Southern African Business Review
Volume 7 Number 1
August 2003

Contents

Research articles

The *Southern African Business Review* is a publication of the Faculty of Economic and Management Sciences, University of South Africa.

IS costing: the case for a reference model

*Frank Bannister, Patrick McCabe & Dan Remenyi**

There are two aspects to any information systems (IS) evaluation: the cost of delivery and the benefit or value received. In the broader literature on IS evaluation, however, surprisingly little attention has been given to the cost side of this equation. In much IS evaluation, research cost is taken as a given without much heed being paid to the metrics and methodologies used to derive it. This paper argues that the cost of IS is a much more complex and contentious issue than it is normally assumed to be, and that it presents the would-be evaluator with several difficulties. IS costing problems are encountered in two forms: conceptual problems and weaknesses in costing systems. These difficulties include questions of cost identification, data capturing, overhead allocation, accounting conventions and disbenefits. These problems are catalogued and explored, and an IS cost reference model is proposed as a potential solution to many of them. It is argued that there is a need for greater emphasis on understanding, accurate measurement and tracking of IS costs, not just as a technique for ensuring IS value, but also as a means of optimising individual and/or organisational investment in and use of IS.

Keywords: IS costs; evaluation; standards; accounting principles, cost modelling.

Introduction

There is an implicit assumption in most of the IS evaluation literature that while IS benefits may sometimes be difficult to estimate, calculating IS costs is straightforward. This assumption is implicit in the remarkable paucity of discussion within the literature of how such costs are measured. This may be because many IS researchers are unaware of how complex IS costing is, or it may be because they consider it to be an accounting problem and the concern of another discipline. However, IS costing is far from straightforward. Even experienced cost accountants are frequently frustrated by the problems they face when it comes to recording and computing the cost of IS. How to account for costs is often a question of interpretation and organisational politics and, on occasion, can be more of an art than a science.

From an evaluation perspective, a further complication is the fact that cost may be seen through the lens of the accountant or the economist – and there are significant differences between these two mindsets. In general, cost accountants accumulate data in order to be able to determine what was actually paid to achieve a specific objective. An economist, or a manager taking an economic perspective, considers the question of investment alternatives and will usually impute opportunity costs. These two positions can be challenging to reconcile.

This paper analyses these costing issues and proposes a possible solution to them in the form of an IS cost reference model. The purpose of this model is to provide a framework within which IS costs can be systematically classified. In order to derive such a model, it is necessary to have a thorough understanding of the problems that arise in IS costing. In arriving at such an understanding, only costs that are material and sufficiently large to have an impact on managerial decisions and their subsequent evaluation need be considered. It does not make sense to waste time and effort chasing minor sums of money or spurious accuracy in the understanding of corporate IS costs.

The remainder of this paper is organised as follows:

- The rationale for calling for greater interest in IS[1] costs is set out.
- A framework for studying the problem of IS costs is presented.

* Frank Bannister and Dan Remenyi are in the School of Systems and Data Studies, Trinity College, Dublin. Patrick McCabe is in the School of Business Studies at Trinity College. E-mail: dan.remenyi@tcd.ie.

1 Although throughout the evaluation and costing literatures, the terms information technology (IT) and information and communications technology (ICT) are widely used, in this paper the more general term 'information systems' (IS) is used to encompass both, except in references to other research that uses specific acronyms.

- A number of conceptual problems in IS costing are identified and explored.
- Some weaknesses commonly found in costing systems are examined.
- A number of problems arising from accounting policies are discussed.
- A cost reference model designed to overcome these problems is proposed.
- Some suggestions for further research are made.

Background – a gap in understanding

It is something of a truism that senior managers are concerned about the substantial amounts that their organisations are spending on IS. This is not a new phenomenon. A typical example was articulated by *Computer Weekly* as long ago as 1991, when it stated:

> Time has run out for IT managers who act like a protected species. Their three-fold failure to understand the business they are supposed to be part of, to communicate with their business colleagues, and to deliver cost effective systems has led to a collapse of faith in the IT department itself.

There is ample survey and other evidence that this concern has persisted over the intervening years (Price Waterhouse 1994, 1996; Wentworth 1996; *ComputerWorld* 1998; Remenyi, Money, Sherwood-Smith & Irani 2000; Willcocks & Graeser 2001; White 2001). While there is evidence of continuing corporate concern about the scale of IS expenditure, another, more fundamental, concern often lurks beneath the surface, namely, that many managers do not even know the cost of their information systems. In their study of IS outsourcing, Lacity & Hirschheim (1993: 190) noted that:

> ... only two of the thirteen companies that participated in the study agree that their IS departments are critical to corporate success. The remaining eleven companies all see their IS departments as necessary, but burdensome, cost pits.

The worry that IS costs are excessive is further supported by Sing (1993), who point out that problems are more abundant than solutions for organisations experiencing rising costs instead of cost reductions. Battles, Mark & Ryan (1996) suggest that senior executives are frustrated by high levels of expenditure. Remenyi et al. (2000) report that many executives have intuitively felt for quite some time that the cost of IS is too high.

A common understanding of cost measurement is important if discussions of, for example, return on investment or value for money are to have meaning throughout the academic, corporate and media worlds (as well as for the effective management of corporate resources). As IS expenditures have continued to climb, understanding IS cost structures and cost behaviour has become progressively more critical and in need of attention.

This gap in knowledge needs to be filled. Even such research on IS costs as there has been, is concerned mainly with visible costs (in other words, those costs that appear in IS financial budgets). A typical example is provided by Lockwood & Sobol (1989), who analysed IS expenditure under a variety of headings in 159 US organisations. However, research shows that up to 40% of IS costs may be incurred outside of the traditional IT budget (Price Waterhouse 1990; Strassmann 1990; Keen 1991; Remenyi et al. 2000). In a typical example of such research, Hochstrasser (1994: 156) estimates that as much 30–50% of IT costs occur outside the official IT budget, commenting that ''a need has been identified to be better aware of the true costs of IT projects''.

There has been little effort to study this missing 40%.[2] The evidence of surveys such as those already cited suggest that only a minority of organisations have any real understanding of their IS costs. Surveys of IS evaluation practice discuss in depth the various types of evaluation methods used and how to choose the best one, but few concern themselves with the basic mechanics of arriving at the actual cost figures against which any benefits should be measured.

It is, of course, quite feasible to evaluate something without any reference to its cost. One can, for example, assess whether an information system has been effective, and possibly even efficient, without knowing how much has been paid for it in monetary terms. One can also use cost proxies such as man-days (Bannister 1995; Strassmann 1997). Another approach is the application of ServQual techniques used to measure user satisfaction as a surrogate measure for IS effectiveness

2 Motta (1988) proposes a taxonomy of IT expenditure, but the taxonomy outlined is at a broad level, design for looking at industry patterns rather than at the expenditure within a given organisation.

(Watson, Pitt & Kavan 1998; Remenyi & Money 1994). However, an accurate knowledge of costs is central to IS evaluation, given the pervasive use of capital budgeting, cost – benefit analysis and variance accounting as IS pre- and post- evaluation methods (Willcocks & Lester 1994; Bacon 1994; Ballantine, Galliers & Powell 1994; Ballantine & Stray 1998). A considerable part of IS evaluation, whether pre- or post-investment, is predicated on knowing what the costs are.

A framework for looking at cost issues

The scope of IS costing is so wide that it is useful to have a framework within which to conceptualise the issues. In order to evaluate anything, it is first necessary to define clearly what is to be evaluated. A review of the literature suggests that IS evaluation is most often applied to one of the following categories:

- Corporate IS expenditure
- IS projects
- Subsets of IS operations.

The latter encompasses such things as:

- A business unit (such as a bank branch)
- A department (such as production)
- A process (such as order processing)
- An application (such as office automation)
- A piece of software (such as the Executive Information System)
- An infrastructural component (such as the network).

While the three categories have much in common, each presents different challenges to the evaluator in defining the expenditure whose outcome or value is to be assessed (see Figure 1).

For example, project costs need to take account of disruption costs, while system costs often give rise to questions of allocation. These differences are identified and discussed in detail in the next section. As will be seen, the various issues that arise in IS costing, and the importance of each issue, vary with the category.

Conceptual problems in costing

Costing problems can be divided into two major groups: conceptual problems and practical problems. While several costing problems exhibit characteristics of both, it is useful to differentiate between the two types. There are several conceptual problems in costing information systems including:

- Defining total investment cost
- Defining total running cost
- Defining system and project boundaries
- Defining temporal boundaries
- Marginal costing
- Opportunity costs
- Disbenefits
- The cost of risk.

Each of these is briefly examined in the sections that follow.

Defining total investment cost

A common question that confronts the evaluator is 'What is the investment cost?' Intuitively, this would seem to be easy to define for the organisation as a whole, perhaps rather less easy for a project, and possibly quite tricky for a subsystem. *Prima facie*, costing of corporate IS requires the identification of all IS-related external expenditure on capital and consumable goods and services, plus internal wage and infrastructural costs incurred in creating the system. There are many lists and categorisations of such costs available in the literature (for example, LAMSAC 1978; Parker & Benson 1988: Appendix C; Hogbin & Thomas 1994; Bannister 1995; Remenyi 1999a, 2000). Unfortunately, as a number of these sources acknowledge, arriving at the investment cost of IS is not quite as straightforward as this. Investment costs include the overheads they attract (for example, learning costs, displace-

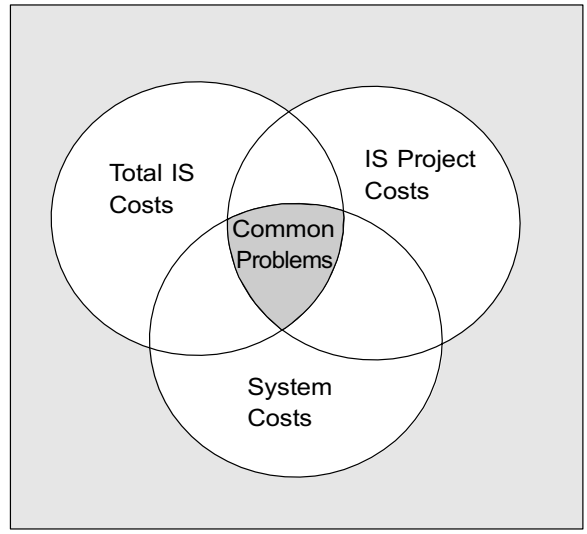


Figure 1: Costing problem classes

ment and administrative costs), which can be difficult to identify. Identifying such overheads is even more problematic with a subsystem because, unlike the organisation as a whole, the boundaries of a system within an organisation are often unclear (Ryan & Harrison 2001). Furthermore, a change in one area can affect several others, and it may not always be obvious how to account for this.

Defining total running cost

A related set of issues arises when one asks the question: 'How much is system X costing per annum?' For most organisations, it is important to know how much the various parts of their IS operation are costing to run, so that divisional and departmental managers can make appropriate decisions about their budgets. This is a particular problem with subsystems in an age of integrated and distributed computing, when defining the boundaries of systems may be difficult. Even if the front-end cost (the costs incurred before a system is commissioned) is known, as soon as a new system is put into production, a series of other costs are immediately incurred. Some of these may be disbenefits. For example, productivity may suffer in the short term as users become familiar with new systems or procedures; there may be disruption of current operations; existing systems may need to be interfaced; or there may be industrial relations problems (Fallon 1993; Hochstrasser 1994; Snyder & Davenport 1997).

One proposed solution to determining running cost is the 'total cost of ownership' (TCO) concept (Gartner Group 1997). Using this, Gartner Group estimated that the typical three-year cost of running a single networked PC was in excess of $27,000 (although this has been widely disputed). TCO has been extended beyond PCs to any identifiable unit (such as software packages, remote access workstations and mainframes). The issues are at their most complex, and are most hotly debated, in the area of networks (*PC Week* 1997; Leach & Smallen 1998; Dudman 1999).

Defining system and project boundaries

Defining the boundaries of systems and projects for costing purposes often presents well nigh intractable problems. In theory, an IS project should be definable, and this should be a major factor in ensuring its professional management. Project management theory states that a well-managed project should have clear boundaries (Turner 1993, 1995; Kerzner 1995; Field & Keller 1998). As Buchanan & Boddy (1992: 8) put it: ''A project is a unique venture with a beginning and an end, conducted by people to meet established goals within parameters of cost, schedule and quality.''

If the project boundaries are clear, the costs should also be clear. Unfortunately, few IS projects are either clear-cut or proceed in isolation (Lock 1996; Remenyi, Sherwood-Smith & White 1997; Remenyi 1999b). They are often part of a much larger IS programme. Furthermore, almost all IS projects use and share existing resources, which immediately raises the question of cost apportionment (as discussed later). IS projects frequently disrupt other activities, and these disruptions push up the organisation's total costs by reducing the efficiency of one or more of the other components of the business system.

Similarly, systems boundaries are not always clear. Many systems are porous, affecting several departments and incurring costs in some or even all of these. The CAD (computer aided design) system may be self contained, but the network and the enterprise resource planning (ERP) system may span every department in the company.

Defining temporal boundaries

This is a feature of all three categories in the framework, but is particularly relevant to investment costs and project costing. A project needs to have a point at which it is deemed to be complete. As already observed, in many IS projects, this point is not clear. Even where termination is achieved, additional post-acquisition or post-commissioning costs almost invariably emerge – for example, where additional resources are necessary to ensure the successful management of the new system, or unanticipated knock-on effects are encountered, which create disbenefits or costs elsewhere.

Some authors and organisations have developed sophisticated IS project cost estimating methodologies to address this shortcoming. Wellman (1992), for example, presents a detailed methodology for estimating the cost of software projects encompassing the concepts of total installation cost (TIC) and life cycle cost (LCC). Wellman takes into account pre- and post-acquisition costs and incorporates such refinements as inflation, exchange rates and the cost of staff replacement into his model. However, indirect costs are only partially considered, and how such costs might be captured is not discussed, except with reference to Block's accomplishment cost procedure (ACP) (Block 1971), although the objectives of ACP are rather

different. While the difficult problem of allocation is not addressed, Wellman's paper does endeavour to address many of the problems referred to.

Marginal costing

The costing of IS projects and systems is frequently distorted by accident or design. Key costs are often overlooked. A common omission is pre-acquisition costs, which are regarded as 'sunk' costs, and not relevant at the point at which the go/no-go decision is made. Pre-acquisition costs can be substantial and include the cost of investigating the feasibility of the project, as well as compiling a thorough business case for the investment. On the one hand, there is no point or purpose in including costs already incurred that will not affect the outcome of the project. On the other, omitting costs such as pre-investment investigations may not reflect the time, effort and expense of the work required to be sufficiently confident to proceed with the particular investment.

Omitting pre-acquisition cost is a traditional marginal or decision cost approach (Horngren, Foster & Datar 1997; Drury 2001). While it employs theoretically sound discounted cash flow (DCF) logic by concentrating on future incremental costs, marginal costing may fail to capture the aggregate long-term impact of an IS investment. An unscrupulous manager can easily manipulate numbers to make the cost of an investment in the short term appear quite low, hiding longer term costs from view. One way of doing this is to keep long-term capital costs relatively low through writing off a disproportionate amount of pre-acquisition cost as current expenditure. Such behaviour is often encouraged at the highest levels within the organisation as it can substantially reduce short-term tax liabilities. It also holds an instant appeal for those managers that are directly responsible for the decision, as the artificially reduced investment enhances the prospects of a favourable evaluation of the decision later on.

The weaknesses of traditional marginal costing are among the factors that have led to the development over the past decade of activity based costing (ABC) (Cooper & Kaplan 1988; Horngren et al. 1997; Atkinson, Banker, Kaplan & Young 2000; Drury 2001). ABC is based on the concept of cost drivers (in other words, that any cost that is a consequence of an activity, even if indirect, is a cost of that activity). It attempts to capture inter-temporal and intra-organisational cost consequences by recognising that there is often a significant lag between decisions or initialising activities and the associated escalation of costs throughout the organisation. The growth in interest in ABC is a recognition of the fact that it may be, to some business executives, a more meaningful measure. The ABC cost model may therefore be far more relevant to IS costing than the marginal costing approach. Even if ABC is not formally institutionalised within an organisation, an ABC mindset facilitates a more comprehensive IS cost measurement.

Opportunity costs

As stated in the introduction, managers and economists tend to view costs differently from accountants. These differences surface in their varying approaches to modelling total cost and marginal cost (see Drury 2001: 46–52). They can also arise in the modelling of investment decisions. Here, the accountant's pre-occupation with revenue and cost flows that are likely to appear in the corporate accounts as a result of a new investment can militate against the inclusion of opportunity costs. These, by definition, will never show up in traditional accrual accounting systems, as these systems capture the costs and revenues only of those activities actually undertaken. Accountants and accounting systems focus on the cost of what is done; opportunity costs are associated with what is not done. Whereas the accountant will frequently ask the cost of a project, a manager will tend to ask the cost of not investing in the project. This approach incorporates the economist's notion of opportunity cost.

Opportunity costs arise in a number of different ways, one of which is when there is little or no organisational slack and where resources need to be diverted from other established activities to the project under evaluation. Any benefits lost from the activity thus impoverished or deferred are regarded as costs of the driving project. Organisations with plenty of slack can often commit considerable resources without impinging on other core activities. Opportunity costs have become more important with the success of business process re-engineering (BPR) in eliminating slack. As organisations become leaner and meaner, such flexibility decreases, with the result that greater attention to opportunity costs as a factor in decision-making now appears to be justified.

Disbenefits

Computer systems can have an adverse impact on company performance, leading to what are sometimes called disbenefits, or soft costs (Parker & Benson 1988; Khalifa, Irani & Baldwin 1999), that

do not generally find their way into the IS cost equation. These can involve losses in productivity for a variety of reasons including:

- Badly designed systems
- Mistakes made due to inadequate training
- Slow system response
- Loss of business during down time
- IS-enabled time-wasting.

Disbenefits are often difficult to identify, let alone quantify. For this reason, they are sometimes disregarded because they are considered too difficult to evaluate. Disbenefits are discussed by a number of commentators (for example, Hochstrasser 1987; Remenyi et al. 2000; Ryan & Harrison 2001). Strassmann (1988) comments that many IS investments are driven by considerations of efficiency rather than effectiveness, and he regards this rather narrow view of the impact of IT as a disbenefit in itself. Chalcraft (1997) discusses the concept of 'uncertainty', and Gurbaxani & Whang (1991) discuss the 'opportunity cost' of poor information as a variation on the disbenefit theme.

Disbenefits can be external or internal, direct or indirect, as illustrated in Figure 2.

External disbenefits arise when investment in technology leads to a loss in benefits/profit from the products or services that the technology is supporting. This gives rise to some difficult questions. Where a new technology alters the structure and the economics of a market, the result may be a much less profitable business, but a company may have no option other than to invest in this technology if it wishes to survive. The only means of accounting for this type of phenomenon may be to use the concept of opportunity costs, but, as already noted, these have no place in the typical corporate accounting system.

	Direct	Indirect
External	Change in Market Economics Profitable Product Displacement	Hacker Attack Loss of Customer Confidence
Internal	Increased Risk of Fraud Disruption	Lower Morale Loss of Productivity

Figure 2: Classification of disbenefits

Disbenefits are not restricted to external competitive effects. The introduction of new technology can lead to more time-wasting by employees, greater stress, new health problems, greater exposure to security problems and so on. The introduction of Web facilities into organisations has led to a significant increase in IT-enabled distractions in the workplace. In Ireland, it was estimated that the television programme *Big Brother* cost small businesses £30 000 a week as employees logged onto the programme Web site during working hours (*Irish Times* 5 September 2000). In the United Kingdom, the figure is estimated to have been £1.4 million per week (*Financial Times* 25 August 2000). Whether such activities merely displace other non-productive activities (such as doing the crossword or gossiping at the water cooler) is an open question and might make for an interesting research project in its own right. In the absence of solid research, intuition suggests that Web availability does lead to its use for some less than productive purposes.

Disbenefits can be subtle and complex to measure, but this is no reason for disregarding them. They impose real costs on a business. Kaplan's (1986) observation that ''conservative accountants who assign zero values to many intangible benefits prefer being precisely wrong to being vaguely right'' can be applied with equal validity to difficult-to-measure disbenefits.

Cost of risk

Computerisation can open up an organisation to a range of risks to which it would not otherwise be exposed, but such risks are not normally converted to a monetary cost. These can range from project-related risks, such as overruns on cost and/or time, to operational risks, such as catastrophic system failure or loss (Remenyi 1999b; Willcocks & Graeser 2001).

The risk associated with any project or system development may be wholly or partially offset by, for example, using an external, reputable supplier and by including penalty clauses in the contract. Suppliers will usually charge for assuming this risk. This additional cost is in effect an insurance premium. In contrast, a company that chooses to go with a low cost bidder may be assuming additional project risk. Like not insuring anything else, this is a hidden cost.

A company can choose to take out insurance to cover the operational risk of major system failure (in

other words, the type of failure that could lead to significant revenue losses (for example in a brokerage or an airline) or have tragic human consequences (for example, in a hospital or a signalling system). Such insurance is typically in the form of a business continuation plan with backup sites, hot standby facilities and so on. This may be supported by limited financial cover provided by an insurance policy. These are clearly identifiable IS costs. However, if a company chooses not to insure against increased risk (or decides against having a contingency plan or backup site), it is, in effect, self insuring. This risk premium is a hidden cost and will not be recognised in the accounting system.[3]

Costing system weaknesses

Introduction

In addition to conceptual problems, there are several potential weaknesses within costing systems themselves. It is convenient to distinguish between two types (Figure 3):

1. Data capture where there can be:
   - Omitted/missed costs
   - Misassigned costs
   - Intertemporal problems.
2. Inadequate or inappropriate costing and accounting policies and methods.

One of the key features of Figure 3 is the fact that cost leakage can occur in all three strands of the model. This is a potentially major concern for management and thus requires well-designed control systems. Even where costs are captured correctly, inappropriate allocation and other accounting policies can lead to cost leakage (in other words, the loss of accurate and relevant cost information from the information system). The issues raised by these data-capture weaknesses will now be considered. Policy issues are considered in the next section.

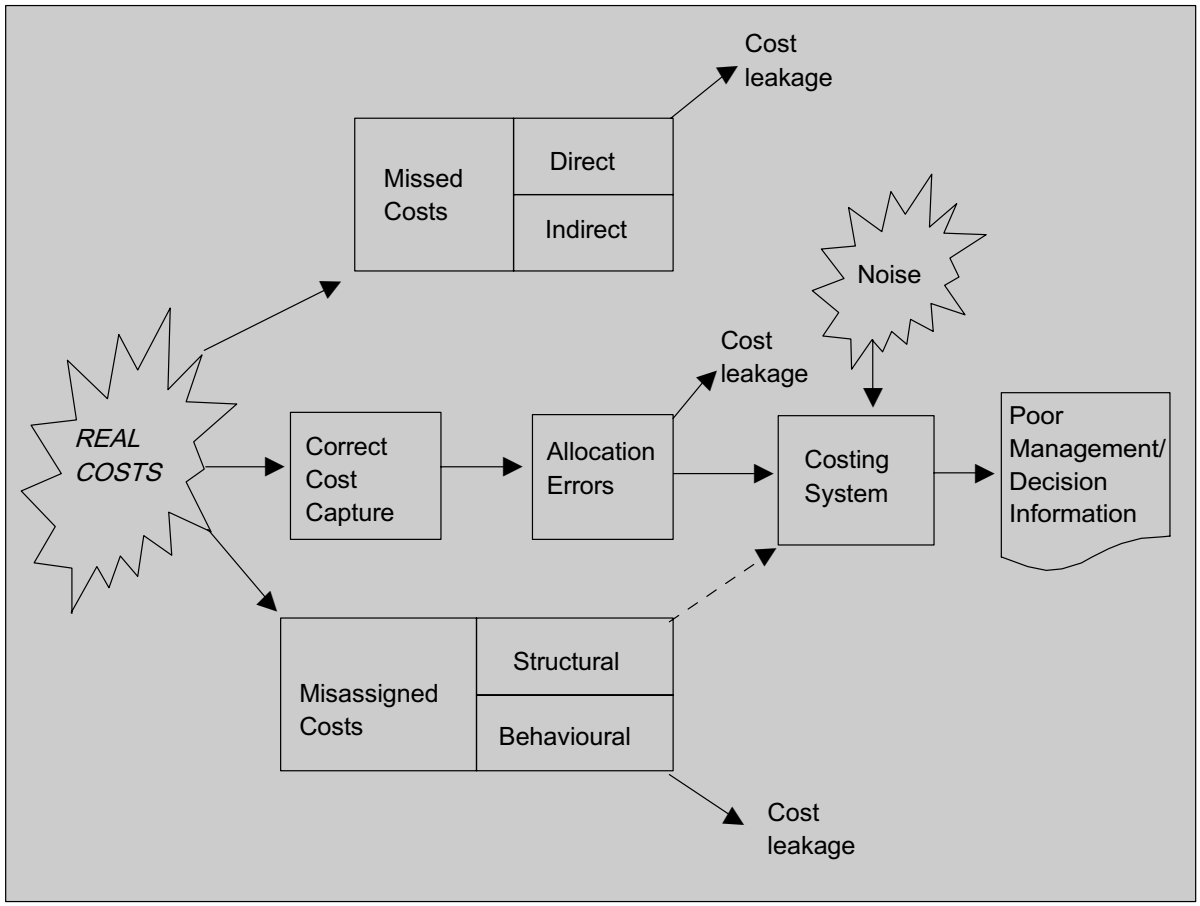


Figure 3: Costing system weaknesses

3 In high profile or listed companies, such unprotected or uninsured risks will almost certainly be recognised in the capital markets and be reflected in the form of higher borrowing costs and/or a lower share price.

Omitted costs

The first type of data capture weakness is omitted costs. 'Omitted costs' are defined as costs that should have been identified and allocated to a particular cost point or centre, but for one or other reason have not been. The question of the cost point with which an item should be identified will depend upon a number of issues, including whether the organisation is following a direct or a full costing approach. For example, in a full costing approach, the payroll costs of non-IS staff engaged in IS-related activities could be regarded as an IS cost, whereas in a direct costing environment, they would not be.

Some costs that are often not identified, and in some cases not even captured, by the typical accounting system include disruption and displacement costs, as well as disbenefits. Kaplan (1986) shows that new systems implementation is often accompanied by a short-term loss in productivity as systems are disrupted and users adapt (Ryan & Harrison 2001). Omitted costs occur because there is no mechanism whereby the accounting system can separate and capture the data at a sufficiently low level of granularity for proper analysis.

Misassigned costs

These are best illustrated by means of an example. A Dublin-based database administrator (DBA) goes to Paris on a three-day course. The cost of the airfare is recorded by the accounting system as a travel expense, the accommodation as a miscellaneous expense and the other expenses as subsistence, when in reality all three are IS training costs. Of course, the airfare is a travel expense. A good accounting system will capture the fact that it is not only a travel expense, but that it is also a departmental expense, a project expense, a training expense and an IS expense. Such data are inherently multidimensional. Unless the accounting system is designed with this multidimensionality in mind, it will not be able to distil out the different facets of the primary data. Weak accounting systems may, for example, capture only the first or the first two of the five dimensions of the administrator's airfare. It is an interesting question how often (where such data are comprehensively captured by sophisticated coding and accounting systems) the resulting richness of this information is used to its full potential in planning, control or decision-making.

Accurate data capture requires two things:

1. A well-designed chart of accounts: If the chart of accounts does not have appropriate headings, costs will be charged to the nearest convenient heading. If an organisation has not changed its chart of accounts in many years, it may be quite unsuitable for the way the business currently operates. This might be called 'structural misassignment'.

2. Using the accounting system correctly: Even where the chart of accounts is well designed, staff sometimes deliberately miscode costs, a process colloquially referred to as 'burying' costs. Sloppy or careless coding is also a problem. A common behaviour is to code costs to a heading where there is still some unused budget rather than to the correct heading, in cases where there has been over-spending on the budget. Another is to deliberately miscategorise goods to circumvent company purchasing rules. This is 'behavioural misassignment'.

Another common source of misassignment is user costs (such as learning costs, errors, consultation time and so on). Even in professional organisations, such as accountancy firms, consultancies and legal practices, which record staff activity, sometimes to the closest minute, accurate computation of user costs requires both an appropriate costing system and a high degree of user discipline. Even good data capture systems are prone to bias and manipulation, especially if budget or actual expenditures are linked to a personal reward system (Lukka 1988). Agency theory also suggests that people massage reported data to meet the criteria by which they perceive that their performance will be judged (Puxty 1985; Morgan 1991; Koford & Penno 1992; Anthony & Govindarajan 1995). From a different perspective, Walker & Johnson (1999) report that slack-building can result from budget incentives. This phenomenon makes the identification of the real cost even more difficult to establish.

Intertemporal data capture problems

The effectiveness of capturing IS costs is also time-related. The further from the period or point of investment that an expenditure occurs, the more difficult it is to tie the cost to the investment. This ripple effect is illustrated in Figure 4.

Over the life of any IS investment, the cost impact diffuses over time to incur more and more remote effects, while the ability of current systems to

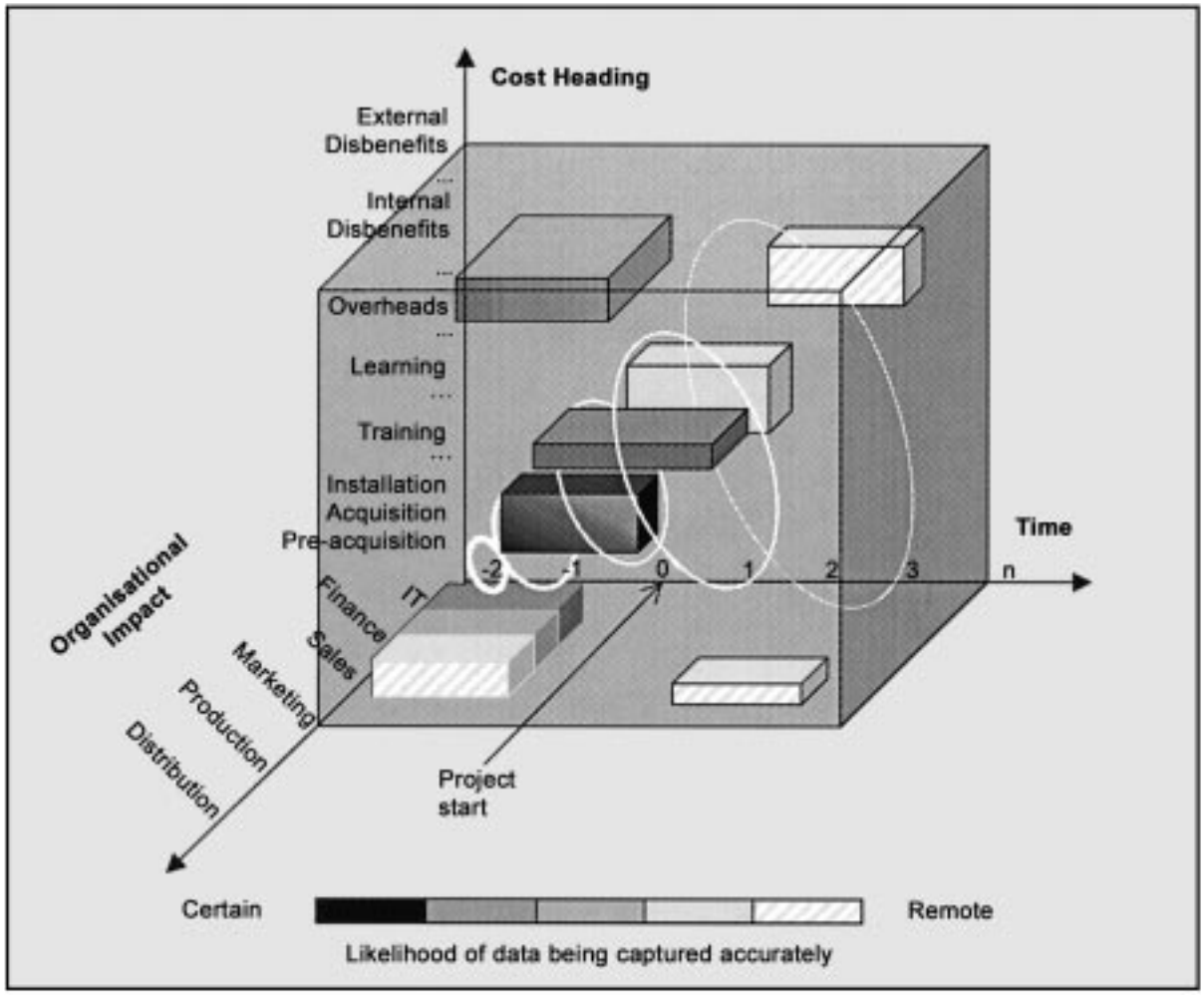


Figure 4: The ripple patern of cost accumulation

capture these costs decreases. The ripple effect occurs in three dimensions: cost type, time and organisation. For example, the impacts of a change in the financial system may be felt initially in the IS and finance department, but may eventually percolate out to departments such as sales, marketing and distribution.

There are two questions that need to be addressed here. At the pragmatic level, the accounting system needs to be able to relate costs over a long time period and in different locations. Project accounting systems can do this, but many organisations either do not have such systems or are unable to capture the data necessary to drive them. At a deeper level, the time horizon for a given cost needs to be determined. It does not make sense to track post implementation costs indefinitely. At some point, the 'project' must cease and the system become part of normal operations. The question is where the cut-over point should be. It should be remembered that many of the disbenefits and risk-related costs associated with projects may show up in the later phases of the project. The risk of late failure always needs to be considered.

Accounting policies

Weaknesses can arise in a costing system from inappropriate accounting policies. This can occur because of failure to apply proper accounting standards or because of simple misjudgement. Among the areas where difficulties can arise are capitalisation/amortisation policies, the early adopter problem, overhead allocation and charge-back.

Capitalisation and amortisation

Two related issues that affect the quantification of IS costs are the capitalisation of expenditure and its subsequent amortisation. Capitalisation refers to the treatment of certain expenditures as acquisition of assets, which then appear on the corporate balance sheet. Amortisation relates to the manner in which this expenditure is subsequently written off

in the income statement over a number of years. Different firms are likely to approach capitalisation in different ways. Some will prefer a conservative approach, capitalising the minimum that is consistent with the accounting standards and the practices in their industry. To this end, the advice of auditors is usually sought. Despite this accounting frame of reference, which is governed by various formal statements of standards, accounting practice can give a significantly erroneous view of the aggregate cost of an IS investment.

The rate at which an organisation amortises its assets depends upon a number of variables. In the first place, the organisation may try to match the amortisation schedule to the useful economic life of the asset. When the organisation is quite profitable and there is a high degree of uncertainty as to the likely duration of an asset's economic life, there may be pressure to accelerate the amortisation schedule. When profits are not adequate, the reverse can happen. Further complications arise from the fact that internal management accounting reports need not comply with external reporting standards. To compound the difficulties, for evaluation techniques that use after-tax cash flows, the amounts allowed for amortisation by the tax authorities are likely to differ from both of these. It is therefore essential that caution be exercised in the use of both capitalised and amortised calculations or measures.

Early adopters and the 'hotel night' problem

Another problem that can distort IS costs is the so-called early adopter or 'hotel night' problem. Because hotels have large fixed overheads, the cost per guest in a hotel falls as the room occupancy increases. Hotels do not charge guests extra because there are fewer of them (on the contrary, they are likely to reduce prices to attract more guests). Perversely, the converse of this logic often occurs when a new system is implemented, and early adopters can end up paying a disproportionate amount of the cost. If initial take-up is slow (or is restricted to a small group) the cost per user is high. As the use of the technology spreads, the cost per user drops. Failure to account properly for this phenomenon (in other words, to use appropriate time horizons), leads to both poor decision-making and corporate game playing. The former can arise when the long-term cost per user of a system is misunderstood, leading to short-sighted decisions. The latter arises when 'clever' managers let other departments absorb the expensive up-front costs of a new system and then 'piggy back' on the apparently lower costs of the system when it settles in.

Overhead allocation

There are few more troublesome problems in cost accounting than that of overhead allocation. Overhead allocation is the subject of both theoretical debate (Ahmed & Scapens 1991) and sometimes intense practical contention. Inappropriate overhead allocation not only results in poor management information, but can also distort user and manager behaviour, leading to counterproductive or sub-optimal actions. Overhead allocation can be the cause of bitter disputes in organisations. Worse, there is seldom an unambiguously right or wrong way to allocate overheads, so judgements made are always open to criticism.

Within the IS function itself, there is a large range of potential overheads, including:

- Shared devices (such as printers, storage and tape drives)
- Shared infrastructure (network, cabling)
- Shared services (helpdesk, system administration, backup)
- Administrative costs (such as insurance and legal expenses)
- IS management
- IS research.

To apportion these costs in a reasonable and fair way, the accountant or manager is faced with the problem of which criterion or criteria to use as the basis of allocation. This is seldom easy, as any choice will typically favour one group and disadvantage others. The list of possible bases for allocation is formidable and includes:

- Headcount
- Number of screens
- Number of devices
- Total capital investment
- CPU usage
- Disk usage
- Network usage.

All of these can give rise to different distortions in information and behaviour.

The following example illustrates the type of problem that can arise. In a large professional services organisation, a small specialist department, comprising fewer than 5% of the staff, had

approximately 8.5% of the organisation's PCs. The cost of central IS services (including the mainframe, the user support group, the help desk and the network) was allocated on the basis of PC count, thus assigning the department an IS overhead cost per head of almost twice that of other departments. Because of the high level of technical expertise in the department, in fact, the staff tended to solve most of their own problems and, pro rata, made less use per capita of help and support services than other departments. Because they were PC oriented, they also made less use of the mainframe, which meant that the cost allocation system was doubly unfair. As the overhead cost was a charge on the department's profitability, and the bonuses of the department's managers were linked to their departmental profit performance, the allocation of IS costs became a personal and highly political issue. As a result, a large amount of time was squandered in meetings, and resentment festered, damaging both productivity and morale.

Charge-back

Cost allocation may be effected in a number of ways. One common method involves the assignment of a periodic charge to user departments. More complex systems of allocation may involve a formal system of charge-back. In some organisations, an internal market is created, whereby the supplying department becomes a profit centre, aiming to maximise revenues, while user departments aim to use the minimum of such services that is consistent with the achievement of their own objectives (Lacity & Hirschheim 1995). These systems are usually more controversial than their more basic counterparts already noted, not least because they aim to influence managerial behaviour and decision-making. Whatever the theory may state, many management textbooks suggest that such goal congruence is hard to achieve, not least because of the continually changing external environment (for example, demand, price and tax rates) and changing relationships between supplying and consuming departments. A good example of this line of argument can be found in Horngren et al. (1997: chapter 25).

Many charge-back systems use a standard cost approach (in that they do not attempt to deal with the 'hotel night' problem). The standard cost is based on a budgeted or anticipated level of usage. Like any standard costing system, standardised charge-back may under- or over-absorb costs. The reason is that standard charge-back normally involves a single charge per 'unit' of consumption that combines the cost of fixed and variable expenses incurred in the supplying department. Standard costing is also unsuitable for activities of a non-repetitive nature (Drury 1999), and a great many IS-related activities are non-repetitive.

Towards an IS cost reference model

The need for standards

The long list of potential problems and difficulties discussed in this paper suggests that there is a need for a standardised approach to costing IS. The purpose of a set of standards would be to ensure that there is a consistent approach to costing information systems investments and expenditure in a manner analogous to that used by the financial accounting profession to ensure consistency of financial reporting. The latter are set out in broad terms as generally accepted accounting principles (GAAPs) and formally enshrined in such publications as *Statements of Standard Accounting Practice* (SSAPs), financial reporting standards (FRSs), and the emerging statements of international accounting standards (IASs). It is not proposed that the standards or guidelines for costing of IS should try to emulate the formal/quasi-legal status of an SSAP (the status of these standards varies between countries; most are self policed by the professional accountancy bodies). It is important to recognise that while the accounting standards bodies can set strict rules for financial accounting and reporting, in the looser world of cost and management accounting, it is much more difficult, if not actually impossible, to be so prescriptive. A certain degree of flexibility is necessary to allow for the great variety of circumstances that can arise in costing generally. Nevertheless, there is a strong case for at least an agreed set of guidelines for computing IS costs, whether such costs are to be used for internal corporate decision-making or for cross-industry benchmarking. The arguments for a model are summarised at the end of this section. First, a proposed approach is outlined.

An IS cost reference model

One approach to creating a set of guidelines for IS costing is to define a reference model. The term 'reference model' is adopted to reflect the fact that such a model would contain a set of standards for good practice to which users could refer and from which they could select the components that are relevant to their circumstances. Such a model should have some or all of the following characteristics:

- A definition of the costs to be included
- Guidelines for data capturing
- Standards and methods for the allocation of overhead costs
- Filters to remove irrelevant data or noise
- Compatibility with existing external accounting rules and standards
- Compatibility with cost accounting practice in comparable areas
- Guidelines for defining temporal boundaries
- A mechanism for defining organisational interdependencies.

An outline of the proposed model is shown in Figure 5.

Each component of this model would contain some or all of the following:

- Recommended policies
- Recommended procedures
- Definitions
- Checklists.

For example, under 'pre-acquisition costs', the nature of such costs would need to be defined and a set of examples provided. In addition, the reference model would provide rules on when pre-acquisition costs may be expensed, and the circumstances under which they should be included in the investment cost carried forward. It should be remembered that the key objective is to define good practice – a code of conduct for effective costing that would, in this case, seek to eliminate the possibility of managers' manipulating such costs to try and achieve a more favourable outcome in a subsequent evaluation.

Example of a component

The purpose of this paper is to propose the framework. It is beyond the scope of the paper to examine any of the components of that framework in detail. However, the following illustrates the type of supporting detail that such a model might involve for 'internal disbenefits'.

There has been considerable research into and discussion of soft benefits and how to measure them (Dempsey, Dvorak, Holen, Mark & Meehan 1998; Waltner 1999; Burn, Marshall & Martin 2002; Willcocks & Graeser 2001). Many of the concepts and ideas developed in this research can be applied to or adapted for disbenefits. It is therefore possible to categorise internal disbenefits and suggest metrics that can be applied to them. One approach to analysing disbenefits is to combine them with the opportunity costs of not taking the action in order to arrive at a balanced measure. Policy in this regard might state that:

- All disbenefits must be identified
- Disbenefits should be classified as 'significant' or 'ignorable'

Applying these rules (for example, to implementing Internet access on all desktops) might result in something like the following:

Disbenefits

- Loss of staff time and productivity as a result of inappropriate surfing

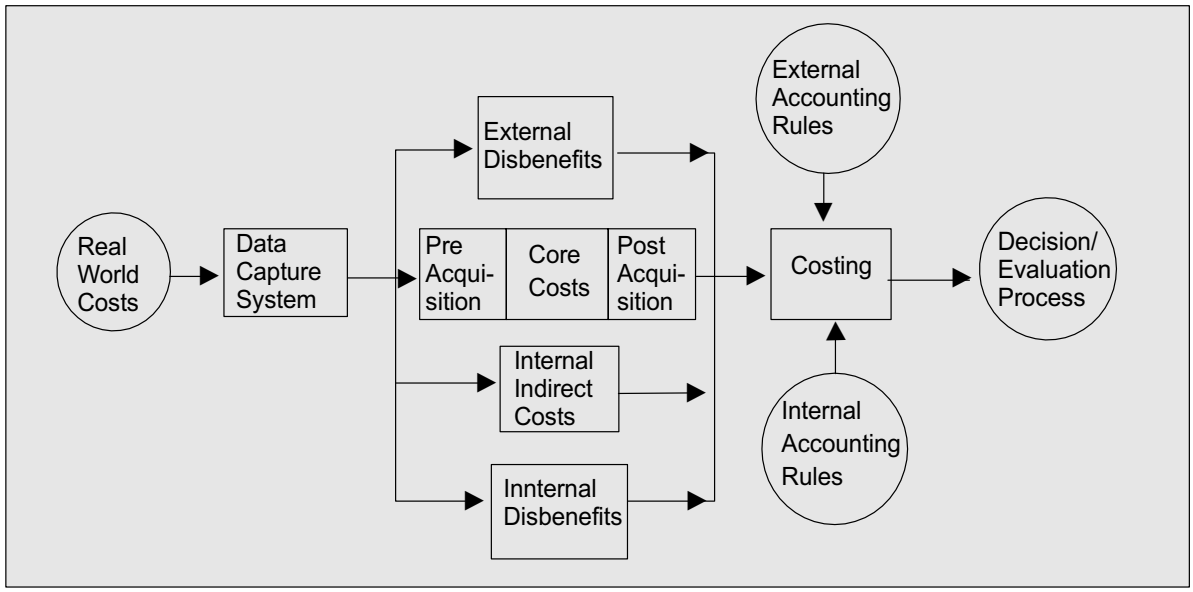


Figure 5: An IS cost reference model

- Loss of productivity by staff using the Internet for other non-work-related purposes
- Risk of staff accessing inappropriate Web sites (such as pornography), with the associated possibility of legal or other exposure for the company
- Risk of staff downloading illegal software
- Risk of staff downloading viruses
- Exposure of the firm to tracking of its activities by outsiders using cookies
- Loss of control over purchasing, particularly of small items.

Significant disbenefits

Of the above:

- Loss of staff time and productivity as a result of inappropriate surfing
- Loss of productivity by staff using the Internet for other non-work-related purposes
- Loss of control over purchasing, particularly of small items.

Cost of disbenefits

The cost might include:

■ Estimated staff time spent on unproductive Internet-related activity per week	3.00 hours
■ Less estimated cost of displaced non-productive activity	1.75 hours
■ Net additional time lost per week per staff member	1.25 hours
■ Number of staff	100
■ Hours lost	125 hours
■ Cost per hour	€ 20
■ Total cost per week	€ 2500

This example is intended only to give an indication of what a detailed model might look like.

The case for a reference model

The arguments in favour of a reference model can be summarised as follows:

1. *There is currently much poor practice that could be improved upon.* It is clear that current costing systems and procedures are often deeply flawed. There may be many reasons for this, including inadequate design, historical inertia, poor understanding or, in rare cases, deliberate deception. A reference model would help users to overcome the first three of these and at least make the fourth more difficult.

2. *There is a need for consistency for evaluation purposes.* Unless costs are broadly correct, evaluation, or at least evaluation that seeks to identify return on financial investment or value for money, may be meaningless. Currently, two organisations in similar circumstances, using the same evaluation techniques but different costing methods, could conceivably arrive at quite different views about the same IT investment.

3. *It would increase transparency.* It was noted at the outset of this paper that many managers feel that they do not fully understand their IT costs. Computing IS costs using a standard approach/model should go a long way towards ameliorating these concerns.

4. *It would make for more useful decisions about IS expenditure.* Under or overstatements of IS cost (as with any other cost) can lead to poor decisions. In a similar manner, poor design of charge-back systems can lead to sub-optimal behaviour at an organisational level as individuals and departments seek to optimise their own positions.

5. *It would save users time and effort.* Like any set of guidelines, an IS costing reference model should provide procedures and policies that can be adopted or adapted by users and managers. This avoids re-inventing the wheel, ensures that what is being done complies with good practice and can save much time and expense.

Further research

This paper has examined a range of theoretical and practical problems in IS costing and suggested a model that addresses these. Without an understood cost basis, comparative financial evaluations of IS are suspect, and individual evaluations may be invalid – in colloquial terms, evaluators could find themselves mixing apples and pears.

This paper has presented a framework for tackling this problem. Work remains to be done to expand this framework into a methodology. Specifically, further work is required in two areas: development of the detailed model and validation of the model structure. The structure outlined is based on an examination of the IS management, IS evaluation

and cost accounting literatures. In addition, the authors have drawn on their professional experience, obtained over many years of working in the field. Nevertheless, the model needs to be validated by applying it in practice in a number of different environments to assess how well it fits.

Clearly, there is much to be gained from a better understanding of IT costs. If organisations and IT professionals were to use a common reference model, such as the one proposed here, there would be an opportunity for the community of IT users, developers, consultants and academics to have a much more useful dialogue about the awkward question of 'how much did that system actually cost?'

References

Ahmed, M.N. & Scapens, R.W. 1991. 'Cost allocation theory and practice: the continuing debate', In D. Ashton, T. Hopper & R.W. Scapens (eds), *Issues in Management Accounting,* pp. 39–60. Hemel Hempstead: Prentice Hall Europe.

Anthony, R.N. & Govindarajan, V. 1995. *Management Control Systems.* Chicago: Irwin.

Atkinson, A., Banker, R.D., Kaplan, R.S. & Young, S.M. 2000. *Management Accounting.* New Jersey: Prentice Hall.

Bacon, J. 1994. 'Why companies invest in information technology', In L. Willcocks (ed.), *Information Management, the Evaluation of Information Systems Investments,* pp. 31–47. London: Chapman & Hall.

Ballantine, J., Galliers, R. & Powell, P. 1994. 'Daring to be different: capital appraisal and technology investments', In L. Willcocks, (ed.), *Information Management, the Evaluation of Information Systems Investments,* pp. 87–97. London: Chapman & Hall.

Ballantine, J. & Stray, S. 1998. 'A comparative analysis of the evaluation of information systems and other capital investments: empirical evidence', In W. Baets (ed.), *Proceedings of the 6th European Conference on Information Systems,* Vol. II: 809–822.

Bannister, F. 1995. *Purchasing and Financing IT.* London: GEE.

Battles, B., Mark, D. & Ryan, C. 1996. 'An open letter to CEOs: How otherwise good managers spend too much on information technology', *McKinsey Quarterly,* 3: 116–127.

Block, E.B. 1971. 'Accomplishment/cost – better project control,' *Harvard Business Review,* May–June: 110–124.

Buchanan, D. & Boddy, D. 1992. *The Expertise of the Change Agent: Public Performance and Backstage Activity.* London: Prentice Hall.

Burn, J., Marshall, P. & Martin, B. 2002. *e-Business Strategies for Virtual Organisations.* Oxford: Butterworth Heinemann.

Chalcraft, C. 1997. *The Justification and Costing of Information Systems.* London: Pitman.

Computer Weekly. 1991. 'Does the IT Department have a future?', 25 July, Sutton, UK.

ComputerWorld. 1998. 'IT budgets rock boat', 20 July, www.computerworld.com/cwi/Printer_Friendly_Version/frame/0,1212,NAV47_STO31897-00.html

Cooper, R. & Kaplan, R.S. 1988. 'Measure costs right: make the right decisions'. *Harvard Business Review,* September–October: 96–103.

Dempsey, J., Dvorak, R., Holen, E., Mark, D. & Meehan, III, W. 1998. 'A hard and soft look at IT investments', *McKinsey Quarterly,* 1: 126–137.

Drury, C. 1999. 'Standard costing: a technique at variance with modern management?', *Management Accounting,* 77(10): 56–58.

Drury, C. 2001. *Management Accounting for Business Decisions.* London: Thomson Learning.

Dudman, J. 1999. 'Counting the cost of networks', *Computing,* 1 July.

Fallon, I. 1993. *The Paper Chase.* London: Harper Collins.

Field, L. & Keller, M. 1998. *Project Management.* London: Open University Press/International Thompson Business Press.

Gartner Group. 1997. *Gartner Group's 1997 PC/LAN TCO Model – The Basics.* Gartner Analytics, 19 December.

Gurbaxani, V. & Whang, S. 1991. 'The impact of information systems on organisations and markets', *Communications of the ACM,* 43(1): 59–73.

Hochstrasser, B. 1987. *Does Information Technology Slow You Down?* London: Kobler Unit, Imperial College.

Hochstrasser, B. 1994. 'Justifying IT investments', In L. Willcocks (ed.), *Information Management, the Evaluation of Information Systems Investments,* pp. 151–170. London: Chapman Hall.

Hogbin, G. & Thomas, D.V. 1994. *Investing in Information Technology: Managing the Decision-Making Process.* Maidenhead: McGraw-Hill.

Horngren, C., Foster, G. & Datar, S.M. 1997. Cost *Accounting – A Managerial Emphasis.* New Jersey: Prentice Hall International.

Kaplan, R. 1986. 'Must CIM be justified on faith alone?', *Harvard Business Review,* March–April: 87–95.

Keen, P.G.W. 1991. *Shaping the Future: Business Design through Information Technology.* Boston, MA: Harvard Business School Press.

Kerzner, H. 1995. *Project Management: A Systems Approach to Planning, Scheduling, and Controlling* (5th edition). New York, London: Van Nostrand Reinhold.

Khalifa, G., Irani, Z. & Baldwin, L. 1999. 'Factors impacting IT/IS evaluation: linking investment barriers to life cycle', In A. Brown & D. Remenyi (eds), *ECITE 99 – Proceedings of the 7th European Conference on Information Systems Evaluation,* pp. 177–192, Brunel University.

Koford, K. & Penno, M. 1992. 'Accounting, principal-agent theory and self interested behaviour', In N. Bowie & R. Freeman (eds), *Ethics and Agency Theory: an Introduction,* pp. 3–22, Oxford University Press, New York.

Lacity, M.C. & Hirschheim, R. 1993. *Information System Outsourcing. Myths, Metaphors and Realities.* Chichester: John Wiley & Sons.

Lacity, M.C. & Hirschheim, R. 1995. Beyond the *Information Systems Outsourcing Bandwagon. The Insourcing Response.* Chichester: John Wiley & Sons.

LAMSAC. 1978. Costing the use of computers in local government. Report of the Computer Panel of LAMSAC. London: Local Authorities Management Services and Computer Committee.

Leach, K. & Smallen, D. 1998. 'What do information technology support services really cost?', *Cause/ Effect,* 21(2): 38–45.

Lock, D. 1996. *Project Management* (6th edition). Aldershot, UK: Gower.

Lockwood, D. & Sobol, M.G. 1989. 'IS spending survey: communications technologies dominate growth areas', *Journal of Systems Management,* December, 31–37.

Lukka, K. 1988. Budgetary biasing in organisations: theoretical frameworks and empirical evidence', *Accounting, Organisations and Society,* 13(3): 281–300.

Morgan, G. 1991. 'The bias factor', *Certified Accountant,* November.

Motta, G. 1988. 'A taxonomy of computerization expenditure in organizations', In N. Bjorn-Andersen & G. Davis (eds), *Information Systems Assessment: Issues and Challenges, Proceedings of the IFIP 8.2 Working Conference on Information Systems Assessment,* pp. 213–236, North-Holland, Amsterdam.

Parker, M. & Benson, R. 1988. *Information Economics: Linking Business Performance to Information Technology.* Englewood Cliffs, NJ: Prentice-Hall.

PC Week. 1997. 'TCO fact or illusion?', 14, 27 June.

Price Waterhouse. 1990. *Information Technology Review (1990/91).* London: Price Waterhouse.

Price Waterhouse. 1994. *International Technology Review (1993/94)*. Menlo Park, CA: Price Waterhouse.

Price Waterhouse. 1996. *Technology Review (1995/96)*. London: Price Waterhouse.

Puxty, A. 1985. 'A critical overview of agency theory', In A. Puxty (ed.), *Issues in Accountability XII: Critiques of Agency Theory in Accountancy*, pp. 1–24. Glasgow: Strathclyde Convergencies.

Remenyi, D. & Money, A. 1994. 'Service quality and correspondence analysis in determining problems with the effective use of computer services', *European Journal of Information Systems,* 3(1): 2–13.

Remenyi, D., Sherwood-Smith, M. & White, T. 1997. *Achieving Maximum Value from Information Systems.* London: John Wiley.

Remenyi, D. 1999a. *IT Investment – Making a Business Case.* Oxford: Butterworth Heinemann.

Remenyi, D. 1999b. *Stop IT Project Failure through Risk Management*. Oxford: Butterworth Heinemann, Oxford.

Remenyi, D., Money, A., Sherwood-Smith, M. & Irani, Z. 2000. *Effective Measurement and Management of IT Costs and Benefits.* Oxford: Butterworth Heinemann.

Ryan, S.D. & Harrison, D. 2001 'Considering social systems costs and benefits in information technology investment decisions: a view from the field on anticipated payoffs', *Journal of Information Systems Management,* 16(4): 11–40.

Sing, S.K. 1993. 'Using information technology effectively: organizational preparedness models', *Information and Management,* 24(3): 133–146.

Snyder, H. & Davenport, E. 1997. *Costing and Pricing in the Digital Age: a Practical Guide for Information Services*. London: Library Association Publishing.

Strassmann, P. 1988. 'Productivity', *ComputerWorld Extra*, 20 June, pp. 10–13.

Strassmann, P. 1990. *The Business Value of Computers.* New Canaan, CA: Information Economics Press.

Strassmann, P. 1997. *The Squandered Computer.* New Canaan, CA: Information Economics Press.

Turner, J.R. 1993. *The Hand Book of Project-Based Management.* Maidenhead, UK: McGraw Hill.

Turner, J.R. 1995. *Commercial Project Manager*. Maidenhead, UK: McGraw Hill.

Walker, K. & Johnson, E. 1999. 'The effects of a budget-based incentive compensation scheme on the budgeting behavior of managers and subordinates', *Journal of Management Accounting Research*, 11: 1–29.

Waltner, C. 1999. 'A hard look at soft benefits', *InformationWeek*, Manhasset, 24 May.

Watson, R., Pitt, L. & Kavan, B. 1998. 'Measuring information systems service quality: lessons from two longitudinal case studies', *MIS Quarterly*, 22(1): 61–69.

Wellman, F. 1992. *Software Costing: an Objective Approach to Estimating and Controlling the Cost of Computer Software.* New York: Prentice Hall.

Wentworth Research. 1996. Assessing IS Performance. Research Report, Wentworth Management Programme, November.

White, T. 2001. *Reinventing the IT Department.* Oxford: Butterworth Heinemann.

Willcocks, L. & Graeser, V. 2001. *Delivering IT and e-Business Value*. Oxford: Butterworth Heinemann.

Willcocks, L. & Lester, S. 1994. 'Evaluating the feasibility of information systems investments: recent UK evidence and new approaches', In L. Willcocks (ed.), *Information Management, the Evaluation of Information Systems Investments*, pp. 49–80. London: Chapman Hall.

When will employees perceive affirmative action as fair?

*Mariëtte Coetzee & L. Vermeulen**

Being fair is a central interest among today's managers concerned about providing equal employment opportunities, fair labour practices and paying a fair day's pay for a fair day's work. The differing perspectives, interests and goals of managers and employees, however, make it difficult to determine what employees regard as fair treatment. The multidimensionality of fairness is evident when one considers how people disagree when asked what is fair. The different answers to questions about the fairness of affirmative action depend on whether the focus is on outcomes, procedures or motives. The fairness of affirmative action should thus be determined by taking the distributive, procedural and interactional components of fairness into consideration.

From a distributive point of view, there is not much an organisation can do about the perceived fairness of a decision to appoint or promote people from previously disadvantaged groups, because legislation, such as the Employment Equity Act (No. 55 of 1998) and the Promotion of Equality and the Prevention of Unfair Discrimination Act (No. 4 of 2000) regulate this issue. There are, however, various ways in which affirmative action decisions can be made and implemented. To increase the perceived fairness of affirmative action decisions, organisations need to reconsider the way they implement affirmative action and treat employees. Research has shown that employees are more inclined to accept unfavourable outcomes if they are treated in a fair and respectful manner.

Introduction

The first democratic and multi-ethnic election on 27 April 1994 brought hope to thousands of South Africans. The government's affirmative action policy enabled workers to visualise a more prosperous future. Before the implementation of affirmative action, people from previously disadvantaged groups that attempted to uplift themselves from poverty through higher education were thwarted by discrimination, prejudice and institutional lag. Those seeking improved economic positions through employment were circumvented by a tradition of preferential treatment for whites. Despite being qualified by skill and competencies, blacks and other groups were forced to bargain in the labour market at a severe disadvantage. Caught in the web of prejudice and legal discrimination, they found that their ethnicity was reason enough for those in charge to deny them social and economic opportunities. When they did manage to secure a job offer, they were consigned to menial positions that paid them less than their white counterparts. These injustices not only led to poverty, but destroyed their spirit.

Since the general election in 1994, attempts to make South Africa a more just society have increased, and topics such as equality and social justice have frequently appeared at the top of all agendas. The government realised that legislation was necessary to guide organisations in promoting justice in the workplace, and the Constitution of the Republic of South Africa (No. 108 of 1996) was therefore used as the primary source of any legislation pertaining to the fundamental right of people to equality.

Some of the most important aims of the Constitution include rectifying injustices of the past and establishing a society based on democratic values, social justice and fundamental human rights (Van Wyk 2002: 42). The need to redress the injustices of the past becomes apparent when one considers the social and economic inequalities that still exist in South Africa, especially those that were generated by apartheid. Any attempts to redress inequalities, however, should be based on upholding the values of human dignity, equality, freedom and social justice in a united, non-racial and non-sexist society

* Mariëtte Coetzee is Senior lecturer: Human Resource Management, School for Business Management, University of South Africa. E-mail: coetzm@unisa.ac.za. Prof L. Vermeulen is Head: Department of Human Resource Management, University of Pretoria.
E-mail: lvermeul@hakuna.up.ac.za.

where all may prosper. If it is not done is this way, reverse discrimination and social and economic inequalities will continue to exist, and any attempts to create a just society, in which all people can live together in peace and harmony, will be doomed to failure.

The purpose of this article is to provide a theoretical overview of what is meant by organisational justice (organisational fairness) and how it relates to the affirmative action domain. It is hoped that managers and human resources practitioners will take note of the fairness principles and reconsider the way they have been implementing affirmative action and treating employees. Doing so may increase employees' perceptions about the fairness of affirmative action and enable the organisation to retain a committed workforce.

Affirmative action

In order to govern the promotion of social justice and eradicate inequalities in the workplace, the government realised that it had to intervene actively not only to prevent further discrimination, but also to purposefully promote the employment and advancement of persons disadvantaged by previous policies. Two acts have been implemented in this regard, namely the Employment Equity Act (No. 55 of 1998) and the Promotion of Equality and the Prevention of Unfair Discrimination Act (No. 4 of 2000). The underlying principle for the intervention of the state was that mere equality of opportunity would not be equitable, as many workers would start with a handicap, and that true equality and equity would be achieved only by strong measures against discrimination and by the purposeful, planned placement and development of persons that were denied equal opportunities in the past. Affirmative action thus became the vehicle for democracy in the workplace (Innes, Kentridge & Perold 1993: 79).

According to Bendix (2001: 435), 'affirmative action' refers to the purposeful and planned placement or development of competent, or potentially competent, persons in, or to, positions from which they were debarred in the past, in an attempt to redress past disadvantages and to render the workforce more representative of the population.

Most of the controversies and problems surrounding affirmative action arise not from the principle, as such, but from the manner in which affirmative action is implemented. Wrong implementation occurs because organisations regard affirmative action as a political imperative with which they have to comply, rather than as a business objective, which includes having as effective and competent a workforce as possible. Consequently, persons are appointed in 'affirmative action positions' merely to window-dress or to fill quotas, usually without due consideration of their suitability for the position or the possibility of support and development. Such arbitrary appointments leave other employees dissatisfied and are unfair to the appointees themselves, since they are either in meaningless positions or cannot handle their tasks, thus confirming the belief that affirmative action appointees are 'no good' (Bendix 2001: 440).

Another problem with affirmative action is the fact that the available pool of previously disadvantaged persons able to fill high-level jobs is extremely small. The result is that a small, highly sought after group of candidates develops, who are continually 'poached' by one organisation from another. Only this elite group thus advances, while the rest of the black African population remains where it was before. Employers should abandon the practice of looking for 'ready-made products' and instead develop persons for upward movement in the organisation. In such cases, affirmative action should be closely linked to the development of employees' skills, abilities and competencies (Thomas 2002: 239).

The most prevalent accusation directed at affirmative action initiatives is that they constitute reverse discrimination. There is a strongly held belief among the general public that affirmative action regulations force employers to appoint underqualified people from previously disadvantaged groups and women at the expense of qualified white males – a result both unfair in concept and detrimental to the competitiveness of the organisation. Affirmative action will become unfair only if previously disadvantaged people are appointed 'at all costs', without granting other people the opportunity to compete. All candidates have to be granted the opportunity to compete and to be assessed against pre-determined criteria, but an additional weighting can be placed on affirmative action aspects. This means that affirmative action candidates are given a slight, but not unreasonable, advantage over the other candidates (Charlton & Van Niekerk 1994: 91).

The success of affirmative action

There has been much advice on how to design and implement affirmative action programmes effectively. Generally, effectiveness is defined as the proportion of employees from previously disadvan-

taged groups that are hired or promoted. Focusing on numbers only is a limited view of effectiveness, however. The perceived fairness of affirmative action practices should also be an important indicator of effectiveness. Social scientists have long recognised the importance of justice as a basic requirement for the effective functioning of organisations and the personal satisfaction of the individuals they employ. One of the most important benefits of organisational justice conceptualisations is that they may be used to explain a wide variety of organisational behaviours. When practices are perceived to be unfair, they cause frustration, threaten employees' self- and social images, and, in some circumstances, produce moral outrage (Greenberg 1996: 8). According to Robinson & Morrison (1995: 291), other adverse effects of violation on employee perceptions of fairness include lowered trust in management and lowered organisational commitment. Hence, just as the injustices associated with selection systems (Gilliland 1993: 711), pay raise decisions (Folger & Konovsky 1989: 123) and other organisational phenomena foster job dissatisfaction and voluntary turnover, the perceived injustices resulting from affirmative action are likely to translate into dislike for a job and to prompt a subordinate to seek alternative employment. Violations of fairness can further result in legal battles, more negative attitudes towards the organisation, decreased job satisfaction, lower self esteem and lower probability that the affirmative action programme will succeed.

South African organisations will continuously be evaluated in terms of how well they meet employment equity targets. This will ensure that the changing nature of society is reflected in the composition of an organisation's workforce. As mentioned earlier, for a programme to be regarded as effective, it should comply with legal requirements, but also satisfy the requirements of fairness. Organisations will thus be under increasing pressure to make use of affirmative action programmes that are technically and morally sound – and can be shown to be so. This is of particular importance if one considers the fact that employees will be more inclined to challenge procedures that they regard as unfair (Cooper & Robertson 1995: 7).

Organisational justice

Being fair is a central interest among today's managers concerned about providing equal employment opportunities and fair labour practices and paying a fair day's pay for a fair day's work. Just as referees should ensure that all participants have a fair chance to compete, managers are responsible for the fair treatment of employees. The differing perspectives, interests and goals of managers and subordinates, however, make it difficult to determine what employees regard as fair treatment. The multidimensionality of fairness is evident when one considers how people disagree when asked what is fair. The different answers to questions about fairness depend on whether the focus is on outcomes, procedures or motives. A further complicating element is the possible interaction of a concern for justice with other motives in social situations (for example, self-esteem, interpersonal attraction) (Folger & Cropanzano 1998: 84). An attempt to describe and explain the role of fairness in the workplace is known as organisational justice (Greenberg 1996: 24). Organisational justice refers to the decisions organisations make, the procedures they use in making decisions and the interpersonal treatment employees receive.

Distributive justice

For many years, the study of fairness in organisations was dominated by a distributive justice orientation, an approach that focused on the fairness of outcomes and decisions. Folger & Cropanzano (1998: 33) defined distributive justice as the ''perceived fairness of the outcomes or allocations that an individual receives''. According to Leventhal (1976: 220) decisions or outcomes are determined by making use of three major rules of justice: the equity rule, the needs rule and the equality rule. The equity rule focuses on contributions, the needs rule is applied for reasons of personal welfare, and the equality rule is used to preserve social harmony. From an affirmative action point of view, the equality rule should thus be used to make decisions.

Procedural justice

As the distributive perspective gained dominance, an independent approach to the study of justice began to develop. Soon studies of reactions to the procedures used to reach decisions were conducted. Researchers became interested in expanding the distributive justice orientation to include a consideration of the methods, mechanisms and processes used to determine outcomes (in other words, adopting a procedural justice orientation). According to Skarlicki & Latham (1996: 164, quoted in Ivancevich & Matteson (2002: 36)), procedural justice refers to the extent to which fair procedures and processes are in place and adhered to, as well as the extent to which individuals regard their

leaders as being fair and sincere and believe that they are logical in what they do, and have a rationale for doing so.

Thibaut & Walker conducted research in 1975 to study employees' reactions to the dispute resolution process, and this led their developing a theory of procedural justice. According to these authors, employees judge the fairness of procedures according to two types of control: the amount of control they, as employees, have over the procedures used to make a decision (referred to as process control) and the amount of control they have over influencing the decision (referred to as decision control) (Thibaut & Walker 1978: 592). People want procedures that allow them to feel that they have participated in developing a decision that will affect them. The opportunity to voice their opinion thus gives them the opportunity to influence others' decisions. Further research revealed that procedures that provided employees with opportunities to influence a decision were perceived as fairer than procedures that denied process control.

Related to the study on control over processes and decisions, Lind & Tyler (1988: 76) developed the self-interest and the group-value models of procedural justice. The self-interest model suggests that people seek decision control because they are concerned with maximising their own outcomes. However, when individuals have to cooperate with others in groups to achieve outcomes, the group-value model comes into play, and the focus shifts from decision control to process control. Procedures are then regarded as more important in attaining fair or favourable outcomes. The group-value model explains the value-expressive effects of process control. Group identity and group procedures are two elements that govern the functioning of groups. People consider procedures that allow them to express their opinions (voice) to be fair, in that they can participate in group processes as valued group members. Even if 'voice' does not produce a favourable outcome, it enhances perceived procedural justice because its value-expressive function confirms the values of group participation and group membership status.

During further research, Tyler & Lind (1992: 137) developed a relational model of authority in groups. According to the authors, three relational concerns with the authority affect procedural justice judgments, namely: trust, neutrality and standing. Trustworthiness can be measured by rating the extent to which the manager behaves fairly and ethically. If the manager behaves fairly and considers the needs and views of the individual, then he/she can be trusted. Trust involves beliefs about the manager's intentions. Neutrality can be judged in terms of the person's unbiased decision-making on the basis of facts. Standing refers to status recognition that is indicated to people by a manager who treats them with dignity, politeness and respect for their rights. By means of the procedures that a manager uses, individuals can judge whether they will be treated fairly on the basis of the manager's relational concerns of trust, neutrality and standing.

Related to Thibaut & Walker's research on the importance of process and decision control in perceptions of fairness perceptions, Leventhal, Karuza & Fry (1980: 86) identify six rules of justice that are used to determine whether procedures are fair. Procedures are regarded as fair to the extent that they:

- Suppress bias
- Create consistent allocations
- Rely on accurate information
- Are correctable
- Represent the concerns of all recipients
- Are based on moral and ethical standards.

The importance of procedural justice is emphasised by the positive impact it has on employees' behaviour. According to Greenberg (1990: 415), the positive consequences of procedural justice include:

- Organisational commitment
- Intent to stay with the organisation
- Organisational citizenship
- Trust in the supervisor
- Satisfaction with decisions made
- Work effort
- Performance.

Interactional justice

As the original conceptualisations of procedural justice have been expanded through research, it has become clear that perceptions of procedural justice are influenced by factors that go beyond the formal procedures used to resolve disputes or allocate rewards. In particular, it has been demonstrated that judgments of procedural justice are influenced by two important factors: the interpersonal treatment people receive from decision-makers and the adequacy with which formal decision-making procedures are explained. Bies & Moag (1986: 44) use the term interactional justice to refer to people's sensitivity to "the quality of interperso-

nal treatment they receive during the enactment of organisational procedures'' and identify the following four attributes of interpersonally fair procedures:

- Truthfulness
- Respect
- Propriety of questions
- Justification.

The first three attributes deal with the nature of the communication while it is occurring. The last one (justification) has to do with removing any discontent following an unfair procedure:

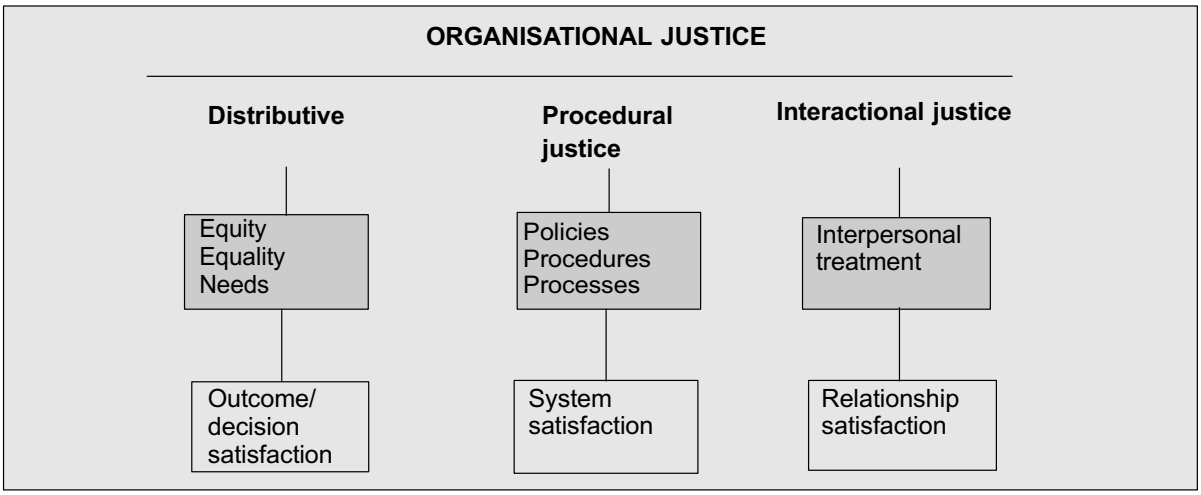
- Truthfulness: Truthfulness has two components: deception and candidness. Employees do not like being deceived and expect to be treated in a forthright manner. Organisations should therefore provide them with realistic and accurate information.
- Respect: Individuals expect to be treated in a polite and respectful manner. This means that insults or discourteous behaviour should be avoided at all cost.
- Propriety of questions: The propriety of questions refers to two aspects. Firstly, questions should not be considered improper by their very nature, and secondly, they should not involve prejudicial statements.
- Justification: Justification comes into play following negative outcomes or unfair treatment. It may be possible to rectify an injustice with an adequate justification. According to Bies & Shapiro (1988: 682), a sense of anger over injustice can be reduced or eliminated by providing the unfairly treated individual with a social account, such as an explanation or an apology. People expect that events affecting them will be explained. If they do not receive an explanation, they doubt whether they have been treated in accordance with a socially rooted expectation for fair processes in human interaction (Weaver 2001: 3). Respect and concern constitute informal social goods, and failing to receive them is regarded as a violation of expectations of justice.

Studies cited by Folger & Cropanzano (1998: 72) show that conflict, low performance and poor attitudes tend to characterise insensitively treated individuals. Figure 1 illustrates the various types of justice and their interrelatedness.

Given that the distinction between distributive, procedural and interactional justice has now been empirically established, the stage is set for researchers to consider how these varieties of justice relate to various organisational variables, such as turnover, commitment, trust and relationships between supervisors and subordinates. According to Chan (2000: 4), injustice can bring about negative consequences, such as reduced job performance (Greenberg 1988: 609; Pfeffer & Langton 1993: 398), less cooperation with co-workers (Pfeffer & Langton 1993: 403), reduced quality of work (Cowherd & Levine 1992: 314), stress (Zohar 1995: 491) and theft (Greenberg 1990: 215). By understanding how perceptions of organisational justice are related to these variables, organisations should attempt to manage employees' perceptions of fairness and influence their performance.

The fairness of affirmative action

In South Africa, no issue has raised more concerns about justice than affirmative action. For many



Source: Adapted from Chan (2000: 5)

Figure 1: Organisational justice

years, affirmative action has been a battleground for competing values, especially competing concepts of distributive justice (Van Jaarsveld 2000: 24). In the USA, after twenty or more years of affirmative action, it has now been admitted that the process has failed. Why? Because affirmative action programmes were not implemented in a fair manner; neither were employees' perceptions regarding the fairness of the affirmative action programmes managed. As soon as employees regard something as unfair, they will reject it, and any further interventions will be doomed to failure. If South Africa wishes to make a success of affirmative action, organisations should understand how perceptions of affirmative action influence employees' attitudes and behaviour and consequently impact on the success of the organisation. Although significant progress has been made in restructuring and transforming South African society and its institutions, inequalities and unfair discrimination remain deeply embedded in social structures, practices and attitudes, undermining the good intentions of the country's constitutional democracy (Van Wyk 2002: 43). A special attempt should thus be made to eliminate discrimination and manage employees' resistance to change.

According to Folger & Skarlicki (1999: 35), employees' negative feelings or resistance to change can be overcome by applying principles of fairness. Such principles provide an opportunity to mitigate some of the adverse organisational consequences of individuals' resentment-based resistance to change. An organisation's obligations towards employees entail more than fair treatment with respect to the salaries and benefits given in exchange for labour (distributive justice), however, and more than fair treatment with respect to the implementation of policies and procedures that determine those levels of compensation (procedural justice). In addition, a moral obligation exists to treat an employee with sufficient dignity as a person (interactional justice). As organisations in globally competitive markets are less able to offer traditional rewards (lifelong employment, promotions, long-term compensation), one of the only means they have for inducing employees to stay is an environment that communicates that it values the employees. If this is the case, interactional justice plays a major role in influencing employees' attitudes and the behaviours required for successful performance, even under conditions of adversity and loss, which is often the case with affirmative action.

If a decision to appoint or promote an affirmative action candidate is done in accordance with the provisions of the Employment Equity Act (No. 55 of 1998), there is not much an organisation can do about the perceived fairness of such a decision. There are, however, various ways in which such a decision can be made, and organisations therefore need to ensure that the procedures, policies and processes they use in making appointment or promotion decisions are procedurally fair, and are viewed as such by employees.

Several principles of procedural justice are cited in the literature. The following list summarises the principles of fairness referred to in seven recent articles on procedural justice (De Witt 1998: 11; Gopinatha & Becker 2000: 72; Harris 2000: 153; Konovsky 2000: 504; Saxby, Tat & Johansen 2000: 211; Simerson, L'Heureux, Beckstein, ZiaMian, Dembowski & Freshman 2000: 448; Tang & Sarsfield-Baldwin 1996: 30; Tata 2000: 264):

- Provide advance notice of intent or decision
- Provide accurate information and adequate feedback
- Support two-way communication
- Explain and justify decisions
- Allow employees to influence the decision
- Consider the interests, views and concerns of all recipients
- Permit appeal, review, reconsideration and correction
- Treat employees with dignity, respect and sensitivity
- Apply administrative procedures consistently.

This list demonstrates the importance of interpersonal relationships in fostering perceptions of fairness. This is understandable if one considers the fact that employees define their relationship with the organisation as one of social exchange. The more an organisation is able to treat its employees in a socially acceptable way, the more employees will accept unfavourable outcomes. A corollary of this implication is that perceptions of fairness based on interactional justice may be the easiest perceptions of fairness to manage. Distribution of outcomes may be constrained by forces outside the manager's control. Similarly, the presence or absence of fair procedures may be a function of organisational policy. In comparison, the fairness of the interactions between managers and employees is often a matter of a manager being truthful, treating employees with respect and providing reasons or explanations for perceived injustices.

Summary

There is clear evidence that people care about justice. Organisations thus have to pursue justice, as measured by reality as well as by perceptions. Unfortunately, the conclusion that it is better to be behaviourally just is too simple. In pursuing principles of distributive justice, it is not possible to achieve all criteria simultaneously. The base criteria of equity, equality and need are incompatible. Thus, organisations are left with a superordinate problem: how to achieve balance among the three principles of distributive justice. One possibility is to determine which of the goals is most important in any given situation.

As regards procedural justice, organisations have to ensure that the procedures, processes and policies they use in making decisions are fair. Employees judge the fairness of procedures according to two types of control they have: the amount of control they have over the procedures used to make a decision, and the amount of control they have over influencing the decision. Procedures are regarded as fair to the extent that they suppress bias, create consistent allocations, rely on accurate information, are correctable, represent the concerns of all recipients, and are based on moral and ethical standards.

A third type of justice, namely interactional justice, refers to the interpersonal treatment employees receive during the enactment of organisational procedures. Procedures are seen as interpersonally fair when they make provision for truthfulness, allow for the respectful treatment of employees, and provide reasons or explanations for perceived injustices.

To thus answer the question: 'When will affirmative action be fair?', organisations need to ensure that any affirmative action appointments are based on valid and fair criteria, that the procedures and processes they use in appointing or promoting affirmative action appointees are applied consistently and are based on accurate, unbiased information, and most importantly, that they treat all employees with dignity and respect.

References

Bendix, S. 2001. *Industrial Relations in South Africa* (4th edition). Lansdowne: Juta.

Bies, R.J. & Moag, J.S. 1986. 'Interactional justice: communication criteria for fairness', In B. Sheppard (ed.), *Research on Negotiation in Organizations*. 1, 43–55.

Bies, R.J. & Shapiro, D.L. 1988. 'Voice and justification: their influence on procedural fairness judgments', *Academy of Management Journal,* 31: 676–685.

Chan, M. 2000. 'Organizational justice theories and landmark cases', *International Journal of Organizational Analysis,* 8(1).

Charlton, G.D. & Van Niekerk, N. 1994. *Affirming Action – beyond 1994: Laying the Foundation for Comprehensive and Effective Affirmative Action.* Kenwyn: Juta.

Cooper, D. & Robertson, I.T. 1995. *The Psychology of Personnel Selection*. London: Routledge.

Cowherd, D.M. & Levine, D.I. 1992. 'Product quality and pay equity between lower-level employees and top management: An investigation of distributive justice theory', *Administrative Science Quarterly*, 37: 302–320.

De Witt, R. 1998. 'The influence of eligibility on employees' reactions to voluntary workforce reduction', *Journal of Management,* September: 1–19.

Folger, R. & Cropanzano, R. 1998. *Organizational Justice and Human Resource Management.* Thousand Oaks, CA: Sage.

Folger, R. & Konovsky, M.A. 1989. 'Effects of procedural and distributive justice on reactions to pay raise decisions', *Academy of Management Journal*, 321: 115–130.

Folger, R. & Skarlicki, D. 1999. 'Unfairness and resistance to change: hardship as mistreatment', *Journal of Organizational Change Management,* 121: 35–50.

Gilliland, S.W. 1993. 'The perceived fairness of selection systems: an organizational justice perspective', *Academy of Management Review,* 18: 694–734.

Gopinatha, C. & Becker, T.E. 2000. 'Communication, procedural justice and employee attitudes: Relationships under conditions of divestiture', *Journal of Management.* 26(1): 63–84.

Greenberg, J. 1988. 'Equity and workplace status: a field experiment', *Journal of Applied Psychology,* 73: 606–613.

Greenberg, J. 1990. 'Organizational justice: yesterday, today, and tomorrow', *Journal of Management,* 16(2): 399–432.

Greenberg, J. 1996. *The Quest for Justice on the Job: Essays and Experiments.* London: Sage.

Harris, L. 2000. 'Procedural justice and perceptions of fairness in selection practice', *International Journal of Selection and Assessment,* 8(3): 148–157.

Innes, D., Kentridge, M. & Perold, H. 1993. *Reversing Discrimination: Affirmative Action in the Workplace.* Cape Town: Oxford University Press.

Ivancevich, J.M. & Matteson, M.T. 2002. *Organizational Behavior and Management* (6th edition). New York: McGraw-Hill.

Konovsky, M.A. 2000. 'Understanding procedural justice and its impact on business organizations', *Journal of Management,* 26(3): 489–512.

Leventhal, G.S. 1976. 'Fairness in social relationships', In J.W. Thibaut, J.T. Spence & R.C. Carson (eds), *Contemporary Topics in Social Psychology.* Morristown, NJ: General Learning Press.

Leventhal, G.S., Karuza, J. & Fry, W.R. 1980. *Beyond Fairness: a Theory of Allocation Preferences. Justice and Social Interaction.* New York: Springer-Verlag.

Lind, E.A. & Tyler, T.R. 1988. *The Social Psychology of Procedural Justice.* New York: Plenum.

Pfeffer, J. & Langton, N. 1993. 'The effects of wage dispersion on satisfaction, productivity, and working collaboratively: Evidence from college and university faculty', *Administrative Science Quarterly,* 38(3): 382–407.

Robinson, S.L. & Morrison, E.W. 1995. 'Psychological contracts and OCB: the effect of unfulfilled obligations on civic virtue behavior', *Journal of Organizational Behavior*, 16: 289–298.

Saxby, C.L., Tat, P.K. & Johansen, J.T. 2000. 'Measuring consumer perceptions of procedural justice in a complaint context', *Journal of Consumer Affairs,* 34(2): 204–216.

Simerson, G., L'Heureux, T., Beckstein, B., ZiaMian, M., Dembowski, J. & Freshman, M. 2000. 'What principles are used to judge the fairness of retrenchment actions?' *Journal of Business and Psychology,* 14(3): 443–458.

Skarlicki, D. & Latham, G. 1996. 'Increasing citizenship behavior within a labor union: A test of organizational justice theory', *Journal of Applied Psychology,* 81(2): 161–169.

Tang, T.L. & Sarsfield-Baldwin, L.J. 1996. 'Distributive and procedural justice as related to satisfaction and commitment', *SAM Advanced Management Journal,* 61(3): 2.

Tata, J. 2000. 'Influence of role and gender on the use of distributive versus procedural justice principles', *Journal of Psychology,* 134(3): 261–269.

Thibaut, J. & Walker, L. 1978. 'A theory of procedure', *California Law Review,* 66: 541–666.

Thomas, A. 2002. 'Employment equity in South Africa: lessons from the global school', *International Journal of Manpower,* 23(3): 237–255.

Tyler, T.R. & Lind, E.A. 1992. 'A relational model of authority in groups', *Advances in Experimental Social Psychology*, 25: 115–191.

Van Jaarsveld, I.L. 2000. 'Affirmative action: a comparison between South Africa and the United States', *Managerial Law,* 42(6).

Van Wyk, M.W. 2002. 'Conceptions of equality and social justice: a philosophical overview with reference to South Africa', *Southern African Business Review,* 6(2): 42–51.

Weaver, G.R. 2001. 'The role of human resources in ethics/compliance management: a fairness perspective', *Human Resource Management Review,* 11(1/2).

Zohar, D. 1995. 'The justice perspective of job stress', *Journal of Organizational Behavior*, 164: 487–495.

South African statutes

Constitution of the Republic of South Africa (No. 108 of 1996).

Employment Equity Act (No. 55 of 1998).

Promotion of Equality and the Prevention of Unfair Discrimination Act (No. 4 of 2000).

Database marketing: application by South African financial institutions

Johannes J. van der Walt & Werner Vermeulen[*]

When this empirical research was carried out, 119 financial institutions were registered with the South African Registrar of Banks and Financial Institutions. The responses from institutions that participated in the research project indicate that South African financial institutions lack the ability and willingness to implement a successful database marketing strategy. The reasons can be found in a misunderstanding of what database marketing entails, as well as a lack of acknowledgement of the important role that database marketing can play in the marketing strategy of financial institutions. Moving the marketing function to a higher level in the organisational hierarchy and appointing database managers with a thorough knowledge of marketing could serve to eliminate this problem.

Keywords: Database marketing, individual clients, individualised information, database management, database technology

Introduction

Database marketing is essentially a unique way for businesses to build long-term business relationships with customers/clients by capturing detailed information about individuals and customising their products to individual needs and tastes. The data are stored in a database, which must be easily accessible to all frontline personnel after being processed into useful marketing information.

Database marketing grew from direct marketing and therefore shares many of the same principles (Nash 1993: 1). It should not be confused with direct marketing, however. The main difference between direct marketing and database marketing is the extreme emphasis on individual clients (Roberts 1997: 27). In the case of direct marketing, a client list is used to select certain clients to contact with an offering. Although the clients may be segmented according to discriminating factors, this still represents a group of people with a common identifier. During database marketing, individual clients are identified first and for practical considerations can be grouped together, but are contacted on an individual basis, if possible. During segmentation, the common characteristics of the members of the segment are emphasised. Segmentation therefore leads to the development of a general product or service that will satisfy most of the needs of the group. During database marketing, the needs of each individual are identified and satisfied on an individual basis. Although some of the elements of the product/service can be generic, it will in almost all cases be possible to customise the product or service to individual needs. It is therefore necessary to be in possession of intimate details of clients, so that the selection of individuals can be based on those details. The detailed information also serves as the basis for customising the institution's offering to clients (Shepard 1995: 14).

It is important for financial institutions to make use of all marketing tools to gain that illusive competitive edge in the very competitive financial services market. Database marketing is a tool that can support the marketing function of financial institutions. On the negative side, the very expensive operation of building a comprehensive database can be wasted if not implemented according to a set of guiding principles.

The South African financial industry

During the years of political and economic isolation, the South African financial industry fell behind developments in the major industrial countries of

* Dr Johannes J. van der Walt is Senior lecturer: Marketing at the School for Entrepreneurship and Business Development, Technikon Free State. E-mail: jvdwalt@tfs.ac.za. Dr Werner Vermeulen is Manager: Quality Improvement, Technikon Free State. E-mail: werner@tfs.ac.za.

the world (Rudolph 1997a), including database marketing developments. According to Chen & Sakaguchi (2000: 66), the ability of banks to gather detailed client information will determine whether they will be able to survive in the modern competitive environment.

South African financial institutions do have national client databases and are therefore ideally positioned to exploit the database marketing opportunity. Historically, South African financial institutions have made use of the branch system, with branches spread all over the country. Each branch has its own database, which is used for local applications (Rudolph 1997b). A complicating factor that has not been addressed in this research is the difficulty of identifying and adding the details of potential clients that live in the underdeveloped areas of South Africa to the database. According to the 1996 census, (Statistics South Africa), 35% of the South African population of 45 million live in informal dwellings or shacks, without addresses or telephones and sometimes without even electricity.

The concept of database marketing is grossly misunderstood, and the key aspect of individuality is therefore mostly ignored in South Africa. Financial institutions spend millions of rands on creating databases, but because of a lack of understanding of the real issues, these are not utilised to their full potential.

The geographical structure creates a problem in supplying real-time information to marketers at the various client contact centres (Talmor 1994: 84). The database should be developed in such a way that it can feed on data from remote centres and dispatch information back to those centres without delay. At the same time, all data should be compiled and processed at a central location to ensure a complete picture to everyone that needs the information (Davis 1993: 39).

Important aspects in the application of database marketing

Database marketing is an important tool to eliminate junk mail. Any unsolicited marketing communication with the client is known as junk mail. It not only annoys clients to receive offerings that are not applicable to their specific circumstances, but also wastes marketing resources. One of the main purposes of database marketing is to have detailed information about clients in the database to enable them to create offerings that meet the specific needs of that client.

Database marketing must therefore include aspects such as proper research into client's needs and circumstances, taking cognisance of privacy matters. The development of an appropriate database, as well as a data processing system that can supply information to support marketing efforts aimed at individual clients, is mandatory. The executive in control of the database marketing programme and the person responsible for the database have an important influence on database marketing and are therefore included in the research.

Purpose of the article

The purpose of the article is to make financial institutions aware of the usefulness of a database as a marketing tool, which is grossly misunderstood and underrated.

Problem statement

The key problem investigated in this research was the ignorance and/or misinterpretation of database marketing by financial institutions of South Africa, even though this industry has all the requirements to utilise databases to their fullest extent as marketing tools.

Methodology

The research focused on all of the locally registered banks and insurance institutions in South Africa, totalling 119 at the time of the survey. It was decided to include all of them, because the numbers are small and it seemed appropriate to establish the stance of each institution. A 32% response was received to the 119 questionnaires sent out. A special and successful attempt was launched to ensure that all the major financial institutions participated in the research, as indicated in Tables 1 and 2.

The respondents selected were the marketing and/or database manager with the highest rank in the institution. It was decided to approach the manager in the highest position, who would have the best overall knowledge of the operations of their respective departments. The specific individuals were identified by telephonic contact with the various institutions.

A structured questionnaire was developed and tested for ambiguity and other possible problems.

Table 1: Market share of some of the banks that participated in the survey

Institution	Investments by institution %	Credit cards %	Lease %	Advance loans %	Mortgages %	Total assets %
ABSA	26	20	20	18	30	21
First National	28	21	30	19	10	18
Standard Bank	22	23	25	19	15	19
***Total %**	**76**	**64**	**75**	**56**	**55**	**58**

* Total percentage represents the percentage market share of the institutions mentioned in the table out of the total of all South African financial institutions

Source: KPMG Banking Survey (2001)

Table 2: Market share of some of the long-term insurance institutions that participated in the research

Institutions	Premium income %	Payouts %
Investec	9	8
Old Mutual	22	27
Sanlam	19	25
***Total %**	**50**	**60**

* Total percentage represents the percentage market share of the institutions mentioned in the table from the total of all South African financial institutions

Source: KPMG Banking Survey (2001)

The questionnaire method was chosen because of the widespread geographical nature of the South African financial institutions, which made it impossible to visit each of them.

The questionnaire contains Yes-No questions as well as Likert-scale and open-ended questions. The main areas tested were the level of importance of the marketing function, the structure and management of the database, the attitude of top management towards database marketing, the technology, and the problems experienced by the database marketers.

Responses from the participants were processed by making use of the Statistical Programme for Social Sciences (SPSS). Frequency tables were drawn up and evaluated for significant deviations. Hypothesis testing was done by means of chi-square tests. In the event of a finding of a p-value of less than 0.1, the nil hypotheses was rejected, and it was accepted that a relationship exists between the factors tested.

Findings

The empirical research revealed the following:

- The use of databases by financial institutions was measured to determine whether they were using a database, and if so, for what functions. As indicated in Table 3, the important support of the database to market products and services is often ignored. For example, only 9% of the financial institutions that participated in the survey use a database to create models to evaluate the susceptibility of clients to use a certain product or service. The database is used by only 59% to stimulate repetitive transactions. Thirty-two per cent did not use the database to identify the specific needs of clients so that

products and services could be tailor-made to their specific needs. These uses of the database form the core of database marketing.

Segmentation of the market is indicated by 82% of the respondents as a use of the information in the database. Although segmentation is a very useful tool in traditional marketing, the aim of database marketing is to target the individual client. Financial institutions should therefore not stop at segmenting the market, but continue to create client profiles to enable the business to identify individuals to target with an offering. Only 45% of the respondents indicated that they continue to develop client profiles after segmentation. After identifying the specific individuals to target, one-on-one communication should take place between the client and the institution to customise the offering and close the deal. Only 4% of the respondents use the database for communication. It is clear from the above information that very few institutions are truly practising database marketing.

- For database marketing to be effective, all relevant information must be available to all contact personnel at all service points. Emphasis is placed on the word 'all'. Decision-makers at service points frequently have to make split second decisions regarding transactions with clients. Examples of this include decisions on making a loan or vehicle finance available to a client. In the past, the approval of client applications tended to be a lengthy process, which clients found annoying. With all information available to personnel at service points, client applications can be approved (or turned down) immediately. A database structure should be used to support this function. In this research project, it was determined whether a database structure was implemented and what problems institutions experienced with it. Table 4 indicates that 65% of financial institutions experience problems with database marketing because they do not have an integrated database. Sixty-one per cent experience problems with database marketing because the data are not always relevant and recent.
- The survey included questions to determine who is the executive responsible for database marketing. The results indicated that financial institutions use information technology experts as database managers, rather than marketing experts. Table 4 shows that 40% of financial institutions have problems with database marketing because they use a database manager without marketing experience. This leads to problems, because the data are not always processed into useful marketing information, but rather into complicated statistical information that does not relate to market products and services.
- Questions relating to the position of the database manager in the hierarchy of the financial institution were included in the survey. The level of the database manager can be related to how important the institution regards database marketing to be. The lower the level, the less importance attached to database marketing. Table 5 indicates a bleak scenario in this regard. In 60% of cases, the databases manager has a lower position than the national marketing manager.

  In very few financial institutions is the position of the database manager located at top-management level. The marketing function

Table 3: Uses of the database

Use	Application: Respondents %
Segmentation of the market	82
Cross-selling of products and services	74
Products/services tailor-made to specific needs	68
Enhancement of repetitive transactions	59
Creation of client profiles for marketing	45
Creation of models for forecasting	9
Communication	4

Table 4: Problems that financial institutions experience with database marketing

Problems with database marketing	Yes Respondents %	No Respondents %
Several databases not integrated with one another	65	35
Problems to keep data relevant and recent	61	39
Database manager is an information technology expert without a marketing background	40	60

in many institutions is not even at the same level as other functional heads. This can be directly attributed to a misunderstanding of what database marketing really entails. If the benefits of database marketing are realised, the spearhead of such an important marketing function should be at top level.

- The budgets of some of the participants also indicate that database marketing is not important to their institutions. Insufficient funds are budgeted for building a database and training all personnel in its application. Going only half way to establishing a good database is as good as going without one. The database must be developed to suit the purposes of database marketing, for which money is needed.

The database

Historically, South African financial institutions have made use of the branch system, with branches spread all over the country. Each branch has its own database, which is used for local applications (Rudolph 1997b). It is very difficult to merge these databases into one and to allow distant client contact centres access to the data. Linked to this problem is that of having different databases for different functional units, with the same merging problems.

The data from the accounting or financial departments were in many cases used to form the basis of the new aggregated database. The content of these data and the format in which they are presented do not suit database marketing purposes. The appointment of a database manager with a marketing background can solve the problem of irrelevant data being captured into the database, as well as ensuring that the data are processed in a form that is suitable for database marketing.

Database management

The management of databases relates to two different problems. Firstly, the management is often situated at a low level in the businesses hierarchy. The database managers do not have sufficient input to the budgets and lack the necessary support to implement a successful database marketing strategy.

Secondly, in almost all institutions, an information technology specialist is appointed to the position of database manager. Direct input from marketing experts is an absolute necessity in the development of a database and its applications. Appointing information technology experts without an in-depth knowledge of marketing leads to the inability of financial institutions to implement database marketing successfully.

Table 5: Position of the database manager in the hierarchy of the institution

Position of the database manager	Respondents %
Similar to national marketing manager	40
Similar to regional marketing manager	10
Similar to branch manager	10
Lower than branch manager	40
Total	**100**

Only general functions are supported by the databases of South African financial institutions, such as segmentation and cross-selling. Supporting functions are usually lacking, such as developing tailor-made services and products to the needs of clients, enhancing repeat selling and building client models to identify clients' susceptibility to buying specific products and services. This can be attributed to the lack of acknowledgement of the importance of database marketing as part of the marketing strategy of financial institutions.

Data source strategy

Several methods should be used to solicit information from clients. These methods include telephone calls, letters, newspapers and magazines, personal interviews, Internet and e-mail. The internal reporting and financial systems of institutions are among the most important information sources for database marketing. Although all these sources are available, they are not necessarily used effectively.

To find all appropriate and useful data, a structured data source strategy should be employed. The basis of such a data source strategy is the concept that all employees should have direct access to the database, so that they can immediately update the data in the database while in contact with the client. If institutions operate with remote client contact centres, they should develop a system to update the data collected from these remote contact centres in real time.

Database software

It seems that South African financial institutions tend to use general database software that has not been developed specifically for the needs of a particular institution. Although less expensive than customised database software, it sacrifices the advantage of developing specific models and other supporting information to exploit database marketing effectively.

Measurement of the success of the campaigns

Very few South African financial institutions measure the effectiveness of database marketing campaigns. Specific success indicators should be identified and closely monitored to ensure that success is achieved. Because the purpose of database marketing is individual contact with clients, the reaction of every client to a specific campaign can be monitored.

Recommendations

Without acknowledging the important role that database marketing can play in the marketing strategy of South African financial institutions, database marketing is doomed in this industry. Only by acknowledging the important role of database marketing will the necessary support and financial resources be made available to develop a proper database marketing system for financial institutions in South Africa. Against the background of this statement, the following recommendations are made:

- A prerequisite for the application of database marketing in financial institutions in South Africa is that senior management make a paradigm shift away from the rigid channels of marketing to accommodate the very radical change in marketing ideas as well as to supply the necessary financial support for the building and implementation of the infrastructure for effective database marketing.
- The whole database marketing system should be market driven and not information technology driven. Information technology experts are apt to focus on the data with insufficient attention to marketing. It is therefore recommended that financial institutions in South Africa recruit or develop their own marketing experts with an information technology background, called a database-marketing manager. A person with skills in both areas (marketing and information technology) will be a crucial element in the successful implementation of database marketing in financial institutions. A person with knowledge in both areas should be in a position to oversee the creation of a database and applications with a marketing supportive output.
- Database-marketing managers should be positioned at top level to ensure thorough attention to the aspects that are most needed. If positioned at a lower level, higher level managers that do not have hands-on experience of database marketing practices can easily end up making the decisions.

Conclusion

Although high costs are involved, it is essential that an integrated database be developed, in which information is categorised specifically for marketing purposes. Every institution must define the specific categories needed for marketing by that specific institution. Once in the correct format for marketing, the information must be made directly and immediately available to all

frontline personnel of that institution, even those in remote contact centres. The frontline personnel must be able to view the information in real-time, while in contact with the client.

References

Chen, L. & Sakaguchi, T. 2000. 'Data mining methods, applications, and tools', *Information Systems Management,* 17(1): 65–70.

Davis, R.A. 1993. 'Banking on databases for financial growth', *Advertising Age,* 15 February, 64(7): 39.

KPMG Banking Survey. 2001. Accessed at www.kpmg.co.za/Modules/Library/detail.cfm?libid=153&year=2001.

Nash, E.L. 1993. *Database Marketing: the Ultimate Marketing Tool.* New York: McGraw-Hill.

Roberts, M.L. 1997. 'Expanding the role of the direct marketing database', *Journal of Direct Marketing,* 11(4): 26–35.

Rudolph, H. 1997a. 'Database marketing and relationship management – Part 1: How to make it work', *Marketing Mix,* January/February: 15.

Rudolph, H. 1997b. 'Database marketing and relationship management – Part 5: A change in direction: adjusting the organisational structure', *Marketing Mix,* June: 21.

Shepard, D. 1995. *The New Direct Marketing* (2nd edition). Chicago: Irwin.

Statistics South Africa. Archive accessed at www.statssa.gov.za/Archives/ArchivesReports.htm.

Talmor, S. 1994. 'Hit and miss', *The Banker,* 144, April: 84–86.

The efficiency of South African banks: a non-parametric approach

J.J.L. Cronje[*]

This paper is the second in a series of two. The first paper[1] *addressed the practical application of data envelopment analysis (DEA) as a functional approach to measuring the technical efficiency of banks. This paper presents the findings of an exercise that applied DEA to 16 South African banks. Nine of the banks were classified as inefficient. The research shows that inefficiency can be ascribed primarily to scale (size) inefficiency. The large banks that were classified as inefficient are operating at decreasing returns to scale. This finding indicates that their efficiency worsens if their scale of operations increases. The smaller banks that were classified as inefficient are operating at increasing returns to scale. The efficiency of these banks should thus improve if their scale of operations increases.*

Introduction

The purpose of this paper is to provide a measure of the efficiency of South African banks. In recent years, increasing environmental pressures have forced banks to reconsider and improve their efficiency in order to maintain their profitability. Governmental pressures on banks to increase their exposure to micro-loans, the Basel Committee recommendations on capital charges regarding the operational risk of banks, extreme currency volatility, and an increase in the number of foreign banks that have either obtained full banking licences in South Africa or have local branch representation are all examples of environmental forces that impact negatively on the profitability of banks. Internal inefficiencies resulting from aspects such as incompetence, negligence, fraud, and internal system failures also impact negatively on the profitability of banks and have been widely reported in the media. However, it is extremely difficult for an external analyst to determine the impact of the individual external and internal factors on the profitability of banks. The reason is that the individual factors cannot be quantified without access to the detailed records of banks.

In this paper, a non-parametric approach is used by applying data envelopment analysis to create a benchmark for the relative efficiency of South African banks. A single aggregate score indicates the performance status of each bank relative to all other South African banks.

Research methodology

Research technique

It was decided to use data envelopment analysis (DEA) for this research. DEA was proposed by Charnes, Cooper & Rhodes in 1978. DEA is a non-parametric linear programming technique that can be used for evaluating the relative efficiency of different decision-making units (DMUs). The technique is used to compute a comparative ratio of outputs to inputs for each unit, which is reported as the relative efficiency score, as discussed in detail in the first paper in this series.

The authors that have applied data envelopment analysis to banking include Sherman & Gold (1985), Rangan, Grabowsky, Aly & Pasurka (1988), Aly, Grabowski, Pasurka & Rangan (1990), Charnes, Cooper & Huang (1990), Elyasiani & Mehdian (1990), Ferrier & Lovell (1990), Yue (1992), Sherman & Ladino (1995), Avkiran (1999), Golany & Storbeck (1999), Thanassoulis (1999), and Stavarek (2002).

Banks assessed

With the ultimate purpose of the research in mind – to provide a measure of the efficiency of South African banks – it was decided to apply DEA to all

* Professor J.J.L. Cronje is in the Department of Business Management, University of South Africa.
E-mail: cronjjjl@unisa.ac.za.

1 Cronje, J.J.L. 2002. 'Data envelopment analysis as a measure for technical efficiency measurement in banking — a research framework', *Southern African Business Review,* 6(2): 32–41.

local registered banks in South Africa for the financial year 2001.

Foreign-controlled banks operating in South Africa were excluded from the research for the following reasons:

- The South African operations of the majority of the foreign-controlled banks represent a very small percentage of their total operations.
- Their scale of operations, as well as the general market conditions and competitive environment in which they operate, is not comparable to local banks.

Only 16 of the 30 local registered banks were included in the research for the following reasons:

- The availability of the financial information was affected by the fact that many of the banks form part of groups, although they operate as separate entities under different trade names. Financial information is provided for groups as a whole and not for the separate trading entities. This can be regarded as a major limitation of the research as it also affects the number of inputs and outputs that could be used for the DEA assessment.
- Six banks were deliberately excluded as they are regarded as too small and specialised in comparison to the other banks included in the survey.
- One bank did not specify one of the inputs required in the research in its published financial information and was therefore excluded from the research.

The 16 banks included in the survey were labelled from A to P in order not to reveal their names.

Specification of inputs and outputs

Previous research on the efficiency of banks was used to identify relevant inputs and outputs. Button & Weyman-Jones (1992) point out that measured DEA efficiency in small samples is sensitive to the difference between the number of DMUs (in this case banks) and the sum of inputs and outputs used. In a typical analysis of banks, each ratio may be associated with a different bank, and the number of such ratios will be the product of the number of inputs and the number of outputs. In an analysis with six inputs and six outputs, there is thus a clear possibility of 36 efficient banks. In general, if there are t outputs and m inputs, we would expect the order of tm efficient banks, suggesting that the number of banks in the set should be substantially greater than tm, in order for there to be suitable discrimination between the banks. This is because the small number of free dimensions remaining increases the chance of each bank being seen as efficient. Moreover, Greene (1993) indicates that a single errant observation can have profound effects on estimates of DEA frontiers.

To cope with the limitations of DEA, the product of the inputs and outputs had to be restricted to 12 to keep it less than the number of DMUs included in the research. This translated to the combined maximum number of inputs and outputs being equal to or less than 7 (inputs + outputs $\leq$ 7).

The research was approached from the view that banks cannot merely be classified as financial intermediaries, the primary business of which is to borrow funds from depositors and lend those funds to others for profit; nor can banks be regarded as institutions that only produce various loans and other investments as outputs from deposits, other funding sources, labour and material. Banks fulfil a more complex role of providing a broad spectrum of financial services to clients. This is especially true in South Africa, considering that all major banking groups comprise commercial, private, investment and retail banking divisions.

With profitability as the ultimate objective of any private enterprise, it was assumed to be appropriate to state the outputs in the form of revenue. In the case of banks, such revenue can be classified into two major groups, namely interest income and non-interest income – the latter being revenue generated from all activities other than interest paid on the different types of deposits that the banks accepts. The use of these outputs also makes it possible to compare banks whose activities may differ.

Inputs represent resources that must be utilised optimally to generate revenue that satisfies the requirements of management and shareholders and compares favourably with that of competitors. As such, the following resources that could be classified as inputs were identified: capital (shareholders' funds); purchased funds (deposits of all types) and labour resources (staff at all levels). It should be noted, however, that the physical number of staff is not generally provided in published financial information. The only available information pertaining to this input is the annual labour costs of banks. Table 1 gives the figures that were used as

Table 1: Inputs and outputs used in the DEA measurement of the efficiency of South African banks

Bank	Inputs			Outputs	
	Capital (R million)	Purchased funds (R million)	Labour (R million)	Interest income (R million)	Non-interest income (R million)
A	332	935	50	170	193
B	13 658	167 736	4 480	22 571	6 408
C	8 672	182 994	1 698	12 114	3 621
D	436	5 763	59	706	40
E	1 204	1 546	104	180	22
F	6 402	50 696	1 089	7 012	2 077
G	2 297	2 236	332	1 855	460
H	1 364	3 044	67	149	559
I	12 321	140 887	3 928	15 185	6 364
J	1 157	17 405	291	2 594	570
K	508	3 141	138	436	120
L	84	623	17	58	13
M	15 729	177 160	2 928	17 186	5 709
N	277	187	79	61	4
O	32 019	242 452	5 242	24 298	8 879
P	141	308	75	48	161

inputs for the DEA measurements. These figures were obtained from the 2001 published financial statements of banks.

The information in Table 1 provides an indication of the relative sizes of the banks included in the research. It is evident from Table 1 that immense differences exist in the sizes of the banks, for example: Capital ranges from R84 million for bank L to R32 019 million for bank O, while interest income amounts to R48 million for bank P and R24 298 million for bank O. The size differences are also indicative of the fact that the activities of banks may differ substantially, as many of the smaller banks can be regarded as niche banks with specialised focus areas, while the larger banks have more comprehensive service offerings. Banks B, C, F, I, M and O are by far the largest banks in terms of inputs and outputs used.

The differences in the sizes of banks can be regarded as a limitation of the research, possibly affecting the efficiency classification, even though differences in the activities of banks were considered during input and output selection. The relationship between the specified outputs (interest income and non-interest income) are especially affected by the nature of the service offerings of banks.

DEA procedure

Two considerations were important in selecting the DEA model for this application. Firstly, since South African banks experienced intense competition and interest rates were increasing during the period being researched, it was decided to use both input minimisation and output maximisation options to identify under-produced outputs and over-utilised inputs. Under input minimisation, potential improvements indicated by DEA focus on the extent to which input quantities can be proportionally reduced without changing the output quantities produced. Similarly, under output maximisation, the results indicate the extent to which output quantities can be proportionally expanded without altering the input quantities used (Coelli 1996).

Secondly, because of the differences in the sizes of banks in South Africa, it was decided to apply both constant returns to scale (CRS) and variable returns to scale (VRS) measurements to decompose the technical efficiency of the banks under assessment into pure technical efficiency and scale efficiency. CRS implies that an increase in input is expected to result in a proportionate increase in output. In many cases, however, the scale of operations can influence a bank's rating. In such cases, the CRS model would not provide accurate measurements. Avkiran (1999) indicates that even in a homogeneous sample, some units that are assessed may be operating at CRS while others would be operating at VRS. CRS implies a proportionate rise in outputs when inputs are increased; thus the scale of operations does not influence the efficiency of the bank. Conversely, VRS implies a disproportionate rise or fall in outputs when inputs are increased. That is, as a bank grows in size, its efficiency will either increase or decrease.

The different linear program equations were solved once for each of the N banks under assessment to make the efficiency of the target bank as large as possible, to determine peer groups, and to determine targets for inefficient banks. (For a detailed discussion of this methodology, refer to Cronje 2002: 32–41.)

The technical efficiency of banks was decomposed into pure technical efficiency and scale efficiency, as the CRS efficiency score represents technical efficiency that measures the inefficiencies as a result of the input/output configuration as well as the size of operations, while the VRS efficiency score represents only pure technical efficiency without scale efficiency. Scale efficiency was calculated by dividing technical efficiency by pure technical efficiency as proposed by Coelli, Rao & Battese (1998).

Once scale efficiency scores were computed, it was possible to determine whether a particular bank operated at optimal returns to scale, increasing returns to scale, or decreasing returns to scale. This required that DEA had to be repeated with non-increasing returns to scale. The scores of banks obtained through this analysis were compared with the scores of banks obtained under CRS and VRS. The results were interpreted as follows (Coelli et al. 1998): if the score for a particular bank under VRS

Table 2: Efficiency scores of banks according to DEA measurements

Bank	Input-orientated CRS formula	Output-orientated CRS formula	Input-orientated VRS formula	Output-orientated VRS formula
A	1.000	1.000	1.000	1.000
B	0.872	0.872	1.000	1.000
C	0.871	0.871	1.000	1.000
D	1.000	1.000	1.000	1.000
E	1.000	1.000	1.000	1.000
F	0.839	0.839	1.000	1.000
G	1.000	1.000	1.000	1.000
H	1.000	1.000	1.000	1.000
I	0.768	0.768	1.000	1.000
J	1.000	1.000	1.000	1.000
K	0.657	0.657	0.724	0.694
L	0.471	0.471	1.000	1.000
M	0.749	0.749	1.000	1.000
N	0.393	0.393	1.000	1.000
O	0.641	0.641	1.000	1.000
P	1.000	1.000	1.000	1.000

equalled the non-increasing returns to scale score, the bank operated under decreasing returns to scale; if the score for a particular bank under VRS was not equal to the non-increasing returns to scale score, this implied a bank that operated under increasing returns to scale; and if the VRS score equalled the CRS score, the bank operated at optimal returns to scale.

Results

Efficiency scores

Table 2 shows the relative efficiency scores of the different banks under assessment. Caution should be exercised in interpreting these results because DEA does not truly rank order DMUs against all other DMUs (Sherman 1988). DEA identifies a DMU as either efficient or inefficient (with varying degrees) against others in its reference set. Theoretically, only those DMUs with identical reference sets can be rank ordered.

The efficiency score is expressed as a number between zero and one. A DMU with a score of less than one is deemed inefficient relative to other DMUs (Avkiran 1999). From the information contained in Table 2, it is evident that nine banks have CRS scores of less than one. On the other hand, only one bank (K) is classified as inefficient with VRS. The results therefore indicate that the pure technical efficiency of all the banks, with the exception of one (bank K), are optimal and that inefficiencies occur mainly as a result of scale inefficiencies. When the CRS information in Table 2 is compared with the information in Table 1, it is clear that all six of the largest banks (B, C, F, I, M and O) are classified as inefficient, compared with three of the smaller banks (K, L and N).

Coelli (1996) states that the input and output orientated efficiency scores of DMUs are equivalent when CRS is used, but may differ when VRS is used. This is confirmed by the findings in Table 2, where the input and output efficiency scores of the banks are the same with the CRS formula, but the efficiency score for bank K differs with the VRS formula, because of this bank's technical inefficiency.

Peer groups

The peer groups identified during the research for each of the inefficient banks appear in Table 3. In solving the equation for a target DMU, the linear program attempts to make the efficiency of the target DMU as large as possible. This search procedure terminates when either the efficiency of the target DMU or the efficiency of one or more other DMUs hits the upper limit of one. Thus, for an inefficient DMU, at least one other DMU will be efficient with the target DMU's set of weights. These efficient DMUs are known as the peer group of the inefficient DMU and hence define efficiency for the inefficient DMU.

The results in Table 3 indicate that:

- The average number of peers for inefficient banks is three.
- Because the six largest banks (B, C, F, I, M and O) were classified as inefficient with CRS, none of them serve as peers for inefficient banks. In fact, the smaller banks that were classified as efficient serve as peers for the large, inefficient banks.

Global leaders

A global leader is determined by identifying the bank that most frequently appears in reference sets. That is, every inefficient bank is compared to a set of efficient banks with a similar input/output configuration; the bank that appears in these reference sets more than others becomes known as the global leader. The benefit of this information is identifying an efficient bank that can be emulated in raising the performance of inefficient banks. The global leaders are reflected in Table 4.

Bank J can be regarded as the global leader according to the information contained in Table 4 (eight times a peer), followed by bank G (six times a peer) in second place and banks A and H (four times a peer) jointly in third place.

Scale efficiency

The scale efficiency of the banks is contained in Table 5. The results indicate that six of the inefficient banks (B, C, F, I, M and O) are operating at decreasing returns to scale, while three (K, L and N) are operating at increasing returns to scale. Those operating at decreased returns to scale are the six largest banks that were classified as inefficient, while those operating at increased returns to scale are the three smallest inefficient banks. This finding is important as it confirms that the inefficiencies can in all instances be ascribed to scale inefficiencies, except in the case of bank K, which is also experiencing pure technical inefficiency (refer to Table 2). The large inefficient banks will therefore become more inefficient if their scale of operations increases, while the inefficient small banks will become more efficient as their scale of operations increases.

Table 3: Peer groups of inefficient banks according to DEA measurements

Inefficient bank	Peers according to input-orientated CRS formula	Peers according to output-orientated CRS formula	Peers according to input-orientated VRS formula	Peers according to output-orientated VRS formula
B	G, P, J	G, P, J		
C	A, H, J	A, H, J		
F	G, H, J	G, H, J		
I	A, J, P	A, J, P		
K	G, J, P	G, J, P	G, J, L, P	G, J, L, P
L	G, J	G, J		
M	A, H, J	A, H, J		
O	A, G, H, J	A, G, H, J		

Table 4: Ranking of global leaders – number of times each bank is a peer for another

Bank	Input-orientated CRS formula	Output-orientated CRS formula	Input-orientated VRS formula	Output-orientated VRS formula
A	4	4	0	0
B	0	0	0	0
C	0	0	0	0
D	0	0	0	0
E	0	0	0	0
F	0	0	0	0
G	6	6	1	1
H	4	4	0	0
I	0	0	0	0
J	8	8	1	1
K	0	0	0	0
L	0	0	1	1
M	0	0	0	0
N	0	0	0	0
O	0	0	0	0
P	3	3	1	1

Improvements required for inefficient banks

For the purposes of determining the improvements in technical efficiency required to increase the performance of inefficient banks to a level that can be regarded as efficient compared with their peers, only CRS input- and output information was used, as VRS information pertains only to pure technical efficiency. The information in Table 6, for example, indicates that for bank B to be classified as efficient with the existing levels of outputs, it should reduce its capital by 23%, its borrowed funds by 13% and its labour costs by 27%, but for bank B to be classified as efficient with the existing levels of inputs, it should increase both its interest income and non-interest income by 15%. The findings are:

Table 5: Scale efficiency of banks according to the different formulations

Bank	Scale efficiency according to input-minimisation formulation	Scale efficiency according to output-maximisation formulation	Scale efficiency classification according to the input-minimisation formulation	Scale efficiency classification according to output-maximisation formulation
A	1.000	1.000	ORS	ORS
B	0.872	0.872	DRS	DRS
C	0.871	0.871	DRS	DRS
D	1.000	1.000	ORS	ORS
E	1.000	1.000	ORS	ORS
F	0.839	0.839	DRS	DRS
G	1.000	1.000	ORS	ORS
H	1.000	1.000	ORS	ORS
I	0.768	0.768	DRS	DRS
J	1.000	1.000	ORS	ORS
K	0.908	0.947	IRS	IRS
L	0.471	0.471	IRS	IRS
M	0.749	0.749	DRS	DRS
N	0.393	0.393	IRS	IRS
O	0.641	0.641	DRS	DRS
P	1.000	1.000	ORS	ORS

Abbreviations used in Table 5:
ORS Optimal returns to scale
DRS Decreasing returns to scale
IRS Increasing returns to scale

- All banks that were classified as inefficient either have to reduce their inputs significantly or increase their outputs significantly in order to become technically efficient.
- The smallest reduction in a single input for any one bank is 13%. The largest reduction is 86%.
- The smallest increase in a single output for any one bank is 15%. The largest increase is 862%.

These vast reductions and increases provide an indication of the enormous differences that exist in the technical efficiency of the banks that were assessed. The results should, however, be interpreted by bearing in mind that, as previously mentioned, the activities of banks may differ substantially as many of the smaller banks can be regarded as niche banks with specialised focus areas, while the larger banks have more comprehensive service offerings. This may have a profound effect on the relationship between the specific inputs and outputs of banks, with the result that not all the error terms can necessarily be ascribed to inefficiencies.

Summary and conclusions

The findings of the research show that all six of the largest banking institutions in South Africa can be classified as technically inefficient. The inefficiency can primarily be ascribed to scale inefficiencies, as the pure technical efficiency assessments of these banks do not reflect inefficiencies. All these banks are operating at decreasing returns to scale.

Table 6: Potential improvements for inefficient banks according to the results of DEA

Inefficient bank	Input-minimisation formulation: Reduction in inputs		Output-maximisation formulation: Increase in outputs	
	Input	**(%)**	**Output**	**(%)**
B	Capital	23	Interest income	15
	Borrowed funds	13	Non-interest income	15
	Labour	27		
C	Capital	13	Interest income	15
	Borrowed funds	54	Non-interest income	15
	Labour	13		
F	Capital	17	Interest income	19
	Borrowed funds	16	Non-interest income	19
	Labour	16		
I	Capital	23	Interest income	30
	Borrowed funds	28	Non-interest income	30
	Labour	23		
K	Capital	34	Interest income	52
	Borrowed funds	34	Non-interest income	52
	Labour	50		
L	Capital	53	Interest income	112
	Borrowed funds	53	Non-interest income	116
	Labour	55		
M	Capital	25	Interest income	33
	Borrowed funds	33	Non-interest income	33
	Labour	25		
N	Capital	73	Interest income	154
	Borrowed funds	61	Non-interest income	862
	Labour	86		
O	Capital	36	Interest income	56
	Borrowed funds	36	Non-interest income	56
	Labour	36		

All the smaller banks that were classified as inefficient are operating at increasing returns to scale, notwithstanding the fact that one of them shows pure technical inefficiencies. The efficiency of these banks should thus improve if their scale of operations increases.

One bank (bank J) that can be emulated in raising the performance of inefficient banks was identified as the global leader.

From the aforementioned, it is evident that the DEA methodology followed in this paper provides crucial information about the technical efficiency of banks. It shows that the DEA framework is extremely flexible, permitting multiple inputs and outputs for evaluation purposes. However, it does not provide an indication of the theoretical efficiency of banks as it is limited to classifying the banks under assessment relative to one another. As such, the measured efficiency of the South African banks included in this research can only be compared to other groups of banks provided that they are all included in the same DEA procedure.

The research results contained in this paper depend largely on the choice of information about banks that was used in the application of the DEA. This is because the measured efficiency is interpreted as the difference between its observed input and output levels and the corresponding optimal values. There is, however, no simple solution to the problem of input and output specification as reasonable arguments can be made for all approaches, and the input and output specifications of researchers depend on how they view banks.

Acknowledgements

I would like to acknowledge the assistance of Mrs A. Mutezo with this project.

References

Aly, H.Y., Grabowski, R., Pasurka, C. & Rangan, N. 1990. 'Technical, scale, and allocative efficiencies in US banking: an empirical investigation', *Review of Economics and Statistics,* May: 211–218.

Avkiran, N.K. 1999. 'An application reference for data envelopment analysis in branch banking: helping the novice researcher', *International Journal of Bank Marketing,* 17(5): 206–220.

Cronje, J.J.L. 2002. 'Data envelopment analysis as a measure for technical efficiency measurement in banking – a research framework', *Southern African Business Review,* 6(2): 32–41.

Button, K.J. & Weyman-Jones, T.G. 1992. 'Ownership structure, institutional organisation and measured X-inefficiency', *American Economic Review,* 82(2): 439–445.

Charnes, A., Cooper, W.W. & Huang, Z.M. 1990. 'Polyhedral cone-ratio DEA models with an illustrative application to large commercial banks', *Journal of Econometrics,* 46: 73–93.

Charnes, A., Cooper, W.W. & Rhodes , E. 1978. 'Measuring the efficiency of decision making units', *European Journal of Operational Research,* 2: 429–444.

Coelli, T. 1996. A guide to DEAP version 2.1: A Data Envelopment Analysis (Computer) Program. Centre for Efficiency and Productivity Analysis Working Paper.

Coelli, T., Rao, D.S.P. & Battese, G.E. 1998. *An Introduction to Efficiency and Productivity Analysis.* Boston, MA: Kluwer Academic Publishers.

Elyasiani, E. & Mehdian, S.M. 1990. 'A nonparametric approach to measurement of efficiency and technological change: The case of large US commercial banks', *Journal of Financial Services Research,* July: 157–168.

Ferrier, G.D. & Lovell, C.A.K. 1990. 'Measuring cost efficiency in banking: econometric and linear programming evidence', *Journal of Econometrics,* October/November, 229–245.

Golany, B. & Storbeck, J.E. 1999. 'A data envelopment analysis of the operational efficiency of bank branches', *Interfaces,* May/June, 29(3): 14–27.

Greene, W. 1993. 'The econometric approach to efficiency analysis', In H.O. Fried, C.A. Knox Lovell & P. Schmidt (eds), *The Measurement of Productive Efficiency: Techniques and Applications.* Oxford: Oxford University Press.

Rangan, N., Grabowsky, R., Aly, H.Y. & Pasurka, C. 1988. 'The technical efficiency of US banks', *Economics Letters,* 28: 169–175.

Sherman, H.D. 1988. *Service Organisation Productivity Management.* Hamilton, Ontario: Society of Management Accountant Accountants of Canada.

Sherman, H.D. & Gold, F. 1985. 'Bank branch operating efficiency: evaluation with data envelopment analysis', *Journal of Banking and Finance,* 9: 297–315.

Sherman, H.D. & Ladino, G. 1995. 'Managing bank productivity using data envelopment analysis (DEA)', *Interfaces,* 25(2): 60–73.

Stavarek, D. 2002. 'Comparison of the relative efficiency of banks in European transition economies', *Proceedings of D.A. Tsenov Academy of Economics 50th Anniversary Financial Conference,* Svishtov, Bulgaria, April, 955–971.

Thanassoulis, E. 1999. 'Data envelopment analysis and its use in banking', *Interfaces,* May/June, 29(3): 1–13.

Yue, P. 1992. Data envelopment analysis and commercial bank performance: A primer with applications to Missouri banks, Federal Reserve Bank of St. Louis, January,/February, 31–45

Small business problems in Gauteng: priorities for entrepreneurial education

*A.A. Ligthelm & M.C. Cant**

The objective of this paper is to establish the extent to which owners and managers of small businesses in a typical South African setting experience selected problems or issues as negatively influencing the success of their small business. Small businesses encounter numerous problems related to environmental, financial, marketing, human resources, social and managerial issues, among others. The paper aims also to investigate whether these problem categories related in any way to the management qualifications of the owner or manager. An awareness of the impact on business success of a wide variety of factors will enable SMEs to be forewarned and proactive in their decision-making and to avoid the main causes of business failure. The results of this study will also enhance the design of syllabuses for management and entrepreneurial training to ensure that they cover issues that impact negatively on business success. Starting a business is risky at the best of times, but the chances of success are enhanced if anticipated problems are understood and addressed before launching a business.

Introduction

The important contribution that SMEs (small and medium-sized enterprises) can make to employment and income generation is increasingly being recognised around the world. In South Africa, this contribution is particularly important, because the non-agricultural formal sectors shed more than a million employment opportunities between 1990 and 2001. Employment in these sectors decreased by 18.2% over the last decade, from 5 694 924 in 1990 to 4 659 330 in 2001 (South African Reserve Bank 2002: S132). This phenomenon, coupled with relatively high population growth, has resulted in an unemployment rate of 30.5% (Statistics South Africa 2003). The rising rates of unemployment are structural in nature, rather than the result of deficient aggregate demand (Hodge 2002: 440). The main causes of the growing unemployment rate include factors such as the low employment elasticity of the South African economy, as well as the high rate of labour force growth.

It is acknowledged that the SME sector is labour intensive and has great potential for increased employment creation. However, this sector has experienced a high degree of business mortality, which impacts negatively on the ability to create sustainable employment opportunities in the long term.

Surveys of business failure suggest that SME entrepreneurs often have good ideas and are competent people but "they do not have a clue of how to run a business and have no underlying appreciation of business fundamentals" (Barron 2000: 1). This article attempts to identify and analyse the reasons for business failure. Training and support programmes can then be designed to equip SME entrepreneurs to survive the ever-increasing competitive and global business environment.

Objective

The objective of the article is to establish the extent to which owners/managers of SMEs in Gauteng experienced selected problems or issues as negatively influencing the success of their businesses. The aim is also to investigate whether there is any relationship between these problems and the possession of management qualifications.

The problem categories identified in the study ranged from macroeconomic and marketing issues to management, human resources, financial and social issues. An awareness of the impact on business success of this wide variety of issues will enable SMEs (both existing and planned) to be forewarned and proactive in their decision-making

* Professor A.A. Ligthelm is at the Bureau of Market Research, University of South Africa. E-mail: ligthaa@unisa.ac.za. Professor M.C. Cant is Head of the Marketing and Retail Section, Department of Business Management, University of South Africa. E-mail: cantmc@unisa.ac.za.

and to avoid the main causes of business failure. The results of the study may also enhance the design of syllabuses for management training to ensure that they cover coverage the issues that impact negatively the success of small businesses.

Problems experienced

Small and medium-sized businesses form an integral part of the national economy and, as such, are influenced by various factors. In this context, the problems that SMEs face in conducting a successful business can be grouped into various categories (Dockel & Ligthelm 2002: 2), namely:

(a) Economy-based problems that are associated with national economic factors related to the state of the economy, such as specific phases in the business cycle and the exchange value of the national currency.
(b) Industry-based problems that are associated with the industry in which the firm operates. Such problems may include the degree of competitiveness, demand and supply factors within the industry, and barriers to entry.
(c) Firm-based problems that refer to internal factors such as the availability of resources, including finance and entrepreneurial skills, as well as the effective use of such resources.

It is sometimes useful to consider the problems caused by factors outside the firm by combining economy-based and industry-based problems as macroenvironmental (or exogenous) problems and distinguishing these from problems over which the individual owner has some control, which can be regarded as firm-based problems and classified as internal-micro (or endogenous).

This classification is relevant for targeting policy and specific problem areas and training needs. If, for instance, the problem is predominantly endogenous in nature, then policy measures should be directed at the firm level, such as providing training and education and small business support structures. If the causes of failure are predominantly exogenous, then policy should be directed towards changing the economic environment in which the firm operates (Dockel & Ligthelm 2002: 2).

An example of an investigation of business problems according to these two categories (exogenous and endogenous) is that of Everett & Watson (1998), which focuses on economy-based factors as causes of small business failure in Australia. The study considered macroeconomic variables such as interest rates, unemployment and inflation rates, but found that firm-based problems were associated with 65% of small business failures. A second study by Peterson, Kozmetsky & Ridgeway (1983) estimated that internal-micro factors were responsible for approximately 70% of small business failures.

This article highlights the types of exogenous and endogenous problems experienced by SMEs. The list of problems experienced by SMEs is contained in Tables 1–8, but lack of space precludes a more detailed discussion.

Research design

Research instrument

In order to fully understand the problems encountered by SMEs and the extent to which they negatively influence the success of small businesses, an empirical survey was conducted in selected areas of Gauteng, enquiring about the personal experiences of entrepreneurs relating to their businesses. In the absence of a comprehensive sample frame of SMEs in the Gauteng province, a judgemental sampling procedure was followed according to specified criteria. The requirements for the sample were that the business:

- Had compiled an annual budget
- Operated from a demarcated business area
- Occupied a permanent structure
- Covered the following economic sectors: retailers, personal services, manufacturers, wholesalers, construction and transportation
- Employed fewer than 50 full-time workers.

These criteria ensured the inclusion of businesses with a reasonable level of sophistication and understanding of concepts such as exchange and inflation rates. In total, 301 businesses were interviewed, mainly in areas occupied by previously disadvantaged individuals.

Analysis

The analysis in this article is based on two statistical procedures. Descriptive data are presented by means of frequency tables and cross tabulations. The opinions of respondents are provided by management qualification as part of the descriptive statistics. This is followed by the presentation of indices calculated according to an analytical hierarchical procedure. The latter calcu-

lations are based on the opinions of both respondents and experts in terms of the relative importance of each variable in influencing the success of a business. The relative importance of each variable (for example, inflation rates as opposed to interest rates, or crime as opposed to exchange rates, or financial issues as opposed to market-related problems) is determined by expert opinion in order to establish a hierarchy of importance of all variables in the questionnaire.

The analysis is based on problems related to macroenvironmental factors as well as marketing, social, human resources and financial issues. It also covers management views on perceived success and the application of management actions.

Survey population

A profile of the survey population is included before the results of the survey are presented in order to establish the background against which responses on issues influencing the success of businesses should be interpreted.

The survey population demonstrates the following characteristics:

- Businesses with a reasonable level of sophistication were included in the sample. Only businesses operating from demarcated business premises, occupying a permanent structure and compiling an annual budget were selected.
- Only business owners or managers were selected as respondents. Almost two thirds (64.5%) of the respondents had acted as managers or had owned their businesses for more than three years.
- One out of every four respondents (23.6%) had a management qualification. Almost half of those with a management qualification had only a basic business certificate, however, and just over a third had a diploma or degree.
- Half (50.8%) the businesses operated as sole ownership enterprises, and the rest were partnerships (12.8%), private companies (16.2%) and close corporations (19.9%).
- Respondent businesses employed an average of 3.8 full-time employees in 2001.
- Four out of every five businesses (79.1%) reported a turnover of less than R500 000 for 2001. Only 7.1% generated a turnover of over R1 million.
- Seven out of every ten businesses (70.3%) was owned or managed by Africans, 14.3% by Asians, 4.0% by coloureds and 11.3% by whites.
- Two thirds of the questionnaires were completed in township areas and one third in the central business districts of Pretoria and Johannesburg.

The profile of respondents confirms that, although an effort was made to select a relatively sophisticated business sample, respondent businesses were generally small to medium enterprises, mainly owned or managed by Africans, with the overwhelming majority located in township areas.

Descriptive data

Table 1 reflects the perceptions of business owners and managers of the possible impact of the macroenvironmental problems on the success of their businesses, by management qualification. The respondents were requested to indicate their experience of macroenvironmental variables that could possibly have a negative influence on the success of their businesses on a five-point scale, ranging from ‘strongly disagree’ to ‘strongly agree’. For the purposes of this paper ‘strongly disagree’ and ‘disagree’ are grouped together, as are ‘strongly agree’ and ‘agree’.

Table 1 clearly shows that managers and owners with management qualifications are far less uncertain (neutral) about the perceived impact of macroenvironmental factors on the success of their businesses. They also tend to agree more with the statement that variables such as high interest rates, inflation, crime, HIV/AIDS and new government legislation would negatively impact on the success of their businesses. It can be deduced that conceptual know-how of the functioning of the macroeconomic environment allows better insight into the impact of exogenous factors on the functioning of businesses within a particular macroeconomic setting, especially in an environment subject to rapid change.

Marketing-related issues

Table 2 reflects the influence of marketing-related variables on the success of businesses by management qualification. Businesses regard marketing-related issues (such as increased competition, limited market size, low demand for products and ineffective marketing) as having the most marked negative influence on the success of their busi-

Table 1: Environmental variables that impact negatively on the success of businesses, by respondents with and without a management qualification

Environmental variable	Management qualification	Disagree/Agree			
		Strongly disagree and disagree %	Neutral %	Agree and strongly agree %	Total %
Interest rates	With	15.3	9.7	75.0	100.0
	Without	11.9	18.1	70.0	100.0
Exchange rate	With	16.7	13.9	69.5	100.0
	Without	11.9	22.5	65.7	100.0
Inflation	With	9.7	4.2	86.1	100.0
	Without	9.2	17.2	73.6	100.0
Unemployment	With	11.0	11.0	78.1	100.0
	Without	9.7	14.6	75.7	100.0
Crime	With	9.8	5.6	84.7	100.0
	Without	12.3	6.6	81.1	100.0
HIV/AIDS	With	30.2	31.5	38.4	100.0
	Without	34.7	40.4	24.9	100.0
Rapidly changing technology	With	40.3	20.8	38.8	100.0
	Without	42.5	17.7	39.8	100.0
New legislation	With	24.4	25.7	50.0	100.0
	Without	33.9	30.8	35.3	100.0

nesses. Table 2 confirms that managers and owners single out increased competition as the major marketing-related issue faced by the respondent businesses. Almost two thirds of the businesses (71.3% with and 67.3% without a management qualification) agreed or strongly agreed with this. Although there is little divergence in the perceptions of those with and those without management qualifications regarding the influence of the various marketing-related issues on the success of their businesses, respondents with management qualifications tend to agree more that issues such as increased competition, lack of knowledge of competitors and ineffective marketing have a negative impact on the success of their businesses. This may suggest that better insight into marketing-related issues is an important factor for business success.

Business skills

The literature often cites a lack of business skills and training as a major cause of business failure (Viviers, Van Eeden & Venter 2001: 11). However, Table 3 shows that the majority of respondents did not regard management training as lacking and thus negatively influencing their businesses. This response was also recorded in other similar studies and is explained as a form of response bias, in that respondents seek social approval in the eyes of the interviewer (Aaker, Kumar & Day 1995: 204). A further possible reason for the survey result was that the screening question allowed only the inclusion of businesses that compiled annual financial statements. Budgeting implies activities such as planning and prioritising, suggesting that respondents already exercise some important

Table 2: Marketing-related issues that impact negatively on the success of businesses, by respondents with and without a management qualification

Market-related variable	Management qualification	Disagree/Agree			
		Strongly disagree and disagree %	Neutral %	Agree and strongly agree %	Total %
Limited market size	With	24.6	26.0	49.3	100.0
	Without	37.3	14.7	48.0	100.0
Increased competition	With	13.7	15.1	71.3	100.0
	Without	19.3	13.2	67.3	100.0
Low demand for product	With	41.7	13.9	44.4	100.0
	Without	37.3	19.6	43.1	100.0
Lack of knowledge of competition	With	28.4	36.5	35.1	100.0
	Without	37.1	31.0	31.9	100.0
Lack of knowledge of market	With	43.8	24.7	31.5	100.0
	Without	46.7	23.1	30.2	100.0
Inability to identify target market	With	46.5	27.4	26.0	100.0
	Without	43.5	25.8	30.7	100.0
Ineffective marketing	With	28.8	28.8	42.4	100.0
	Without	31.7	28.6	39.6	100.0
Poor location	With	55.4	16.2	28.4	100.0
	Without	49.3	18.1	32.6	100.0

business management functions. They may therefore not regard management skills as lacking in their businesses.

Table 3 reflects the views of respondents on the importance of business training by management qualification. Although the majority of respondents with and without a formal management qualification maintain that a lack of business skills does not impact negatively on their businesses, a larger percentage of those without management qualifications agreed that the lack of management training and skills impacted negatively on the success of their businesses.

Management functions

Management functions, as described in the questionnaire, generally refer to the tasks of planning, organising, leading, coordinating and control. Table 4 confirms the importance of implementing effective management functions as a means of improving the success of SMEs.

Respondents in the possession of management qualifications generally attached greater priority to the execution of management functions. Of particular importance in this regard are sensitivity towards change as an integral part of running a business, effective planning, organising, leading and control. The importance of this finding is that it

Table 3: Issues related to business skills that impact negatively on the success of businesses, by respondents with and without a management qualification

Business skills	Management qualification	Disagree/Agree			
		Strongly disagree and disagree %	Neutral %	Agree and strongly agree %	Total %
Lack of technical skills	With	63.0	12.3	24.7	100.0
	Without	68.0	12.4	29.7	100.0
Insufficient experience and knowledge of the field of business in which I operate	With	54.0	10.8	35.2	100.0
	Without	65.1	12.8	22.1	100.0
A lack of management training	With	66.2	8.1	25.7	100.0
	Without	54.9	15.0	30.1	100.0
A lack of management skills	With	64.9	9.5	25.7	100.0
	Without	57.6	12.4	30.1	100.0

affirms that management training not only transfers management skills to learners but also reinforces the lateral perception that the application of skills contributes to business success.

Social problems

Social issues refer to the problems experienced by SME owners, in that their work tends to invade their personal lives. Long hours spent at work place pressure on relationships with family members and friends. The pressure stems mainly from the dual roles that have to be performed by many owners as both diligent income providers and caring spouses.

Table 5 reflects that slightly fewer respondents in possession of a management qualification suffered from a negative impact of their business activities on their family and social life (see 'Agree and strongly agree' column).

Human resources problems

The human resources problem category reflects the extent to which SME owners experienced certain human resources issues as negatively influencing the success of their businesses. Generally, respondents do not regard human resources issues as a serious problem that jeopardises the profitability of their businesses. The two most serious human resources problems, mentioned by approximately one third of respondents (38.9% of those with and 34.5% of those without a management qualification), were the introduction of new labour laws and the inability to find and attract suitable staff (28.2% and 34.1% respectively). The least important of the human resources issues was poor labour relations (15.0% and 21.4%), poor staff planning (15.5% and 22.2%), high labour turnover (19.5% and 23.9%) and low labour productivity (23.6% and 27.8%).

Table 6 confirms that the impact of human resources problems is more severe in businesses managed or owned by persons without management experience. Between 4% and 6% more businesses managed or owned by a person without management qualifications indicated human resources problems as a cause of failure. The largest divergence is in relation to issues such as staff planning, labour relations and the inability to appoint suitable staff.

Table 4: Management function issues that impact negatively on the success of businesses, by respondents with and without a management qualification

Management functions	Management qualification	Disagree/Agree			
		Strongly disagree and disagree %	Neutral %	Agree and strongly agree %	Total %
I set time aside each day/ week to plan and prioritise activities for the day/week	With	17.8	13.7	68.5	100.0
	Without	16.9	15.2	67.9	100.0
I tend to neglect planning due to time pressure	With	48.6	20.3	31.1	100.0
	Without	52.9	24.9	22.2	100.0
Daily routine/operational tasks tend to take up most of my time	With	33.8	12.2	54.1	100.0
	Without	36.0	17.3	46.7	100.0
Long-term goals and objectives are not meaningful considering the pace of change in my industry	With	32.0	26.4	41.7	100.0
	Without	40.2	29.2	30.6	100.0
I am able to spend my time more effectively if I plan better	With	12.3	11.0	76.7	100.0
	Without	16.0	13.7	70.4	100.0
I view failure as a valuable learning experience	With	12.9	14.3	72.8	100.0
	Without	13.2	14.1	72.7	100.0
I constantly try to involve employees in planning and decision making	With	22.2	8.3	69.5	100.0
	Without	18.3	13.8	68.0	100.0
I prefer to do most of the work as I want to be in control of what is happening in my business	With	32.5	5.4	62.1	100.0
	Without	25.4	14.2	60.4	100.0
I regard change as an integral part of running a business	With	6.8	8.1	85.1	100.0
	Without	6.7	13.3	80.1	100.0

Table 5: Social issues that impact negatively on family life and the success of businesses, by respondents with and without a management qualification

Social variables	Management qualification	Disagree/Agree			
		Strongly disagree and disagree %	Neutral %	Agree and strongly agree %	Total %
My family/friends regularly complain that I spend too much time at work	With	32.4	21.6	46.0	100.0
	Without	36.8	11.1	52.2	100.0
Since having my own business. my social and family life has suffered due to time pressure	With	39.2	18.9	41.9	100.0
	Without	37.6	15.5	46.9	100.0
My business consumes my whole life	With	29.7	27.0	43.2	100.0
	Without	35.8	19.5	44.6	100.0
I regularly suffer from ill health	With	72.6	11.0	16.4	100.0
	Without	68.6	18.6	12.8	100.0

Financial problems

Financial problems that had the most widespread impact were difficulties in obtaining finance or credit (48.6% with and 55.5% without a management qualification) followed by high operating expenses (61.1% and 48.8%) (Table 7). As highlighted in the literature, the granting of consumer credit also presents significant challenges for small businesses. No fewer than 36.5% of respondents with and 43.6% without management qualifications agree that bad debt poses a serious problem to their businesses. The following financial problems were highlighted as the least serious problems (mentioned by approximately a quarter of businesses): overinvestment in inventories (24.0% and 21.6%) and failure to analyse financial information (24.7% and 22.7%). Approximately one third of respondents agreed to problems with inadequate credit management (27.4% and 28.5%), insufficient knowledge of bookkeeping (37.8% and 26.3%), failure to conduct financial planning (27.1% and 32.0%) and poor cash flow management (31.1% and 36.0%).

Table 7 confirms that owners and managers with a management qualification experienced fewer financial problems than those without management qualifications. However, respondents without qualifications reported a less severe impact with regard to issues such as the burden of high operating expenses and insufficient knowledge of bookkeeping.

Business success

Selected factors were used to establish whether small business owners perceived their businesses as being successful. Respondents were asked to indicate how they view their businesses in terms of growth in turnover, profits and employees. In general, almost half the owners were of the opinion that their businesses are successful in terms of turnover and profits.

Table 8 shows that businesses with owners or managers with management qualifications experience a slightly higher success rate than those without. Half the owners and managers with management qualifications rate their businesses as very successful, compared to 48.0% without management qualifications.

Table 6: Human resources issues that impact negatively on the success of businesses, by respondents with and without a management qualification

Human resources issues	Management qualification	Disagree/Agree			
		Strongly disagree and disagree %	Neutral %	Agree and strongly agree %	Total %
An inability to attract and find suitable staff	With	53.6	18.3	28.2	100.0
	Without	52.9	13.0	34.1	100.0
Low labour productivity	With	45.8	30.6	23.6	100.0
	Without	43.5	28.7	27.8	100.0
New labour laws	With	40.2	20.8	38.9	100.0
	Without	35.4	30.0	34.5	100.0
High labour turnover	With	45.8	34.7	19.5	100.0
	Without	35.6	40.5	23.9	100.0
Poor labour relations	With	54.8	30.1	15.0	100.0
	Without	46.4	32.3	21.4	100.0
Poor staff planning	With	63.4	21.1	15.5	100.0
	Without	57.9	19.9	22.2	100.0
Poorly trained employees	With	59.5	14.9	25.7	100.0
	Without	54.0	20.7	25.3	100.0

Relative importance of business management functions of SMEs

The analytical hierarchical procedure

The relative importance of the various business functions is established by considering the opinions of the business owners and managers, as well as the opinions of business management experts in terms of the relative importance of each variable that may influence the success of a business. The latter is determined through the analytical hierarchical procedure, developed by Saaty (1980). It allows a systematic evaluation of qualitative variables by assigning numerical values to subjective judgements about the relative importance of items. The result of the analytical hierarchical procedure is a mathematical weighting that reflects the relative importance of each of the variables when compared with all other variables (Ghyoot 2002: 33). The relative weighting is then expressed as an index (out of 100). Space limitations preclude a full exposition of the applied methodology (see Saaty (1980) for an exposition of the model, and Ghyoot (2002: 33) and Ligthelm & Cant (2002) for an application of the model).

Findings

By combining the rating by business management experts of the relative importance of business management functions through the analytical hierarchical method with the Likert rating of the business survey, an index is constructed, portraying the relative importance of the various business management functions (such as marketing, human resources, and general management). The index is expressed out of 100, implying that the higher the index, the greater the importance of a specific function.

Figure 1 shows the importance of the various business management functions as perceived by respondents and weighted according to expert opinion. Marketing, with an index of 58.0 is

Table 7: Financial issues that impact negatively on the success of businesses, by respondents with and without a management qualification

Financial issues	Management qualification	Disagree/Agree			
		Strongly disagree and disagree %	Neutral %	Agree and strongly agree %	Total %
Difficulty in obtaining finance/credit	With	36.5	14.9	48.6	100.0
	Without	30.5	13.8	55.5	100.0
Failure to do financial planning/budgeting	With	60.8	12.2	27.1	100.0
	Without	49.8	18.2	32.0	100.0
Failure to analyse financial information	With	64.3	11.0	24.7	100.0
	Without	52.2	25.0	22.7	100.0
Inadequate bookkeeping	With	58.9	13.7	27.4	100.0
	Without	57.9	14.8	27.3	100.0
Insufficient knowledge of bookkeeping	With	54.1	8.1	37.8	100.0
	Without	54.5	19.2	26.3	100.0
Overinvestment in inventory	With	46.6	29.6	24.0	100.0
	Without	43.7	34.7	21.6	100.0
Heavy operating expenses	With	25.0	13.9	61.1	100.0
	Without	33.8	17.3	48.8	100.0
Burdensome debt	With	50.0	19.4	30.6	100.0
	Without	34.7	23.0	42.4	100.0
Poor cash flow management	With	44.6	24.3	31.1	100.0
	Without	42.2	21.8	36.0	100.0
Poor credit management	With	50.7	21.9	27.4	100.0
	Without	45.8	25.8	28.5	100.0
Bad debts	With	47.3	16.2	36.5	100.0
	Without	42.7	13.8	43.6	100.0

regarded as the most important business management function contributing to the success of businesses. Problems such as a lack of demand for products and services, increased competition, lack of knowledge of the market, inability to identify the correct target market, ineffective marketing and poor location of businesses are perceived as major problems that have to be addressed to ensure a

Table 8: Success factors in businesses, by respondents with and without a management qualification

Success factor	Management qualification	Disagree/Agree			
		Strongly disagree and disagree %	Neutral %	Agree and strongly agree %	Total %
My business has experienced growth in turnover in the last two years	With	32.9	27.1	40.0	100.0
	Without	29.1	30.9	40.0	100.0
My business has experienced growth in employees in the last two years	With	48.6	18.6	32.9	100.0
	Without	43.2	29.5	27.2	100.0
My business is very profitable	With	20.3	39.9	39.9	100.0
	Without	21.5	35.9	42.6	100.0
I regard my business as very successful	With	13.5	36.5	50.0	100.0
	Without	16.0	36.0	48.0	100.0

profitable endeavour. This is to be expected, in view of the severe competition and the continuous establishment of new businesses in the survey areas. New business operators often establish businesses in a poor location, utilising family as employees and own finance, and frequently creating a situation of overtrading in a limited market.

This implies that financial and human resources issues are less of a problem than limited demand for products and services in a limited market.

Figure 1 also indicates that marketing-related issues are followed by financial (index 56.3) and human resources problems (51.8) as important areas to be addressed. Macroenvironmental issues

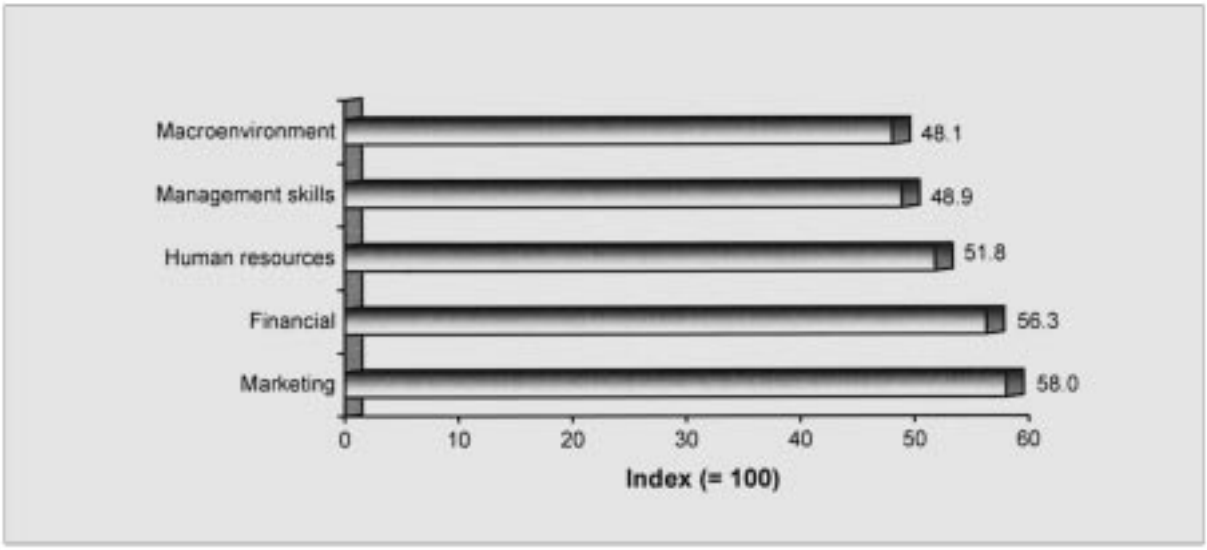


Figure 1: Indices of relative importance of business management functions

(48.1), such as inflation, unemployment and interest rate levels, that fall outside the direct sphere of influence of entrepreneurs, are perceived as the least important for ensuring business success.

By differentiating the relative importance attached to business management functions by the socio-demographic characteristics of owners and managers, the following emerges:

- The relative importance of the various business functions by management status (in other words, owner only or manager only or both) shows no significant variation, which implies that management status does not significantly alter the relative importance attached to the various business functions shown in Figure 1. The only exception is that owners attach greater importance to financial issues than managers (indices of 60.2 and 51.9 respectively). This is to be expected, since an owner would naturally regard the financial outcome of their businesses as more important than a person acting only as manager. Profitability enhances an owner's income, while managers of SMEs often receive fixed remuneration.
- Length of experience as an owner or manager of an SME shows that owners and managers with less than two years' experience tend to attach greater importance to all management functions than the average index shown in Figure 1. The relative importance of the various functions in terms of contributing to business success remains the same, notwithstanding the length of experience – marketing remains the top priority, followed by financial issues.
- Some minor variations are evident regarding the importance attached to the various management functions by type of enterprise. Sole traders emphasised financial issues as being of particular importance for business success, while private companies attached less importance to financial issues. Marketing and macro-environmental issues are of particular concern for private companies. Partnerships and close corporations rate human resources issues relatively high.
- The human resources function is of less concern to small businesses without any employees, while all management functions are regarded as relatively important to medium-sized businesses (the largest category in the survey).
- SMEs with an annual turnover of R5 million or more (the largest category in the survey) attached greater importance to management functions being instrumental in success than did smaller enterprises. The larger firms placed particular emphasis on marketing, general management and human resources issues.

Summary

The sample population, consisting of relatively small yet sophisticated businesses in township and central business districts in Gauteng, emphasised that the following issues are of particular concern for the success of their businesses:

- Marketing-related issues
  - Increased competition (overtrading of the market)
  - Limited market size
  - Ineffective marketing
  - Lack of knowledge of competitors
  - Poor location

- Financial issues
  - Difficulty in obtaining finance or credit
  - High operating expenses
  - Management of consumer credit
  - Poor cash flow management
  - Lack of financial planning

- Human resources issues
  - New labour laws
  - Inability to attract and maintain suitable staff
  - Low labour productivity
  - Poorly trained employees
  - High labour turnover

- Management functions
  - Adapting to the changing business environment
  - Time management
  - Delegation and cooperative management
  - Planning and prioritising
  - Effective control

- Microenvironmental issues
  - Crime and corruption
  - Inflation
  - Unemployment (limited market)
  - Interest rates
  - Exchange rates

The respondents confirm that they regard all management functions as important for the success of their businesses. Slight variations were detected in the importance attached to problem areas. Respondents singled marketing-related issues out as the most important for business success,

followed closely by financial issues, general management functions and, lastly, the impact of macro-environmental aspects on business activities.

Training programmes in business management and entrepreneurship should endeavour to accommodate the indicated problem areas in their syllabuses. Addressing the various issues in general in training material is not sufficient. The contents should focus on the specific problems experienced by SMEs and allow for the practical application of concepts and procedures. This is the challenge for management and entrepreneurship training.

References

Aaker, D.A., Kumar, V. & Day, G.S. 1995. *Marketing Research* (5th edition). New York: John Wiley & Son.

Barron, C. 2000. 'Brilliant ideas but spectacular flops', *Sunday Times Business Times*, 9 April, p. 1.

Dockel, J.A. & Ligthelm, A.A. 2002. 'Factors that contribute to small business survival' [submitted for publication]. Pretoria: University of South Africa.

Everett, J.E. & Watson, J. 1998. 'Small business failure and external risk factors', *Small Business Economics,* 11: 371–390.

Ghyoot, V.G. 2002. 'Using management science in site selection: a case study in office-site selection for a regional authority in South Africa', *Management Dynamics*, 11(3): 33–41.

Hodge, D. 2002. 'Inflation versus unemployment in South Africa: Is there a trade off?' *South African Journal of Economics*, 70(3): 417–443.

Ligthelm, A.A & Cant, M.C. 2002. Business success factors of SMEs in Gauteng: a proactive entrepreneurial approach. Pretoria: Bureau of Market Research, Unisa. (Research Report No. 311.)

Peterson, R.A., Kozmetsky, G. & Ridgeway, N.M. 1983. 'Perceived causes of small business failures: a research note', *American Journal of Small Business,* 8(1): 15–19.

Saaty, T.L. 1980. *The Analytical Hierarchy Process.* New York: McGraw-Hill.

South African Reserve Bank. 2002. *Quarterly Bulletin.* Pretoria.

Statistics South Africa. 2003. *Labour Force Survey.* September 2002. Pretoria.

Viviers, S., Van Eeden, S. & Venter, D. 2001. *Identifying small business problems in the South African context for proactive entrepreneurial education.* Paper delivered the 11th Global Int-Ent Conference, Kruger National Park, 2–4 July 2001.

Informal retail structures in South Africa: an exploratory study[1]

*A.A. Ligthelm**

The current climate of widespread unemployment and high levels of poverty has resulted in a burgeoning microenterprise economy in South Africa. The proliferation of small enterprises, particularly in the retail sector, is also thriving on the demand profile of households with limited disposable income. Their needs for commodities and services such as clothing, food, housing, transport and personal services can appropriately be supplied by the informal sector in terms of affordability, type and quality.

The aim of this article is to contribute towards a better insight into informal retail trade activities, focusing on spaza trade in South Africa. It is estimated that spaza retailers captured approximately 2.7% of South Africa's retail trade, amounting to more than R7 billion, in 2000. Approximately 70% of microbusinesses are concentrated in the trade sector. Of particular importance in this regard are spaza shops, defined as small retail businesses operating from a residential stand or home and engaged in the trading of consumer goods.

Introduction

It is estimated that over a quarter of workers in the world operate in the informal sector. In developing countries, informal activity mobilises between 30% and 80% of the workforce, mainly in cities, which are experiencing a large influx of people from the countryside. In Africa, it is estimated that this sector accounts for a significant (or even major) part of urban employment (two out of every three people derive their livelihoods from the informal sector) and it is estimated to be growing at an annual rate of 7% (Karl 2000: 53).

The emergence of the informal sector is largely attributed to the divergence between the growth in especially the urban population and employment growth in the formal economy. Job creation in the formal sector frequently occurs at a far lower rate than growth in the labour force. The shortage of productive employment opportunities therefore compels people to fend for themselves.

The needs of people with limited disposable income to attend to such things as feeding and clothing themselves and their families, finding accommodation and moving around generates a very significant demand for a variety of products and services that can be supplied by informal enterprises. Products and services provided by the informal sector, such as transport and motor repairs, are often more affordable than those supplied by the formal sector, and their type and quality more appropriate (Bugnicourt 2000: 55). It should, however, be recognised that branded products sold at small convenience shops may be more expensive than in large chain stores.

The informal sector comprises a variety of activities ranging from small commercial activities to small production and service enterprises. In most countries, including South Africa, the majority of informal businesses are located in the retail trade sector. This article focuses on informal businesses in this sector.

Objective and outline

The size and nature of the informal sector are largely unrecorded. The objective of this article is to determine the extent, role and characteristics of spaza trade in South Africa. An in-depth profile of and insight into the dynamics of spaza trade will allow roleplayers in the manufacturing, wholesale

* Professor A.A. Ligthelm is in the Bureau of Market Research, University of South Africa.
E-mail: *ligthaa@unisa.ac.za*.

1 This article is based on a comprehensive research report entitled *Characteristics of Spaza Retailers: Evidence from a National Survey,* published by the Bureau of Market Research, University of South Africa.

and retail sectors to mirror their activities against the overall size and activities of the spaza market segment.

The article seeks first to provide a brief exposition of concepts such as the *informal sector* and *spaza retailers* in order to develop a conceptual framework within which spaza trade can be considered. The article then goes on to describe briefly the survey methodology and to highlight the main findings of the research and concludes with calculations of the size of spaza retail trade in South Africa.

The informal sector

The term 'informal sector', as used in economic literature, was never proposed as a definition of a homogeneous group. Instead, the term was identified with a subset of economic activities in mind and intended to describe a domain as, for instance, rural or urban sector, with the purpose of merely focusing and demarcating research activities and policy development. As in the case of the urban or rural sector, the information sector needs to be disaggregated in order to draw meaningful conclusions (Sethuraman 1997).

Given that the informal sector encompasses numerous features ranging from type of activity to size, regulatory requirements and legal status, it is acknowledged that no single definition of the informal sector exists, nor is one anticipated (Naidoo 2002: 13). However, the International Labour Organisation's (ILO) International Conference of Labour Statisticians (ICLS) attempted to describe the informal sector for the purposes of delineating the sector for national accounting purposes (ILO 1993: 25). The resolution states that the informal sector as a subset of the household sector may be broadly characterised as consisting of units engaged in the production of goods and services with the primary objective of generating employment and incomes to the persons concerned. These units typically operate at a low level of organisation, with little or no division between labour and capital as factors of production and on a small scale. Labour relations, where they exist, are based mostly on casual employment, kinship or personal and social relations rather than contractual arrangements with formal guarantees.

On the basis of the above, informal enterprises display the following characteristics, among others:

- As far as legal identity is concerned, informal enterprises do not have corporate status.
- They do not maintain a complete set of accounting books. From a production point of view, at least some portion of production output is destined for the market.
- Enterprises are non-registered units in terms of taxes, labour or any other regulatory frameworks.

Although the resolution of the ICLS can be regarded as very general, it provides a perspective of the nature of enterprises considered to be part of the informal sector and is also applicable to the small spaza retailers discussed in this article.

Naidoo (2002: 31) maintains that activities undertaken in the informal sector could further be categorised into two main groups. On the one hand, people find employment in the informal sector as a result of a survival or employment strategy (working on casual jobs, having temporary or unpaid jobs, or holding more than one job) or, on the other hand, they could be involved in illegal or unofficial earning strategies (actively participating in tax evasion, avoidance of labour regulation and other government or institutional regulations, or operating an unregistered business). Generally, the majority of people in developing countries (and to some extent, those in transitional countries) resort to the informal sector as the only means of survival, while illegal participation in the informal sector tends to be encountered in more developed countries (Tansel 2000: 10). The informal sector plays an important and controversial role. While it offers opportunities for the unemployed, jobs in this sector tend to be low-paid, with little job security. Moreover, entrepreneurial activity is promoted, but at the price of non-compliance in respect of tax, labour and other regulations.

This article concentrates on only one subset of businesses operating in the informal economy, namely spazas or tuck shops.

Spaza shops

As already mentioned, enterprises in the informal sector embrace all sectors ranging from manufacturing to personal services and trade. This study covers only one business type within the retail trade sector, namely spaza shops. Prior to defining a spaza shop, the relative importance of the trade sector in the informal economy will be indicated.

In a World Bank (1993: 1) census of microenterprises in Mamelodi (near Pretoria) and KwaZakhele (near Port Elizabeth), it was found that approximately 70% of black microbusinesses are concen-

trated in the commerce and trade sector. The contribution of the trade sector as a percentage of total informal sector activity amounted to approximately one third (Ligthelm 1995: 12; Naidoo 2002).

The informal trade sector consists of various types of enterprises, ranging from spazas or tuckshops to hawkers and catering enterprises. Spazas represent an important component of the trade sector and form the focus of this study. For the purposes of this study, a spaza or tuck shop is defined as a shop/business operating in a section of an occupied residential home or in any other structure on a stand in a formal or informal township which is zoned (or used) for residential purposes and where people permanently live. The business practice of spazas entails ordinary retailing (the buying of consumer goods from manufacturers, wholesalers, et cetera and the selling of goods to customers). The terms 'spaza shop' and 'tuck shop' are used interchangeably as referring to the same type of enterprise.

Survey methodology

Scope of the survey

The survey units consisted of individual spaza shops located in South Africa, participating in the loan scheme of Quatro Trading.

Quatro Trading, which forms part of the NAIL (New Africa Investment Limited) group of companies, is a microlending organisation specialising in the niche market of informal trade finance and operating as a franchisor. At the time of the survey in 2000, Quatro Trading had 262 franchisee branches in South Africa and other countries in southern Africa. Quatro Trading franchisee branches are located on the premises of wholesalers to the informal enterprises and they supply trade-related finance to informal traders, particularly spaza shops. Quatro Trading had supplied credit to an estimated 15 000 to 20 000 informal traders at the time of the survey. The approximately 6 000 spaza retailers that were actively utilising Quatro Trading's loan facilities at the time of the survey were used as a sample frame for selecting spazas for interviewing (Ligthelm 2002: 7).

Since the sample frame comprised only Quatro Trading clients, the question still remained whether there is any difference between spaza retailers that are clients of Quatro Trading and those that do not make use of Quatro Trading's credit facilities. During the planning of the survey, the inclusion of both clients and non-clients of Quatro Trading was proposed. For practical reasons (especially in view of the national coverage of the survey and the limited time available to franchisees who acted as interviewers), non-clients were excluded from the survey. Quatro Trading's selection criteria for extending loans to spazas are very lenient, and the only enterprises that do not qualify are those that have been in existence for less than six months. However, Quatro Trading inspects applicants' businesses to ensure that they operate from some form of physical facility, thereby excluding hawkers from their client base. The relatively lenient approach is motivated by empowerment considerations, especially in view of the fact that Quatro Trading supplies business finance to the marginalised and under-banked sector of the South African economy. The distribution network is comprehensive, covering a range of areas, from rural to metropolitan. It can therefore be assumed that the clients of Quatro Trading are fairly representative of spaza shops in South Africa. Only the newly established (operating for less than six months) and larger informal spazas that do not need Quatro Trading credit are thus excluded from the survey population. This implies that the two extremes on the spaza spectrum are excluded (namely, small newly established spazas, and those that are in a position to negotiate loans or credit directly from wholesalers or financial institutions, for example).

Questionnaire

A prestructured questionnaire was used for collecting data. Most of the questionnaire had been used in a previous survey of spazas, implying the use of a well-tested questionnaire. The questionnaire enquired about the following issues, among others: profile of the spaza owner, physical characteristics of the spaza shop, financing, employment, procurement of merchandise, relationship with suppliers, transport, hostilities towards spazas, business problems, advertising and promotion, and location.

Sampling

A list of active clients of Quatro Trading formed the sample frame. It contained information on all franchisee branches and numbers of active customers. A random selection of spazas was drawn from this list. Quatro Trading supplied interviewers with the names and addresses of the spazas to be interviewed. If a spaza owner refused to participate in the survey, the spaza was replaced with a spaza retailer on a separate list in the possession of the project leader.

Response

A total of 340 spaza retailers were interviewed. They were located in all nine provinces, ranging from 26 in KwaZulu-Natal to 64 in the North-West province. Most of the survey was conducted during the latter part of 2000, excluding the holiday season, and a few questionnaires were completed during the first quarter of 2001.

Survey results and findings

The spaza entrepreneur

Most people in the economic and political world agree that promoting a spirit of entrepreneurship is the key to creating jobs and improving competitiveness and economic growth (Antipolis 2000: 7). Entrepreneurship combines innovation (such as the risk-taking provision of 'new' goods and services) and individual initiative, resulting in organisational renewal (in other words, the improvement of existing businesses and the establishment of new ones) (Cross 1995: 4; Research Centre of Entrepreneurship 2002: 2).

In their study of SMMEs in 37 countries, the 2002 Global Entrepreneurship Monitor (Reynolds, Bygraves, Autio, et al. 2002: 6) found that age and gender have a very stable relationship with entrepreneurial activity. Men are twice as likely to be involved as women, and those between the ages of 25 and 44 are most likely to be involved in all types of entrepreneurial activity.

The findings in the local spaza survey are closely aligned with international evidence. Just fewer than two thirds of spaza owners (64.0%) are male and 36.0% female. Almost eight out of every ten (78.5%) are aged between 25 and 49. The educational level shows that just over one third (36.9%) have a grade 12 (12 years of schooling) or higher (post-school) qualification. Only 4.2% have no formal schooling.

General firm characteristics

Starting a business

The response in a global entrepreneurship study to the question, 'Why do people become entrepreneurs?' (Reynolds et al. 2002: 6), shows that about two thirds of the adults engaged in entrepreneurial activities are voluntarily pursuing an attractive business opportunity, while about one third are engaged in entrepreneurship out of necessity, in that they can find no other suitable work. Opportunity-motivated entrepreneurs are dominant in developed countries, while necessity-motivated entrepreneurs comprise up to half of those involved in entrepreneurship in developing countries.

The local survey among spaza retailers indicates that just fewer than one in every four retailers can be described as opportunity-motivated entrepreneurs – only 6.2% of business owners chose to start their own firm when they perceived a lucrative business opportunity, while a further 17.6% of owners joined an already existing business, which is evidence of some form of family business culture present in the survey communities. The necessity-motivated entrepreneurs are dominant in the spaza retail trade. The survey shows that just over half (52.1%) the spaza owners were unemployed when they started their own businesses. Since the majority of the unemployed today comprise either young aspirant entrants to the market, who have never held a first job, or the long-term unemployed with non-transferable skills, this entrepreneurship pattern suggests that the majority of entrants have limited business experience when starting their own businesses. A further 17.9% wish to supplement their income, and 5.6% wish to work from home. These motivational forces do not necessarily ensure an effective business endeavour, however.

Business survival

The age of a business reflects its market experience and affects its ability to grow and move to the next size class. The market experience and size of a business are also important variables that determine a business's access to financial sources. A quarter (25%) of spazas have been in operation for more than seven years. A further 14.7% of businesses have been trading for between five and seven years. This is an encouraging finding against the backdrop of the high unemployment rate in South Africa, indicating that spazas may be regarded as a more permanent option for those unable to find work in the formal economy (39.7% of spazas have been in business for five years or more). However, the figure also confirms the relatively high rate of establishment of new businesses. Almost one in every four businesses (23.0%) have been in operation for less than two years and 39.5% for less than three years. International experience shows that births and deaths among small, medium and micro enterprises (SMMEs) tend to be concentrated among younger and smaller firms (Picot & Dupuy 1995).

Physical characteristics and infrastructure

Efficient service delivery, good infrastructure and safety and security are some of the leading factors

that contribute to a productive business climate and influence an investor's decision to locate in a certain area. This section highlights the locational aspects, as well as the spazas' access to infrastructure.

Physical structure

By definition, spazas are located in residential areas, either attached to or inside the main residential structure, or operated from a separate structure on the residential stand. Spaza operate from the following types of structures:

- Brick building in backyard: 28.8%
- Room/garage attached to house: 20.5%
- Inside main house: 19.9%
- Shack on stand: 13.1%
- Metal container on stand: 2.4%
- Other: 15.5%.

Location of spazas

As already mentioned, Quatro Trading's clients are spread across all types of areas. About one in three (29.9%) are located in rural areas, 23.4% in informal residential areas, and 44.9% in formal residential areas. Just under 2% were located on hostel premises. The majority of spaza retailers advertise the location of their shops by means of signboards.

Water and electricity

A substantial percentage of spazas confirm the availability of basic infrastructural services. Just over three quarters (76.5%) indicated that they have electricity, and more than three in every five (62.5%) have access to water on the stand from which the spaza operates.

Equipment

Spaza shops are operated with only basic equipment and amenities. Only 17.1% reported access to refrigeration facilities and 22.2% access to deep-freeze facilities. Fewer than one in every ten confirmed access to a telephone and a cash register.

Trading aspects

The majority of spaza clients are served over a counter (76.4%) or through a window (17.4%). Only 6.2% of spaza retailers allow self help.

Other findings of the survey include:

- Most respondents (86.1%) confirmed that merchandise is displayed in the shop.
- Most respondents (89.4%) display the selling price of merchandise either on the items themselves or on the shelves.
- Three out of every five spaza shops advertise the location of the shop by means of a signboard.

Entrepreneurial finance

Sources of capital

The survey findings indicate that the majority of spazas finance their start-up capital requirements through private savings and loans from family and relatives. The following were evident:

- Private savings, consisting of savings of owners and household members, are used by 62.6% of firms as the main source of finance for establishing their businesses. This is followed by a further 20.0% that use loans from relatives and friends. Private savings and loans from family and friends therefore constitute the main start-up capital source for 82.6% of firms.
- For 17.1% of respondents, retrenchment payments were a source of finance for starting up their businesses.
- A relatively small percentage of spazas rely on formal credit, such as loans from formal financial institutions (3.8%) and from informal organisations (1.8%).

This pattern correlates with international experience, in that most new small firms receive their initial financial support from informal investment made by family, friends, business associates and other personal contacts. An extremely small portion of the most promising firms (perhaps 1 in 10 000) receives funding from venture capital sources (Reynolds et al. 2002: 32). Worldwide, informal investment is thus a crucial component of the entrepreneurial process.

Capital investment

Capital requirements for the establishment of a spaza are limited. On average, start-up investment amounts to R4 058. This contrasts with the initial investment in spazas in Tembisa in 1998, which amounted to only R1 082 (Ligthelm & Van Zyl 1998: 65) and to the initial investment of R27 800 for establishing SMEs in Tshwane in 2001 (Ligthelm & Morojele 2001: 23).

Figure 1 shows the spectrum of spazas by investment category. Just under half (46.1%) were established with less than R1 500, and a third

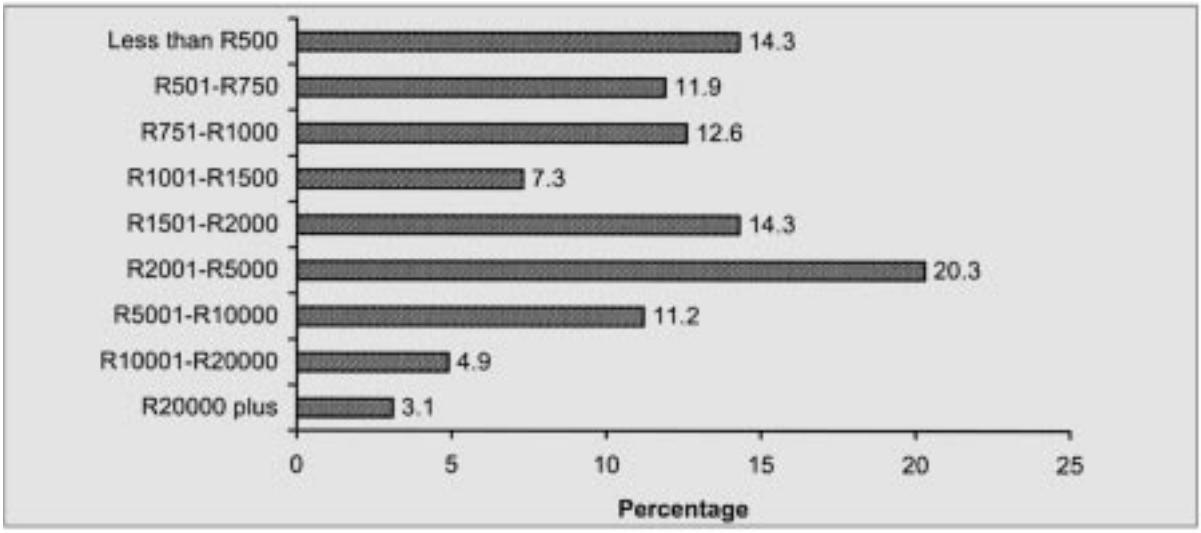


Figure 1: Amount of capital invested at start-up of the business by invesment category

(34.6%) with between R1 501 and R5 000. Only 8% invested in excess of R10 000 as start-up capital.

International experience confirms that SMMEs across the world are often established with limited capital. For instance, according to an analysis of America's fastest growing private companies in 2000, 16% started with less than US$1 000, 42% with US$10 000 or less, and 58% with US$20 000 or less, and fewer than 5% started with venture capital. Small investments, primarily by family and friends, are therefore crucial in funding not only microcompanies but also future superstars (Reynolds et al. 2002: 34).

Labour

Average employment

Recent policy indications by government highlight the notion that small business development can become an important source of employment and income generation. However, average employment of spaza retailers is very limited and is constantly eroded by a high degree of mortality in this sector. New firms enter on a regular basis, but fewer than half survive beyond five years (Ligthelm & Morojele 2001: 21).

Table 1 shows that the average employment of spaza retailers was 2.92 employees per business. This figure correlates closely with a spaza survey in Tembisa in 1998, which showed an average employment of 2.35 (Ligthelm & Van Zyl 1998: 27). The table also shows that spazas can be described as family-run businesses, because household members represent three-quarters (75.7%) of all employees. This is true for both full-time and part-time employees. It is also evident from the table that just over 80% of household members (80.1%) and other employees (80.3%) are involved in running the spaza shop on a full-time basis.

Labour remuneration

Respondents were asked to disclose their total payroll for the month preceding the interview. On average, spaza retailers spent R1 558 on labour remuneration, including the remuneration of the owner, in 2000. This figure averaged R533 per

Table 1: Average number of employees by type of employment

Type	Full-time	Part-time	Total
Members of household	1.77	0.44	2.21
Non-members	0.57	0.14	0.71
Total	**2.34**	**0.58**	**2.92**

employee per month in 2000, which is substantially less than the minimum wage of R800 (for a 45 hour week) prescribed for domestic workers in metropolitan areas since November 2002, despite the extended trading hours of spaza retailers. Total labour remuneration as a percentage of turnover amounted to 34.9%.

Relationship with suppliers

Type of suppliers

In response to the question, 'Indicate type of suppliers from whom merchandise is purchased', the majority of spaza owners (96.8%) indicated that they procure their merchandise from wholesalers, followed by the fresh produce market (23.2%) and manufacturers or producers (10.0%). Only 4.7% procure some of their merchandise from supermarkets and hypermarkets, and 3.8% from other retail outlets.

Support received from suppliers

Suppliers are becoming more aware of the size and growing importance of the spaza market. As a result, they are increasingly providing support to spazas. Figure 2 shows the percentage of spaza retailers that have received some form of support from suppliers. The most important is the provision of credit (47.7%), discount to spazas (40.9%), delivery of merchandise (21.8%) and discount prices (18.2%).

Product delivery

Respondents mentioned the following types of merchandise, in order of importance, as the most often delivered by suppliers: bakery products, soft drinks and dairy products.

Maize, personal care products (such as soap), gas, cigarettes and tobacco are also delivered in some areas.

Relationship with customers

Advantages and disadvantages of spazas

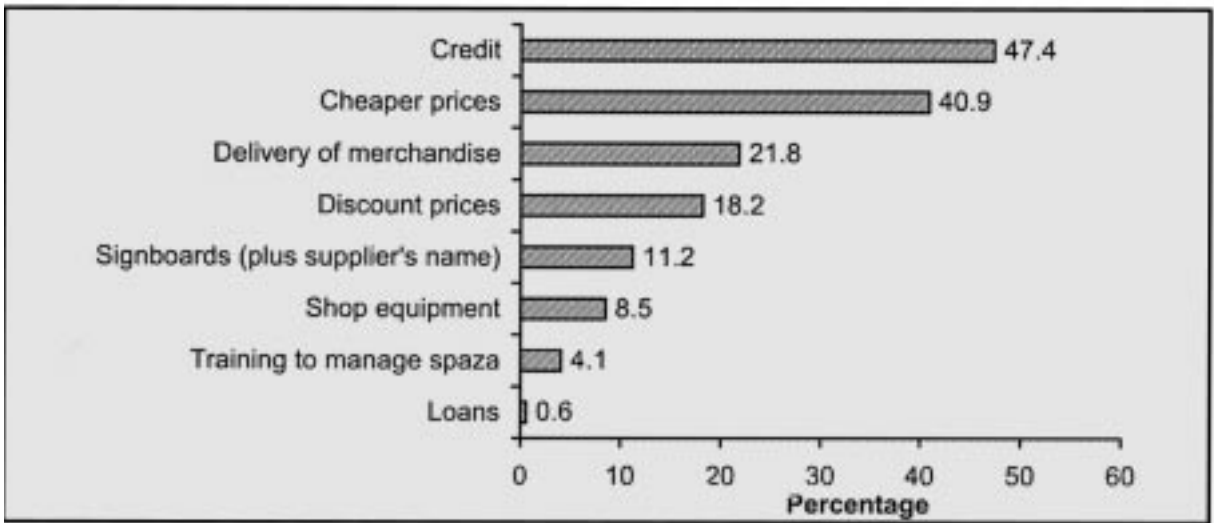
Respondents were asked about the advantages and disadvantages to the buying public of shopping at spazas. The three most important advantages of spazas, as perceived by owners, are that they are close to or within walking distance of customers' homes (51.8%), that they offer good, friendly service (40.0%) and they are open for long hours or at all times. The extension of credit was also mentioned by 20.9% of respondents as an important advantage that spazas offer the public.

The high prices of branded products was mentioned by 62.6% of spaza owners as a major disadvantage of spazas. Other disadvantages include stock shortages (39.1%), poor customer service (26.9%), dirty shop/environment (22.4%), poor product quality (16.2%) and the reluctance of some owners to offer credit to customers (14.4%).

Number and composition of customers and credit

The estimated average number of households (not individuals) that regularly buy at a particular spaza amounted to approximately 44 households per week. Just over two in every five spaza shops (41.2%) serve between 21 and 50 households regularly. Just over a third (34.2%) reported fewer

Figure 2: Type of support received from supplies



than 20 households, and 9.2% reported between 50 and 75 regular customers (households).

The gender and age compositions of spaza customers are as follows:

- On average, 42.4% of customers are male and 57.6% female.
- On average, 44,0% of customers are adults and 56.0% children.

As indicated, the extension of credit by spazas is perceived as an important service to the community. Credit is extended to 81.7% of the regular customers (households) of spazas. In reply to the question on how frequently customers settle their accounts, spaza owners indicated that almost three in every five (59.1%) settle their accounts on a monthly basis and 21.7% on a weekly basis.

Constraints encountered by spazas

Owners were asked to indicate the three most serious problems that they experience. Limited trading stock, resulting from a lack of finance, tops the list as the most prominent constraint experienced by spaza retailers (38.8%). This is followed by the high crime rate, mentioned by 25.0% of the respondents. Further constraints mentioned by respondents are severe competition resulting from overtrading (20.6%), the lack of (or expense of) transport to procure merchandise (19.7%), allowing too much credit resulting in bad debt (17.1%) and a lack of water and electricity (13.5%).

Respondents were also asked to mention the three most important means of improving their businesses. Maintaining sufficient stock levels was mentioned by almost a third of respondents as an important means of improving profitability. This was followed by more competitive prices (17.4%), financial assistance and loans (16.5%) and advertising and promotion (15.9%).

Turnover and products traded

The spazas were asked to indicate their monthly turnover according to specified class intervals. The calculations in this section are based on the averages of the class intervals. Figure 3 shows the monthly turnover of spaza retailers according to income group. Just under one third of the businesses (31.1%) recorded a turnover of between R1 000 and R2 999, while just over a third reported monthly turnover figures of between R3 000 and R9 999. A total of 16.3% experienced a turnover of less than R1 000 per month, and 17.6% more than R10 000.

The average monthly turnover reported by spaza retailers was R4 480. This turnover figure correlates

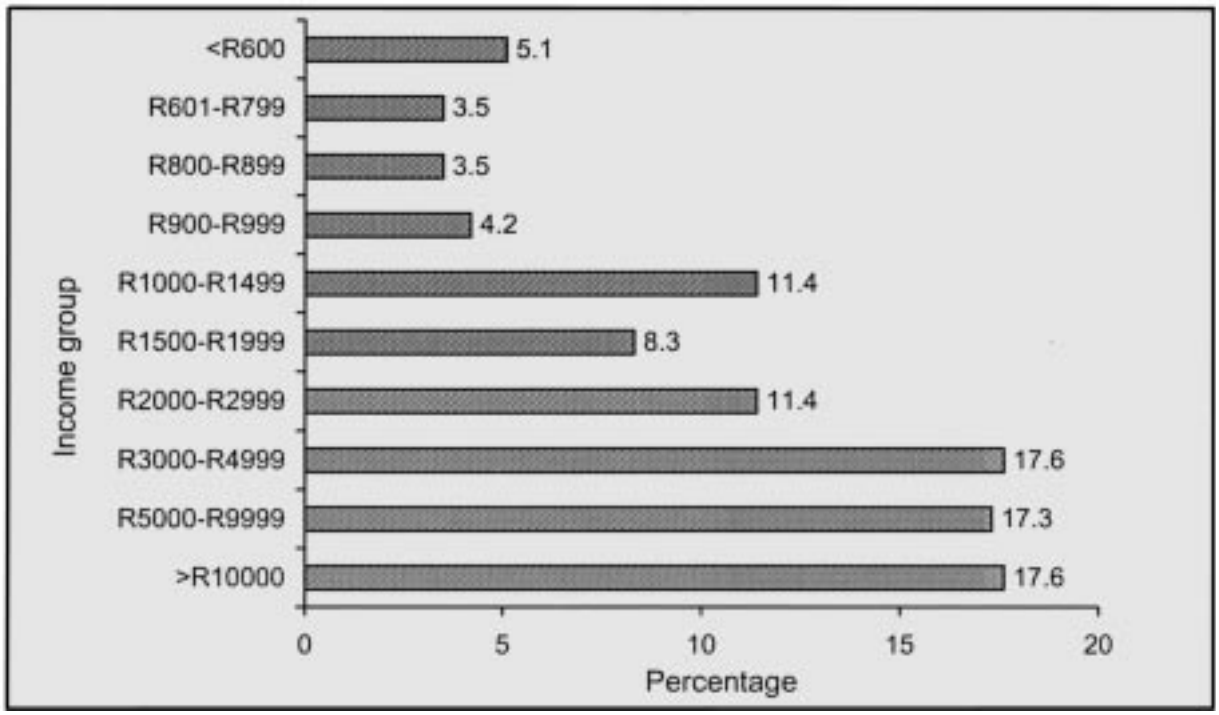


Figure 3: Monthly turnover of spaza retailers by income group

closely with the R4 350 monthly turnover reported by spaza retailers in Tembisa in 1998 (Ligthelm & Van Zyl 1998: 141).

Respondents were also asked to list the products that contribute the most to their monthly turnover. The following seven products were mentioned most often:

	% of spazas
Soft drinks	57.1
Cigarettes	55.9
Paraffin/candles	41.5
Maize meal	38.5
Alcolholic beverages	32.6
Bread	27.6
Sugar	20.9

Summary and conclusions

Although spaza retailers are often depicted in the literature as survivalist enterprises operating at bare survival level (Rogerson 1996; Reuters et al. 1994; Horn, Levin & Sofisa 1993), empirical evidence shows clear signs of spazas becoming not only a permanent phenomenon in the South African economic arena, but also more sophisticated and closely linked with the rest of the economy.

Signs of sophistication, permanency and linkages with the formal economy include the following:

- In reply to the question, 'Will you accept a job in the formal sector if offered today?', two out of every five respondents (41.1%) answered in the negative. It seems, therefore, that these owners perceive their spazas as a permanent career path and not as a stopgap for survival.
- Although a relatively high establishment rate prevails among spazas, the survey confirmed that 25.0% of spazas have been in operation for more than seven years and a further 14.7% have been in operation for between five and seven years.
- The majority of spazas have electricity and running water.
- In nine out of every ten spazas, goods are displayed with their prices marked on the items or shelves.
- Three out of every five spaza retailers advertise the location of their shops by means of signboards.
- Although the establishment of spazas requires only a small initial investment (R9 140), two thirds of spaza owners recorded additional capital needs once the business was in operation.
- The majority of spaza owners use bank facilities, such as savings and cheque accounts, in running their businesses.
- A large percentage of owners consult posters, pamphlets, catalogues and magazines when deciding what merchandise to procure.
- Ninety-six per cent of spaza retailers buy their stock from wholesalers.
- An ever-increasing number of producers and suppliers realise the importance of the spaza market and deliver goods to spaza shops.
- Spaza owners increasingly monitor marketing communication trends. For example, a large percentage confirmed an increase in turnover for a specific product during a radio advertising drive of the specific product.

The change in the composition of the South African economy towards a larger share of informal retailers in the market requires considerable adjustment in market views and production planning in various sectors of the economy. It is estimated that spaza retailers captured approximately 2.7% of retail trade (Martins 2001, 2002), amounting to just over R7 billion in 2000. The importance of this market segment is highlighted by the fact that the turnover of spaza retailers is larger than the combined turnover of the so-called branded superettes, including Kwikspar, 8 Till Late, Seven Eleven, Friendly Grocer, Foodies, OK Foods, Score, Rite Valu, Shield and Sentra. The turnover of spaza retailers constitutes just over 20% of the combined turnover of hypermarkets and supermarkets, including Shoprite/Checkers, Pick 'n Pay, Spar, Clicks and Woolworths Food Stores (A.C. Nielsen 2002).

References

Nielsen, A.C. 2002. *Grocery Universe*. Johannesburg.

Antipolis, N.C. 2000. *European Forum: Training for Entrepreneurship.* Brussels: European Union. Accessed at *www.europa.eu.int/comm/entrepreneurship*.

Bugnicourt, J. 2000. 'A hope ignored: the informal or popular economy', *The Courier: Africa-Caribbean-Pacific-EC,* (178), December/January 1999–2000.

Cross, J.C. 1995. 'Entrepreneurship and exploitation: measuring independence and dependence in the informal economy', *International Journal of Sociology and Social Planning,* 17(3/4): 37–63.

Horn, G.S., Levin, M. & Sofisa, T.N. 1993. 'The informal sector in Mamelodi and KwaZakhele', *Africa Insight,* 23.

ILO. 1993. *Resolution concerning the international classification of status of employment.* Fifteenth International Conference of Labour Statisticians, International Labour Office.

Karl, K. 2000. 'The informal sector', *The Courier: Africa-Caribbean-Pacific-EC,* 178, December/January 1999–2000.

Ligthelm, AA. 1995. *The informal sector of the South African Economy.* Pretoria: Bureau of Market Research, University of South Africa. (Research Report No. 222.)

Ligthelm, AA. 2002. *Characteristics of spaza retailers: evidence from a national survey.* Pretoria: Bureau of Market Research, University of South Africa. (Research Report No. 305).

Ligthelm, A.A. & Morojele, M. 2001. *Small scale enterprise development in the Tshwane metropolitan municipality: problems and future prospects.* Pretoria: Bureau of Market Research, University of South Africa. (Research Report No. 291).

Ligthelm, A.A. & Van Zyl, S.J.J. 1998. *Profile study of spaza retailers in Tembisa.* Pretoria: Bureau of Market Research, University of South Africa. (Research Report No. 249).

Martins, J.H. 2001. *Expenditure of households in Gauteng by expenditure item and type of outlet, 2000.* Pretoria: Bureau of Market Research, University of South Africa. (Research Report No. 289).

Martins, J.H. 2002. *Expenditure of households in the Cape Peninsula by expenditure item and type of outlet, 2001.* Pretoria: Bureau of Market Research, University of South Africa. (Research Report No. 304).

Naidoo, G.P. 2002. *An investigation into linkages between the formal and informal sectors in South Africa using the 1993 input-output table (1993).* Pretoria: Vista University.

Picot, G. & Dupuy, R. 1995. *Job creation by size class: Recent evidence for Canada*. Workshop on SMMEs, Washington DC.

Reynolds, P.D., Bygraves, W.D. & Autio, E., et al. 2002. *Global Entrepreneurship Monitor: 2002 Executive Report.* London: London Business School.

Research Centre of Entrepreneurship. 2002. *Mission statement and strategy.* Brussels: EHSAL Katolieke Universiteit.

Rogerson, C.M. 1996. *Rethinking the informal economy of South Africa.* Midrand: Development Bank of South Africa.

Sethuraman, S.V. 1997. *Urban poverty and the informal sector: a critical assessment of current strategies. New York*: United Nations Development Programme.

Tansel, A. 2000. 'Wage earners, self employed and gender in the informal sector in Turkey', Background paper to the conference on 'Barriers to participation: The informal sector in emerging democracies'.

World Bank. 1993. *Characteristics of and constraints facing black businesses in South Africa: survey results.* Washington.