, etc.)</td>
<td>1,816,300</td>
<td></td>
</tr>
<tr>
<td>5222</td>
<td>Non-depository credit intermediation (credit cards, mortgage lending, etc.)</td>
<td>659,930</td>
<td></td>
</tr>
<tr>
<td>5223</td>
<td>Activities related to credit intermediation (brokers for lending)</td>
<td>294,910</td>
<td></td>
</tr>
<tr>
<td>5231</td>
<td>Securities and commodity contracts intermediation and brokerage</td>
<td>516,010</td>
<td></td>
</tr>
<tr>
<td>5232</td>
<td>Securities and commodity exchanges</td>
<td>8,010</td>
<td></td>
</tr>
<tr>
<td>5239</td>
<td>Other financial investment activities (mutual funds, etc.)</td>
<td>344,950</td>
<td></td>
</tr>
<tr>
<td>5241</td>
<td>Insurance carriers</td>
<td>1,258,050</td>
<td></td>
</tr>
<tr>
<td>5242</td>
<td>Agencies, brokerages, and other insurance-related activities</td>
<td>907,880</td>
<td></td>
</tr>
<tr>
<td>5251</td>
<td>Insurance and employee benefit funds</td>
<td>47,730</td>
<td></td>
</tr>
<tr>
<td>5259</td>
<td>Other investment pools and funds</td>
<td>41,190</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total employees</td>
<td>5,916,470</td>
<td></td>
</tr>
</tbody>
</table>

Table 2 A sample of Financial Service Operations Job Titles, Responsibilities, and Products

<table>
<thead>
<tr>
<th>Titles</th>
<th>Products</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vice President, Operations, Operations Manager, Financial Operations Supervisor, Foreclosure and Bankruptcy Operations Manager, Risk Operations Team Manager, Team Manager Ops Control-Fixed Income, National Director Operations, Hedge Fund Operations Specialist, VP/ Director Of Operations</td>
<td>Premiums, Claims, Refunds, Cash flow and treasury management, Customer statements, Loan servicing and support, Trade confirmations, Reconciliations, Tax reporting, Security settlements, Mortgages</td>
</tr>
</tbody>
</table>

Nature of work/responsibility

Brokerage operations, Improve customer service, resolves customer issues, Review security pricing, Vendor support, Authorize net settlement, Hand-off of data, ensure data integrity, Verifying transactions, Tracking missing transactions, Leverage technology, Maintain ops controls, update policies, procedures, Back-office support, Understand regulations, Ensure compliance, Attain profit and revenue benchmarks, Reduce risk, Improve quality, Six sigma, Operational processing efficiency, Problem solving, Ensure best practices, Streamline activities, Manage key expenses, Work management tools, Monitors work flow, Production/testing

Source: Monster.com jobs listings during the week of March 8, 2009.

Even though services account for about 84% of the total employment in the economy, only about 4% of this workforce is employed in financial services. This might come as a surprise to some because financial services transactions in one form or another are so ubiquitous in our lives. Not surprisingly, however, the number of financial services firms is about 7% of the total non-farm firms and contributes about 13% of total non-farm sales. Only wholesale trade has a similar employment and number of firms with a larger contribution to sales.

Table 1 provides employment information for the sub-codes within financial services. As can be seen, retail banks, insurance companies, and insurance brokers together employ about two-thirds of the financial services workforce. A cursory look at the table gives a sense of the diversity of the services sector. Clearly operations management problems and approaches used to solve them have to be customized for particular types of services—we already know that what works for manufacturing may not work for services, but by looking at Table 1 we can also realize that what works for retail trade or recreational services may not work for financial services. A quick glance through Monster.com’s job openings for operations managers in financial service firms shows a wide variety of titles, responsibilities, and “products” related to such jobs. This is shown in Table 2, and gives a sense of the wide swath of topics that could be covered in academic research on financial services operations. As financial services are such an important segment of the services economy, we wish to explore whether operations in financial services are indeed unique, or share several characteristics with services in general. That is the motivation for this special issue.

1.2. Distinctive Characteristics of Operations in Financial Services

There are several unique operational characteristics that are specific to the financial services industry and that have not been given sufficient attention in the general treatment of services in the extant literature. We list below a number of these unique operational characteristics and elaborate on them in what follows:

- Fungible products with an extensive use of technology
- High volumes and heterogeneity of clients
- Repeated service encounters
- Long-term contractual relationships between customers and firms
- Customers’ sense of well-being closely intertwined with services
- Use of intermediaries
- Convergence of operations, finance, and marketing.

1.2.1. Fungible Products with an Extensive Use of Technology. One obvious difference between operations in financial services and operations in manufacturing and in other service industries is that the “widgets” in financial services are money, or related financial instruments. As there is a declining use of the physical vestiges of money such as coins, currency,
bond, or stock certificates, much of the transactions are in the form of bits and bytes. Thus inventory is fungible and can be transported, broken up, and reconstituted (facilitating securitization, e.g.) in malleable ways that are simply not possible in manufacturing or in other service industries (see, e.g., the recharacterization of bank reserves in Nair and Anderson 2008).

The increased use of online transactions (in brokerages, credit card payments, retail banking, and retirement accounts, e.g.) are forcing fundamental changes in the way operations managers think about capacity issues (for statement mailing and remittance processing, or for transfer of ownership in securities, e.g.). The fact that adoption of online transactions is still growing and has not yet matured and leveled off makes capacity planning a big challenge. Yet we are aware of very little research that would help managers deal with this issue.

1.2.2. High Volumes and Heterogeneity of Clients. Financial services are characterized by very high volumes of customers and transactions. Furthermore, customers are not all alike. In many firms, a small fraction of the customers generate most of the profits, giving the firms an incentive to view them differently and provide differential treatment, given the firms’ limited resources. For example, high net worth individuals may be treated differently by asset management firms; banking clients who keep high balances in checking accounts and transact heavily may be handled differently from depositors who keep almost all their funds in savings accounts and certificates of deposits (CDs) and transact minimally; revolvers (i.e., customers who carry over balances from one month to the next) may be regarded differently from transactors (customers who do not revolve balances) by a credit card firm. In most non-financial services, because of a limited number and sporadic interactions with customers (e.g., in restaurants and amusement parks), one customer is considered, for the most part, similar to the next one in terms of margins and attention required.

1.2.3. Repeated Service Encounters. In contrast to other service industries, where research typically focuses on a single encounter (“the moment of truth,” “when the rubber hits the road”), financial services are characterized by repeated service encounters or potential encounters between the firm and its customers due to regular monthly statements, year-end statements, buy/sell transactions, insurance claims, money transfers, etc. Anecdotal evidence from the brokerage and investment advisory industry suggests that clients with low asset balances and transaction volumes contribute the least to firm revenue and the most to operational cost through calls for customer service. One online bank discouraged calls to customer service because it found that just a few calls by a client could wipe out all the profit from the client’s savings account. Very little research in service operations management has focused on this issue. New customers constitute another major group that is more likely to make calls with billing questions or inquiries regarding their statements. Should billing and statementing to new customers be handled differently, perhaps with more care, than to existing customers? Obviously, if this observation is true, differential handling can reduce the traffic to call centers. Existing call center research usually assumes the call volume to be a given, for the most part, and the focus is on “managing” the traffic. This is akin to traditional manufacturing where it was assumed that large setup times were a given, and a good way to “manage” would be to use an optimal batch size. One learns to live with such a constraint. Not until just in time (JIT) manufacturing came along did managers question why setup times were large and what could be done to reduce them.

At one credit card company, operations managers struggled for years to cope with volatile demand in bill printing, mailing, remittance processing, and call center operations. Daily volumes could fluctuate easily between half a million and one and a half million pieces of mail in remittance processing, and managers were reconciled to high overtimes and idle times because they felt they had no control over when customers mailed in their checks, and reducing float was important. Call volumes at the call centers were similarly volatile, as were volumes in the bill printing and mailing operations. This situation continued until someone recognized that all these problems were interconnected and to a large extent within the control of the firm. As it happens, the portfolio of current customers is distributed into about 25 cycles, one for each working day of the month. For example, customers in the 17th cycle are billed on the 17th working day of the month. Care is taken to ensure that customers in the same zip code are put in the same cycle so volume-mailing discounts from the US postal service can be obtained, and the cycles are level loaded. However, this allocation to cycles was carried out several years back and over time some customers had closed their accounts while new customers had been added to existing cycles, resulting in large differences in the numbers of customers in the various cycles, and in a wide variability in the printing and mailing of monthly statements. On the remittance side, an analysis found that there was less randomness in customer payment behavior than one would expect. There were broadly four cohorts of customers: one that sent in payments on receipt of the statement, a second that mailed checks based on due date, a third that acted based on salary payment date, and a fourth that acted randomly. The first two
cohorts, the largest ones, were in fact dependent on the cycles that the firm had set up many years back. Similarly, billing calls were also heaviest soon after the statement was received by the customer, again traffic that was determined by the cycles created by the firm.

If the cycles could now be level loaded, many of these problems would disappear (similar to what happened in manufacturing when setup times were dramatically reduced thereby enabling lean operations). But there was a problem—the firm needed to inform each customer if their cycles were moved, for good reason because customers needed to plan their finances. However, this notification was not necessary if the move was to be within ±3 days from their current cycle. An optimization model found that this constrained move was sufficient to take care of the vast majority of moves that were initially thought to be necessary.

This example illustrates how stepping back and taking a broader view of the situation and collaborating across processes can have a major impact on financial service operations, something that is lacking in the current literature.

1.2.4. Long-Term Contractual Relationships Between Customers and Firms. Connected to the previous characteristic of repeat encounters is the recognition that, unlike in other services, in financial services the firm and the customer have a relatively long-term contractual arrangement. However, technology and information availability makes comparison shopping easy, resulting in easy switching between firms, and therefore high attrition. This loss of customers makes the acquisition process very important to the continued growth and profitability of the firm. Similarly, loyalty programs (such as rewards and balance transfer programs in the credit card business) are important to stanch the bleeding. The design and execution of these programs are based on complicated processes that need to consider risks, costs, redemptions, incremental sales, scheduling and sequencing of offers, etc. Researchers in financial services operations, by not making their presence felt in these areas, are missing the boat with regard to issues that are the most important (“must do” activities) for the firm, and may be paying instead too much attention to relatively mundane and low-impact issues (“good to do” activities).

Just as the above processes aim at increasing revenue, there may be other processes that are put in place to reduce unnecessary costs. In the insurance business, for example, the claims processes may primarily revolve around a call center, which has attracted sufficient attention in the literature as we will see later. But unnecessary costs can be reduced by fraud prevention and detection, and subrogation activities (money the firm pays out but is owed to it by other carriers). Timely intervention can avoid expiry of opportunities to collect dollars owed, and more attention could be paid to even small opportunities. There is an extensive literature in risk and insurance journals on scoring for fraud prevention and detection, but leveraging that information in the claims process can benefit from an operations perspective.

Another example from the insurance industry concerns worker’s compensation claims, where the process for handling workplace injury can have long tails spanning several years before the claim is closed. The process is complicated with interactions between the worker, the employer, medical practitioners, hospitals, state authorities, and lawyers (both on the staff of the insurance firm and panel counsel, i.e., lawyers who are hired on an ad hoc basis). There are several opportunities here (Jewell 1974) to speed up the process (and speed up the workers’ return to work, which is in the insurance firm’s interest), reduce costs, detect fraud, ensure that review triggers are not overlooked, increase the utilization of staff counsel compared with panel counsel by better scheduling of appearances for hearings, etc. We are unaware of any recent operations management literature in this area.

1.2.5. Customers’ Sense of Well-Being Closely Intertwined with Services. Along with the ease of manipulating the putty at the core of the financial services process comes the responsibility of working with something that is so close to the customer’s sense of well-being and worth. Poor operations management that results in delays, quality issues, or slowness can and will attract regulatory scrutiny and unfavorable publicity, and will generate immediate rebukes from the customer in the form of calls, complaints, and because the account can be easily moved around, customer attrition. At least two factors make the detection of errors due to operational faults and their exposure to the clients relatively easy in financial services:

(i) the amount, frequency, and detail of communication and disclosure as required by regulation, and

(ii) the clients’ heightened propensity and incentive to check for error in something so closely linked to their livelihood and sense of security.

Because of the above, tolerance for error is significantly lower than in other industries. For example, faulty processes resulting in incorrect calculations of interest amounts in savings, mortgage loan, or credit card accounts, or in inappropriate handling of stock
dividend payments, become obvious immediately after monthly statements are sent out.

The above customer service issues are distinct from the perceived quality of the performance of individual customer brokerage portfolios, retirement accounts, annuities, mutual funds, and interest accrued in retail banking. With the plethora of information available comparing a customer’s firm with others in the same space, moving an account to the competition is only a few clicks away. Even though performance may depend on the economy, the stock market, investment research, and fund manager performance, recognizing the costs (Schneider 2010) and capacity issues (the increased transactions during market turbulence, e.g.) are important operations management concerns that have received little attention.

1.2.6. Use of Intermediaries. This is an important aspect of the financial services industry. In some cases a direct-to-consumer approach is used (credit cards); in other cases most of the customer facing work is done by intermediaries (financial advisors, insurance agents, annuity sales through banks, etc.); and in still other cases the firm’s employees and its agents have to collaborate with one another (insurance). Working through an intermediary entails a set of issues not normally seen in other services that function without intermediaries. For example, financial product and service design and delivery get filtered through the prism of what the agent feels is in his or her own best interest. At times, the relationship between the firm and the intermediary is not exclusive, hereby adding a layer of complexity because the customer may choose between products from competing firms. Therefore, what gets planned in the corporate offices of the financial services firms and what is seen by the customer may be quite different. The operations management literature, to our knowledge, has not paid attention to product and service design in such situations, because the implications on customer lifetime interactions with the firm go much beyond initial pricing, product features, and the inventory of brochures left with the agents.

1.2.7. Convergence of Operations, Finance, and Marketing. There is probably no other industry where this convergence is more pronounced. These functions are supported and enabled by a healthy dose of statistics, technology, and optimization. By focusing only on back-office operations such as call centers, researchers in service operations are leaving a lot on the table. There is very little research in the service operations literature that leverages this convergence, which requires a choreography, as described by Voss et al. (2008), who put it in a more limited context of operations and marketing. For example, the client acquisition process in full-service retail brokerage and investment advisory firms begins at the corporate level, where it draws resources from marketing, strategy, information technology, and operations, and is ultimately implemented through the sales force of brokers/financial advisors. Customer acquisition at a credit card company is a competitive differentiator and a complex process focused on direct mail campaigns. At many large firms, the budgets for direct mail run into several hundred million dollars annually. By focusing on the billing mailroom, collections, call centers, and billing call centers, researchers in service operations are working on a problem akin to quality inspectors at the end of the production line—by then it is too late, the volumes of mail and calls are baked in during the mailing campaign creation, while their skills could have made the mailings more effective and targeted (given the minuscule response rates), resulting in fewer delinquent accounts (requiring fewer outbound collection calls), and perhaps also fewer billing calls to inbound call centers.

At a more sophisticated level, very few credit card firms use contact history in mailing solicitations, which may result in repeated mailings to chronic non-responders. The managers developing campaign strategies may not have the analytical background that a researcher in service operations can bring to bear, and the cycle time for campaign creation is typically so long and complicated that much attention gets expended on scoring for credit and response, file transfers from credit bureaus and data vendors, scrubbing of data, etc. These complicated processes leave little time to incorporate experience from a previous mailing because a reading of the results of that mailing takes time (prospective customers may not respond immediately, even if they do respond), and file structures may not have been designed to carry information about previous contacts and the response to them.

Another indication of this convergence is that the majority of the undergraduate and MBA hiring at an investment bank in the greater New York City area from one of the region’s business schools is in the COO’s operation, whether the student concentrations are in finance, marketing, information systems, or operations.

The foregoing does not imply that no significant work has been done in the research on financial service operations, just that areas of work have had a narrow focus. The purpose of this article and this special issue is to begin to expand that focus and encourage research in neglected or emerging areas in financial service operations. We will survey existing research next, not only where the attention is only on financial service operations, but also where research in service industries in general has substantial application in financial service operations. In Appendix
A, we provide an overview of the various operations processes in financial services and highlight the ones that have been addressed in operations management literature.

This survey paper is organized as follows. Sections 2 through 9 go over eight research directions that are of interest from an operations management point of view. The first couple of sections consider the more general research topics, whereas the later sections go into more specific topics and more narrowly oriented research areas. Section 2 focuses on process and system design in financial services, while section 3 considers performance measurements and analysis. Section 4 deals with forecasting, because forecasting plays a major role in virtually every segment of the financial services industry. The next section focuses on cash and liquidity management; this section relates cash management to classic inventory theory. Sections 6 and 7 deal with waiting line management and personnel scheduling in retail banking and in call centers. Even though these two topics are strongly dependent on one another, they are treated separately; the reason being that the techniques required for dealing with each one of these two topics happen to be quite different from one another. Section 8 focuses on operational risk in financial services. This area has become very important over the last decade and this section describes how this area relates to other research areas in operations management, such as total quality management (TQM). Section 9 considers product pricing and revenue management issues. The last section, section 10, presents our conclusions and discusses future research directions.

2. Financial Services System Design

Service systems design has attracted quite a bit of attention in the academic literature. It is clear that service design has to be as rigorous an activity as product design, because the customer experiences the service first hand, much like a product, and comes away with impressions regarding the quality of service. Although the quality of service delivery depends on a number of factors, such as associate training, technology, traffic, neighborhood customer profile, access to the service (channel access), and quality of resource inputs, the service experience gets baked into the process at the time of the service design itself, and therefore a proper service design is fundamental to the success of the customer experience.

2.1. Aspects of Service Design

Service research has usually focused on capacity management (type of customer contact, scheduling, and deployment) and the impact of the response to variability on costs and quality. For long the nature of customer contact has influenced service design thinking by creating front-office/back-office functions (Sampson and Froehle 2006, Shostack 1984). Shostack also pioneered the use of service blueprinting for identifying fail points where the firm may face quality problems. She illustrated this methodology for a discount brokerage and correctly identified that many of the operational processes are not seen by the customer; she then focused on the telephone communication step, the only one with client contact. This focus on client contact tasks, whether in the front office or in the back office, is widespread in services research in general and in research on financial services operations in particular. One reason may be that service researchers have found it necessary to motivate their work by differentiating services from products (whether it is service marketing vs. product marketing, or service design vs. product design), and client contact is an obvious differentiator.

From the outset, it has been clear that service processes are subject to a significant amount of randomness from various sources. Frei (2006) discusses the various sources of randomness in service processes and how firms react to them in the design of their services. She identifies five types of variability—customer arrival variability, request variability, customer capability variability, customer effort variability, and customer preference variability. She states that firms design services to factor in this variability by trying either to accommodate the variability at a higher cost (cross training of employees, increased automation, variable staffing) or to reduce the variability with a view to increasing efficiency rather than cost (off peak pricing, standard option packages, combo meals).

2.2. Focus on Single Encounters

Much of the services literature, however, focuses on single service encounters, which are common in services such as fast food. Even if a customer repeatedly visits the same restaurant, there is not the kind of stickiness to the relationship as can be found in financial services. Retail banking seems to have attracted the most attention among financial services with respect to service design, but here again the focus is on disparate single visits to the branch or Automated Teller Machine (ATM), rather than as part of a life cycle of firm–customer interactions. Other than meeting the branch manager when opening an account, there is usually no other recognition of the stage of the relationship in the delivery of service. Perhaps this will change with time as more firms start experimenting with their service delivery design as Bank of America has been doing (Thomke 2003).

2.3. Descriptive vs. Prescriptive Studies of Financial Services

Several descriptive studies have focused on retail banking (Menor and Roth 2008, Menor et al. 2001),
substitution of labor with information technology (Fung 2008), the use of customer feedback to improve customer satisfaction (Krishnan et al. 1999), the use of distribution channels (Lee et al. 2004, Xue et al. 2007), self-service technologies (such as ATMs, pay at the pump, see Campbell and Frei 2010a,b, Meuter et al. 2000), online banking (Hitt and Frei 2002), and e-services in general (see Boyer et al. 2002, Ciciretti et al. 2009, Clemons et al. 2002, Furst et al. 2002, Menor et al. 2001). These studies talk about the types of customers who use the various different channels and how firms have diversified their delivery of services using these new channels as newer technologies have become available. However, they are usually descriptive, rather than prescriptive, in that they speak about how existing firms and customers have already adopted these technologies, rather than what they should be doing in the future. For example, there are few quantitative metrics to measure a product (e.g., its complexity vis-a-vis customer knowledge), a process (e.g., face to face vs. automated), and proximity (on-site or off-site) to help a manager navigate financial service operations strategies from a design standpoint based on where her firm is now. In that sense, financial service system design still has ways to go to catch up with product design (product attributes, customer utility, pricing, form and function, configuration, product development teams, etc.) and manufacturing process design (process selection, batch/line, capacity planning, rigid/flexible automation, scheduling, location analysis, etc.). Because batching and lot sizing issues have been of considerable interest in the history of the study of manufacturing processes, and because online technologies have made the concept of batching considerably less important, it would be interesting to see how research in service systems design unfolds in the future. One paper with prescriptive recommendations for service design in the property casualty insurance industry is due to Giloni et al. (2003).

3. Financial Services Performance Measurement and Analysis

3.1. Best Practices and Process Improvement

Many service firms are measuring success by factors other than profitability, using such factors as customer and employee loyalty, as measured by retention, depth of relationship, and lifetime value (Heskett et al. 1994). Chen and Hitt (2002), in an empirical study on retention in the online brokerage industry, found that ease of use, breadth of offerings, and quality reduce customer attrition. Balasubramanian et al. (2003) find that trust is important for online transactions, because physical appearance of branches, etc. no longer matter in such situations. Instead, perceived environmental security, operational competence, and quality of service help create trust.

In general, service quality is difficult to manage and measure because of the variability in customer expectations, their involvement in the delivery of the service, etc. In general, there may be two different measures of service quality that are commonly used: the first refers to and measures the actual service provided (e.g., customer satisfaction, resolution, etc.), the second may refer to the availability of service capacity/personnel (e.g., service level, availability, waiting time, etc.). The first type of quality measure is not as nebulous in financial services where the output is generally related to monetary outcomes. If there is an error in the posting of a transaction, or if quarterly returns from a mutual fund are below industry performance, there is an immediate customer reaction and the points in the service design that caused such failures to occur is apparent, whether it is in remittance processing or in the hiring of a fund manager. Quality in financial services is not influenced by such matters as the mood of the customer, as may be the case in other services. This makes ensuring quality in financial services more doable and one of the foci of the research in operational risk management which we will discuss later.

Roth and Jackson (1995) found that market intelligence and imitation of best practices can be an effective way of improving service quality, and that service quality is more influenced by service process choices and the cumulative impact of investments than by people’s capabilities. Productivity measurement in services is also a challenge (Sampson and Froehle 2006). Bank performance as a result of process variation has been studied by Frei et al. (1999).

This current special issue of Production and Operations Management provides some interesting new cases of process improvement in financial services. The paper by Apte et al. (2010), “Analysis and improvement of information-intensive services: Evidence from insurance claims handling operations,” presents a classification of information-intensive services based on their operational characteristics; this paper proposes an empirically grounded conceptual analysis and prescriptive frameworks that can be used to improve the performance of information- and customer contact-intensive services. The paper by De Almeida Filho et al. (2010) focuses on collection processes in consumer credit. They develop a dynamic programming model to optimize the collections process in consumer credit. Collection processes have been the Cinderella of consumer lending research, because psychologically lenders do not enjoy analyzing their mistakes, and also once an accounting loss is ascribed to a defaulted loan, there had been little incentive for senior managers to keep track of how
The paper by Buell et al. (2010) investigates why self-service customers are more reluctant to change their service provider. This paper’s primary contribution is to investigate how satisfaction and switching costs contribute to retention among self-service customers. This is a particularly important issue in the financial services industry where considerable investments have been made in developing self-service distribution channels, and migrating customers to them.

3.2. An Example of Best Practices: Asset Management

Asset management provides an interesting example of an area within the financial services sector that has been receiving an increasing amount of research attention with regard to best practices from various operations management perspectives. The body of research on operations management in asset management is growing, however, not always produced by operations management researchers, but often by those in the finance world (Black 2007, Brown et al. 2009a, b, Kundro and Feffer 2003, 2004, Stulz 2007), who examine operational risk issues in hedge funds. A collection of operations management research papers in asset management can be found in a recent book by Pinedo (2010). Alptuna et al. (2010) present a best practices framework for the operational infrastructure and controls in asset management and argue that it is possible to effectively implement such a framework in organizations that enjoy a strong, principle-based governance. They examine conditions under which the cost-effective strategy of outsourcing asset management operations can be successful for asset managers and their clients. Figure 1, which has been adapted from Alptuna et al. (2010), shows the multiple constituent parts that must work together in order for a typical asset management organization to function effectively. Figure 2, also adapted from Alptuna et al. (2010), lists the functions in the investment management process according to their distance from the end client. Typically, the operations-intensive functions reside in the middle and back offices; accordingly, the untapped research potential of operations in asset management must be sought there. One can create a similar framework, as shown in Figures 1 and 2, for a typical retail bank, credit card issuer, mortgage lender, brokerage, trust bank, asset custodian, life or property/casualty insurer, among others, none of which is less complex than an asset manager. Outsourcing operations adds to the complexity by introducing elements of quality control for outsourced pieces and coordination between the main organization and the third-party provider (State Street 2009).


**Figure 1  Typical Structure of an Asset Management Organization**

<table>
<thead>
<tr>
<th>Asset management</th>
<th>Independent internal oversight functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>o Investment research, management, and execution</td>
<td>o Compliance, legal and regulatory manager</td>
</tr>
<tr>
<td>o Sales and client relationship management</td>
<td>o Controllers</td>
</tr>
<tr>
<td>o Product development</td>
<td>o Credit and market risk management</td>
</tr>
<tr>
<td>o Marketing</td>
<td>o Internal audit</td>
</tr>
<tr>
<td></td>
<td>o Valuation oversight</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Internal support teams</th>
<th>External service providers</th>
</tr>
</thead>
<tbody>
<tr>
<td>o Billing</td>
<td>o Brokerage, clearing and execution</td>
</tr>
<tr>
<td>o Human resources</td>
<td>o Custody and trust services</td>
</tr>
<tr>
<td>o Operations</td>
<td>o Fund administrator</td>
</tr>
<tr>
<td>o Operational risk</td>
<td>o Prime brokerage and financing</td>
</tr>
<tr>
<td>o Performance</td>
<td>o Reputable auditor</td>
</tr>
<tr>
<td>o Tax</td>
<td>o Valuation (reputable third-party valuation firm)</td>
</tr>
<tr>
<td>o Technology</td>
<td></td>
</tr>
</tbody>
</table>
agement programs at asset management firms should be strategic and tactical (see also Cruz and Pinedo 2009). Nordgard and Falkenberg (2010) give an IT perspective on costs in asset management. Campbell and Frei (2010a) examine cost structure patterns in the asset management industry. Amihud and Mendelson (2010) examine the effect of transaction costs on asset management, and study their implications for portfolio construction, fund design, trade implementation, cash and liquidity management, and customer acquisition and development strategies.

3.3. Performance Analysis Through Data Envelopment Analysis (DEA)

There are numerous studies on performance and productivity analyses of retail banking that are based on DEA. DEA is a technique for evaluating productivity measures that can be applied to service industries in general. It compares productivity measures of different entities (e.g., bank branches) within the same service organization (e.g., a large retail bank) to one another. Such a comparative analysis then boils down to the formulation of a fractional linear program. DEA has been used in many retail banks to compare productivity measures of the various branches with one another. Sherman and Gold (1985), Sherman and Ladino (1995), and Seiford and Zhu (1999) performed such studies for US banks; Oral and Yolalan (1990) performed such a study for a bank in Turkey; Vassiloglou and Giokas (1990), Soteriou and Zenios (1999a), Zenios et al. (1999), Soteriou and Zenios (1999b), and Athanasopoulos and Giokas (2000) for Greek banks; Kantor and Maital (1999) for a large Mideast bank; and Berger and Humphrey (1997) for various international financial services firms. These papers discuss operational efficiency, profitability, quality, stock market performance, and the development of better cost estimates for banking products via DEA. Cummins et al. (1999) use DEA to explore the impact of organizational form on firm performance. They compare mutual and stock property liability companies and find that in using managerial discretion and cost-efficiency stock companies perform better, and in lines of insurance with long payouts mutual companies perform better.

Cook and Seiford (2009) present an excellent overview of the DEA developments over the past 30 years and Cooper et al. (2007) provide a comprehensive textbook on the subject. For a good survey and cautionary notes on the pitfalls of improper interpretation and use of DEA results (e.g., loosely using the results for evaluative purposes when uncontrollable variables exist), see Metters et al. (1999). Zhu (2003) discusses methods to solve imprecise DEA (IDEA), where data on inputs and outputs are either bounded, ordinal, or ratio bounded, where the original linear programming DEA formulation can no longer be used. Koetter (2006) discusses the stochastic frontier analysis (SFA) as another bank efficiency analysis framework, which contrasts to the deterministic DEA.
4. Forecasting

Forecasting is very important in many areas of the financial services industry. In its most familiar form in which it presents itself to customers and the general public, it consists of economic and market forecasts developed by research and strategy groups in brokerage and investment management firms. However, the types of forecasting we discuss tend to be more internal to the firms and not visible from the outside.

4.1. Forecasting in the Management of Cash Deposits and Credit Lines

Deposit-taking institutions (e.g., commercial banks, savings and loan associations, and credit unions) are interested in forecasting the future growth of their deposits. They use this information in the process of determining the value and pricing of their deposit products (e.g., checking, savings, and money market accounts, and also CDs), for asset-liability management, and for capacity considerations. Of special interest to these institutions are demand deposits, more broadly defined as non-maturity deposits. Demand deposits have no stated maturity and the depositor can add to the balance without restriction, or withdraw from "on demand," i.e., without warning or penalty. In contrast, time deposits, also known as CDs, have a set maturity and an amount established at inception, with penalties for early withdrawals.

Forecasting techniques have been applied to demand deposits because of their relative non-stickiness due to the absence of contractual penalties. A product with similar non-stickiness is credit card loans. Jarrow and Van Deventer’s (1998) model for valuing demand deposits and credit card loans using an arbitrage-free methodology assumes that demand deposit balances depend only on the future evolution of interest rates; however, it does allow for more complexity, such as macroeconomic variables (income or unemployment), and local market or firm-specific idiosyncratic factors. Janosi et al. (1999) use a commercial bank’s demand deposit data and aggregate data for negotiable order of withdrawal (NOW) accounts from the Federal Reserve to empirically investigate Jarrow and Van Deventer’s model. They find demand deposit balances to be strongly autoregressive, i.e., future balances are highly correlated with past balances. They develop regression models, linear in the logarithm of balances, in which past balances, interest rates, and a time trend are predictive variables. O’Brien (2000) adds income to the set of predictive variables in the regression models. Sheehan (2004) adds month-of-the-year dummy variables in the regressions to account for calendar-specific inflows (e.g., bonuses or tax refunds), or outflows (e.g., tax payments). He focuses on core deposits, i.e., checking accounts and savings accounts; distinguishes between the behavior of total and retained deposits; and develops models for different deposit types, i.e., business and personal checking, NOW, savings, and money market account deposits.

Labe and Papadakis (2010) discuss a propensity score matching model that can be used to forecast the likelihood of Bank of America’s retail clients bringing in new funds to the firm by subscribing to promotional offerings of CDs. Such promotional CDs carry an above-market premium rate for a limited period of time. Humphrey et al. (2000) forecast the adoption of electronic payments in the United States; they find that one of the reasons for the slow pace of moving from checks to electronic payments in the United States is the customers’ perceived loss of float. Many electronic payment systems now address this, by allowing for payment at the due date rather than immediately.

Revolving credit lines, or facilities, give borrowers access to cash on demand for short-term funding needs up to credit limits established at facility inception. Banks typically offer these facilities to corporations with investment grade credit ratings, which have access to cheaper sources of short-term funding, for example, commercial paper, and do not draw significant amounts from them except:

(i) for very brief periods of time under normal conditions,
(ii) when severe deterioration of their financial condition causes them to lose access to the credit markets, and
(iii) during system-wide credit market dysfunction, such as during the crisis of 2007–2009.

Banks that offer these credit facilities must set aside adequate, but not excessive, funds to satisfy the demand for cash by facility borrowers. Duffy et al. (2005) describe a Monte Carlo simulation model that Merrill Lynch Bank used to forecast these demands for cash by borrowers of their revolver portfolio. The model uses industry data for revolver usage by borrower credit rating, and assumes Markovian credit rating migrations, correlated within and across industries. Migration probabilities were provided by a major rating agency, and correlation estimates were calculated by Merrill Lynch’s risk group. The model was used by Merrill Lynch Bank to help manage liquidity risk in its multi-billion portfolio of revolving credit lines.

Forecasting the future behavior and profitability of retail borrowers (e.g., for credit card loans, mortgages, and home equity lines of credit) has become a key component of the credit management process. Forecasting involved in a decision to grant credit to a new borrower is known as “credit scoring,” and its origins in the modern era can be found in the 1950s. A discussion of credit scoring models, including related public policy issues, is offered by Capon (1982). Fore-
casting involved in the decisions to adjust credit access and marketing effort to existing borrowers is known as “behavioral scoring.” The book by Thomas et al. (2002) contains a comprehensive review of the objectives, methods, and practical implementation of credit and behavioral scoring. The formal statistical methods used for classifying credit applicants into “good” and “bad” risk classes is known as “classification scoring.” Hand and Henley (1997) review a significant part of the large body of literature in classification scoring. Baesens et al. (2003) examine the performance of standard classification algorithms, including logistic regression, discriminant analysis, k-nearest neighbor, neural networks, and decision trees; they also review more recently proposed ones, such as support vector machines and least-squares support vector machines (LS-SVM). They find LS-SVM, and neural network classifiers, and simpler methods such as logistic regression and linear discriminant analysis to have good predictive power. In addition to classification scoring, other methods include:

(i) “response scoring,” which aims to forecast a prospect’s likelihood to respond to an offer for credit, and
(ii) “balance scoring,” which forecasts the prospect’s likelihood of carrying a balance if they respond.

To improve the chances of acquiring and maintaining profitable customers, offers for credit should be mailed only to prospects with high credit, response, and balance scores. Response and balance scoring models are typically proprietary. Trench et al. (2003) discuss a model for optimally managing the size and pricing of card lines of credit at Bank One. The model uses account-level historical transaction information to select for each cardholder, through Markov decision processes, annual percentage rates, and credit lines that optimize the net present value of the bank’s credit portfolio.

4.2. Forecasting in Securities Brokerage, Clearing, and Execution

In the last few decades, the securities brokerage industry has seen dramatic change. Traditional wirehouses charging fixed commissions evolved or were replaced by diverse organizations offering full service, discount, and online trading channels, as well as research and investment advisory services. This evolution has introduced a variety of channel choices for retail and institutional investors. Pricing, service mix and quality, and human relationships are key determinants in the channel choice decision. Firms are interested in forecasting channel choice decisions by clients, because they greatly impact capacity planning, revenue, and profitability. Altschuler et al. (2002) discuss simulation models developed for Merrill Lynch’s retail brokerage to forecast client choice decisions on introduction of lower-cost offerings to complement the firm’s traditional full-service channel. Client choice decision forecasts were used as inputs in the process of determining the proper pricing for these new offerings and for evaluating their potential impact on firm revenue. The results of a rational economic behavior (REB) model were used as a baseline. The REB model assumes that investors optimize their value received by always choosing the lowest-cost option (determined by an embedded optimization model that was solved for each of millions of clients and their actual portfolio holdings). The REB model’s results were compared with those of a Monte Carlo simulation model. The Monte Carlo simulation allows for more realistic assumptions. For example, clients’ decisions are impacted not only by price differentials across channels, but also by the strength and quality of the relationship with their financial advisor, who represented the higher-cost options.

Labe (1994) describes an application of forecasting the likelihood of affluent prospects becoming Merrill Lynch’s priority brokerage and investment advisory clients (defined as clients with more than US$250,000 in assets). Merrill Lynch used discriminant analysis, a method akin to classification scoring, to select high quality households to target in its prospecting efforts.

The trading of securities in capital markets involves key operational functions that include:

(i) clearing, i.e., establishing mutual obligations of counterparties in securities and/or cash trades, as well as guarantees of payments and deliveries, and
(ii) settlement, i.e., transfer of titles and/or cash to the accounts of counterparties in order to finalize transactions.

Most major markets have centralized clearing facilities so that counterparties do not have to settle bilaterally and assume credit risk to each other. The central clearing organization must have robust procedures to satisfy obligations to counterparties, i.e., minimize the number of trades for which delivery of securities is missed. It must also hold adequate, but not excessive, amounts of cash to meet payments. Forecasting the number and value of trades during a clearing and settlement cycle can help the organization meet the above objectives; it can achieve this by modeling the clearing and settlement operation using stochastic simulation. A different approach is used by de Lascurain et al. (2011): they develop a linear programming method to model the clearing and settlement operation of the Central Securities Depository of Mexico and evaluate the system’s performance through deterministic simulation. The model’s formulation in de Lascurain et al. (2011) is a relaxation of a mixed integer
programming (MIP) formulation proposed by Güntzer et al. (1998), who show that the bank clearing problem is NP-complete. Eisenberg and Noe (2001) include clearing and settlement in a systemic financial risk framework.

4.3. Forecasting of Call Arrivals at Call Centers
Forecasting techniques are also used in various other areas within the financial services sector; for example, the forecasting of arrivals in call centers, which is a crucial input in the personnel scheduling process in call centers (to be discussed in a later section). To set this in a broader context, refer to the framework of Thompson (1998) for forecasting demand for services. Recent papers that focus on forecasting call center workload include a tutorial by Gans et al. (2003), a survey by Aksin et al. (2007), and a research paper by Aldor-Noiman et al. (2009).

The quality of historical data improves with the passage of time because call centers become increasingly more sophisticated in capturing data with every nuance that a modeler may find useful or interesting. Andrews and Cunningham (1995) describe the autoregressive integrated moving average (ARIMA) forecasting models used at L.L. Bean’s call centers. Time series data used to fit the models exhibit seasonality patterns and are also influenced by variables such as holiday and advertising interventions. Advertising and special calendar effects are addressed by Antipov and Meade (2002). More recently, Soyer and Tarimcilı (2008) incorporate advertising effects by modeling call arrivals as a modulated Poisson process with arrival rates being driven by customer calls that are stimulated by advertising campaigns. They use a Bayesian modeling framework and a data set from a call center that enables tracing calls back to specific advertisements. In a study of FedEx’s call centers, Xu (2000) presents forecasting methodologies used at multiple levels of the business decision-making hierarchy, i.e., strategic, business plan, tactical, and operational, and discusses the issues that each methodology addresses. Methods used include exponential smoothing, ARIMA models, linear regression, and time series decomposition.

At low granularity, call arrival data may have too much noise. Mandelbaum et al. (2001) demonstrate how to remove relatively unpredictable short-term variability from data and keep only predictable variability. They achieve this by aggregating data at higher levels of granularity, i.e., progressively moving up from minute of the hour to hour of the day to day of the month and to month of the year. The elegant textbook assumption that call arrivals follow a Poisson process with a fixed rate that is known or can be estimated does not hold in practice. Steckley et al. (2009) show that forecast errors can be large in comparison to the variability expected in a Poisson process and can have significant impact on the predictions of long-run performance; ignoring forecast errors typically leads to overestimation of performance. Jongbloed and Koole (2001) found that the call arrival data they had been analyzing had a variance much greater than the mean, and therefore did not appear to be samples of Poisson distributed random variables. They addressed this “overdispersion” by proposing a Poisson mixture model, i.e., a Poisson model with an arrival rate that is not fixed but random following a certain stochastic process. Brown et al. (2005) found data from a different call center that also followed a Poisson distribution with a variable arrival rate; the arrival rates were also serially correlated from day to day. The prediction model proposed includes the previous day’s call volume as an autoregressive term. High intra-day correlations were found by Avramidis et al. (2004), who developed models in which the call arrival rate is a random variable correlated across time intervals of the same day. Steckley et al. (2004) and Mehratra et al. (2010) examine the correlation of call volumes at later periods of a day to call volumes experienced earlier in the day for the purpose of updating workload schedules.

Methods to approximate non-homogeneous Poisson processes often attempt to estimate the arrival rate by breaking up the data set into smaller intervals. Henderson (2003) demonstrates how a heuristic that assumes a piecewise constant arrival rate over time intervals with a length that shrinks as the volume of data grows produces good arrival rate function estimates. Massey et al. (1996) fit piecewise linear rate functions to approximate a general time-inhomogeneous Poisson process. Weinberg et al. (2007) forecast an inhomogeneous Poisson process using a Bayesian framework, whereby from a set of prior distributions they estimate the parameters of the posterior distribution through a Monte Carlo Markov Chain model. They forecast arrival rates in short intervals of 15–60 minutes of a day of the week as the product of a day’s forecast volume times the proportion of calls arriving during an interval; they also allow for a random error term.

Shen and Huang (2005, 2008a, b) developed models for inter-day forecasting and intra-day updating of call center arrivals using singular value decomposition. Their approach resulted in a significant dimensionality reduction. In a recent empirical study, Taylor (2008) compared the performance of several univariate time series methods for forecasting intra-day call arrivals. Methods tested included seasonal autoregressive and exponential smoothing models, and the dynamic harmonic regression of Tych et al. (2002). Results indicate that different methods perform best under different lead times and call volumes levels.
Foretelling other aspects of a call center with a significant potential for future research include, for example, waiting times of calls in queues, see Whitt (1999a, b) and Armony and Maglaras (2004).

5. Inventory and Cash Management

5.1. Cash Inventory Management Under Deterministic and Stochastic Demand

Organizations, households, and individuals need cash to meet their liquidity needs. In the era of checks and electronic transactions, an amount of cash does not have to be in physical currency, but may correspond only to a value in an account that has been set up for this purpose. To meet short-term liquidity needs, cash must be held in a riskless form, where its value does not fluctuate and is available on demand, but earns little or no interest. Treasury bills and checking accounts are considered riskless. Cash not needed to meet short-term liquidity needs can be invested in risky assets whereby it may earn higher returns, but its value may be subject to significant fluctuations and uncertainty, and could become wholly or partially unrecoverable. Depending on the type of risky asset, its value may or may not be quickly recoverable and realizable at a modest cost (as with a public equity that is listed in a major stock exchange). Determining the value of certain types of risky assets (e.g., private equity, real estate, some hedge funds, and asset-backed fixed income securities) may require specialized valuation services, which could involve significant time and cost. Risky assets can also be subject to default, in which case all or part of the value becomes permanently unrecoverable.

Researchers have produced over the last few decades a significant body of work by applying the principles of inventory theory to cash management. We review the cash management literature from its beginnings so we can put later work in context, and we have not identified an earlier comprehensive review that accomplishes this purpose. Whistler (1967) discussed a stochastic inventory model for rented equipment that was formulated as a dynamic program; this work served as a model for the cash management problem. One of the early works produced an elegant result that became known as the Baumol–Tobin economic model of the transactions demand for money, independently developed by Baumol (1952) and Tobin (1956). The model assumes a deterministic, constant rate of demand for cash. It calculates the optimal “lot sizes” of the risky asset to be converted to cash, or the optimal numbers of such conversions, in the presence of transaction and interest costs. Tobin’s version requires an integer number of transactions and therefore approximates reality more closely than Baumol’s, which allows that variable to be continuous.

The concept of transactions demand for money, addressed by the Baumol–Tobin model, is related to, but subtly different from, precautionary demand for cash that applies to unforeseen expenditures, opportunities for advantageous purchases, and uncertainty in receipts. Whalen (1966) developed a model with a structure strikingly similar to the Baumol–Tobin model, capturing the stochastic nature of precautionary demand for cash. Sprenkle (1969) and Akerlof and Milbourne (1978) observed that the Baumol–Tobin model tends to under-predict demand for money, partly because it fails to capture the stochastic nature of precautionary demand for cash. Sprenkle’s paper elicited a response by Orr (1974), which in turn prompted a counter-response by Sprenkle (1977).

Robichek et al. (1965) propose a deterministic short-term financing model that incorporates a great degree of realistic detail involved in the financial officer’s decision-making process, which they formulate and solve as a linear program. They include a discussion on model extensions for solving the financing problem under uncertainty. Sethi and Thompson (1970) proposed models based on mathematical control theory, in which demand for cash is deterministic but does vary with time. In an extension of the Sethi–Thompson model, Bensoussan et al. (2009) allow the demand for cash to be satisfied by dividends and uncertain capital gains of the risky asset, stock.

In what became known as the Miller–Orr model for cash management, Miller and Orr (1966) extended the Baumol–Tobin model by assuming the demand for cash to be stochastic. The cash balance can fluctuate randomly between a lower and an upper bound according to a Bernoulli process, and transactions take place when it starts moving out of this range; units of the risky asset are converted into cash at the lower bound, and bought with the excess cash at the upper bound. Transaction costs were assumed fixed, i.e., independent of transaction size. In a critique of the Miller–Orr model, Weitzman (1968) finds it to be “robust,” i.e., general results do not change much when the underlying assumptions are modified.

Eppen and Fama (1968, 1969, 1971) proposed cash balance models that are embedded in a Markovian framework. In one of their papers, Eppen and Fama (1969) presented a stochastic model, formulated as a dynamic program, with transaction costs proportional to transaction sizes. Changes in the cash balance can follow any discrete and bounded probability distribution. In another one of their papers, Eppen and Fama (1968) developed a general stochastic model that allowed costs to have a fixed as well as a variable component. They showed how to find optimal policies for the infinite-horizon problem using linear programming. In their third paper, Eppen and Fama (1971) proposed a stochastic model with two risky assets, namely “bonds”
and “stock”; the stock is more risky but has a higher expected return. They also discussed using “bonds” (the intermediate-risk asset) as a “buffer” between cash and the more risky asset. Taking a similar approach, Daellenbach (1971) proposed a stochastic cash balance model using two sources of short-term funds. Girgis (1968) and Neave (1970) presented models with both fixed and proportional costs and examined conditions for policies to be optimal under different assumptions. Hausman and Sanchez-Bell (1975) and Vickson (1985) developed models for firms facing a compensating-balance requirement specified as an average balance over a number of days.

Continuous-time formulations of the cash management problem were based on the works of Antelman and Savage (1965), and Bather (1966), who used a Wiener process to generate a stochastic demand in their inventory problem formulations. Their approach was extended to cash management by Vial (1972), whose continuous-time formulation had fixed and proportional transaction costs, and linear holding and penalty costs, and determined the form of the optimal policy (assuming one exists). Constantinides (1976) extended the model by allowing positive and negative cash balances, determined the parameters of the optimal policy, and discussed properties of the optimal solution. Constantinides and Richard (1978) formulated a continuous-time, infinite-horizon, discounted-cost cash management model with fixed and proportional transaction costs, linear holding and penalty costs, and the Wiener process as the demand-generating mechanism. They proved that there always exists an optimal policy for the cash management problem, and that this policy is of a simple form. Smith (1989) developed a continuous-time model with a stochastic, time-varying interest rate.

5.2. Supply Chain Management of Physical Currency

Physical cash, i.e., paper currency and coins, remains an important component of the transactions volume even in economies that have experienced a significant growth in checks, credit, debit and smart cards, and electronic transactions. Advantages of cash include ease of use, anonymity, and finality; it does not require a bank account; it protects privacy by leaving no transaction records; and it eliminates the need to receive statements and pay bills. Disadvantages of cash include ease of tax evasion, support of an “underground” economy, risk of loss through theft or damage, ability to counterfeit, and unsuitability for online transactions.

Central banks provide cash to depository institutions, which in turn circulate it in the economy. There are studies on paper currency circulation in various countries. For example, Fase (1981) and Boeschoten and Fase (1992) present studies by the Dutch central bank on the demand for banknotes in the Netherlands before the introduction of the Euro. Ladany (1997) developed a discrete dynamic programming model to determine optimal (minimum cost) ordering policies for banknotes for Israel’s central bank. Massoud (2005) presents a dynamic cost minimizing note inventory model to determine optimal banknote order size and frequency for a typical central bank.

The production and distribution of banknotes, and the required infrastructure and processes, have also been studied. Fase et al. (1979) discuss a numerical planning model for the banknotes operations at a central bank, with examples from pre-Euro Netherlands. Bauer et al. (2000) develop optimization models for determining the least-cost configuration of the US Federal Reserve’s currency processing sites given the trade-off between economies of scale in processing and transportation costs. In a study of costs and economies of scale of the US Federal Reserve’s currency operations, Bohn et al. (2001) find that the Federal Reserve is not a natural monopoly. Opening currency operations to market competition and charging fees and penalties for some services provided for free by the Federal Reserve at that time could lead to more efficient allocation of resources.

The movement of physical cash among central banks, depository institutions, and the public must be studied as a closed-loop supply chain (see, e.g., Dekker et al. 2004). It involves the recirculation of used notes back into the system (reverse logistics), together with a flow of new notes from the central bank to the public through depository institutions (forward logistics). The two movements are so intertwined that they cannot be decoupled. Rajamani et al. (2006) study the cash supply chain structure in the United States, analyze it as a closed-loop supply chain, and describe the cash flow management system used by US banks. They also discuss the new cash recirculation policies adopted by the Federal Reserve to discourage banks’ overuse of its cash processing services, and encourage increased recirculation at the depository institution level. Among the practices to be discouraged was “cross shipping,” i.e., shipping used currency to the Federal Reserve and ordering it in the same denominations in the same week. To compare and contrast new and old Federal Reserve policies for currency recirculation, Geismar et al. (2007) introduce models that explain the flow of currency between the Federal Reserve and banks under both sets of guidelines. They present a detailed analysis that provides optimal policies for managing the flow of currency between banks and the Federal Reserve, and analyze banks’ responses to the new guidelines to help the Federal Reserve understand their implications. Dawande et al. (2010) examine the conditions that
can induce depository institutions to respond in socially optimal ways according to the new Federal Reserve guidelines. Mehrotra et al. (2010), which is a paper in this special issue of Production and Operations Management, address the problem of obtaining efficient cash management operating policies for depository institutions under the new Federal Reserve guidelines. The mixed-integer programming model developed for this purpose seeks to find “good” operating policies, if such exist, to quantify the monetary impact on a depository institution operating according to the new guidelines. Another objective was to analyze to what extent the new guidelines can discourage cross shipping and stimulate currency recirculation at the depository institution level. Mehrotra et al. (2010a) study pricing and logistics schemes for services such as fit-sorting and transportation that can be offered by third-party providers as a result of the Federal Reserve’s new policies.

5.3. Other Cash Management Applications in Banking and Securities Brokerage

US banks are required to keep on reserve a minimum percentage (currently 10%) of deposits in client transaction accounts (demand deposits and other checkable deposits) at the Federal Reserve. Until very recently, banks had a strong incentive to keep funds on reserve at a minimum, because these funds were earning no interest. Even after the 2006 Financial Services Regulatory Relief Act became law, authorizing payment of interest on reserves held at the Federal Reserve, banks prefer to have funds available for their own use rather than have them locked up on reserve. Money market deposit accounts (MMDA) with checking allow banks to reduce the amounts on reserve at the Federal Reserve by keeping deposits in MMDA accounts and transferring to a companion checking account only the amounts needed for transactions. Only up to six transfers (“sweeps”) per month are allowed from an MMDA to a checking account, and a client’s number and amount of transactions in the days remaining in a month is unknown. Therefore, the size and timing of the first five sweeps must be carefully calculated to avoid a sixth sweep, which will move the remaining MMDA balance into the checking account. Banks have been using heuristic algorithms to plan the first five sweeps. This specialized inventory problem has been examined by Nair and Anderson (2008), who propose a stochastic dynamic programming model to optimize retail account sweeps. The stochastic dynamic programming model developed by Nair and Hatzakis (2010) introduces cushions added to the minimum sweep amounts. It determines the optimal cushion sizes to ensure that sufficient funds are available in the transaction account in order to cover potential future transactions and avoid the need for a sixth sweep.

The impact of the sequence of transaction postings on account balances and resultant fees for insufficient funds, similar to the cost of stock-outs in inventory management, has been studied by Apte et al. (2004). They investigate how overdraft fees and non-sufficient funds (NSF) fees interact in such situations.

Brokerage houses make loans to investors who want to use leverage, i.e., to invest funds in excess of their own capital in risky assets, and can pledge securities that they own as collateral. In a simple application of this practice, known as margin lending, the brokerage extends a margin loan to a client of up to the value of equity securities held in the client’s portfolio. The client can use the loaned funds to buy more equity securities. Calculating the minimum value required in a client’s account for a margin loan can become complex in accounts holding different types of securities including equities, bonds, and derivatives, all with different margin requirements. The complexity increases even more with the presence of long and short positions and various derivative strategies practiced by clients. Rudd and Schroeder (1982) presented a simple transportation model formulation for calculating the minimum margin, which represented an improvement over the heuristics used in practice. A significant body of subsequent work has been published on this problem, especially by Timkovsky and collaborators, which is more related to portfolio strategies and hedging. We believe that the approach in the paper by Fiterman and Timkovsky (2001), which is based on 0–1 knapsack formulations, is methodologically the most relevant to mention in this overview.

6. Waiting Line Management in Retail Banks and in Call Centers

6.1. Queueing Environments and Modeling Assumptions

In financial services, in particular in retail banking, retail brokerage, and retail asset management (pension funds, etc.), queueing is a common phenomenon that has been analyzed thoroughly. Queueing occurs in the branches of retail banks with the tellers being the servers, at banks of ATM machines with the machines being the servers, and in call centers, where the operators and/or the automated voice response units are the servers. These diverse queueing environments turn out to be fairly different from one another, in particular with regard to the following characteristics:

(i) the information that is available to the customer and the information that is available to the service system,
(ii) the flexibility of the service system with regard to adjustments in the number of servers dependent on the demand,
(iii) the order of magnitude of the number of servers.

Even though in the academic literature the arrival processes in queueing systems are usually assumed to be stationary (time-homogeneous) Poisson processes, arrival processes in practice are more appropriately modeled as non-homogeneous Poisson processes. Over the last couple of decades, some research has been done on queues that are subject to non-homogeneous Poisson inputs (see, e.g., Massey et al. 1996). The more theoretical research in queueing has also focused on various aspects of customer behavior in queue, in particular abandonment, balking, and reneging. For example, Zohar et al. (2002) have modeled the adaptive behavior of impatient customers and Whitt (2006) developed fluid models for many-server queues with abandonment.

In all three queueing environments described above, the psychology of the customers in the queue also plays a major role. A significant amount of research has been done on this topic, see Larson (1987), Katz et al. (1991), and Bitran et al. (2008). As it turns out, reducing wait times may not always be the best approach in all service encounters. For example, in restaurants and salons, longer service time may be perceived as better service. In many cases, customers do not like waiting, but when it comes their turn to be served, would like the service to take longer. In still others, for example, in grocery checkout lines, customers want a businesslike pace for both waiting and service. The latter category, which we may call dispassionate services, are more common in financial service situations, though the former, which we may call hedonic services, are also present—for example, when a customer visits their mortgage broker or insurance agent, they would not like to be rushed. In the following subsections, we consider the various different queueing environments in more detail.

6.2. Waiting Lines in Retail Bank Branches and at ATMs

The more traditional queues in financial services are those in bank branches feeding the tellers. Such a queue is typically a single line with a number of servers in parallel. There are clearly no priorities in such a queue and the discipline is just first come first served. Such a queueing system is typically modeled as an M/M/s system and is discussed in many standard queueing texts. One important aspect of this type of queueing in a branch is that management usually can adjust the number of available tellers fairly easily as a function of customer demand and time of day. (This gives rise to many personnel scheduling issues that will be discussed in the next section.)

In the early 1980s retail banks began to make extensive use of ATMs. The ATMs at a branch of a bank behave quite differently from the human tellers. In contrast to a teller environment, the number of ATMs at a branch is fixed and cannot be adjusted as a function of customer demand. However, the teller environment and the ATM environment do have some similarities. In both environments, a customer can observe the length of the queue and can, therefore, estimate the amount of time (s)he has to wait. In neither the teller environment, nor the ATM environment, can the bank adopt a priority system that would ensure that more valuable customers have a shorter wait. Kolesar (1984) did an early analysis of a branch with two ATM machines and collected service time data as well as arrival time data. However, it became clear very quickly that a bank of ATMs is capable of collecting some very specific data automatically (e.g., customer service times and machine idle times), but cannot keep track of certain other data (e.g., queue lengths, customer waiting times). Larson (1990), therefore, developed the so-called queue inference engine, which basically provides a procedure for estimating the expected waiting times of customers, given the service times recorded at the ATMs as well as the machine idle times.

6.3. Waiting Lines in Call Centers

Since the late 1980s, banks have started to invest heavily in call center technologies. All major retail banks now operate large call centers on a 24/7 basis. Call centers have therefore been the subject of extensive research studies, see the survey papers by Pinedo et al. (2000), Gans et al. (2003), and Aksin et al. (2007). The queueing system in a call center is actually quite different from the queueing systems in a teller environment or in an ATM environment; there are a number of major differences. First, a customer now has no direct information with regard to the queue length and cannot estimate his waiting time; he must rely entirely on the information the service system provides him. On the other hand, the service organization in a call center has detailed knowledge concerning the customers who are waiting in queue. The institution knows of each customer his or her identity and how valuable (s)he is to the bank. The bank can now put the customers in separate virtual queues with different priority levels. This new capability has made the application of priority queueing systems suddenly very important; well-known results in queueing theory have now suddenly become more applicable, see Kleinrock (1976). Second, the call centers are in another aspect quite different from the teller and the ATM environments. The number of servers in either a teller environment or an ATM environment may typically be, say, at most 20,
whereas the number of operators in a call center may typically be in the hundreds or even in the thousands. In the analysis of call center queues, it is now possible to apply limit theorems with respect to the number of operators, see Halfin and Whitt (1981) and Reed (2009). Third, the banks have detailed information with regard to the skills of each one of its operators in a call center (language skills, product knowledge, etc.). This enables the bank to apply skills-based-routing to the calls that are coming in, see Gans and Zhou (2003) and Mehrotra et al. (2009).

In call centers, typically, an enormous amount of statistical data is available that is collected automatically, see Mandelbaum et al. (2001) and Brown et al. (2005). The data that are collected automatically are much more extensive than the data that are collected in an ATM environment. It includes the waiting times of the customers, the queue length at each point in time, the proportion of customers that experience no wait at all, and so on.

Lately, many other aspects of queueing in call centers have become the subject of research. This special issue of Production and Operations Management as well as another recent special issue contain several such papers. For example, Örncel and Aksin (2010) focus on the effects of cross selling on the management of the queues in call centers. They focus on the operational implications of cross selling in terms of capacity usage and value generation. Chevalier and Van den Schriek (2008), and Barth et al. (2010) consider hierarchical call centers that consist of multiple levels (stages) with a time-dependent overflow from one level to the next. For example, at the first stage, the front office, the customers receive the basic services; a fraction of the served customers requires more specialized services that are provided by the back office. Van Dijk and Van der Sluis (2008) and Meester et al. (2010) consider the service network configuration problem. Meester et al. (2010) analyze networks of geographically dispersed call centers that vary in service and revenue-generation capabilities, as well as in costs. Optimally configuring a service network in this context requires managers to balance the competing considerations of costs (including applicable discounts) and anticipated revenues. Given the large scale of call center operations in financial services firms, the service network configuration problem is important and economically significant. The approach by Meester and colleagues integrates decision problems involving call distribution, staffing, and scheduling in a hierarchical manner (previously, these decision problems were addressed separately).

Even though most call center research has focused on inbound call centers, a limited amount of research has also been done on outbound call centers. Rising delinquencies and the importance of telemarketing has increased the need for outbound calling from call centers. Outbound calling is quite different from inbound calling, because the scheduling of calls is done by the call center rather than the call center being at the mercy of its customers. This presents unique challenges and opportunities. Bollapragada and Nair (2010) focus in this environment on improvements in contact rates of appropriate parties.

7. Personnel Scheduling in Retail Banks and in Call Centers

7.1. Preliminaries and General Research Directions

An enormous amount of work has been done on workforce (shift) scheduling in manufacturing. However, workforce scheduling in manufacturing is quite different from workforce scheduling in service industries. The workforce scheduling process in manufacturing has to adapt itself to inventory considerations and is typically a fairly regular and stable process. In contrast to manufacturing industries, workforce scheduling in the service industries has to adapt itself to a fluctuating customer demand, which in practice is often based on non-homogeneous Poisson customer arrival processes. In practice, adapting the number of tellers or operators to the demand process can be done through an internal pool of flexible workers, or through a partnership with a labor supply agency (see Larson and Pinker 2000).

As the assignment of tellers and the hiring of operators depend so strongly on anticipated customer demand, a significant amount of research has focused on probabilistic modeling of arrival processes, on statistical analyses of arrival processes, and on customer demand forecasting in order to accomplish a proper staffing. For probabilistic modeling of customer arrival processes, see Stolletz (2003), Ridley et al. (2004), Avramidis et al. (2004), and Jongbloed and Koole (2001). For statistical analyses of customer arrival data, see Brown et al. (2005) and Robbins et al. (2006). For customer demand forecasting, see Weinberg et al. (2007) and Shen and Huang (2008a, b).

From a research perspective, the personnel scheduling problem has been tackled via a number of different approaches, namely, simulation, stochastic modeling, optimization modeling, and artificial intelligence. The application areas considered included the scheduling of bank tellers as well as the scheduling of the operators in call centers. Slepicka and Sporer (1981), Hammond and Mahesh (1995), and Mehrotra and Fama (2003) used simulation to schedule bank tellers and call center operators. Thompson (1993) studied the impact of having multiple periods with different demands on determining the employee requirements in each segment of the schedule; see also Chen and Henderson (2001). Green et al. (2007) and
Feldman et al. (2008) have addressed the problem from a stochastic point of view and have developed staffing rules based on queueing theory. So et al. (2003) use better staff scheduling and team reconfiguration to improve check processing in the Federal Reserve Bank.

### 7.2. Optimization Models for Call Center Staffing

Many researchers have considered this problem from an optimization point of view. An optimization model to determine the optimal staffing levels and call routing can take a more aggregate form or a more detailed form. Bassamboo et al. (2005) consider a model with $m$ customer classes and $r$ agent pools. They develop a linear program-based method to obtain optimal staffing levels as well as call routing. Not surprisingly, a significant amount of work has also been done on the more detailed optimization of the various possible shift structures (including even lunch breaks and coffee breaks). In large call centers, there are many different types of shifts that can be scheduled. A shift is characterized by the days of the week, the starting times at the beginning of the day, the ending time at the end of the day, as well as the timings of the lunch breaks and coffee breaks. Gawande (1996) (see also Pinedo 2009) solved this staffing problem in two stages. In the first stage of the approach, the so-called solid-tour scheduling problem is solved. (A solid tour represents a shift without any breaks; a solid tour is only characterized by the starting time at the beginning of the day and by the ending time at the end of the day.) There are scores of different solid tours available (each one with a different starting time and ending time). The demand process (the non-homogeneous arrival process) usually can be forecast with a reasonable amount of accuracy. The first stage of the problem is to find the number of personnel to hire in each solid tour such that the total number of people available in each time interval fits the demand curve as accurately as possible. This problem leads to an integer programming formulation of a special structure that actually can be solved in polynomial time. After it has been determined in the first stage what the actual number of people in each solid tour is, the placement of the breaks is done in the second stage. Typically, the break placement problem is a very hard problem, and is usually dealt with through a heuristic.

Gans et al. (2009) applied parametric stochastic programming to workforce scheduling in call centers. Gurvich et al. (2010) have developed a chance constrained optimization approach to deal with uncertain demand forecasts. Brazier et al. (1999) applied an artificial intelligence technique, namely co-operative multi-agent scheduling, on the workforce scheduling in call centers; their work was done in collaboration with the Rabobank in the Netherlands. Harrison and Zeevi (2005) developed a staffing method for large call centers based on stochastic fluid models.

Nowadays another aspect of personnel management in call centers has become important, namely call routing. As the clients may have many different demands and the operators may have many different skill sets, call routing has gained a significant amount of interest, see Mehrotra et al. (2009) and Bhulai et al. (2008). Optimal call routing typically has to be considered in conjunction with the optimal cross training of the operators, see Robbins et al. (2007).

### 7.3. Miscellaneous Issues Concerning Workforce Scheduling in Financial Services

Very little research has been done on workforce scheduling in other segments of the financial services industry. It is clear that workforce scheduling is also important in trading departments or at trading desks (equity trading, foreign exchange trading, or currency trading). Such departments have to keep track of information on trades, such as the number of trades done per day for each trader, amounts per trade and total, and trades’ impact with regard to risk limits of each trader or the entire desk. Automated systems exist for recording such information, but operational risk mitigation practices necessitate the involvement of skilled personnel in populating and reviewing of the data. In the case of a broker/dealer, relevant pieces of information may include the commission charged for each trade, whether the firm acted in a capacity of principal or agent, and trading venue (e.g., an exchange, a dark pool, or an electronic communication network). Institutional and retail brokerage and asset management firms need to maintain databases with account-level, client-level, and relationship-level information. Typically such information is taken from forms filled out by clients on hard copy or electronically, and/or from legal agreements signed by the client and the firm. Skilled personnel must perform the back-office tasks of reviewing documents, and populating the appropriate fields through user interfaces of databases. Given the regulatory implications of errors in this process (e.g., Sarbanes-Oxley 404 compliance), the personnel assigned to such tasks must have received a rigorous specialized training. A factor that may add complexity to workforce scheduling is the emerging trend of outsourcing such back-office tasks, especially to offshore locations in Asia or Europe. In offshore outsourcing, time-zone differences and high turnover of trained professionals can be serious issues to contend with.

As stated earlier, workforce scheduling and waiting line management are strongly interconnected. A larger and better trained workforce clearly results in a better queueing performance. However, workforce scheduling is also strongly connected to operational
risk. A larger and better trained workforce clearly results in a lower level of operational risk (especially in trading departments, where human errors can have a very significant impact on the financial performance of the institution). This topic will be discussed in more detail in the next section on operational risk.

8. Operational Risk Management

8.1. Types of Operational Failures
Operational risk in financial services started to receive attention from the banking community as well as from the academic community in the mid-1990s. Operational risk has since then typically been defined as the risk resulting from inadequate or failed internal processes, people, and systems, or from external events (Basel Committee 2003). It covers product life cycle and execution, product performance, information management and data security, business disruption, human resources and reputation (see, e.g., the General Electric Annual Report 2009, available at www.ge.com). Failures concerning internal processes can be due to transactions errors (some due to product design), or inadequate operational controls (lack of oversight). Information system failures can be caused by programming errors or computer crashes. Failures concerning people may be due to incompetence or inadequately trained employees, or due to fraud. Actually, since the mid-1990s a number of events have taken place in financial services that can be classified as rogue trading. This type of event has turned out to be quite serious because just one such an event can bring down a financial services firm, see Cruz (2002), Elsinger et al. (2006), Chernobai et al. (2007), Cheng et al. (2007), and Jarrow (2008).

8.2. Relationships between Operational Risk and Other Research Areas
It has become clear over the last decade that the management of operational risk in finance is very closely related to other research areas in operations management, operations research, and statistics. Most of the methodological research in operational risk has focused on probabilistic as well as statistical aspects of operational risk, for example, extreme value theory (EVT) (see Chernobai et al. 2007). The areas of importance include the following:

(i) Process design and process mapping,
(ii) Reliability theory,
(iii) TQM, and
(iv) EVT.

The area of process design and process mapping is very important for the management of operational risk. As discussed in earlier sections of this paper, Shostack (1984) focuses on process mapping in the service industries and, in particular, process mapping in financial services. Process mapping also includes an identification of all potential failure points. This area of study is closely related to the research area of reliability theory.

The area of reliability theory is very much concerned with system and process design issues, including concepts such as optimal redundancies. This area of research originated in the aviation industry; see the classic work by Barlow and Proschan (1975) on reliability theory. One major issue in financial services is the issue of determining the amount of backups and the amount of parallel processing and checking that have to be designed into the processes and procedures. Reliability theory has always had a major impact on process design.

Procedures that are common in TQM turn out to be very useful for the management of operational risk, for example, Six Sigma (see Cruz and Pinedo 2008). However, it has become clear that there are important similarities as well as differences between TQM in the manufacturing industries and operational risk management in financial services. One important difference is that TQM in manufacturing (typically referred to as Six Sigma) is based on statistical properties of the Normal (Gaussian distribution), because any parameter that is being measured may have an equal probability of having a deviation in one direction as in the other. However, in operational risk management the analysis of operational risk events focuses mainly on the outliers, i.e., the catastrophic events. For that reason, the distributions used are different from the Normal; they may be, for example, the Lognormal, or the Fat Tail Lognormal.

The statistical analysis often focuses on the occurrences of rare, catastrophic events, because financial services in general are quite concerned about being hit by such rare catastrophic events. It is for this reason that EVT has become such an important research direction, see De Haan and Ferreira (2006). EVT is based on a limit theorem that is due to Gnedenko; this limit theorem specifies the distribution of the maximum of a series of independent and identically distributed (i.i.d.) random variables. These extreme value distributions include the Weibull, Frechet, and Gumbel distributions. They typically have three parameters, namely one that specifies the mean, one that specifies the variance, and a third one that specifies the fatness of the tail. The need for being able to parameterize the fatness of the tail is based on the fact that the tail affects the probabilities of catastrophic events occurring.

8.3. Operational Risk in Specific Financial Sectors
There are several sectors of the financial services industry that have received a significant amount of
attention with regard to their exposure to operational risk, namely,

(i) Retail financial services (banking, brokerage, credit cards),
(ii) Trading (equities, bonds, foreign exchange),
(iii) Asset management (retail, institutional).

There are several types of operational risk events that are common in retail financial services. They include security breaches, fraud, and systems breakdowns. Security breaches in retail financial services often involve clients’ personal information such as bank account numbers and social security numbers. Such events can seriously damage an institution’s reputation, bringing about punitive regulatory sanctions, and, in some well-known instances, becoming the basis for class-action lawsuits. There has been an extensive body of research in the literature that has focused on understanding the risk of information security (see Bagchi and Udo 2003, Dhillon 2004, Hong et al. 2003, Straub and Welke 1998, Whitman 2004). Garg et al. (2003) have made an attempt to quantify the financial impact of information security breaches. Fraud as a cause of an operational risk event is a major issue in the credit card business. An enormous amount of work has been done on credit card fraud detection, mainly by researchers in artificial intelligence and data mining; see, for example, Chan et al. (1999). Upgrades and consolidations of information systems may be major causes of operational risk events associated with systems breakdowns.

Operational risk events in trading may be due to human errors, rogue trading (i.e., either unauthorized trades or illegal trades), or system breakdowns. A number of such events have occurred in the last two decades with catastrophic consequences for the institutions involved. Several rogue traders have caused their institutions billions of dollars in losses, see Netter and Poulsen (2003).

Asset management has also been one of the areas within the financial services sector that have recently received a significant amount of research attention as far as operational risk is concerned. During the financial crisis of 2007–2009, some of the most catastrophic losses suffered by investors resulted from failures to properly address operational risk, for example, in the fraud perpetrated by Bernard Madoff (see Arvedlund 2009). Operational risk can be effectively addressed by implementing robust operational infrastructures and controls in organizations that enjoy strong governance, as presented by Alptuna et al. (2010). Operational risk is most salient in the loosely regulated domain of hedge funds. Several studies have shown that many of the hedge funds that have gone under had major identifiable operational issues (e.g., Kundro and Feffer 2003, 2004). Brown et al. (2009a) propose a quantitative operational risk score *ω* for hedge funds that can be calculated from data in hedge fund databases. The purpose of the *ω* score is to identify problematic funds in a manner similar to Almank’s *z* score, which predicts corporate bankruptcies, and can be used as a supplement for qualitative due diligence on hedge funds. In a subsequent study the same authors, Brown et al. (2009b), examine a comprehensive sample of due diligence reports on hedge funds and find that misrepresentation, as well as not using a major auditing firm and third-party valuation, are key components of operational risk and leading indicators of future fund failure.

8.4. Other Research Directions

There are a number of other important research directions that already have received some research attention and that deserve more attention in the future. First, how can we analyze the trade-offs between operational costs (productivity) and operational risk? In the manufacturing and in the services literature, some articles have appeared that discuss the trade-offs between costs and productivity on the one hand and quality on the other hand; see, for example, Jones (1988) and Kekre et al. (2009). However, this issue has not received much attention yet as far as the financial services industry is concerned. Second, a fair amount of research has been done with regard to mitigation of operational risk. The financial services industry has thoroughly studied what the aviation industry has been doing with regard to operational risk, see Cruz and Pinedo (2008). In particular they have considered near-miss management practices of operational risk, see Muermann and Oktem (2002). Two more recent approaches for mitigating operational risk include insurance (in particular with regard to rogue traders, see Jarrow et al. 2010) and securitization (e.g., catastrophe bonds), see Cruz (2002). However, these issues concerning mitigation require more study. Third, it has been observed that there is a significant interplay between operational risk, market risk, and credit risk. For example, when the markets are very volatile, it is more likely that human errors may be made or systems may crash. These correlations have to be analyzed in more detail in the future.

The growing body of research on operational risk in financial services presents interesting cross-fertilization opportunities for operations management researchers, given the increasing visibility of operational risk and the potential losses in financial services. For banks, the Basel Capital Accord in a recent revision began to require risk capital to be reserved for potential loss resulting from operational risk. Wei (2006) developed models based on Bayesian credibility theory to quantify operational risk for firm-specific capital adequacy calculations.
9. Pricing and Revenue Management

9.1. Pricing of Financial Services: Background and Academic Research
Financial services organizations expend serious efforts and resources on pricing and revenue management. Applications are diverse; they include the setting of:

(i) interest rates (APR) on deposits and credit products,
(ii) trading commissions,
(iii) custody fees,
(iv) investment advisory fees,
(v) fund fees (which for hedge funds can be a function of assets and performance), and
(vi) insurance policy premia.

Pricing and revenue management are intertwined with many operations management functions in large financial services firms, because pricing strongly affects consumer demand for products and services, and customer attrition. Complicated pricing mechanisms can increase the volume of billing questions to call centers. All of these can have significant implications on how these products and services are best delivered (e.g., capacity issues, quality issues) as well as on cash (inventory) management.

In addition to market mechanisms, pricing in financial services may be driven by other factors or constraints that may complicate or simplify it. For example, usury laws specify maximum interest rates to be charged for lending to consumers or businesses by banks, credit cards, or pawn shops. Insurance regulations vary by jurisdiction. Fees in open-ended mutual funds offered to US investors are governed by the Investment Company Act of 1940. The Employee Retirement Income Security Act of 1974 (ERISA), which regulates US pension plans, specifies that plan trustees and investment advisors are considered “fiduciaries” who should act in the best interests of plan participants. Among the duties of fiduciaries is to ensure that the plan pays reasonable investment expenses, including fund, advisory, and custody fees, and trading commissions. Preferred pricing provisions, known as Most Favored Nation (MFN) clauses, are typically included in investment management agreements of pension plans governed by ERISA in the United States and by similar laws elsewhere. By the fiduciary standard of ERISA, transactions in accounts that hold plan assets must reflect the best value for the services received. Such services include execution of trades, research, investment advice, and anything else that may be paid through these transaction costs.

A body of literature exists in economics and finance on some aspects of pricing in financial services, and vendors offer pricing services and software, but little academic research with an operations management orientation has been published. In the rest of this section, we discuss the economic foundations of price formation and attempt to link them to research in finance on pricing practices specific to insurance, credit, and hedge funds. We also review the meager operations management literature on pricing in financial services, examine the challenges faced by researchers, and propose links to other research that might help remedy some of the issues.

9.2. Theory of Incentives and Informational Issues in Pricing of Financial Services
By the neoclassical economic assumption of rational individual behavior in perfect market competition, prices for financial products and services should be formed by:

(i) firms’ efforts to maximize profit, i.e., revenue minus cost, and
(ii) consumers’ desire to maximize their utility when faced with exogenous prices.

In this elegant model, information is perfectly known and shared by all economic agents. In a setting that more closely approximates reality, information is incomplete and asymmetrically shared, making price formation a more complex process. The theory of incentives addresses the informational issues that arise in the principal–agent economic relationship. In such a relationship, a “principal” delegates a set of tasks to an “agent” who possesses special competencies to perform the tasks and the two may have conflicting interests. Informational issues present in a principal–agent relationship include:

(i) moral hazard, whereby the agent has private information that can be used to take actions to serve its interests; such actions may work against the interests of the principal, who has to assume some of these actions’ adverse consequences, and
(ii) adverse selection, where an agent uses private information about its own characteristics to gain advantage in selecting a contract offered by the principal.

Non-verifiability is a third issue that may arise when the principal and the agent share information that cannot be verified by a third party, for example, a court of law. The related free-rider problem refers to asymmetric sharing of benefits and costs in resource usage among economic agents. The principal–agent model addresses the design of contracts with appropriate incentives to best align the interests of principal and agent in the presence of the informational issues of moral hazard, adverse selection, and non-verifiability. Laffont and Martimort (2002) focus on the situation of
a principal dealing with a single agent, to whom the principal offers a take-it-or-leave-it contract without negotiation. Their book begins with a review of the literature on incentives in economic thought since Smith (1776) examined incentives in agriculture, and Hume (1740) explicitly defined the free-rider problem. Contract negotiations can be addressed by game theory; Camerer (2003) wrote a very accessible text on behavioral game theory in which he cites all relevant literature. The text by Tirole (1988) contains thorough discussions on price formation.

Moral hazard and adverse selection are responsible for anomalies in insurance and credit product pricing. Akerlof (1970) noted that aggregate risk in segments of the insured population will be an increasing function of insurance premium paid. This is the case because private knowledge of individuals’ state of health, drinking or smoking habits, driving behavior, stability of employment, etc., leads them to accept or reject higher premia offered for health, life, automobile, or mortgage insurance. The same holds true for credit, as Stiglitz and Weiss (1981) noted. Higher interest rates offered would be a lot more likely to be accepted by borrowers whose private self-assessment of their default probability is high. These borrowers would then become even more likely to default because of the high credit cost burden. Insurance and credit are therefore rationed, because there is no price high enough to be profitable with certain customers. That means, as Rothschild and Stiglitz (1976) found, that equilibria in competitive insurance markets under imperfect and asymmetric information needed to be specified by both price and quantity of contracts offered. Stiglitz (1977) examined differences in the role of imperfect information in insurance pricing under a monopoly regime compared with perfect competition.

9.3. Pricing in Asset Management, Securities Trading and Brokerage, and Credit Cards

Pricing in the asset management industry consists primarily of fees charged on a client’s assets under management (AUM) in investment vehicles such as mutual funds, hedge funds, or separately managed accounts (SMAs). Fees can be

(i) fixed, regardless of AUM,
(ii) asset-based, i.e., a percentage of AUM, or
(iii) performance-based, i.e., dependent on AUM’s return.

Pricing structures reflect costs of different vehicles, are formed in the process of bringing investor demand into equilibrium with each vehicle’s capacity, and attempt to address the principal-agent issues between investor (principal) and investment manager (agent). Capacity of an investment vehicle refers to:

(i) operational infrastructure, which tends to be more sophisticated and expensive for hedge funds than for mutual funds, and has additional complexities for SMAs, and
(ii) implementation, whereby profitable opportunities become scarcer as investment vehicles become larger; and the effectiveness of execution of a vehicle’s investment strategies depends to a large degree on the market liquidity of the securities traded by the investment manager in the vehicle’s portfolio.

Mutual funds and similarly managed SMAs typically charge asset-based fees on a calendar basis. As assets grow due to good performance and inflows of new funds resulting from this good performance, the manager gets rewarded for skill and effort. As securities traded by mutual funds are typically liquid, implementation capacity is rarely an issue. In contrast, hedge funds and like-managed SMAs often face capacity constraints that limit their size due to diminishing returns to scale, which makes the asset-based fee an inadequate incentive for hedge fund managers. Hedge fund pricing, therefore, constitutes a more representative application of the principal–agent model because it must use more complete incentive mechanisms to minimize informational issues between fund manager (agent) and investor (principal). A typical pricing structure for hedge funds may consist of

(i) a base (or management) fee, which can be a percentage of AUM paid on a calendar schedule, and covers operating costs of the fund, and
(ii) an incentive (or performance) fee, which allows the manager to keep a share of the value created for the investor during agreed-on time intervals, to ensure that the interests of the two parties are aligned.

The manager earns an incentive fee if value is created for the investor according to an agreed-on metric usually based on either monetary units or rates of return (see Bailey 1990). Typically, incentive fee arrangements have asymmetric payoffs, i.e., they reward gains and do not penalize losses. However, they often require that an investment’s value be at or above a historical maximum, called a high water mark (HWM), before an incentive fee becomes payable to the manager. This implies that prior losses, if any, must have been recovered. Earning high incentive fees depends on fund manager skill; hedge funds with highly skilled managers have higher revenues. Such hedge funds can afford the higher expense of setting up and maintaining a robust operational infrastructure and control framework, as discussed by Alptuna et al. (2010). Incentive fees are also common in private equity, where they are typically known as carried
interest and paid by investors on the liquidation of a partnership and distribution of proceeds.

A new rule adopted by the Securities and Exchange Commission under Section 205 of the Investment Adviser’s Act of 1940 permitted the use of performance fees by registered investment advisers. Several published articles then examined performance fee schemes, with emphasis on inherent moral hazard and its mitigation. Record and Tynan (1987) described incentive fee arrangements, and examined the basic issues involved. Davanzo and Nesbitt (1987) analyzed incentive fee structures and their impact to the business of investment management. Kritzman (1987) proposed ways to deal with moral hazard issues, and to reward the manager’s skill rather than investment style or chance; these issues were also addressed by Grinblatt and Titman (1987). Grinold and Rudd (1987) added another perspective by examining fee structures that would be appropriate for the investor and for the manager according to the latter’s investment skills. Bailey (1990) demonstrated that incentive fee metrics based on value added in monetary units serve investors’ interests better than metrics based on rates of return. Lynch and Musto (1997) examined how incentive fees compare with asset-based fees, especially in extracting value-creating effort by the manager. Anson (2001) valued the call option implicit in incentive fees by Black–Scholes analysis. The non-linear optionality in incentive fees was found by Li and Tiwari (2009) to be optimal even for mutual funds. Golec and Starks (2004) examined the reduction in risk levels of mutual funds after the option-like incentive fee option was prohibited by an act of Congress in 1971.

More recent works have studied the role of HWMs in incentive fee contracts. In a seminal paper, Goetzmann et al. (2003) valued incentive fee contracts with HWMs using models with closed-form solutions. Anson (2001) and Lee et al. (2004) examine the free-rider problem in hedge funds that offer all investors the same HWM: late investors can avoid paying incentive fees if they enter after losses suffered by earlier investors. This issue was addressed in practice by offering investors a different HWM based on their time of entry.

Pricing for security transactions is determined in a deregulated and competitive market. It has been experiencing a downward trend driven by the dramatic cost reductions brought by technological innovations such as electronic trading (see Bortoli et al. 2004, Levecq and Weber 2002, Stoll 2006, 2008, Weber 1999, 2006). Transaction commissions typically pay for the costs of brokerage, clearing and execution, trading, research, and investment advice before the brokerage firm can make a profit. Economies of scale are very important, because a brokerage’s operational infra-

9.4. Revenue Management in Financial Services: Challenges and Opportunities

Revenue management principles can be applied to financial services pricing with some adaptation. The framework can address key considerations, such as rationing in credit and insurance (see, e.g., Akerlof 1970, Phillips 2010, Stiglitz and Weiss 1981). Evidence of rationing is reflected in protecting capacity for airline fare classes (Talluri and Van Ryzin 2005). Another key consideration is consumer behavior (Talluri and Van Ryzin 2004). Some revenue management concepts, for example, capacity, can be quite different in financial services than in the transportation and hospitality industries. It can also be idiosyncratic to each financial product, for example, a CD vs. a mutual fund, or a hedge fund. Boyd (2008) discussed the challenges faced by pricing researchers. Identification problems exist (see Koopmans 1949, Manski 1995) that make it hard to build demand curves from data and...
therefore figure out demand elasticities, which are key to pricing. These problems are especially acute in industries where sales and client relationship persons have a lot of information on customers, including demand and price elasticities. Such information remains private and not centrally shared, as is often the case in some areas of financial services. For example, firms have databases with prices of closed sales but no information on prices of refused offers; the latter remains in the field. Remedies have been proposed and can be considered, such as Gonik's (1978), which Chen (2005) analyzed and compared with a set of linear contracts. Tools used by Athey and Haile (2002) may also be considered.

10. Concluding Remarks and Future Research Directions

In this paper, we have attempted to present an overview of operations management in the financial services industry, and tried to make the case that this industry has several unique characteristics that demand attention separate from research in services in general. We have identified a number of specific characteristics that make financial services unique as far as product design and service delivery are concerned, requiring an interdisciplinary approach. In Appendix A, we provide an overview table of the various operational processes in financial services and highlight the ones that have attracted attention in operations management literature. From the table in Appendix A, it becomes immediately clear that many processes in the financial services industries have received scant research attention from the operational point of view and that there are several areas that are worthy of research efforts in the future. These include each step in the financial product and service life cycle as well as in the customer relationship life cycle.

Much work remains to be done on the design of financial products so that they are

(i) easier to understand for the customer (resulting in fewer calls to call centers),
(ii) easier to use (better online and face-to-face interactions, with less waiting),
(iii) less prone to operational risks induced by human errors,
(iv) easier to forecast and arrange the necessary operational resources for, and
(v) able to take advantage of pricing and revenue management opportunities.

Service designs need to recognize the fact that financial services are relatively sticky, involve long-term relationships with customers, and are at the same time prone to attrition due to poor performance or frustrating service encounters. Anecdotal evidence suggests that it is six times more expensive to acquire a new client than to service an existing one, making operations a really important factor in financial services.

As described in this paper, there is an extensive literature on traditional service operations research topics such as waiting lines, forecasting, and personnel scheduling that are applicable to financial services as well. Inventory models have been successfully applied to cash and currency management. Operational risk management is an emerging area that is attracting quite a lot of attention lately. We expect researchers to branch out and address other non-traditional operational issues in financial services, some of which we have highlighted here. Many of these are likely to be cross-disciplinary, interfacing information technology, marketing, finance, and statistics.

Another area with a potentially significant payoff is the optimization of execution costs in securities trading. Amihud and Mendelson (2010) and Goldstein et al. (2009) examine transaction costs in recent studies, and Bertsimas and Lo (1998) develop trading strategies that optimize the execution of equity transactions. A significant body of work exists in algorithmic trading, market microstructure, and the search for liquidity in securities markets (see, e.g., O'Hara 1998). Operations management researchers can examine the problem and develop solutions by synthesizing elements from this diverse set of disciplines and points of view.

Another interesting area of research could be the integration of the various objectives for improving operations in financial services in which interactions among components can be viewed and modeled holistically. For example, most financial services are quite keen on improving the productivity of their processes. However, one has to keep in mind that there are strong relationships between the productivity of the processes, the operational risk encountered in these processes, and the quality of the services delivered to the clients. When one reduces headcount, productivity may indeed go up; however, the operational risk may increase and the quality of service may go down. It is of great interest to the financial services industry that these interdependencies are well understood.

Pricing and revenue management of financial services could be an area that is ripe for academic research with a potential short-term payoff that may be large. Operations management researchers could leverage related work in economics, finance, and may adapt revenue management principles to develop novel pricing methodologies for financial services.

Operations management research may also be key in enabling visionary ideas for reforming corporate governance. In the principal-agent relationship between corporate shareholders and management, proxy voting is the most important tool available to
shareholders for ensuring that management acts according to their interests. This tool is seldom effective due to shareholder apathy and obstacles to voting through institutional ownership of shares. Holton (2006) proposed proxy aggregation mechanisms, such as a proxy exchange, that could make the process more effective in improving corporate governance. Such mechanisms would ensure that the voting of proxies is handled by informed entities acting in the best interests of shareholders, with thorough knowledge of the issues to be voted on. Implementation of proxy aggregation mechanisms is almost certainly expected to face complex logistical issues, which operations management can examine and address.

Acknowledgments
The authors thank five anonymous referees, an editor, and other readers for constructive comments that helped revise an earlier version of the paper. The presentation, content, flow, and readability of the paper has benefited as a result. The views expressed in this paper are those of the authors and do not necessarily reflect those of Goldman Sachs.

Appendix A: Operations Management Processes in Financial Services

<table>
<thead>
<tr>
<th>Process realm</th>
<th>Operational processes</th>
<th>Strategic processes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acquisition/origination</td>
<td>Current customer portfolio management</td>
</tr>
<tr>
<td>Retail banking</td>
<td>Campaign management</td>
<td>Cash, currency, and deposit management</td>
</tr>
<tr>
<td></td>
<td>Deposit product operations</td>
<td>Teller, ATM, and online operations</td>
</tr>
<tr>
<td></td>
<td>Retail loan origination</td>
<td>Treasury operations</td>
</tr>
<tr>
<td></td>
<td>Loan syndication</td>
<td>Credit agreement maintenance</td>
</tr>
<tr>
<td></td>
<td>Lending operations</td>
<td>Loan syndicate management</td>
</tr>
<tr>
<td></td>
<td>Credit risk management</td>
<td>Risk management</td>
</tr>
<tr>
<td></td>
<td>Insurance</td>
<td>Policy maintenance</td>
</tr>
<tr>
<td></td>
<td>Acquisition planning</td>
<td>Claims investigation and processing</td>
</tr>
<tr>
<td></td>
<td>Solicitation management</td>
<td>Worker’s comp case tracking</td>
</tr>
<tr>
<td></td>
<td>Application processing</td>
<td>Property disposal (auto auctions, etc.)</td>
</tr>
<tr>
<td></td>
<td>Underwriting</td>
<td>Risk management</td>
</tr>
<tr>
<td></td>
<td>Credit cards</td>
<td>Statement operations</td>
</tr>
<tr>
<td></td>
<td>Campaign creation and</td>
<td>Remittance processing</td>
</tr>
<tr>
<td></td>
<td>management</td>
<td>Call center management</td>
</tr>
<tr>
<td></td>
<td>Credit risk management</td>
<td>Plastic printing and mailing operations</td>
</tr>
<tr>
<td></td>
<td>Plastic printing and</td>
<td>Risk management</td>
</tr>
<tr>
<td></td>
<td>mailing operations</td>
<td>Credit, operational</td>
</tr>
</tbody>
</table>

Continued


References


Appendix A (Continued)

<table>
<thead>
<tr>
<th>Process realm</th>
<th>Operational processes</th>
<th>Strategic processes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acquisition/origination Current customer portfolio</td>
<td>Collections</td>
</tr>
<tr>
<td>Mortgage banking</td>
<td>Management mortgage servicing operations</td>
<td>Product offerings</td>
</tr>
<tr>
<td></td>
<td>Securitization Remittance processing</td>
<td>Term, APR</td>
</tr>
<tr>
<td></td>
<td>Secondary market operations Call center management</td>
<td>Processing and call</td>
</tr>
<tr>
<td></td>
<td>Credit risk management Risk management Credit, operational</td>
<td>centers</td>
</tr>
<tr>
<td>Brokerage/investment</td>
<td>Client prospecting Client at risk handling</td>
<td>Collections operations</td>
</tr>
<tr>
<td>advisory</td>
<td></td>
<td>Account offerings management</td>
</tr>
<tr>
<td></td>
<td>Cold calling, referrals Client litigation</td>
<td>Outsourcing/insourcing</td>
</tr>
<tr>
<td></td>
<td>Anti-money laundering Commission management</td>
<td>Capacity issues</td>
</tr>
<tr>
<td></td>
<td>Margin lending Trust services</td>
<td>Technology issues</td>
</tr>
<tr>
<td></td>
<td>Channel lending Risk management Credit, operational Anti-money laundering</td>
<td>Trading platform design</td>
</tr>
<tr>
<td>Asset management</td>
<td>Investment manager due diligence Pricing structures</td>
<td>Collections operations</td>
</tr>
<tr>
<td></td>
<td>Pricing structures Custodian operations</td>
<td>New product/fund design</td>
</tr>
<tr>
<td></td>
<td>New account setup processes Fund administration</td>
<td>Organization design</td>
</tr>
<tr>
<td></td>
<td>Operational risk management Fund accounting</td>
<td>Location of sales and client</td>
</tr>
<tr>
<td></td>
<td>Anti-money laundering Valuation operations</td>
<td>relationship management</td>
</tr>
<tr>
<td></td>
<td>Operational risk management Anti-money laundering</td>
<td>Outsourcing/insourcing</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Capacity issues</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Technology issues</td>
</tr>
</tbody>
</table>

Processes in bold have been addressed in operations management literature.
Processes in italics have received less attention in the operations management literature.


Gawande, M. 1996. Workforce scheduling problems with side constraints. Presentation at the semi-annual INFORMS meeting in Washington, DC.


