Why do we believe in economy of scale?

Professor John Seddon, managing director Vanguard.

July 2010

‘Scale economy... is a concept that should be discarded.’
H Thomas Johnson, Professor of Accounting, Portland State University

I was motivated to write this by a recent experience. I bombed at a presentation to a group of county council chief executives. It was just before the election. I had started by remarking on my disappointment at the Conservative Party’s hiring of the people behind New Labour’s public-sector factories; Sir Peter Gershon and Martin Read were now telling them that billions could be saved by sharing services across Whitehall. I asked the chief executives to help me by explaining why it is that we believe in economies of scale. After some poor shots at responding, I realised that they did not know. I made the mistake of describing this as ‘pathetic’. An error, I know, but that is how I felt. These highly paid executives are spending large amounts of money on creating huge public-sector service factories without any conscious rationale for doing so. Lacking knowledge of the arguments for economies of scale, they just assume the theory has some basis in fact.

As they didn’t know, it is reasonable to assume that the people in central government, whose bidding the chief executives were doing, wouldn’t know either. So for everyone who doesn’t know, this is a guide to why we believe in economies of scale and why we shouldn’t.

Adam Smith

On the back of a twenty-pound note there is a picture of Adam Smith and the words:

‘The division of labour in pin manufacturing: (and the great increase in the quantity of work that results)’

This much many people know. But before we leap to the conclusion that the division of labour is universally valuable and economic, let’s take a broader look back. Adam Smith was a moral philosopher and recognised as the father of modern economics. We need to explore the way economists have thought about economy of scale, for the idea has had such a strong influence on politicians and managers alike.

Judging from the economics literature, the arguments for economy of scale are a mix of a) the plausibly obvious and b) a little hard data, brought together to produce two broad assertions, for which there is little hard factual evidence.

a) The plausibly obvious
To run a business you need a phone. You may grow your business while you only have the one phone, so your (phone) overheads fall as a proportion of your revenue. It is generally true that, for the same inputs (people, machines, lights, phones), you will be able to get more output up to a point where you have to add more of some input. Economists turn this obvious truth into a theory that can be illustrated with a graph: the long-run average cost curve.

Let’s look at the long-run average cost curve.

![Long-Run Average Cost (LAC) Curve](https://example.com/graph.png)

The graph describes the relationship between unit cost and the volume of output for a firm. The economists tell us that as we move from the far left towards the point ‘A’ we are experiencing economies of scale. At point ‘A’, where the curve is at its lowest, we are making the most of the resources at hand and performance is optimal. Think, for example, of centralising telephone calls in a service organisation. We will be able to pick up more calls, because we have more available resource (people ready to take calls). Compared with more, smaller, units, a centralised, larger-scale service might need fewer managers and might take advantage of only one telephony system and one building. So the unit cost per call falls. This is an argument for sharing services.

If we keep moving to the right from the point ‘A’, we start to experience diseconomies of scale. Economists talk of rising costs as organisations become more complex, bureaucratic and so on; in extraction industries they point to the higher costs of exploiting more-difficult-to-get-to resources. Again, obvious truths, and we might think of a parallel in the public sector, where the cost of the thousands of specifiers, inspectors and ‘improvement’ agencies that supposedly ‘drove’ public-sector reform under the last government, was not only a cost in itself, but created greater costs through the public sector’s compliance with ill-judged ideas.

But is there proof for the LAC theory? While it isn’t difficult to find anecdotal evidence to support the idea, there are no empirical studies showing it has predictive capability. It is an
idea: a reasonable idea, even a plausible one, but it is not knowledge. You cannot use it to predict where you will be on the curve or to know whether you are moving towards or away from optimisation. As a theory for practicing managers, it is about as useful as a chocolate fireguard.

Building on the theory of the long-run average cost curve, economists argue there are three reasons for economies of scale:

**Indivisibilities** This is the phone argument above. You need to have inputs to create your outputs and as you can’t have half a phone or half a manager, some costs are ‘indivisible’. Although this is obvious, its relevance will vary enormously between organisations. It does raise practical questions but it cannot be used in any predictive way by managers. It may explain phenomena, but it is not a theory: it is not based on knowledge. It is no more than an observation.

The observation will enable managers to speculate about the relationship between costs (inputs) and ends (outputs), but this will not be the same as having knowledge of the means. And only knowledge of the means will lead to sustainable improvement.

**Machinery** Large scale is presumed to be necessary to take advantage of investment in large machines. In manufacturing the justification for larger machinery is lower unit costs as bigger machines can turn out more things faster. In service organisations, large scale is thought to be necessary to take advantage of investment in computers as the means for the control and management of the work. The argument is that the investment in telephony and information technology systems necessitates larger scale – doing more work – to reap the benefit.

While economy through investment in larger machinery might be plausibly obvious, I will show later that it is not universally true. To put it another way: low total costs are not necessarily achievable through low unit costs; in fact, as we shall see, they are perfectly compatible with relatively high costs per unit.

**Specialisation** This is the Adam Smith pins argument: greater productivity results from breaking tasks down into their parts. This is something that is proven, so we have to ask about its generalisability – will it always be true? It is specialisation that has provided us with the little hard data.

— **b) A little hard data**

Adam Smith’s results cannot be argued with. Nor can those of Henry Ford. On the Model T production line in the early 20th century, specialisation of tasks led to great leaps in productivity (although in any full accounting part of the gain would be offset by the enormous human cost). For Ford, specialisation was coupled with standardisation. The parts on Model Ts are universal; any part from any year fits any other. This was standardisation at its zenith and to this day it is of enormous value to people who restore Model Ts. But this is hardly the hallmark of modern-day manufacturing.
And that is about the sum of the arguments for economies of scale. Despite the evidence being limited, it is a mainstream idea that is treated as an axiom by renowned management experts like Michael Porter, who claimed:

‘Scale economies can be present in nearly every function of a business, including manufacturing, purchasing, research and development, marketing, service network, sales force utilization, and distribution.’\(^1\)

Like the economists, Porter provides no evidence. In one sense he is saying that anything can be improved and that is true, but he equates improvement with scale, and to be sure about that we need evidence.

So to summarise, the arguments for economies of scale boil down to two broad assertions: the organisation design argument and the savings argument.

1. The organisation design argument: specialisation and standardisation will lead to lower costs and greater productivity
2. The savings argument: cost savings will be made through common IT systems, less buildings and fewer managers.

Both arguments are concerned with the management of costs. Indeed, everything we have discussed thus far has been concerned with costs. It seems that managing costs has become managers’ de-facto purpose. But should that really be their only or even number one concern?

**Industrialisation of the public sector**

Because this is the route that most mass production followed, does it follow that public services should be specialised, standardised and centralised – so that they can become industrialised in the same way? Should we treat a requirement from a customer as something to be dissected into parts and ‘reassembled’? For this is exactly what service factories do. They take a requirement, break it into its constituent parts and send the parts to specialised workers who are expected to complete them in ‘standard’ times. The workers are measured as production resources, monitored and managed with activity targets (so many units per hour or day), service standards (how quickly something is done) and standard times (the time given to do it).

To manage these systems, organisations need layers of management above the work whose purposes are to plan, monitor and report on the use of resources and the achievement of the measures-in-use. Management is primarily concerned with managing activity in the assumption that activity represents cost. But managing costs causes costs – it is the Achilles heel in the argument for economies of scale.

**Managing costs causes costs**

Before we explore how managing costs causes costs in service organisations, we need to consider an important advance in manufacturing, where this counterintuitive truth was first discovered.

Taiichi Ohno was the man who developed the Toyota Production System: a system designed to make vehicles not at the rate the machines demanded in order to achieve economies of scale, but at the rate of customer demand. He took Ford’s ideas a massive leap further. While Ford was concerned with unit cost, Ohno concentrated on total cost. He took the view that cost was in flow – how smoothly and economically the parts were brought together in the final assembly – not just the aggregation of unit costs. How long the part is in the system is also part of its real cost. So Ohno’s factory had parts delivered to the manufacturing line at the rate the line required them – ‘pulled’ in his language (by kanban or ‘just-in-time’).

While ‘scale’ manufacturing plants are usually replete with stock – inventory at steps throughout the manufacturing process and finished goods at the end, created because managers focus on running machines, keeping them busy, to achieve low unit costs – Ohno minimised stock throughout the process, his ideal batch size being one. Whereas most manufacturers still focus on unit costs (and employ accountants for whom it is central to their doctrine: for example, inventory counts as ‘value’ on the balance sheet), Ohno focused on the flow of the work, confident that better flow would lead to lower overall costs. And so it did. His system would tolerate higher unit costs; it was not dependent on low costs per unit. What was critical was the availability of the part, not the cost – an affront to convention. Ohno was the first to demonstrate that greater economy comes from flow rather than scale.

His second and more profound challenge to convention was to put variety into the line, making different models in the same production line. A corollary of economies of scale is that a system must be large and standardised to deliver high volume, and thus low cost; but by definition it can’t deliver variety at the same time. Variety = low volumes = high cost. The Toyota System disproved this axiom and broke new ground. It showed that it is possible to manufacture small volumes in high varieties at lower costs. It was Toyota’s ability to achieve high variety, high quality and low cost at the same time that caused Tom Johnson to propose that the concept of economies of scale had outlived its usefulness and should be abandoned.

It is because he was developing a manufacturing system to produce a variety of vehicles at the rate of customer demand (get the order, make the car) that Ohno employed the Japanese term ‘muda’ to describe the types of waste that he could see in the Toyota factories. His insight was that anything not connected to the ‘pull’ of the customer added cost. Excess inventory would slow the whole system down and inhibit the ability to produce variety at the rate of demand. It would also result in cars being produced that did not correspond to customer demand and whose price would have to be cut in order to shift them – more muda. Managers of conventional mass-production organisations might consider it kamikaze manufacturing: what if the necessary part was not available? But Ohno knew relentless attention to flow would make sure that it was. And putting the workers in control of the line – they could stop it any time they saw a defect – would ensure flow was treated as paramount.
Perhaps more has been written about Toyota than any other manufacturer in history, with the possible exception of Ford’s first assembly line. Yet despite its fame among manufacturers (and leaving aside the recent fall from grace, which by Toyota’s own admission was due to moving away from these principles), the Toyota System has had no influence on mainstream economic thinking.

The secret of Ohno’s method was studying the work, as a system. His favourite word was ‘understanding’. He would turn in his grave if he knew that his system was being sold to managers as a set of ‘lean’ tools; he taught managers to find out what their real problems were, as they were often different from the ones they thought they had. The tools developed in the Toyota System were developed to solve the problems associated with designing a system to produce cars at the rate of demand. In service organisations we have different problems to solve.

**Economy of scale in service organisations**

So let us now turn to what the drive to create economies of scale has meant in practice in the public sector. What actions have managers of service organisations taken in pursuit of scale economies and do they deliver better economic results? As we do that I shall make observations on what is learned when we study what actually happens in the work, as opposed to economic theory.

**Centralising call-handling**

The first action taken by service managers is to centralise call-handling. They do this in pursuit of the obvious truth that bringing together work currently carried out over a number of units with small groups of people will allow them to take more calls and make better use of resources (agents will take more calls in any working period). As we have already noted, a single large unit may economise on managers and equipment – a single telephony system, for instance. It also promises to provide a better service, particularly if we think of service as measured by ‘service level’ – how quickly we pick up the phone. But, of course, picking up the phone tells us nothing about the service a customer receives. We will come back to that.

Centralised call-handling was enabled by the invention of the bedrock of call-centre telephony: the automatic call-distribution system. However, today’s technology enables the same feature – distribution of calls – to be applied across sites. Which brings us back to the long-run-average cost discussion: physically centralising calls is not necessarily a ‘no-brainer’ – the same can be achieved virtually – other factors may also have a bearing on the right answer for any particular circumstance.

**Moving more work on to phones**

Moving work on to phones is the second and distinctively different idea favoured by scale thinkers. Telephone interaction costs less than face-to-face service. Is this an obvious truth or potentially misleading? It is certainly plausible. It is a distinctively different idea because it
makes the assumption that the work that was done face-to-face can be done on the phone. But it can create problems, as I illustrate by example.

The movement of ‘telephone work’ to call centres from bank branches in the 1980s caused an explosion in the volumes of demand – many more phone calls. So the banks built more call centres. Why? It was the result of an egregious error made by most service organisations in the private sector (which doesn’t like talking about mistakes). Designating some elements of service as ‘telephone work’ failed to take into account the relationship between the customer and service-provider. The call-centre agent is not the person you were talking to in the mortgage department. The rise in call volumes was attributable to the creation of ‘failure demand’ (demand caused by a failure to do something or do something right for the customer, Seddon 2003). The same phenomenon duly occurred more recently in the public sector. When, complying with central requirements, local authorities and housing associations established call centres and moved telephone work from their various departments into them, demand was much greater than expected or planned for. However, the increasing demand was not the result of success but of failure. Our studies show that failure demand in such call centres can run as high as 80% of the total. In other words, the vast majority of these expensive call centres’ efforts go into doing work that shouldn’t be done at all.

It is important to note that management’s failure was a failure to know what was going on in the work and to make the mistake of assuming all calls are ‘work to be done’; something I need to explain before we move on.

**Economy of scale: the management equation**

Managers of service organisations are taught to be preoccupied with solving the following equation: How much work is coming in? How many people do I have? And how long do they take to do things? The calculation is then used to plan and manage resources. It has been embedded in ‘best practice’ guidance for the public sector. However, the first question – how much work is coming in? – begs a more important one: what is the nature of the work coming in?

Failure demand is an idea that is easily understood and, bizarrely, it became the focus of a government target (NI 14). The reasoning is clear: failure demand represents massive costs, so removing it will reduce costs; setting targets is the way to remove it. As with all target regimes, the consequence was mis- and under-reporting. It is impossible remove failure demand without understanding and rooting out its causes, something to which we will return.

For now, the point is that moving work to a call centre without understanding the work and determining whether the call centre can provide that service leads to more (failure) demand. This is why many of our private-sector clients are now repatriating outsourced work from India. If one call (even at lower costs) creates more calls, then costs go up not down. Costs are in flow, not in activity.

---

Of course managers endeavour to organise the service to solve callers’ problems, but always fail to achieve complete success. These are two common tactics:

- **Specialisation:** By making workers specialize in certain kinds of work, managers can cut training costs. Specialisation also allows computers to route calls (press one for this and two for that).
- **Graded response:** This means having generalists to pick up the calls with two levels of specialists (‘level two and level three’) behind them. The idea is to concentrate the most expensive resource on the most demanding calls.

These measures sound plausibly sensible, but in practice neither of them works very well. They create fresh problems of: operational definitions of work types, accurate sorting, accurate allocating, reconnecting fragmented tasks. To solve these problems managers create their own factory, swathes of managers whose job is it to solve these problems, yet they unwittingly exacerbate the problems; their vain attempts to achieve control send the services further out of control.

At the heart of the problem is the issue of variety. A key difference between service and manufacturing is the nature of demand. Studying customer demands into service organisations always reveals high variety. It is variety that stymies management’s attempts to control the work and deliver high quality service. The problem is amplified by the division of work between a front and a back office.

**The back office**

The ‘back office’ was invented to ‘de-couple’ the service work from the customer. The argument goes like this: the job of management is to get the most from resources (people), service workers are often interrupted by customers, so having a front office to find out and describe to the back office what the customer needs enables the latter to work without interruption and thus optimise the use of its time and effort. Result: economies of scale.

In practice, what happens is that customers’ demands become fragmented. In the private sector it is not unusual for one customer request to be turned into as many as nine sub-tasks to be completed by different people in back-office functions. As in other mass-production systems, managers believe that by breaking down and standardising tasks they will gain economies of scale. They take as given that the work will arrive in the right places, be done in the standard times and returned within the service levels. Careful study of the work shows that this rarely happens. The fragmentation of work creates waste in the back office and failure demand in the front office.

These designs are usually institutionalised with IT systems. Thus, workflow systems used to connect front and back offices turn work into electronic work objects for transmission between computers. Citizens’ records can then be tracked through ‘customer relationship management’ (CRM) software. In practice CRM mostly serves to bureaucratisse lots of value-less activity: passing messages, progress-chasing and so on. If the service worked, there
would be a lot less traffic on the CRM system. If failure demand is high, CRM systems are full of institutionalised waste.

Meanwhile, look at what is being reported to management: more activity from the workers and lower transaction costs – the promise of scale economies is delivered! Yet the system is evidently creating costs. The first signal – as it was for the banks in the 1980s – is rising volumes of work, but as organisations don’t take historic measures of work volumes, and the planning and contractual measures associated with new public-sector factories are all concerned with service levels and unit costs, it is not readily apparent. The recent studies by Advice UK³ highlight the enormous volumes of knock-on failure demand into voluntary services caused by a failure of two primary agencies – the DWP and HMRC – to deliver effective services where people need them.

Similar alarming costs of failure are evident in all public-sector factories. If you doubt this claim spend a few hours studying customer demand into public-sector services.

One cost that is apparent is sickness, absenteeism and staff turnover. Being treated as a ‘resource’ to be ‘optimised’ is not motivating. Nor is the realisation that it is impossible to help people solve their problems because of the need to work to the internal arbitrary measures. In some respects life in modern public-sector factories is little different to the conditions that created Ford’s ‘five-day man’⁴. Both HMRC and NHS Direct currently report low levels of staff morale. It is a design problem that managers mistakenly treat as a ‘people’ problem.

**Shared services**

The argument is as follows: we all do HR, IT and finance, so why don’t we share them, creating centralised units to provide these services and gain economies of scale? We may grant the ‘cost-savings’ argument for fewer managers, IT systems and buildings. But we should have strong doubts about the ‘organisation design’ argument – simplify, standardise, centralise – for fear it will have an adverse impact on the service. Is this a sound method? From a systems view it is not, since this approach will prevent the system absorbing variety. In fact, that is specifically what it is designed to do: remember Henry Ford’s famous, ‘Any customer can have a car painted any color that he wants, so long as it is black’.

To get a measure of the consequences, consider some of the fantastically expensive failures of shared services projects, such as the shared services programme for the UK Research

---


⁴ Henry Ford’s mass production plants had an astronomical turnover of workers. Newly hired workers stayed an average of only 3 months. Many walked off the job without any formal notification and were presumed to have quit after missing five days of work: the notion of the ‘five day man’ was born and at one point accounted for 70 per cent of the workers leaving Ford (Buchanan and Huczynski 2004 ‘Organizational Behaviour: An Introductory Text’ Pearson Education: Harlow p441)
Councils (sharing IT, HR and finance) which was bought at £40m and is now forecast to cost £120m, and it ain’t over till the shared services work (which they don’t and probably won’t).

Much the same has been reported with the shared services initiative at the Department for Transport: sharing HR and finance was supposed to save the taxpayer £57m; but it is now on track to cost £81m. These projects had the added complication of being driven by an IT system: in simple terms the IT system was created and then forced on to the work.

As with the economists’ arguments for economies of scale, the focus of all of the above activity is on cost – pushing down transaction costs – driven by Treasury economic theories, not empirical knowledge.

**From observations to theory**

Just as economists make observations and turn them into theory, I shall, presumptively, do the same. We have made the following observations and know that they are predictive:

- If you move work to the telephone without understanding whether and how you can deliver the service by telephone, you will create failure demand
- If you specialise and/or standardise work, you will create more waste (failure demand and waste in flow)
- If you control the work and workers through arbitrary measures (service level agreements, targets, standards) you will create more waste (failure demand and waste in flow)

The second and third observations are grounded in the knowledge that specialisation and standardisation are the causes of waste and failure demand in service organisations. Yet they featured in the last government’s reform regime’s ideas about ‘best practice’; compliance created and continues to create waste and higher costs, not improvement. It is particularly ironic that failure demand became a target. Management’s focus needs to switch management from cost to value. Following Ohno, managing value will cause costs to fall. We can say that:

- Any waste in a system represents, by definition, a failure to provide value for customers
- Studying (value and failure) demand is the most powerful lever for improvement
- The design of work should focus on providing value, and that alone
- A focus on managing value will drive costs out of a system
- Control needs to be located where the work is done (controlled by the workers)

These principles enable the system to absorb variety; hence we can predict that service will improve as costs fall. We can theorise these principles together as a different way to design a service: we could describe this principle as ‘economy of flow’.

There is a growing body of evidence to support the idea of economy of flow. Work designed to manage value – serve customer needs – provides better service at much lower costs, and transforms morale at the same time. In the private sector, moving from a ‘specialised’ service

---

5 The Independent 19/01/2010 ‘Labour’s computer blunders cost £26bn’

6 For more on how these principles translate to service design see “Re-thinking Lean Service”, Seddon, O’Donovan and Zokaei: http://www.systemsthinking.co.uk/6-brendan-jul09.asp
centre to one trained to handle all customer issues releases an average of 20% of the operating costs. Removing the front/back-office split releases a similar amount. Both tactics also deliver improvements to the service. Further, often greater, savings are made by removing the now redundant management activity that formerly sat above the old design, trying to solve the wrong problems. In the public sector the gains from redesigning services are much greater than any subsequent savings from ‘sharing’, which are marginal by comparison.

As with the TPS in manufacturing, service design based on economy of flow challenges convention. It is local: expertise is provided at the point of transaction for the bulk of the work, with further expertise ‘pulled’ by the customers’ demands as necessary.

Because flow designs are simple, they make much less use of IT than conventional scale-based systems. Before we consider how, let us return to Ohno’s discoveries in manufacturing. While conventional wisdom has it that bigger machines produce greater volumes at lower costs, Ohno taught his managers to get massive improvements from older and smaller machines. His insight was that if flow the most important attribute of a machine is to be ready when the part arrives. This may mean that at other times it is more efficient for machines to stand idle – anathema to conventional thinking.

In much the same way, flow designs use ordinary desktop software for much of the local control of work, and new IT systems are not introduced without first understanding and improving the work (without changing the IT). If IT is then required its precise requirement is better understood. The consequence is generally much lower spend on IT and much more value from it.

To give one example: Comserv/MTS, a private-sector provider working with Portsmouth City Council, has developed a housing repairs service that provides repairs on the day and time the tenants request. If BT could do that we would all cheer. The design developed by the players delivers the service at half the original cost. It represents an economic benchmark; an extraordinary example of improvement. Having developed the organisation design manually they then brought IT skills in to automate the features required. The new IT system, supporting this economic benchmark, cost all of £3,000! Most IT systems bought for managing repairs work cost upwards of £100,000 – to measure and control the wrong things.

**The knowledge opportunity**

For how long have we heard politicians and economists say we live in a knowledge economy? In contrast to the rhetoric, the recent past has seen these actors conspire to dumb down public services. By contrast, massive performance benefits are being realised by those who have had the courage to ‘smarten up’ – redesigning their service to place human expertise at the point of transaction to service customer demands. Putting people in control

---

actually improves the control within a system and is intrinsically motivating for all those involved. It also underlines that while the division of labour among workers is often a mistake, a much bigger one is the division of labour between managers (as decision-makers) and workers.

In short, we should maintain a healthy scepticism for generalised claims about management derived from economics. One authoritative commentator put it this way:

‘Orthodox economics is in many ways an empty box. Its understanding of the world is similar to that of the physical sciences in the Middle Ages. A few insights have been obtained which will stand the test of time, but they are very few indeed, and the whole basis of conventional economics is deeply flawed… Increasingly, the subject is taught not as a way of learning how the world might operate, but as a set of discovered truths about how the world does operate… It cannot be stated too often that very little of the content of (economic) textbooks is known to be true, in the sense that many of the statements on, say, engineering are known to be true.’

The management gurus have also had their doubts. One of the central threads in Tom Peters’ first book (‘In Search of Excellence’1982) is that scale models seem to apply only in very limited ways to human beings. One factor may be that as the number of people in organisations increases arithmetically, possible interactions between them – thus complexity, thus the difficulty of managing them – goes up geometrically. At any event, Peters and his co-writer Waterman were struck by the repeated finding that effectiveness seemed to go with much smaller, less tidy units than conventional wisdom proposed.

‘Whatever happened to economies of scale?... The excellent companies understand that beyond a certain surprisingly small size, diseconomies of scale seem to set in with a vengeance’

When he was at the Cabinet Office Geoff Mulgan commissioned a study into the evidence for economies of scale in public-sector reform. The study found none. It was not published.

While the plausibly obvious ‘cost-savings’ argument remains relevant but something that should not preoccupy managers, the organisation design argument, being based on cost thinking, is the fundamental flaw in scale thinking. Two complementary strands of evidence grow: the costs of failure from industrialisation and the profound improvements being achieved through managing value and flow.

Economy of scale is a myth. Economy comes from flow.

---