Evaluating Fund Capacity: Issues and Methods

Dr Michael O’Neill
Portfolio Manager, Investors Mutual
Adjunct Associate Professor, Bond Business School

Dr Geoff Warren
Research Director, CIFR

Synopsis
We examine the issues and methods involved in evaluating the size that a fund might attain before it becomes unable to create additional value for investors. We discuss how capacity is defined; identify ten drivers; and outline methods for conducting capacity analysis. Models that predict capacity assuming a fund adjusts the manner in which it trades or constructs portfolios as funds under management increase are detailed. We also provide an overview of transaction cost modeling, which is integral to predicting capacity. Implications for fund managers, as well as asset owners and other fund investors, are highlighted. This report is primarily written as an aid for investment industry participants who wish to evaluate the capacity associated with a given investment signal.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Introduction</td>
<td>2</td>
</tr>
<tr>
<td>1.1 Report Summary</td>
<td>2</td>
</tr>
<tr>
<td>1.2 Implications</td>
<td>4</td>
</tr>
<tr>
<td>2. Defining Capacity</td>
<td>5</td>
</tr>
<tr>
<td>3. Fund Size, Performance and the Drivers of Capacity</td>
<td>7</td>
</tr>
<tr>
<td>3.1 Literature on Fund Size and Performance</td>
<td>7</td>
</tr>
<tr>
<td>3.2 Drivers</td>
<td>8</td>
</tr>
<tr>
<td>3.3 Implications for Capacity Analysis</td>
<td>15</td>
</tr>
<tr>
<td>4. Analyzing Capacity</td>
<td>16</td>
</tr>
<tr>
<td>4.1 Approaches</td>
<td>16</td>
</tr>
<tr>
<td>4.2 Literature: Capacity Analysis for Specific Strategies or Signals</td>
<td>21</td>
</tr>
<tr>
<td>5. Predictive Models and their Component Parts</td>
<td>23</td>
</tr>
<tr>
<td>5.1 Implementation Shortfall and Related Concepts</td>
<td>24</td>
</tr>
<tr>
<td>5.2 Frameworks for Predictive Models</td>
<td>24</td>
</tr>
<tr>
<td>5.3 Predictive Models Appearing in the Literature</td>
<td>29</td>
</tr>
<tr>
<td>5.4 Modeling of Execution Costs</td>
<td>32</td>
</tr>
<tr>
<td>5.5 Expected Alpha and Opportunity Costs</td>
<td>37</td>
</tr>
</tbody>
</table>

References | 40 |
1. Overview

This report examines the evaluation of capacity in an investment management context, outlining the key issues and various methods of analysis. We address the following question: “how large can a fund get before it is unable to create additional value for its investors?” In doing so, we frame the discussion under the assumption that an active fund forms a portfolio based on a particular investment ‘signal’, which should be read as a general term for the information or process by which investment opportunities are identified. A signal may comprise indicators, forecasts, viewpoints, a strategy, and so on. The issue being considered is how far a signal can be leveraged through additional funds under management (FUM), before management should consider either closing the fund to new money, or changing the signal and thus the investment process.

This report is written as a reference document that we trust will be of value to the investment management community. It addresses the drivers and analysis of capacity, and reviews the related literature. The presentation mostly revolves around capacity for equity funds, which has been the focus of the publically available research. This report also provides the foundation for forthcoming research. Two additional outputs are planned. One is an opinion piece discussing capacity from the perspective of institutional asset owners, including the management of capacity in a multi-asset context and related agency issues when using external managers. Another report will illustrate the various methods for measuring capacity, and investigate the sensitivity of capacity estimates to how it is modeled.

The remainder of this overview comprises two parts. The first summarizes the key messages from the report by section. The second sets out the implications for fund managers, and then asset owners and other fund investors.

1.1. Report Summary

**Section 2** discusses the **definition** of capacity. We observe that multiple definitions exist, distinguished by the approach used to identify the fund size beyond which a fund loses the ability to create additional value. For the purpose of this report, the definition we embrace is **the FUM beyond which a fund can no longer deliver the return required by investors using a particular investment signal**. The latter may include a component to compensate investors for both active risk and fees. This definition views capacity from the investor’s perspective. It is differentiated from other definitions that focus on the maximization of total value-add in dollar terms, or the point at which excess returns are eroded entirely.

**Section 3** identifies **ten drivers** of capacity. These provide a mix of positive and negative effects, many of which are interconnected. Three drivers act to enhance capacity: **economies of scale** related to fixed costs; **economies of scope**, such as improved access to opportunities and information networks for larger funds; and **flexibility to adjust the implementation** of a signal as a fund grows. The latter can help mitigate the adverse effects of fund size. Of the drivers acting to limit capacity, **diseconomies in trading and portfolio construction** are a key element. These relate to the increasing difficulty in capturing the returns arising from a signal as FUM grows, due to problems in trading assets in the desired size at attractive prices. Key factors are that larger trades result in greater execution costs due to market impact, coupled with the possibility that a fund may be blocked from taking some positions due to encountering trade or holding constraints. In essence, a limited quantum of value can be extracted from most signals, and this inevitably gets diluted as FUM grows beyond some level. We also identify three other limiting factors on capacity: **organizational diseconomies; staff effects**, notably any loss of focus by key individuals as a fund grows; and **other investors using similar strategies or signals**. Two drivers act as influences on potential capacity: the investment approach, and the investment universe. **Investment approach** relates to the nature of the signal and how it is implemented, including: portfolio concentration; the nature of the opportunities generated by the signal (e.g. their distribution, and the
magnitude and timing of associated returns); whether the approach demands or supplies liquidity; investment style; and dealing capability. The *investment universe* in which a signal is being applied can influence capacity through both the potential to trade in volume with high liquidity, and the nature and range of opportunities being generated. These aspects can change over time, leading to capacity varying with market conditions. Our tenth driver relates to *asymmetry in accumulating versus liquidating*, and the idea that exiting large positions can be more difficult than building them. This asymmetry links to capacity via the risk that large funds could encounter difficulty in exiting under certain circumstances.

Section 4 discusses the analysis of capacity. We observe that capacity analysis falls into four broad groups. First is *rules of thumb*, where capacity is defined with regard to simple measures such as the percentage of market segment being addressed or days to exit. Second is *ex-post analysis*, under which selected portfolio metrics are examined for signs that capacity constraints may be emerging. Metrics that might be monitored include: number of positions; signs of convergence between the portfolio and its benchmark; execution costs and time to complete trades; realized portfolio returns versus a target portfolio; and trends in active returns. Third is *simulation analysis based on an existing portfolio or signal*. This involves re-evaluating performance or implementation under conditions where FUM is ‘scaled up’, allowing for portfolio or trading constraints, or perhaps the greater transaction costs associated with larger trades. Simulation analysis works from the assumption that the fund aims to implement a signal in a similar way, and evaluate the costs or constraints encountered in doing so. The fourth approach, which we call *predictive models*, goes a step further by assuming that a fund adjusts either its trading strategy or portfolio construction as FUM increases. This approach is the most advanced, and arguably realistic as it is reasonable to assume that funds will adjust the way they manage as FUM increases. However, it is also the most difficult to apply, taking the analysis into the territory of dynamic optimization under some techniques. Predictive models are also exposed to model and parameter uncertainty, thus outputs need to be interpreted with care.

Section 4 also reviews the literature on capacity analysis. We overview the research that examines the capacity associated with particular signals, which is often conducted using simulation analysis. We also summarize the literature on predictive models. The main finding is that the results of capacity analysis appear to be highly dependent on the model employed and how the analysis is undertaken.

Section 5 takes a close look at predictive models and their components. The discussion is aimed at readers wanting to delve into the modeling of capacity and transaction costs in some depth. (Others may wish to skip this section.) A key feature of the predictive models is that they embed the trade-off that a fund faces as FUM and hence the required trade size increases. Specifically, trading in larger volume gives rise to greater potential execution costs, but this may be partly mitigated by delaying trades in order to spread the market impact. However, delaying trades runs the risk that the return arising from a signal might be lost if the market price adjusts. Hence predictive models *trade off execution costs against the opportunity costs* of missed returns. Predictive models estimate capacity on the assumption that the fund manages this trade-off with respect to its particular signal. In describing such models, we commence by explaining the concept of *implementation shortfall*, which expresses the reduction of returns arising from the sum of execution cost and opportunity cost. We outline how predictive models may allow for the trade-off between execution costs and opportunity costs in either scheduling trades, or more generally in the construction of the portfolio itself. Section 5 also discusses two key inputs into predictive models: a *transaction cost model*, and an *accrual profile for expected returns* arising from a signal. The extensive literature on transaction cost modeling is summarized, noting the lack of consistency in how transaction costs are modeled by researchers. We observe that existing research has paid little attention to the return accrual profile associated with differing signals.

Section 6 rounds off by offering some concluding comments, and discussing ‘what is next?’ This section outlines the plans for additional work under this project, and notes areas for further research.
1.2. Implications

For Fund Managers

This report provides a companion for fund managers who are interested in evaluating capacity for a fund that aims to exploit a specific signal. Below we draw out some of the more notable aspects that arise, and offer some advice.

- **Capacity analysis is an inexact science** – Both intuition and a review of the literature suggests that capacity estimates can vary significantly with how the analysis is conducted. Further, capacity is dynamic, and may fluctuate with aspects such as market conditions and how the fund manager or other players respond under increasing FUM. Accordingly, capacity is best evaluated from a range of perspectives; and estimates should be interpreted using judgment and not taken literally.

- **Adding rigor into the estimation of capacity is probably worth the effort.** There are three areas where a more limited analysis might yield inaccurate estimates of capacity. First is that capacity can be overestimated by failing to allow for both transaction costs and any holding or trading constraints. In particular, it is important to accommodate the possibility that some desired positions may become infeasible at higher FUM. Second, capacity might be mis-estimated if the modeling is based on broad portfolio averages in circumstances where the distribution matters. For example, if a fund that invests across the market generates the bulk of its returns from small-caps, conducting analysis on a segmented or even stock-by-stock basis becomes necessary. Third, capacity can be underestimated by failing to account for the latitude to adjust how a signal is implemented as FUM increases. The latter suggests that using predictive models may be worthwhile.

- **Ex-post analysis should be undertaken regardless** – Ex-post analysis of portfolio metrics provides a means of keeping a finger on the pulse. While examining historical outcomes may not indicate the capacity that a fund might ultimately target, it can flag when capacity problems are emerging. Further, model-based analysis is never completely reliable, and always subject to both model and parameter uncertainty. For instance, the modeling of how transaction costs increase with trade size cannot be undertaken with precision. Capacity analysis based on simulations or predictive models should always be accompanied by ex-post analysis to provide confirmatory evidence.

- **Setting up to analyze capacity** – Funds intent on undertaking rigorous capacity analysis should maintain records that allow construction of a target (i.e. paper) portfolio, as well as perhaps data on the signals and associated expected returns. Comparison between the actual and target portfolio returns provides a direct estimate of implementation shortfall, which is one of the better ex-post metrics. Applying predictive models also requires estimates of the magnitude and accrual profile of expected returns from the signal. While this could be estimated by observing the realized returns from target portfolio positions or the underlying signal, data on the signal itself and the return expectations on which the target portfolio was based can support a more direct analysis.

- **More to capacity than just trading costs and constraints** – Quantitative capacity analysis only tells part of the story. Other influences on capacity should be considered, even if in a qualitative manner. Thought should be given to whether there are any significant economies of scale or scope; and whether organizational or staff constraints may matter given the way that a fund is managed. It is also critical to gauge the implications arising from other funds using the signal or similar strategies, and whether these funds are attracting inflows.

For Asset Owners, and Other Fund Investors

We intend to address capacity from the perspective of asset owners, such as pension funds, in a forthcoming companion piece. Nevertheless, the current report should assist such investors in their role as users of investment management services, both through external managers and via internal asset
management. With regard to external managers, it highlights the issues related to fund capacity that asset owners (or their consultants) need to consider when selecting managers and evaluating their performance. It provides guidance on the questions to ask managers; and the sort of analysis that might be requested to demonstrate capacity. The issues and methods raised should also assist in the management of multi-manager portfolios, including manager transitions.

In addition, there is a trend towards larger asset owners becoming directly involved in investment management by bringing assets in-house (see Gallagher et al., 2016). One consideration in doing so is to address perceived capacity limits in using external active managers, specifically the mandate size that managers will accept. Our discussions hold various implications for in-house management. First, to the extent that capacity relates to investment signals rather than funds, in-house management may not alleviate capacity constraints where the in-house team pursues a similar investment approach to external managers. To address capacity issues, a complementary or different approach should be adopted. Second, our analysis provides direction on investment approaches that might be most ‘scalable’ for a large asset owner. Scalable approaches might: focus on larger, liquid assets; limit the trading required through either adopting less concentrated portfolios, or investing for the long term; or be liquidity supplying, rather than liquidity demanding. Examples of potentially effective approaches include: establishing a ‘core’ portfolio where modest active bets are taken in large liquid securities, possibly based on quantitative signals; adopting strategic positions in assets, perhaps coupled with close engagement; thematic investing; long-term, contrarian investing; and standing ready to provide funds in support of large capital raisings. Gallagher et al. (2016) relay examples of Australian superannuation funds following some of these approaches in managing in-house. Finally, asset owners might bear in mind that capacity may relate to organizational aspects, or constraints on what key employees can handle. These elements need to be managed. They also may prove the limiting factor in many of the alternative asset classes.

2. Defining Capacity

Our focus is how investment performance changes as FUM increases. For these purposes, we define ‘capacity’ as the FUM beyond which a fund can no longer deliver the return required by investors using a particular investment signal. This general definition takes the perspective of the investor. Applying it requires establishing the boundaries for FUM and the signal. It also involves specifying whether capacity is measured on a ‘normal course of business’ or ‘liquidation’ basis.

Vangelisti (2006) proposes the three ‘definitions’ of capacity outlined below. Each implies a different measure of the ‘value’ by which capacity is evaluated.

- **Threshold capacity** – FUM beyond which a fund can no longer achieve its stated objective, e.g. a target excess return or alpha. This measure of capacity can accommodate making allowance for sufficient returns to compensate investors for active risk and the fees they pay.

- **Wealth-maximizing capacity** – FUM that maximizes the amount of wealth created, e.g. FUM that maximizes the product of alpha and FUM. This creates the greatest amount of value to be shared between the fund manager and investors. The distribution of the available value will depend on the fees that investors pay to the fund manager.

- **Terminal capacity** – FUM at which net excess return is reduced to zero. Under certain conditions this may amount to the FUM that maximizes revenue for the fund management company, even though the investor may reap no benefit.
The threshold capacity definition is most appropriate where the concern is delivering adequate value to end-investors. This is the definition that we adopt. Applying this measure requires establishing a return or alpha threshold in accordance with investor objectives. Note that this measure does not imply maximizing the magnitude of returns or alpha, which might occur at very low levels of FUM. Rather, it amounts to a promise to at least deliver the outcomes that investors require.

One boundary issue that emerges when defining FUM is: ‘whose FUM?’ Capacity is often viewed from the perspective of an individual fund; and we also follow this line. The key limitation of this approach is that capacity does not necessarily attach to a ‘fund’ per se, but rather to investment signals. There is likely to be overlap in signals used by various investors, which effectively means that a fund is probably sharing the available capacity. Any capacity analysis needs to establish whether the scope is a particular fund, overlapping funds operating within an investment organization, or all investors who are utilizing a signal. The problem with the latter is that it extremely difficult to apply. Further, capacity analysis is often undertaken in practice by specific funds, or perhaps funds using similar signals within an investment organization, with a view to answering the question of how much FUM should be accepted. This report is written from this perspective, in full knowledge that defining capacity with respect to a fund imposes a somewhat unnatural limit on the scope of analysis which may lead to capacity being overestimated. The difficulties in defining the scope of FUM exploiting a signal only reinforce the notion that capacity analysis is an inexact science.

A second boundary issue relates to the scope of the ‘signal’, and the investment process that aims to exploit it. This has implications for what actions are admitted under capacity analysis. The following points present a progressive expansion in scope with regard to the signal and the related process:

- The fund is assumed to trade and form portfolios in exactly the same manner;
- The fund attempts to trade and form portfolios in exactly the same manner, but allowance is made for being unable to do so due to hard constraints such as limits on asset holdings or volume that can be traded in the market;
- The fund forms the same target portfolio based on the signal, but adjusts the way that it implements as FUM increases by scheduling trades to limit expected transaction costs;
- The fund adjusts the portfolio that it targets based on the signal as FUM increases, including the way that it implements the trades, by taking into account expected transaction costs;
- The fund adjusts the signal itself as FUM increases.

Each expansion in scope implies capacity is being evaluated for a somewhat broader definition of ‘the signal’. For example, if transaction costs are taken into account, then the cost of transacting effectively becomes endogenous and arguably a component of the signal on which expected returns and hence portfolios are assumed to be formed. In this report, we draw the line at the fourth dot point, thus excluding making changes to the base signal as FUM increases. This implies that capacity is evaluated with respect to a particular signal that generates forecasts for expected returns, while allowing adjustments to be made to how the signal is implemented. Limiting the scope of analysis to a particular signal – and ruling out changes to the signal itself – broadly accords with the idea that investors may be concerned with the extent to which the returns arising from particular investment process can be reproduced under higher FUM. Widening the scope to include changes in the signal would probably amount to analyzing a substantially different investment process.

---

1 It is very hard to determine the extent to which a particular signal is driving investment decisions across the industry. Investment processes often differ to some degree, so that any overlap will probably be partial. Investment processes typically mix a range of signals with differing weights, and this is hard to observe.
Capacity can also be estimated under the assumption of ‘normal course of business’ or ‘liquidation’. The normal course of business definition assumes that the fund is expected to continue in its current form, and is concerned about evaluating capacity in order to establish the scope to accept additional inflows. We write this report from this perspective, noting that this accords with the usual motivation for undertaking capacity analysis in the first place. The alternative is estimating capacity under the assumption that the fund is being liquidated, either in whole or in part. The reason to undertake such an analysis would be to address how FUM might interact with the downside risks that can arise from the need to meet redemptions. While this is useful to know in order to understand the full risk-return profile associated with a fund as it grows, we will not be delving into this approach in any depth. However, we recommend that any evaluation of capacity include consideration of the downside risks associated with becoming too large, specifically around the ability to liquidate positions in order to meet redemptions in response to a change of view.

Once capacity is attained with respect to a particular signal, it is envisaged that a fund would give consideration to one of two courses of action. The first would be closing the fund to new money. The second would be adjusting the signal itself, and hence moving to a new investment process. This second course of action should not be undertaken lightly, particularly where fund investors expect the investment process that they invested in originally.2

3. Fund Size, Performance and the Drivers of Capacity

The concept of capacity assumes that there exists some underlying relation between fund size and performance. In this section, we explore the nature of such a relation, and hence the drivers of capacity. We commence with an overview of the academic literature that examines whether there is any observed relation between FUM and performance. We highlight that the evidence is mixed, and discuss why this is not inconsistent with the existence of an optimal capacity. We then list and discuss ten drivers of the relation between fund size and performance, and by implication capacity. The main message is that the relation between fund size and performance is far from straightforward; and that a multitude of potential drivers of capacity exist. Further, not all drivers of capacity are directly related to FUM, or may do so in a highly contextual and dynamic manner.

3.1. Literature on Fund Size and Performance

If capacity constraints exist in investment management, one might initially expect to observe a negative relation between FUM and performance – at least beyond some minimum level of efficient size. However, this need not necessarily hold. No relation between FUM and performance could be observed in the data under certain conditions. Assume that all funds would generate similar excess returns when operating at the optimal FUM (i.e. capacity) for their particular investment approach; but that the optimal FUM varies with approach. Under the case where all funds operate at capacity, one might observe variation in FUM but no variation in performance, so that there would appear to be no relation between FUM and performance. This situation would be consistent with Berk and Green (2004), who hypothesize that excess returns will not be observed in equilibrium as either fund inflows or fees will adjust until excess returns are eroded. Note that this hypothesis assumes that investors respond efficiently to evidence that a manager is skilled, so that the best funds gather FUM quickly. On the other hand, a relation between FUM and performance may arise if the fund management market did not converge on the equilibrium described by Berk and Green (2004), due to frictions that permit some funds to operate at other than optimal FUM (i.e. capacity), or result in a lagged response. For example, a negative relation between FUM and performance might be observed if some funds accepted more

2 In some instances, investors expect innovation in the investment process, e.g. some quant funds and hedge funds.
FUM than was optimal; or it took time for fund flows to react so that performance for smaller funds held up for a period before attracting enough investors to erode the excess. A relation between size and performance might also be observed if optimal FUM was correlated with excess returns in some way.

Against this background, the academic evidence on the relation between FUM and performance is, perhaps unsurprisingly, quite mixed. Chen et al. (2004), Yan (2008), Chan et al. (2009) and Vidal-Garcia (2016) find that smaller funds outperform larger funds, and provide evidence that the relation is strongest for funds with greater demand for liquidity due to aspects such as focusing on small caps, less liquid securities, or their investment or trading style. Fung et al. (2008) document that fund flows act to reduce alpha for hedge funds. These studies all hint that increases in FUM are associated with reduced performance. On the other hand, Indro et al. (1999) find that performance increases with fund size up until the 9th decile. This is suggestive that there exists an optimal capacity, and that not all funds operate at the optimal level. Another body of research fails to find any significant relation between FUM and performance, e.g. Grinblatt and Titman (1989), Gallagher and Martin (2005) and Elton et al. (2012).

Evidence also emerges from the literature that differences in portfolio construction, trading strategy and fees help to mitigate the impact of larger FUM on performance, through limiting the costs associated with size. For example, Chan et al. (2009) find that larger fund size is associated with more securities, less small stocks, lower bet sizes and less trading. Grinblatt and Titman (1989) and Christoffersen et al. (2007) provide evidence that larger funds incur lower transaction costs; and Elton et al. (2012) mention that larger funds have lower expense ratios. These results underpin the notion of capacity as a dynamic concept. They suggest that fund managers change how they manage as FUM increases.

### 3.2. Drivers

This section discusses the potential influences on the relation between size and performance, and henceforth capacity, from the perspective of an individual fund. We identify ten drivers, some positive, some negative, and some acting as influences on the potential level of capacity. The drivers we discuss are listed in Figure 1. As the drivers interact, they should be viewed as a body of influences rather than separate items. A key message is that the relation between fund size and performance is conditional on the circumstances of the fund and the investment approach that it employs. The remainder of this section discusses each driver, and leaves the implications for capacity analysis to Section 3.3.

<table>
<thead>
<tr>
<th>Driver</th>
<th>Effect on Capacity Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i) Economies of scale arising from fixed costs</td>
<td>Positive</td>
</tr>
<tr>
<td>(ii) Economies of scope</td>
<td>Positive</td>
</tr>
<tr>
<td>(iii) Organizational diseconomies</td>
<td>Negative</td>
</tr>
<tr>
<td>(iv) Staff effects</td>
<td>Mixed, possibly negative</td>
</tr>
<tr>
<td>(v) Diseconomies in trading and portfolio construction</td>
<td>Key negative</td>
</tr>
<tr>
<td>(vi) Investment approach</td>
<td>Influence on potential capacity</td>
</tr>
<tr>
<td>(vii) Investment universe</td>
<td>Influence on potential capacity</td>
</tr>
<tr>
<td>(viii) Other investors using similar strategies or signals</td>
<td>Negative</td>
</tr>
<tr>
<td>(ix) Flexibility to adjust the implementation</td>
<td>Mitigating effect</td>
</tr>
<tr>
<td>(x) Asymmetry in accumulating vs. liquidating</td>
<td>Negative; relates to risk</td>
</tr>
</tbody>
</table>

3 Foster and Warren (2015) describe such a process.
(i) Economies of scale arising from fixed costs

Economies of scale related to the spreading of fixed costs can help to boost net returns as FUM grows, and hence work to enhance capacity. Fixed costs in fund management relate to aspects like administration, systems and research. The existence of service providers such as custodians and administrators reduces the potential magnitude of this effect, to the extent that they allow smaller funds to share in the economies of scale by out-sourcing their back office and other administrative functions.

(ii) Economies of scope

Economies of scope refer to the fact that large funds may benefit from greater access to opportunities and information networks, as a consequence of their larger capital, influence and resources (Scobie, 2013; RARE, 2015). Larger funds have the resources to analyze a wider range of opportunities, including being more able to invest in unlisted assets. They can build deeper internal and external networks to assist in the gathering and interpretation of information. In addition, larger funds are often more relevant to service providers such as stockbrokers, simply because they pay more brokerage than small funds. This can generate preferable access to deal flow and information. Some of these economies of scope may relate to fund management organizations (e.g. fund families – see Chen et al., 2004) rather than specific funds, which adds complexity to the relation between performance and FUM at a fund level. In some asset classes – particularly unlisted markets – the ability to source assets can be an important constraint on capacity. Larger funds can sometimes have an advantage in sourcing and managing such as due to their greater resources and broader networks. Indeed, size can be an advantage in assets such as unlisted infrastructure or property for this reason.

(iii) Organizational diseconomies

Larger size can give rise to bureaucracy and co-ordination problems, as the number of employees and complexity increase. However, the impact may depend on how the organization is structured. Larger funds can be less flexible at times, and hence less able to respond quickly to opportunities. This can especially be the case where multiple approvals are required – although this problem can be mitigated by delegating the authority to act. Fund managers within larger organizations can be called on to give greater attention to business matters including marketing, with an associated lessening of focus on investment. Eggins (2008) compares Australian start-up or ‘boutique’ equity managers with larger ‘institutional’ managers over the period 1996 to early-2008. He finds that boutique managers initially generate relatively high risk-adjusted returns, which progressively erode over the 5 years after start-up. He attributes this pattern in part to managers initially focusing predominantly on investment issues, which then shifts to managing the development and risk of their business as the fund grows and matures. On the other hand, larger funds can mitigate these problems, at least in part, by appointing investment product specialists. The latter can ensure that fund managers remain focused on investment matters by absorbing the load of addressing business issues and marketing.

In some situations, organizational diseconomies may relate to the number of positions or transactions that a fund pursues, rather than FUM per se. Lopez-de-Silanes et al. (2015) provide evidence that diseconomies of scale in private equity are related to the number of simultaneous investments, rather than FUM. They link their findings to organization theories of knowledge transfer and communications, and the diseconomies that can arise as the number of required actions increases. This suggests that capacity may relate to the number of positions or transactions in circumstances where the constraining resource is the ability of an organization to coordinate multiple tasks. Such situations may arise where either intensive management is required, or where there exists a range of delegated tasks to manage. This is more likely to occur for unlisted alternative assets where intensive management of investments

---

Eggins (2008) also refers to small managers benefiting from the flexibility afforded by low FUM.
is critical for success. Broadening the capability of an organization to invest in such assets can hence be a balancing act between achieving economies of scope and limiting any co-ordination diseconomies.

(iv) **Staff effects**

An increase in fund size can give rise to a range of effects related to staff. These effects are mixed, and may depend on aspects such as how the organization and remuneration is structured. On balance, staff effects are probably more likely to limit capacity. On the positive side, larger funds have access to greater resources that could be used to attract skilled staff, or add more staff as required. They can also offer career progression. These positives are pitched against some important negatives.

Skilled managers often prefer not to work for large funds. Many are keen to remain focused on investing, and have a strong desire to outperform. Such personality types may value the flexibility of more modest FUM; are reluctant to get drawn into business issues; and do not value career progression within large organizations. If anything, they are likely to start up their own fund, rather than remain with a large employer.

Further, Pollet and Wilson (2008) suggest that there may be limits to the human capital that can be productively added to a fund, perhaps due to hierarchy costs. The issue is that focus can be diluted as an investment team grows, as it gives rise to the need to organize, relate to, and train other team members. While this concept is closely related to the organizational diseconomies discussed above, the additional point is that success in investment management can rely heavily on the skill of a particular individual or a small group of key staff. Where this is the case, capacity may be related to the time constraints placed on the key individuals. If growth in FUM distracts these key individuals due to the need to deal with other staff members, this could be to the detriment of performance. Adding more staff does not help, and might only make things worse. Whether the time of key individuals is an important constraint will vary across funds. For example, it is more likely to be an issue for a fundamental stock-picker, and less of a problem for a quantitative process that is applied across a stock universe. It may also depend on the extent that the organization is structured to allow managers to focus primarily on investment matters.

(v) **Diseconomies in trading and portfolio construction**

Diseconomies related to trading and portfolio construction is the most widely recognized source of potential deterioration in performance as FUM grows, and are indeed a key driver of capacity. In broad terms, if there is a limit to the maximum dollar value that can be extracted from a signal, it is inevitable that returns per dollar invested must decline as FUM grows beyond some point. The mechanism by which return erosion occurs is via a decline in the ability to extract returns from a signal under larger required positions and hence trade sizes, which become either inaccessible, or only able to be accessed on less attractive terms due to market impact or deferral of trades. In this way, rising FUM can result in greater implementation shortfall: a concept initially raised by Perold (1988), and which we describe in depth in Section 5.1.

Deterioration in the ability to effectively invest on a signal as FUM increases occurs for a number of inter-connected reasons:

- **Holding constraints** – Constraints may exist on the holdings that can be accessed as a proportion of the size of individual securities or market segments. For example, a fund may be constrained to hold no more than 5% of the free float of any particular stock under its mandate, or because becoming a substantial holder in any one company is considered imprudent. Such holding constraints are encountered progressively as FUM grows, becoming binding from smaller assets upwards.
• **Trading constraints** – In some instances, it may not be possible to trade because the volume is either unavailable, or a self-imposed trading constraint is encountered.\(^5\) While it might be argued that volume is *always* available at some price, in practice this concept is tested only in exceptional circumstances. Trades are often implemented by ‘participating’ in the order flow available in the market. A fund attempting to trade might expect to be able to secure (say) 10%-40% of the available volume on average, depending on the trading strategy. If the order flow is absent, trades do not occur. Trading constraints are encountered progressively as FUM grows, becoming binding from more illiquid assets upwards.\(^6\)

• **Market impact** – It is well documented that the prices at which transactions occur deteriorate as trade size increases.\(^7\) This occurs for two reasons. First, larger price concessions may be required to encourage another party to provide the liquidity by taking the other side. In part, this is tied to asymmetric information and adverse selection: the counterparty can become increasingly concerned over the risk that they might be trading with an informed investor as trade size increases. It could also reflect sloped supply and demand curves for assets, at least in the short run, perhaps due to limits on the capital available to market makers. The second source of market impact is that the information on which the fund is trading becomes incorporated into the price as a consequence of trading. That is, other investors infer the information from the trades or orders. To the extent that larger trades are more likely to be driven by information,\(^8\) the price will tend to adjust to a greater degree. Further, larger funds may be more exposed to market impact to the extent that they draw more attention to their activities: elements of game theory can come into play when a large fund is attempting to move its portfolio holdings. In sum, the nature of market impact means that the trades that do occur will tend to do so at less attractive prices as FUM increases.

• **Deferring trades** – Trades may be deferred for reasons of either trading constraints, or a decision to break up the trade in order to limit market impact. The potential risk in doing so is that the price moves in an adverse direction, leading to the opportunity cost of missing out on the available returns. Given the impact of both trade constraints and market impact increase with FUM (as discussed above), it follows that exposure to opportunity cost from deferring trades also increases with FUM.

• **Foregoing trades** – In addition, increasing FUM makes it more likely that some trades will be foregone altogether because they become unattractive. This can occur as a consequence of the price fully incorporating the information that is embedded in a signal, either due to market impact or the information becoming revealed. An example of the latter would be where a manager correctly anticipates better-than-expected earnings, but is unable to secure sufficient volume prior to the earnings report due to a lack of order flow. As the probability of failing to trade prior to information being revealed is greater for larger trades, it becomes more likely that some returns will be left on the table entirely as FUM increases.

• **Investing in second-best opportunities** – An implication of not being able to implement a signal in full, due to either constraints or deferring and/or foregoing some trades, is that the portfolio weighting must be allocated to alternative investments. It is probable that the manager will end up ‘going down the list’ to opportunities that are not as attractive, thus diluting expected performance.

---

\(^5\) Trading constraints might be imposed as a means of limiting the potential for excessive market impact. They may also be useful when the use of available capacity needs to be budgeted.

\(^6\) A related issue is whether larger funds may be less able to extract excess returns from participating in capital raisings such as initial public offerings. This could occur to the extent that the more attractive new issues are concentrated among smaller companies, and the amount of stock available to a fund declines with FUM in relative terms. On the other hand, in some circumstances larger funds may be favored by issuers on the basis that they present as a substantial long-term investor. Investment banks may also favour their larger clients with allocations.

\(^7\) Related literature is discussed in Section 5.4.

\(^8\) For this reason, larger orders are often broken up into smaller orders so the investor ‘hides their tracks’.
(vi) Investment approach

Capacity varies with how and where a fund invests. For example, capacity is likely to be higher for funds that focus on liquid large-cap stocks and trade irregularly and patiently; and lower for funds where the approach relies on concentrated positions in small-caps that are turned over regularly. Here we address the role of investment approach, which interacts with capacity on various interconnected dimensions as discussed below (focusing on the partial impact, holding all else constant). The role of the investment universe to which an approach is applied is discussed as point (vii). The following dimensions can influence the capacity that arises under an investment approach:

- **Concentration or breadth** – More concentrated approaches, where fewer assets are held in the portfolio, tend to have lower capacity, as this implies larger positions and bigger trades when portfolios are rebalanced. For instance, the average position and trade size of a concentrated stock-picking fund with 20 positions will be in excess of 5-times greater than that for a fund that holds 100 positions and constrains itself to deviating from index weights by no more than a few percent (so that a portion of the portfolio is never turned over).

- **Payoff timing and turnover** – Investment approaches where the returns accrue over longer periods – that is, where the signal decays slowly – will tend to have lower turnover and higher capacity. For example, a fundamentally-based approach where the payoff occurs progressively over three years will have a turnover of about one-third, as well as not being as adversely impacted by deferring trades in order to limit market impact. This type of strategy should have much higher capacity than an approach that trades on reversals occurring over three months, which implies constant trading and turnover of about 400% p.a.

- **Magnitude and distribution of opportunities** – Opportunities that offer higher excess returns have greater capacity due to a larger buffer. Consider a situation where capacity is defined with respect to a required (excess) return of 3% p.a. An investment approach that is expected to generate a baseline return of (say) 8% can suffer larger return erosion through trading frictions and the like before reaching capacity, than an approach where the return is only 4%. The distribution of opportunities also matters. Capacity is likely to be lower where the best opportunities are concentrated in small, illiquid stocks. In addition, a low ‘hit rate’ might reduce capacity if it brings exposure to more losing positions, to the extent that these are harder to exit at attractive prices than winning positions (as discussed below).

- **Liquidity-supplying versus liquidity-demanding** – Investment approaches that are liquidity-supplying by their nature are likely to have higher capacity. Rather than incurring market impact, part of their return generation may actually rely on providing liquidity to other investors who are willing to pay for it. Capacity in this instance depends in part on the extent to which the market is populated by liquidity-demanders. Value investing and other contrarian approaches can be liquidity-supplying, as the fund will often be trading ‘against the market’. Another example is an enhanced index fund that is willing to deviate from the benchmark to supply liquidity when the opportunity arises, and then rebalancing back once prices have reverted. Liquidity-demanding strategies are more likely to be those with short payoff timing, as discussed above, as they are under pressure to get set quickly.

- **Style** – While capacity differs across investment styles, this often stems directly from the aspects discussed above, i.e. concentration, payoff timing, and the nature of opportunities, including whether they are liquidity supplying versus demanding. The investment universe from which opportunities are drawn (e.g. large or small cap) is also relevant, and is discussed below. In general, higher capacity tends to be associated with long-sighted fundamentally-based approaches, as they involve lower turnover and the timing of trading is often less critical. These include value investing, certain applications of growth investing, and investing on themes that extend over a cycle or the long term. Value investing and other contrarian approaches may be resilient to delay in trading on a signal.
Indeed, at times such delays can work to their advantage, to the extent that timing trades is difficult and signals arising under these styles can suggest moving too early. On the other hand, capacity tends to be lower for styles that are based on signals that are short in duration, and are laden with price-relevant information that increases the potential for market impact or opportunity cost from deferring trades. Examples include styles that focus on momentum, unrevealed information, or news flow.

Another consideration is the extent to which concentrated positions in specific securities are required (implying low breadth), versus potential to substitute between securities in order to obtain exposure to a particular factor. The latter can help to mitigate the effects of short horizons and high turnover for quant funds, for instance.

**Dealing capability** – Given the importance of market impact, the dealing capability of a fund can influence its capacity. An efficient trading process and skilled dealing desk can enhance capacity. Anand et al. (2012) provide evidence that institutional trading performance is economically important, persistent over time, and contributes significantly to portfolio returns. The ability to secure large blocks of an asset at relatively attractive prices can also matter.9

**(vii)** Investment universe

The investment universe to which a signal is applied can influence capacity through both the general nature of the universe itself, and via changes in market conditions over time. Capacity is likely to be greater where a signal is applied to a universe that supports trading in large volume with high liquidity, and presents a wider range of opportunities. In general, this will tend to occur in larger markets. Some obvious examples are that capacity will be higher in large-caps relative to small-caps; and for global markets relative to smaller individual countries. In hedge funds, global macro funds that trade in equity indices, bonds and currencies using derivatives or ETFs will have much higher capacity than convertible arbitrage funds, where both the investment universe and liquidity are limited. Broader investment universes can also support higher capacity to generate alpha by giving rise to more potential opportunities. In addition, alpha-generating capacity should be enhanced where the universe is less concentrated, and its components behave independently. For example, capacity in the Australian REIT sector is hampered by the dominance of a handful of larger securities and the comparative homogeneity of the sector. A further issue is the degree to which asset ownership is concentrated versus diffuse. Investment universes where assets are owned by a variety of investors are more likely to offer higher liquidity and greater capacity than where ownership is concentrated in the hands of few large players.

Another consideration is whether a fund includes short selling in its armoury. Short selling can be capacity-increasing to the extent that it expands the universe of opportunities. However, it gives rise to some specific issues, such as borrowing capacity, stock lending, short interest and risk management. (These issues are all highly specialized, so we merely flag them here.)

Capacity can also vary with changes in market conditions in two main ways. First is via liquidity conditions, which influence the ability to exploit a signal through the ability and cost to trade. Some authors have estimated capacity to have improved over time due to increased liquidity by, e.g. Landier et al. (2015), Novy-Marx (2016). However, increased liquidity has another side. It may assist in reducing the magnitude of expected returns in the first instance, which itself could reduce capacity. For example, Chordia et al. (2014) find evidence of the attenuation of returns from 12 anomalies, and relate this to increased scope for arbitrage in part as a consequence of lower market frictions.

Second, the magnitude and availability of investment opportunities to exploit varies with market conditions. To some extent, this can reflect an interaction with liquidity conditions. Liquidity is

---

9 We received feedback from one market participant that market impact in Australian small cap stocks was often less than implied by transaction cost models due to the ability to secure large blocks on many occasions.
typically greater in rising markets, which may increase capacity for signals where implementation is liquidity-demanding. On the other hand, capacity associated with a signal that relies on supplying liquidity – such as acting as the buyer of last resort or contrarian investing – may encounter higher capacity during weak market states when forced sellers are abundant. Another example is that markets where valuation dispersion is wider can provide more opportunities for value investors. Finally, capacity in unlisted markets such as infrastructure may vary with the supply of investible assets, which in turn might reflect aspects such as government policy.

(viii) Other investors using similar strategies or signals

Capacity may relate not just to an individual fund, but also to how much capital is directed at exploiting the signal on which it invests, as well as the intensity of competition to implement based on the signal. The main implication is that unique investment approaches should have greater capacity. If a range of investors are using similar strategies or signals, this will reduce capacity for each. It will raise the competition for the available volume, while at the same time decreasing the number of investors who could potentially be willing to trade on the other side. This can lead to greater market impact, and decrease the chances of being able to secure trades in the desired volume. This issue can relate to both other investors operating in the market, and different funds housed within the same investment management organization. In the latter case, it may be appropriate to define the ‘fund’ more broadly to account for demands from all products that are aiming to exploit a particular signal.

A number of authors find evidence that increased FUM erodes the opportunity to generate excess returns for hedge funds. Chordia et al. (2014) find that returns from 12 anomalies have attenuated over time, and relate this in part to higher FUM for hedge funds as well as lower market frictions. Similarly, Kokkonen and Souminen (2015) find that the level of mispricing is negatively associated with flows and FUM managed by hedge funds. Akbas et al. (2015) also find that flows into hedge funds (but not mutual funds) attenuate the returns to certain anomalies, particularly those related to growth, momentum and accruals.

(ix) Flexibility to adjust the implementation

Capacity can be influenced by the latitude for a fund to change how it invests on a signal in response to increasing FUM. In particular, the diseconomies related to trading and portfolio construction involve a trade-off between trading greater amounts immediately at less attractive prices, versus deferring the trade and running the risk that the price adjusts and hence incurring opportunity costs. How this trade-off is managed as FUM grows can influence the capacity that can be effectively extracted from a signal; and is what predictive models aim to capture. An example would be a value fund that responds to increased FUM by extending its intended holding period, recognizing that long-term positions of a more strategic nature may be preferable when larger stakes are being contemplated. A further example would be for a fund to change the investment universe from which opportunities are drawn, e.g. excluding micro-caps once FUM reaches a certain level. All of these changes do not represent major changes to the signal itself, but might be expected to enhance capacity at the margin.

(x) Asymmetry in accumulating vs. liquidating

An asymmetry can exist with respect to the ability to accumulate relative to the ability to liquidate positions that has relevance for capacity. In simple terms, it can sometimes be easier to set, than get out. Consider an equity fund that acquires in excess of (say) 5% of a smaller company, commencing from a zero holding. The process of accumulating and liquidating a position are not necessarily the

---

10 Northfield recognizes crowded trades, and suggests that one response might be to ‘increase the transaction cost coefficients’ when modeling execution costs.
A related asymmetry arises with respect to inflows and outflows into a fund. Inflows may be invested in a measured way (assisted through parking un-invested amounts in either cash or futures). This can help to limit the market impact. Outflows, however, need to be immediately satisfied by cash, providing a fund with limited ability to manage the market impact. This can also create a link with fund lifecycles (Scobie, 2013), whereby no apparent capacity constraints occur as inflows are received in the early stages, but subsequently capacity issues emerge as the fund matures and inflows then taper off or even turn to outflows. Further, to the extent that greater FUM is associated with potential for outflows of a larger magnitude, the risk of loss through market impact if outflows occur increases with FUM. This reflects the notion that large, non-discretionary trades tend to have higher market impact.

The implication of these asymmetries is that capacity does not only relate to the ability to invest successfully as FUM grows. Another issue is whether the potential costs of liquidating could increase disproportionately with FUM. This is tied to the probability of having to liquidate relatively large positions with limited trading discretion, such as might occur under major changes in view with respect to relatively large holdings, or situations leading to substantial fund outflows. To a large extent, this tenth driver has implications for the downside risk that can be associated with managing greater FUM. It basically asks what could go wrong if things do not go to plan.

### 3.3. Implications for Capacity Analysis

The discussion in this section holds three implications for capacity analysis. First, capacity should be evaluated from a variety of angles, and not just through the lens of increased difficulty in implementing a target portfolio at attractive prices as FUM grows. While this can be central, it is not the entire story. The drivers of capacity are many and varied. Other factors to consider include economies of scale and scope; and organizational and staff constraints. In some circumstances, the latter may matter more than the cost and ability to trade. For example, where the key resource is the skill of a particular individual, or the ability to handle multiple positions or transactions, then capacity needs to be examined through a different lens. The risk of encountering difficulties in liquidating large positions might also be given consideration. Second, capacity should be viewed as a dynamic concept. Funds can change the manner in which they manage in response to increases in FUM. Capacity can fluctuate with market conditions, due to aspects like variation in liquidity or the availability of opportunities. It is entirely reasonable to base capacity analysis around how the fund currently approaches investment under ‘average’ market conditions. However, it should always be borne in mind that things evolve; and that capacity is not fixed and may fluctuate over time. Third, evaluating capacity from the perspective of an individual fund is tenuous, as the amount of funds that are aiming to collectively exploit a signal matters. Whereas the intent of capacity analysis is often to estimate how large a particular fund can get before it should consider closing (or making changes to its process), the fact is that capacity will always operate at the overall signal level to some degree. It is important that this dimension be considered, even if informally.
4. Analyzing Capacity

This section outlines the various approaches that can be used to analyze capacity, as well as the related literature. The methods and research discussed largely focus on equities, where researchers have directed their attentions. Section 4.1 describes four broad groups of approaches to capacity analysis. Section 4.2 overviews the literature examining the capacity associated with particular strategies or signals, most of which investigates capacity associated with equity market anomalies.

4.1. Approaches

The aim in this sub-section is to provide a broad overview of the types of approaches that are available for evaluating the capacity of a fund that is implementing a given signal. We draw both from the literature, and what we have been able to discover about approaches used in the industry. The coverage of industry practice is not comprehensive, but we believe it is indicative. We attempted to discover how the investment industry evaluates capacity by asking fund managers and consultants, and collecting practitioner research where available. This process left us with the impression that capacity analysis is not well-developed across much of the industry, particularly in Australia. Most analysis appears based on observed portfolio holdings or trades; and involves rules of thumb, calculation of simple metrics, and sometimes simulation analysis. Predictive models that estimate capacity under the assumption that the trading strategy or portfolio construction changes as FUM increases do not seem to be broadly used. However, the analysis of Frazzini et al. (2012) that we discuss in Section 4.3 points to at least one large global manager that appears to be doing so, and there may be others.

Capacity analysis falls into four broad groups, as listed below in increasing levels of sophistication:

(i) **Rules of thumb** – Rules are imposed with respect to the limits placed on certain portfolio metrics that are considered symptomatic of capacity.

(ii) **Ex-post analysis** – Selected portfolio metrics are monitored for signs that capacity constraints may be emerging.

(iii) **Simulation analysis based on existing portfolios or signals** – The change in performance or other attributes with increasing FUM is evaluated by running simulations that ‘scale up’ an existing portfolio (or a portfolio implied by a given signal) and its implied trades, accounting for the impact of constraints on portfolio holdings or trading and perhaps greater transaction costs.

(iv) **Predictive models** – Capacity is estimated by predicting portfolio performance using a model that integrates the signal and transaction costs, allowing for either trading and/or portfolio construction to be adjusted as FUM increases. This approach assumes that the fund changes the way it implements based on a signal as FUM increases.

Below we discuss each of the approaches in broad terms. Section 5 then delves into predictive models in some depth. Before proceeding, we offer some general points about capacity analysis. First, it is an inexact science. Accurate estimates of capacity are unlikely to be forthcoming; and capacity is better thought of as sitting within a wide and indicative range. Second, the approaches to evaluating capacity should not be considered as mutually exclusive, and using a combination seems sensible. In particular, as simulations and predictive models are subject to model and estimation uncertainty, it is highly advisable that they be combined with ex-post analysis. The latter provides a ‘finger on the pulse’, and can help to confirm whether any predictions about capacity are on track as a fund grows. Third, capacity should be approached as a relative and potentially dynamic concept, rather than as an absolute. Where possible, capacity analysis should involve relating FUM or position size to the value of the
assets being addressed, or relating trade size to the volume available in the market. This will facilitate adjusting readily to the evolution in market size and liquidity over time.

(i) Rules of Thumb

We have identified two rules of thumb used in capacity analysis:

- **Limits on percentage of market capitalization** – The notion is that capacity constraints are encountered once FUM exceeds a certain portion of the market capitalization in the segment in which the fund invests. In equities, a typical limit would be 0.5%-1% of the market capitalization (see Scobie, 2013; Iverson and Gregory, 2015; RARE, 2015; Zenith, 2016).

- **Days to exit** – Estimates are formed of how long it would take to liquidate a portfolio, given the stocks it holds and their average daily turnover. Analysis requires making an assumption about the ‘participation rate’, defined as the percentage of traded volume that will be available to the fund. A number of variations exist on this type of analysis, including: days to liquidate the entire portfolio (e.g. IML, 2012; Scobie; 2013); days to trade back to benchmark (Scobie, 2013); weighted average days to liquidate (Zenith, 2015); and estimation of the portion of the portfolio that can be liquidated across various time frames (RARE, 2015). Days to exit measures are effectively indicators of fund liquidity, and allude to the concept of liquidation capacity. Judgment is applied in gauging whether days to exit is acceptable. For instance, we understand that one guideline used by asset consultants is a requirement that equity funds can be substantially liquidated within 3 months.

Rules of thumb can be relatively easy to calculate, and hence provide a ready measure. The main disadvantage is that they offer only a rough guide when based on broad portfolio averages, which can be misleading when what matters is encountering constraints or costs in the tails of the holdings distribution. For instance, the average percentage of capitalization held or days to exit across the portfolio do not directly address the FUM at which capacity constraints would be encountered for a fund that extracts most of its alpha from small cap equities. Rules of thumb can thus lack the granularity required to account for the wide range of drivers of capacity discussed in Section 3.2. It can also be difficult to calibrate the level of the metric at which capacity problems might emerge, which will vary across funds according to aspects like their investment approach or investment universe.

(ii) Ex-Post Analysis

Ex-post analysis entails monitoring trends in selected portfolio metrics and performance. The aim is to look for signs of diminishing ability to effectively implement a signal, which may indicate that the fund may be encountering capacity constraints. Portfolio attributes\(^\text{11}\) that might be monitored include:

- **Number of positions in the portfolio** – An increase in the number of positions being held may signal greater difficulty in fully exploiting a signal by investing only in the most-preferred opportunities, thus causing the fund to ‘go down the list’ of potential investments to build its portfolio.

- **Convergence on the benchmark** – Increased difficulty in exploiting a signal through the most-preferred opportunities may also manifest in the fund’s portfolio converging towards its benchmark in terms of its attributes or performance patterns. Measures of benchmark convergence include tracking error,\(^\text{12}\) active share (Cremers and Petajisto, 2009), and the correlation between portfolio returns and benchmark returns.

---

\(^{11}\) Some of these metrics are outlined in RARE (2015).

\(^{12}\) Predicted tracking error might be a better measure than realized tracking error.
• **Trade execution costs** – Post-trade analysis could be conducted to gauge whether execution costs are rising as FUM increases due to greater market impact. Nevertheless, when examined in isolation, execution costs need to be interpreted carefully as a measure of capacity issues. First, execution costs can fluctuate for a range of reasons, including market conditions and the type of securities traded. Execution costs should be evaluated with controls for such influences, perhaps through segmentation by security sector, or by benchmarking against expected execution costs or the market. Second, implementation difficulties may also manifest in deferral or even failure to trade, which is not captured by execution costs.

• **Implementation shortfall** – Implementation shortfall provides a more direct and complete measure of the erosion of potential returns arising from a signal, as it embeds both execution costs and opportunity costs. Implementation shortfall can be estimated by comparing realized portfolio performance with that for a target (or ‘paper’) portfolio (Perold, 1988). The target portfolio is the portfolio that arises from implementing as indicated by the signal at the time of decision at zero cost. Implementation shortfall can be divided into execution costs and opportunity costs, thus revealing whether the shortfall is stemming from greater trading costs versus deferring or foregoing trades. Collins and Fabozzi (1991) discuss the attribution of transaction costs relative to a benchmark (i.e. target) portfolio in detail. Estimating implementation shortfall requires maintaining records for the target portfolio in parallel to the actual portfolio.

• **Active returns** – Analysis of active returns for a portfolio may also yield clues on whether capacity is being tested. Erosion of active returns may be a sign of capacity problems, especially if it arises from security selection and is accompanied by evidence from other ex-post indicators as discussed above. Perold (1988) notes that performance erosion in the absence of increasing implementation shortfall is more likely a signal of problems with the investment process rather than capacity issues.

• **Days to enter and exit** – The trading activity of a fund may be examined for signs of increasing time to enter and exit positions, which could provide an indication of greater difficulty in executing trades in response to a signal. This is the ex-post equivalent of the days to exit rule of thumb.

(iii) **Simulation Analysis Based on Existing Portfolios or Signals**

Another approach to capacity analysis is to conduct simulations to gauge what happens to fund performance or selected portfolio attributes when observed positions and/or changes in those positions are ‘scaled up’. The aim is to estimate the hypothetical impact of attempting to implement the same portfolio and/or the associated trades under increasing FUM. Simulations can be conducted in a variety of ways. Key aspects of any analysis include: how capacity is measured and defined; the formulation of portfolio holding or trading constraints; and whether increased execution costs are taken into account as trade size increases. A number of applications of this type of analysis can be found in both practice and the literature, distinguished by their approach to these aspects. We provide a sense of the range of approaches by summarizing five pieces of analysis below. In addition, Section 4.2 overviews the literature where the capacity associated with specific anomalies is evaluated, much of which is based on simulation analysis. While only one analysis among those listed below incorporates transaction costs, this is a feature of most of the research summarized in Section 4.2.

• Iverson and Gregory (2015) describe how NZ Super re-evaluates the portfolio performance of its New Zealand equity managers as FUM increases, under the assumption that they are only able to trade 10%-15% of available volume in any month. The simulation delivers an estimate of alpha erosion as a consequence of trading constraints.

• IML (2012) estimates how increases in FUM could impact on the portion of the portfolio that does not meet three constraints: minimum position size of 0.5%, holding limit of 10% of market
capitalization, and 10 days or less to liquidate a position assuming 25% participation in average daily volume (ADV). They also examine how execution costs, including market impact, could increase with FUM and hence percentage of ADV traded, based on the type of stocks held in their portfolios. Capacity is evaluated through judging the return implications of the portfolio being constrained in its stock holdings, or incurring higher execution costs.

- **RARE (2015)** estimates how increases in FUM impact on the portion of their portfolio that becomes constrained under a holding limit of 7.5% of free float; as well as the portion of their portfolio that would take over 10 days to liquidate given observed market volume. They then define capacity by placing a limit of 10% on the portion of the portfolio that must be reallocated under the constraint test; and a limit of 50% on the portion of the portfolio that takes over 10 days to liquidate.

- **Cantara et al. (2016)** describe how MFS models capacity by examining the percentage of an existing portfolio that can be replicated under constraints on share ownership as a percentage of both shares outstanding and average daily trading volume. Hallmarks of the analysis include a ‘product share constraint’ that allows for the need to allocate shares available to MFS as an organization across various funds; and generating a range of capacity estimates based on different input assumptions.

- **Seto et al. (2015)** describe how Parametric estimates capacity for an emerging market strategy by simulating the portfolio with constraints on holdings of illiquid stocks, and the amount that is permitted to be invested in small markets. Capacity is estimated under the requirement that the simulated portfolio has a minimum level of overlap with the target portfolio.

- **Vangelisti (2006)** conducts simulations at differing levels of FUM for a signal that combines value and momentum characteristics, with limits imposed on holdings as a percentage of free float and portfolio concentration. Capacity is based around achieving a minimum alpha target or threshold.

(iv) **Predictive Models**

Simulation approaches as described above estimate capacity under the assumption that a fund attempts to implement a signal regardless, but in doing so encounters either constraints or greater execution costs as FUM increases. Predictive models go a step further by allowing for the possibility that a fund may change the way that it trades and/or constructs portfolios based on a particular signal as FUM increases. While changing the trading or portfolio construction process need not be envisaged by a fund manager, it would seem more reasonable to assume that they adjust their implementation, than to assume that they will just act in the same manner regardless of the implications. Indeed, it likely that manager would react to any adverse effects arising from greater FUM in order to protect performance, given its central role in how they are evaluated and rewarded. Section 3.2 mentioned some papers that find evidence that funds do indeed change how they manage as FUM increases.

We explore predictive models and their component parts in depth within Section 5. Here we provide a general overview. Predictive models of the type we are considering fall into two groups.

- **Trade optimization only** – The target portfolio is assumed to remain unchanged, reflecting the portfolio that optimizes the expected payoff from the signal ignoring transaction costs.\(^\text{13}\) Analysis is then undertaken under the assumption that implementation occurs by scheduling trades to minimize implementation shortfall. The optimization requires addressing the trade-off between the expected increase in execution costs that arise from trading quickly, against the potential opportunity cost of deferring or failing to undertake trades. Risk may be taken into account, and is typically defined with

\(^{13}\) The target portfolio is formed to maximize the trade-off between risk and return, subject to any constraints that relate to the portfolio itself, e.g. maximum position sizes relative to the benchmark, tracking error, etc. Basically, it is the portfolio that the fund would prefer to hold regardless of FUM and transaction costs.
regard to uncertainty over future prices, specifically the probability that they move in an adverse direction. The consequence of the latter is higher execution costs and/or opportunity costs.\(^{14}\)

- **Portfolio optimization** – The portfolio is formed by optimizing the trade-off between the expected return arising from the signal, transaction costs and risk; subject to any portfolio or trading constraints. The concept of a target portfolio that is uniquely identified with the signal assuming zero transaction costs no longer exists. While implementation shortfall may be estimated, it is no longer being optimized in isolation of the target portfolio and hence is interpreted differently to that suggested by Perold (1988). Risks related to the assets held in the portfolio, and uncertainty over trade prices, become integrated under the modeling. Borkovec et al. (2009) provide three worked examples to demonstrate that joint optimization of the trading schedule and portfolio construction can have a significantly positive effect on portfolio performance. This suggests that incorporating transaction costs into portfolio construction may make a significant difference to capacity estimates.

Predictive models effectively extend the boundaries of the signal to incorporate the transaction costs associated with implementation, in addition to the underlying expected return and risk associated with the investments. Predictive models imply a multi-period optimization, either through applying dynamic control or simulation techniques. However, as will be seen in Section 5, analysis may be collapsed down to a single period problem where turnover acts as the lever, and the rate at which expected returns accrue is embedded in the return expectation for the period.

Predictive models can be used to estimate capacity by modeling portfolio performance allowing for the fact that required holdings, and hence trade size, become larger as FUM grows. The trade schedule and/or portfolio are optimized taking into account the fact that larger trades are associated with greater market impact, while deferring trades will lead to increases in both potential opportunity cost and transaction cost risk. Effectively, they model an optimal response to deterioration in the expected trade-off between execution costs and opportunity cost as FUM increases. Incorporation of constraints on trading or holdings into the analysis would probably preclude closed form solutions, and require solving using numerical techniques (e.g. simulations, or constrained optimization techniques).

One key advantage of predictive models is that they incorporate the time dimension for accrual of expected returns in a rigorous manner. Returns arising from investment signals accrue at differing rates, and this can be integral to capacity. For instance, returns to momentum signals are generally perceived to be relatively ‘fast’ as much of the value is captured over months, and possibly even weeks. In contrast, returns to value signals tend to be ‘slow’ as they accrue over longer periods. This makes it more important to trade quickly under momentum strategies than value strategies, in order to harvest as much of the potential gains as possible. The implication is that increases in FUM more readily bring momentum strategies up against the trade-off between paying higher execution costs to get set quickly, versus deferring trades and incurring opportunity costs. Predictive models provide the mechanism for incorporating differences in how funds might respond into the estimation of capacity.

An example may help. Consider a momentum-based signal that is expected to generate excess returns of (say) 9% accruing evenly over 6 months, or at the rate of 1.5% per month. In this case, the opportunity cost associated with deferring trades may be meaningful. It hence may be optimal for a momentum-based fund to trade quickly regardless, in order to capture the available returns before they dissipate. Predictive models will probably reveal that execution costs swiftly eat into returns as FUM rises, which will act to restrict capacity. Nevertheless, capacity may still be marginally higher than that arising under the assumption that all trades are undertaken regardless of costs, to the extent that at some level it may be optimal to defer some trading. In contrast, consider a value-based signal that is expected to deliver a 15% gain over two years, which initially accrues at a modest rate, with the bulk of return occurring in

---

\(^{14}\) Trading more quickly reduces risk as it reduces the scope for adverse price movements to occur.
year 2. In this case, the opportunity cost of deferring trades for a few months will be low. Under these circumstances, predictive models would assume that the value fund optimally responds to increasing FUM by deferring trades. This is likely to lead to much higher estimates of capacity than under a simulation approach where it is assumed that observed trades are implemented over the same time horizon if at all possible.

In this way, predictive models support the analysis of capacity in a manner that mimics how a fund might logically respond when faced with such trade-offs in practice. Section 5.4 notes some of the evidence around institutional trading, observing that trading strategy varies with investment style and that institutions trade strategically. In contrast, simulations based on attempts to implement existing portfolios or signals in the same fashion do not allow for the possibility that fund managers can change their behaviour. Failing to allow for how managers might respond to the constraints imposed by higher FUM could lead to significantly under-estimating the capacity associated with an investment signal.

While Section 5 discusses the application of predictive models in depth, it is worth observing here one point of distinction. Some authors employ a bottom-up approach by aggregating from the security level (e.g. Serbin et al., 2009; Brandes et al., 2009). Others analyze portfolio averages (e.g. Kahn and Shaffer, 2005; Coppejans and Madhavan, 2007). The latter tend to work within the ‘fundamental law of active management’ framework (see Grinold, 1989), focusing on aspects such as optimal turnover and breadth. Bottom-up approaches are more complex, but offer three advantages. First, they permit the relation between expected returns from a signal and liquidity to be exploited, noting that illiquid stocks may offer greater potential alpha. Second, they can be more precise when key variables such as transaction costs and expected returns vary across securities in non-linear ways. In this case, an aggregated solution may differ from that based on weighted average inputs. Third, they can accommodate imposing constraints at the security level, e.g. limits on the percentage of market capitalization that may be held.

4.2. Literature: Capacity Analysis for Specific Strategies or Signals

It is worth initially mentioning that a number of authors examine whether some return anomalies survive after adjustment for transaction frictions, without directly addressing the question of whether an optimal capacity exists. This research is closely aligned with the use of simulations to estimate capacity by analyzing existing portfolios or signals, as was described above under point (iii) of Section 4.1. For example, Loeb (1991) re-evaluates the excess returns from small caps after adjusting for estimated transaction costs as well as risk.\(^{15}\) Returns from momentum signals have been re-examined after adjusting for estimated transaction costs by Knez and Ready (1996)\(^ {16}\) and Lesmond et al. (2004); and after imposing both bid-ask spreads and trading or holding constraints by Bettman et al. (2009) and Bettman et al. (2010). Beckers and Vaughan (2001) use simulations to examine how performance deteriorates as the portion of market capitalization owned increases, for portfolios formed based on combined value, price momentum and earnings revision signals. They allow for both transaction costs and portfolio holdings constraints. Unsurprisingly, these studies document how anomalies can become attenuated after allowance is made for trading frictions.

A number of researchers extend this line of research by providing estimates of optimal capacity for particular fund styles or signals. This research generally finds that capacity is lower for faster, high turnover signals such as momentum, relative to slower, low turnover signals such as value. However, the reported capacity estimates vary considerably across studies. This variation appears to reflect differences in sample data and method, including: how transaction costs are estimated; how portfolios

---

15 Loeb (1991) corrects for beta, based on estimates with adjustment for thin trading.
16 Knez and Ready (1996) examine strategies to exploit momentum in small caps, and cross-correlation between small and large caps.
are constructed; whether trading discretion is permitted; and how performance is evaluated. The results thus signal that capacity estimates are highly reliant on how it is being modeled and measured. Of particular importance appears to be whether the analysis incorporates constraints on holdings or trading; as well as whether trades or portfolio construction is optimized allowing for transaction costs. Below we summarize\textsuperscript{17} some of the notable papers, converting any reported capacity estimates into equivalent June 2016 values.\textsuperscript{18}

- Indro et al. (1999) use a regression model of the relation between fund size and performance to estimate the optimal FUM for three fund styles in 1993-1995. They estimate that optimal FUM is higher for growth funds ($1.4-$1.5 billion; equivalent to $2.3-$2.4 billion in June 2016) than value funds (about $0.5 billion; equivalent to $0.8 billion); but highest for blend funds ($1.9-$2.0 billion; equivalent to $3.1-$3.2 billion).

- Korajczyk and Sadka (2004) examine the profitability of long-only momentum strategies in the US allowing for transaction costs. They estimate that excess returns can be earned for a liquidity-weighted strategy up to FUM of $4.5-$5.0 billion in 1999 (equivalent to $6.4-$7.1 billion in June 2016); although the estimated wealth-maximizing capacity is lower at around $2.5 billion ($3.6 billion).

- Chen et al. (2005) estimate decay in excess returns for book-to-market, size and momentum strategies in the US over the period 1963-2002, allowing for both transaction costs and constraints on both trade size and portfolio holdings at 1% and 5% of market capitalization respectively. They report capacity estimates in 2002 dollars for a range of signals and investment approaches, based on the point at which excess returns are eroded to zero. Their largest reported capacity estimate sits at $21.5 billion (equivalent to $28.3 billion in June 2016) for one long-only momentum strategy; although the long-short version tops out at $1.7 billion ($2.2 billion). Maximum capacity for investing on small-cap signals is about $0.4 billion ($0.5 billion). For book-to-market, the maximum break-even capacity is estimated at $4.6 billion ($6.1 billion). However, capacity is much lower than these maximums when based on other approaches to implementing the signals.

- Frazzini et al. (2012) examine the performance and capacity of four long-short strategies based on size, value, momentum and short-term reversal signals across 19 countries, after adjusting for transaction costs only. A unique feature of their analysis is that they have access to the intended and actual trades from a large institutional manager, which allows them to consider both opportunity costs as well as calibrate execution costs using realized trades (i.e. implementation shortfall). They also report estimates for both the traditional ‘Fama-French’ style factors, and for optimized versions where trading costs are minimized subject to a tracking error constraint. Most of their capacity estimates are considerably larger than those reported by other authors. Estimated break-even FUM for the long-short version of size, value, momentum and reversals factors are $103, $83, $52 and $10 billion respectively in the US market; and $156, $190, $89 and $13 billion globally. The equivalent trade-optimized estimates are $1,807, $811, $122 and $17 billion globally. (June 2016 values are about 6% greater.) Two features contribute to the high capacity estimates of Frazzini et al. (2012). First, they do not appear to apply any constraints on trading or holdings (other than tracking error). They thus assume all targeted positions are accessible, which may not be achievable in practice. This appears to be the key reason why their small-cap strategies are estimated to have the greatest capacity. Second, their transaction cost estimates are notably lower than those reported elsewhere in the literature. The authors suggest that this is because most transaction cost models are estimated for all trades, including those where immediacy is demanded; whereas an institutional investor that manages their trades will incur lower costs. This point, along with the larger capacities for the

---

\textsuperscript{17} As describing the plethora of samples and methods would be quite involved, we only aim to provide a general sense for the findings from this body of research.

\textsuperscript{18} The quoted estimates are converted to values as at June 2016 by adjusting for changes in the US CPI.
optimized strategies, underlines how the assumption that trading is optimized can have an important influence on capacity estimates.

- Landier et al. (2015) estimate capacity for signals associated with the four anomalies of book-to-market (value), low volatility, cash flow to assets (‘quality’) and share repurchases. Their analysis is based on Gârleanu and Pedersen (2013), who develop a predictive model involving portfolio optimization. Indicative capacity estimates are provided for the US in two periods – 1990-2001 and 2002-2013 – based on differing inputs for large and mid-cap stocks. Capacity is defined as the level of FUM that delivers a Sharpe ratio of 0.3. Estimated capacity is highest for the cash flow signal, at as much as $79 billion over the period 2002-2013 ($91 billion in June 2016 dollars). This is due to a combination of both a high Sharpe ratio before costs and a persistent signal (i.e. a slow accrual rate). Low volatility has the next highest capacity ($13 billion in 2002-2013; $15 billion in June 2016), followed by share repurchases ($1.8 billion; $2.1 billion in June 2016). Value had zero capacity, in a large part due to a low Sharpe ratio before costs over the period. The estimates suggest that capacity was substantially higher in the second period, e.g. capacity for the cash flow signal was only $6.6 billion in the 1990-2001 period, versus $79 billion in 2002-2013.

- Novy-Marx and Velikov (2016) analyze the reduction in returns as a consequence of transaction costs for 23 anomalies. They provide ‘back of the envelope’ estimates of break-even capacity in 2012, based on average attributes for portfolios formed on the signals. They also estimate alternative capacity estimates under an asymmetric trading rule for opening and closing positions, which is designed to limit turnover. In general, they find that capacity decreases as turnover increases. While reporting a range of capacity estimates, they summarize their findings by noting that: the size and gross profitability signals could attract “hundreds of billions of dollars of new arbitrage capital”, book-to-market has capacity “roughly half as large”; momentum could attract just over $10 billion; and that only a few of the high frequency strategies remain profitable after accounting for effective spreads, and these have capacities in the order of only a few hundred million dollars. Again these authors do not appear to apply any constraints on trading or portfolio holdings, which may overstate capacity with respect to strategies that rely on small caps.

5. Predictive Models and their Component Parts

In this section, we examine predictive models and their component parts in some depth. The discussion is quite technical, and aimed at readers wanting to delve into the details of modeling of capacity and transaction costs. (Others not interested in this level of detail may wish to skip this section.) We start by describing the concept of implementation shortfall in Section 5.1, which is integral to the development of the class of models being discussed. Section 5.2 outlines the framework under which predictive models operate. This is followed in Section 5.3 by an overview of the literature on optimal trading and portfolio construction under both transaction costs and signals that decay, and their application to capacity analysis. Finally, we drill down into the two key component parts of predictive models. Section 5.4 reviews the literature on transaction cost models, which is extensive but yields no clear consensus around the preferred model. Section 5.5 discusses the profiling of expected returns (or alpha) arising from specific strategies or signals. Few studies have examined the profile of alpha in a manner that is suitable for use in evaluating capacity under predictive models.

---

19 Large-caps are those US stocks ranked in the top 500, and mid-caps 501 to 1,500.
20 Estimates in mid-2016 dollars would be about 4% greater.
21 Novy-Marx and Velikov (2016) apply their strategies to all listed stocks in the US using the CRSP database.
5.1. Implementation Shortfall and Related Concepts

The concept of implementation shortfall (IS) is proposed by Perold (1988), and is central to analysis of the capacity associated with a particular signal using predictive models. Perold describes IS as a measure of the “degree to which you are unable to exploit your stock selection skill”, which is estimated as the difference between the performance of a ‘real’ and a ‘paper’ portfolio. IS comprises two components: execution cost (EC) which “relates to the transactions you actually execute”, and opportunity cost (OC) “which relates to the transactions you fail to execute”. EC is the sum of commissions and transfer taxes, plus market impact (MI). (The latter is also referred to as price impact.) Perold describes MI as “the difference between the price you could have transacted at on paper (the average of the bid and ask at the time of the decision to trade) and the price you actually transacted at”. MI in turn arises as a function of the cost of demanding liquidity or immediacy,22 and the possibility that your information gets incorporated into the price before you are able to trade (also see Treynor, 1981). These two effects are associated with transitory and permanent MI respectively, although this is not discussed by Perold (1988). OC is simply the returns forgone by failing to transact either immediately before the price adjusts, or perhaps failure to transact at all. The latter might occur because some constraint is encountered (e.g. limits on the percentage of an asset that may be held, or lack of available volume), or due to a decision not to trade because it is no longer considered profitable. Perold presents IS in a manner that accords with the ex-post analysis of the effectiveness by which a signal is implemented. However, he does not address how expectations about IS (i.e. E[IS]) may feed back into optimal trade scheduling or portfolio construction.

Arnott and Wagner (1990) and Collins and Fabozzi (1991) build on the concept of IS. Collins and Fabozzi provide a useful set of definitions of the components of transaction costs. They recognize that transaction costs can be divided into fixed and variable costs, and that EC may stem from price movements arising from the trades of other investors as well as MI arising from the investor’s own trading. Both sets of authors discuss how trading costs can be minimized by identifying the optimal trade-off between EC and OC, and how investment strategies can have differing OC functions. This presages the literature on optimal trading, which we review in Section 5.3.

5.2. Frameworks for Predictive Models

We now provide a framework for understanding predictive models, and how they might be used to estimate capacity for a particular signal.23 We commence by describing predictive models in their most general form, under which transaction frictions are taken into account in portfolio construction. We then present a more limited case under which only the trade schedule is optimized taking into account transaction frictions for a given target portfolio, but where the latter is formed independently of transaction costs and trading constraints. This more limited case accords with the IS framework initially proposed by Perold (1988).

Figure 2 presents a general framework for capacity analysis using predictive models that allow for optimal portfolio construction under trading frictions. The framework is presented in the form of a schematic that describes a decision process for optimally capturing the returns arising from an investment signal that accrues over time, under conditions where trading is costly and may be subject to constraints. The presentation implicitly assumes a one-off investment in a portfolio which is held for the horizon over which returns to the signal accrue. Thus no consideration is given to dynamics portfolio

---

22 The cost of demanding liquidity relates to a price concession offered to another party as inducement to take the other side of the trade.

23 By implication, these models address informed trading, rather than rebalancing or other informationless trades.
management, under which signals are being continually updated and portfolios are constantly rebalanced.24

Applying the framework entails specifying expectations for three key elements:

- **Returns to the signal** – Both magnitude of expected returns arising from the signal (or the expected signal return, $E[SR]$) and its accrual profile are required. The latter determines the expected opportunity cost ($E[OC]$) arising from deferring trades, or failure to trade.

- **Execution costs ($E[EC]$)** – Expected execution costs are estimated using a transaction cost function.

- **Portfolio risk** – A definition of portfolio risk is specified, accompanied by a risk aversion coefficient ($\gamma$). The definition of risk would stem from the portfolio objectives, and may take various forms, depending on the context.

Inputs should ideally be estimated at the security level, with portfolio construction allowing for covariation across securities in estimating risk while ensuring any constraints are satisfied.

The box appearing at the top left in Figure 2 captures the signal. $E[SR]$ is the return that is expected to arise from trading on the signal immediately at zero cost. It hence establishes a baseline against which $E[IS]$ can be measured. The accrual profile is specified in order to characterize $E[OC]$. The presentation of the signal and associated expected returns is intended to be general – we are not proposing any particular form for the source of the signal and its accrual profile. However, we note that $E[SR]$ could arise from a wide range of sources, such as exposure to factors or characteristics, deviations from intrinsic value, analyst projections, and so on. It could potentially follow any profile in theory, and should not be viewed as limited to a constant decay rate. The presentation allows for the possibility that tax effects might be incorporated into $E[SR]$ and its accrual profile, such that any tax effects associated with deferral of trades become included in $E[OC]$.

The box at the top right in Figure 2 captures execution costs. $E[EC]$ reflects the expected cost of transacting in the market, which in turn is expected to increase with trade size and hence the rate at which trades are undertaken in order to establish a position. Section 5.4 provides an overview of transaction cost modeling, noting that many candidate formulations appear in the literature. Figure 2 lists some of the typical components appearing in these models. Most models separate fixed and variable costs. Fixed cost ($FC$) reflects the baseline cost of transacting stemming from aspects such as commissions and fees. Bid-ask spreads are included in $FC$ in Figure 2, although this is debatable to the extent that spreads can vary across securities and market conditions. Variable costs largely relate to market impact ($MI$). This is typically modeled conditional on variables like trade size, trading strategy, security characteristics and market conditions; and may have permanent and temporary components. The relation between $MI$ and FUM via its impact on potential trade size is central to capacity analysis.

The presentation of the approach to portfolio construction appearing in the middle box of Figure 2 is intentionally general. It is standard in the sense that it contemplates maximizing the trade-off between expected portfolio return and ‘risk’ subject to portfolio constraints. Nevertheless, it embeds some important elements that are specific to the class of predictive models being discussed here. The key aspect is that expected portfolio return ($E[PR]$) is defined as $E[SR – OC – EC]$, implying that portfolios are constructed with consideration for both the signal return and the expected loss of those returns arising from implementation shortfall ($E[IS] = E[–OC – EC]$). Under predictive models, $E[IS]$, including $E[EC]$, becomes endogenous under the portfolio construction process.25

---

24 Such dynamic elements add considerable complexity, and require imposing additional structure.
25 Tax effects would also become endogenous where they are incorporated in $E[SR]$ and $E[OC]$. 
Figure 2: Capacity Analysis under Predictive Model with Optimal Portfolio Construction

**Signal**
- Expected Signal Return ($E[SR]$)
  - Return arising from immediately trading at decision price at zero cost
- Expected return accrual profile
  - Basis for estimation of Expected Opportunity Cost ($E[OC]$) = change in return from deferral of trades
- Tax effects might be incorporated in the return signal, hence estimated $E[SR]$ and $E[OC]$

**Execution Costs**
- Expected Execution Cost ($E[EC]$) is estimated using a transaction cost function, and comprises:
  - Fixed Cost ($FC$)
    - Commissions and fees
    - Spread
  - Variable Cost: Market Impact ($MI$)
    - Trade size
    - Trading strategy
    - Security characteristics, e.g. volume, market cap, volatility
    - Market conditions
    - Permanent vs. temporary effects

**Portfolio Construction**
- $\text{MAX} \{E[SR – OC – EC] – \gamma \text{Risk}\} \quad s.t. \text{constraints}$
- Optimization => schedule of trades and portfolio positions
- Constraints may relate to position size, holdings (e.g. maximum % of market cap) and trading (e.g. volume participation)
- Definitions:
  - $PR = SR – OC – EC$
  - $OC = PR – SR$
  - $EC = FC + MI$
  - Implementation Shortfall (IS) = $–OC – EC$
  - $\gamma$ = risk aversion parameter
  - Risk is defined relative to objectives, e.g. standard deviation, tracking error, shortfall; ideally allowing for return and cost component covariance

**Expected Return & Capacity Evaluation**
- Expected Excess Return = $E[XR] = E[PR] – \text{Required Return (RR)}$
  - where:
  - RR is the threshold return, taking into account returns required to compensate for risk and management fees
- **Capacity** defined as FUM where $E[XR] = 0$
Of the variety of constraints that may be imposed when constructing portfolios, three are highlighted within Figure 2:

- **Constraints on portfolio positions** – These relate to limits on portfolio weights. They might include maximum deviations from benchmark weights, sector exposure limits, and so on. Note that this constraint is not a function of FUM.

- **Constraints on holdings** – These relate to specific investments. An example would be a limit on holding no more than 5% of market capitalization of any particular stock.

- **Constraints on trading** – These relate to expected limits on the physical ability to trade. An example would be to assume that participation in average daily volume cannot exceed some level, say 50%.

The output from the optimization would comprise a schedule of portfolio positions and associated trades over the investment horizon. While we do not specify any particular functional form for estimating the portfolio weightings and trades, one approach might be to optimize with respect to the expected portfolio value and its risk at the end of the analysis period over which \( E[SR] \) is expected to accrue.

The bottom box in Figure 2 describes the evaluation of expected returns and capacity. The evaluation is undertaken through focusing on expected excess return (\( E[XR] \)), defined as the expected portfolio return (\( E[PR] \)) less a required return (\( RR \)). The latter is the return required to adequately compensate investors for risk and the fees they pay. It may also include an additional minimum return hurdle that allows investors to reap some net benefit, if so desired. This approach evaluates capacity under a ‘threshold’ capacity definition (see Vangelisti, 2006, and Section 2). Capacity analysis is undertaken by running the model at differing FUM. Elements that are bolded and underlined are affected by changes in FUM, and hence are the key levers in capacity analysis. They include trade size (which impacts on \( E[EC] \)), and any holding or trading constraints.

Figure 3 presents a framework for capacity analysis under a predictive model with trade optimization only. Given this is a limited case of the general framework outlined in Figure 2, we only highlight selected elements and do not describe each component in full. Under this approach, the trade schedule is optimized for a given target portfolio under trading frictions. Trading frictions comprise execution costs, and any trade or holding constraints that may be encountered when forming the actual portfolio. The target portfolio represents the portfolio that the manager would hold in the absence of any trading frictions, incorporating any portfolio constraints such as position limits. The target portfolio is entirely a function of \( E[SR] \), asset risk and portfolio constraints; and accords with the paper portfolio concept of Perold (1988). The optimal trade schedule is determined by trading off \( E[IS] \) against aversion to variability in \( IS \). The latter relates to the fact that deferring trades leaves an investor exposed to uncertainty over future prices, and hence at risk of encountering either greater \( EC \) or greater \( OC \) if prices move in an adverse direction. By contrast, the act of trading limits this risk by crystallizing both \( EC \) and \( OC \). Trade optimization delivers a schedule of trades, which if combined with the target portfolio generates a schedule of portfolio positions. \( E[PR] \) can be estimated either directly from the schedule of portfolio positions, or indirectly by adjusting the expected return on the target portfolio (\( E[SR] \)) by the expected implementation shortfall (\( E[IS] = E[-EC-OC] \)). The indirect method is reflected in the dashed lines appearing in Figure 3. Capacity analysis is again undertaken by running the model at differing FUM, with the elements that are affected by changes in FUM being bolded and underlined.
Figure 3: Capacity Analysis under Predictive Model with Optimal Trade Scheduling

- **Signal**
  - $E[SR]$
  - Accrual profile

- **Target Portfolio**
  - $\text{MAX } (E[SR] – \gamma \cdot \text{Risk})$
  - s.t. portfolio constraints (e.g. position size)

- **Opportunity Cost**
  - Function of $E[SR]$
  - accrual profile

- **Transaction Cost Model**
  - $EC$ as function of trade size

- **Trade Schedule**
  - $\text{MAX } E[\text{Terminal Portfolio Value (TPV)}]$
  - $= \text{CUMULATION of } E[SR – OC – EC]$
  - s.t. Aversion to variability in IS
  - Trade constraints (participation, holdings)

- **Schedule of Portfolio Positions**

- **Expected Return & Capacity Evaluation**
  - **Expected Excess Return** $= E[XR] = E[PR] – RR$
    - where:
      - $RR$ is the threshold return
  - **Capacity** defined as FUM where $E[XR] = 0$
5.3. Predictive Models Appearing in the Literature

We now provide a review of the literature on predictive models, against the background of the frameworks discussed in Section 5.2. The class of models we address here are distinguished in that they allow for both execution costs and expected returns from a signal that decays over time. The discussion occurs in three parts. First, we provide an overview of models of optimal trading in the context of transaction costs and a decaying signal. These models were the first to be developed in this class; and broadly accord with the framework presented in Figure 3. We then discuss selected models of optimal portfolio construction under transaction costs and a decaying signal. This literature is more recent, and moves in the direction of the general framework presented in Figure 2. However, the available models typically impose a limiting structure or simplifying assumptions, in order to generate closed form solutions or tractable analysis. Finally, we discuss how this class of models has been applied to capacity analysis, thus constituting the ‘predictive model’ concept that we address. It is worth noting that some commercial providers offer the means to identify optimal trading schedules or portfolios based on models similar to those outlined below, e.g. ITG, J.P.Morgan.

Models with Optimal Trading

Arnott and Wagner (1990) were among the first to place some structure around the problem of determining trading strategies under transaction costs and decaying expected returns arising from an investment signal. They use graphs to describe the trade-off between ‘immediacy cost’ and OC, where the latter accrues at a different rate over time for ‘fast’ versus ‘slow’ ideas. This work presages more formal models involving dynamic optimization, which have been developed by Bertsimas and Lo (1998), Almgren and Chriss (2001) and Huberman and Stanzl (2005). These authors address the issue of optimally scheduling trades over a given horizon to minimize transaction costs, while maximizing the probability of capturing the available expected returns. Bertsimas and Lo (1998) minimize transaction costs in a manner that effectively amounts to minimizing IS, but do not consider risk. The models of Almgren and Chriss (2001) and Huberman and Stanzl (2005) essentially trade off net realized prices (i.e. after EC) against ‘risk’, defined as the variance of expected trade prices. As discussed, the intuition is that an investor will be averse to the uncertainty over price arising from deferring trades, as adverse price movements could lead to either greater EC or OC. Capturing this uncertainty through the variance of prices implicitly assumes that the risks are normally distributed around the expected price path. Introducing a risk aversion parameter allows the optimal strategy to be identified, and supports tracing out an efficient frontier.

While these models of optimal trading are framed in relatively general terms, their implementation requires the specification of the dynamics for the ‘unperturbed’ expected (i.e. benchmark) price in the absence of trading, and how trading affects prices and hence execution costs. The effect of trading on execution costs comprises both the impact of trading on realized net prices at the time of the trade (i.e. temporary MI plus fixed costs); and any flow-through effect into subsequent prices (i.e. permanent MI). The functional form for these inputs determines the nature of the model, and dictates how it might be solved, i.e. closed form, approximation, or numerical estimation such as simulations. All authors initially present their model under the assumption that the unperturbed price (or excess return) follows a random walk, before considering extensions where prices (or returns) are predictable under very simple structures. Bertsimas and Lo (1998) adjust the price level for an information component that decays with time as an AR(1) process; while Almgren and Chriss (2001) extend their model for a linear ‘price...
drift’ term and serial correlation. Huberman and Stanzl (2005) make the point that price predictability is problematic as it may admit arbitrage; although they present a version of their model with correlated orders from noise traders that imply predictable MI effects. With regard to MI, all authors address the possibility of temporary and permanent MI. There is also recognition that MI need not be linear, in particular the temporary component. The MI functions selected by these authors for the purpose of solving the model or constructing examples are typically basic.  

Models with Optimal Portfolio Construction

More recently, authors have extended this class of models by allowing expected transaction costs and return decay to influence portfolio construction as well as trading. This amounts to an integration of trading and portfolio construction, leading to an ‘optimal’ schedule of trades and portfolio positions over time. Engel and Ferstenberg (2007) present such a model in the mean-variance tradition with a finite horizon that can accommodate the incorporation of alpha expectations into the trajectory of expected prices. They solve their model for MI dynamics following Almgren and Chriss (2001), including temporary and permanent components. Gârleanu and Pedersen (2013) present what is currently the most advanced model, taking into account expected returns and their future evolution, as well as MI. Under their formulation, the investor trades towards an ‘aim’ portfolio, which provides a moving target that may never be reached. They derive closed form solutions under the assumptions of: excess returns that decay at a constant rate (i.e. AR(1)); a quadratic function for temporary MI; and additive distortions to account for permanent MI. Sivaramakrishnan et al. (2015) develop a two-period model as a simplification of Gârleanu and Pedersen (2013), and apply it to portfolio construction and simulation under two alpha signals that decay at differing rates.

Application of Predictive Models to Capacity Analysis

Perold and Salomon (1991) appear to be the first to directly contemplate the concept of evaluating capacity using predictive models. Working within the implementation shortfall framework of Perold (1988), they describe the trade-off between excess return, the cost of execution, and the opportunity costs arising as a consequence of deferring or forgoing trades. They present a graphical analysis that focuses on the wealth-maximizing amount of FUM, illustrating how it varies with assumptions about the magnitude of expected alpha and transaction costs.

Both Kahn and Shaffer (2005) and Serbin et al. (2009) extend Perold and Salomon (1991) by modeling capacity in a mean-variance framework that aligns with the ‘fundamental law of active management’ of Grinold (1989). These authors maximize net alpha conditional on: alpha prior to costs; active risk relative to the benchmark; an expression for transaction costs; and FUM. This is done by solving for the optimal level of turnover. Under these models, the reduction in net alpha that occurs with FUM is partly

30 Obizhaeva and Jiang (2013) argue that MI is a complex function of supply/demand dynamics around the limit-order book. They model an optimal trading strategy that takes these dynamics into account, and derive substantially different solutions to those emerging from the more traditional models of optimal trading.

31 Coppejans and Madhavan (2007) also present a mean-variance model of optimal portfolio formation under an alpha signal and transaction costs, within the fundamental law of active management framework of Grinold (1989). However, their model is for a single period, and only generates predictions for optimal turnover per period rather than a schedule of trades over time.

32 The applications presented by these authors differ in how they characterize the relation between trade size and execution costs. Kahn and Shaffer (2005) using a square root function of trade size; and Serbin et al. (2009) use the ITG Agency Cost Estimator (ITG ACE) which is described further below. Kahn and Shaffer (2005) impose an expected alpha level, while Serbin et al. (2009) simulate under the assumption that the investor predicts the actual returns over the forthcoming month with error.
mitigated by reducing turnover, which arises endogenously. However, no allowance is made for the time period over which alpha accrues.\textsuperscript{33}

Coppejans and Madhavan (2007) extend the analysis within a fundamental law of active management framework via a model where the trade-off between alpha and transaction costs is embedded within the transfer coefficient (see Clarke et al., 2002). Through defining transaction costs as implementation shortfall, and allowing for these costs to be ‘amortized’ at a potentially differing rate to alpha decay, they effectively embed a link to the timing of alpha accrual within a single period model. They apply their model to capacity analysis by expressing costs as a function of FUM, with components to capture both market impact as a function of stock variance and economies of scale from fixed costs.\textsuperscript{34} This model generates a FUM at which the information ratio is maximized; and allows for the negative effects of increased FUM to be partly mitigated by reductions in turnover and breadth. Nevertheless, their model still generates only an estimate of optimal turnover, rather than a schedule of optimal trades over time. It is also highly stylized, particularly in the way in which market impact is modeled.\textsuperscript{35}

The papers described so far focus on turnover as the ‘lever’ through which transaction costs are controlled. Brandes et al. (2009) extend the analysis by Serbin et al. (2009) by forming optimal portfolios that allow for expected transaction costs to vary across stocks, in line with the ITG Agency Cost Estimator (ITG ACE\textsuperscript{\textregistered}) model. Their application involves simulations with monthly rebalancing. They demonstrate that taking transaction costs into account when forming portfolios can have a significant impact on returns and hence capacity estimates (as well as realized turnover), relative to a situation where portfolios are formed in isolation and implemented accordingly.

Deriving an optimal schedule of trades subject to an accrual profile for expected returns requires moving from a single-period to a multi-period analysis, and embracing dynamic optimization. Amihud and Mendelson (2013) present a model that moves in this direction, describing ‘investment profit’ as a function of the multi-period sum of expected returns less transaction costs over an analysis horizon, where transactions costs are expressed as a quadratic function of trade size plus a fixed cost. They derive an expression for the FUM that optimizes profit, including a simplified version under the assumption of equally-weighted holdings. While Amihud and Mendelson (2013) only address estimation of capacity using the simplified version of their model, the full model could feasibly provide a foundation for locating the optimal trade schedule in conjunction with capacity analysis. However, their formulation includes only basic terms for expected returns based on alpha and transaction costs, and does not explicitly account for the opportunity cost associated with deferring or failing to trade.

Landier et al. (2015) provide one example of capacity analysis using multi-period dynamic optimization, drawing on the model of Gârleanu and Pedersen (2013). Capacity is modeled as a function of: baseline expected performance in the absence of transaction costs, described by a Sharpe ratio; a quadratic transaction cost function; an alpha signal with a constant rate of decay; and a target net

\textsuperscript{33} Vangelisti (2006) also refers to modeling capacity using an optimization that trades off: alpha based on country, value and momentum characteristics; risk of deviating from the benchmark along a number of dimensions; the cost of trading, and penalties for illiquid positions. However, the model is left unexplained.

\textsuperscript{34} Coppejans and Madhavan (2007) express market impact as a linear function of FUM in excess of a threshold, where the coefficient is a linear function of stock variance. Economies of scale are modeled as a declining function of FUM. They appear to impose an assumed expected alpha level via a baseline information ratio.

\textsuperscript{35} A number of authors present similar models to Coppejans and Madhavan (2007) based on the fundamental law of active management, e.g. Grinold (2007), Qian et al. (2007), Sneddon (2005, 2008). These models generate optimal portfolios and turnover allowing for transaction costs, and are derived under conditions where the horizon of an investment signal is profiled by either the autocorrelation of the signal, or the correlation between the signal and returns over a forecast horizon. However, none of these authors apply their models to analysis of capacity.
Sharpe ratio target of 0.3. These assumptions support a closed form solution, although their suitability might be questioned.\textsuperscript{36} The findings from this research were discussed in Section 4.3.

5.4. Modeling of Execution Costs

This section describes the concepts and methods for estimating $E[EC]$. The discussion fleshes out the aspects identified in the box at the top right of Figure 2. The definition of $EC$ we discuss here is the difference between the net realized price at which trades occur, and the (mid) price prevailing at the time the decision to trade was made (the ‘decision price’).\textsuperscript{37} The net realized price comprises the weighted average price at which the security is traded on the market (the ‘traded price’), adjusted for any commissions and fees. This definition is chosen as it accords with the concept of implementation shortfall, which is suitable for analyzing the capacity where the key concern is erosion of returns that can be extracted from a particular signal.

Differences between the traded prices and decision prices incorporate a range of potential influences, including:

- **Spread** – There is typically a minimum price concession because trades occur at the bid or ask price.

- **Liquidity provision cost** – This is the price concession required to attract a liquidity provider where more volume is sought than is available at the bid or ask price, i.e. trading into the order book. Costs associated with liquidity provision are often considered temporary,\textsuperscript{38} with the expectation that prices will revert once the order is completed.

- **Information leakage** – The traded price may adjust because the information contained in the trade becomes impounded into the price. In this case, the adjustment should be permanent.

- **Other security-specific price adjustments** – The traded price may adjust due to the arrival of new information, or as a consequence of orders placed by other investors.

- **General market movements** – The traded price may adjust with movements in the overall market. This can lead to asymmetric costs between buys and sells, depending on market direction.

The above suggests a workable formulation for modeling $E[EC]$ for the purposes of capacity analysis using predictive models. Commissions, fees and the spread might be designated as fixed costs ($E[FC]$), on the basis that they amount to the baseline cost of trading. Note that these cost components may vary across securities and time, and are only ‘fixed’ in the sense that they set a minimum cost for trading a particular security at a certain point in time. Expected variable cost ($E[VC]$) might be modeled as a function of the cost of liquidity provision and information leakage, and will be denoted as the ‘market impact’ (MI) for our purposes. Other security-specific price adjustments can be ignored for modeling $E[EC]$ on the assumption that they have an expected value of zero; although clearly they matter for price risk related to both $E[EC]$ and $E[OC]$. General market movements might be excluded under certain assumptions, such as where buy and sell trades are being matched, provided that both sides of the trade have equivalent market exposure. This leaves the following as a feasible, basic structure:

\textsuperscript{36} Landier et al. (2015) find that their predicted performance is not substantially different to simulated results from trading based on the signal, suggesting that the model may be robust to the constant signal decay assumption.

\textsuperscript{37} This follows Perold (1988), but differs from Collins and Fabozzi (1991), who define $EC$ as the sum of price impact and market timing costs.

\textsuperscript{38} The assumption that liquidity-motivated trades are always temporary can be debated, to the extent that sustained price adjustments can arise from supply and demand effects. Bouchard et al. (2004) present a model along these lines, and provide empirical evidence consistent with orders having permanent MI in the absence of information.
**For Buy Trades:**

\[ E[EC] = \text{Net Realized Price} - \text{Decision Price} = E[FC] + E[VC] \]

\( (>0 = \text{cost}; <0 = \text{benefit}) \)

where:

\[ FC = \text{Fixed Costs} = \text{Minimum Cost to Trade} \]
\[ = \text{Commissions + Fees + Spread} \]

\[ VC = \text{Variable Cost} = \text{Market Impact} \]
\[ = \text{Liquidity Provision Cost (Temporary) + Information Leakage (Permanent)} \]

The question arises regarding how components of \( E[EC] \) should be characterized in practice, in particular \( E[MI] \). Our aim in the remainder of this section is to overview the models appearing in both the academic literature and those used in practice. These models appear either in the form of empirical analysis of the determinants of \( EC \), or as models that can be used to estimate \( E[MI] \). The main message is that there exists a wide range of formulations, and that there is no commonly-accepted transaction cost model. \( MI \) models differ in attributes such as functional form, the range of determinants considered, and the parameterization of coefficients.

Borkovec and Heidle (2010) provide an overview of transaction cost modeling, noting that there are two major classes of models. The first approach is *empirical*. This involves analyzing realized transaction costs for securities arranged into groups by order size, stock characteristics, etc. The attributes for these groups are then used as predictors of \( E[EC] \) and its distribution for securities pertaining to each group. The second approach involves *analytical models*. These are usually in the form of structural models based on theoretical determinants of transaction costs, which are estimated using observed data. Borkovec and Heidle (2010) point out that most of the structural models used in practice are based on papers by Bertsimas and Lo (1988), Almgren and Chriss (2000) and Huberman and Stanzl (2005). These models estimate transaction costs conditional on the ‘optimal’ strategy that trades off \( E[MI] \) against \( E[OC] \). We overviewed these models in Section 5.3.

**Model Variables**

Listed below are the variables that appear in various models of \( MI \), either in the form of studies analyzing the determinants of \( MI \) for observed institutional equity market trades or orders, or transaction cost models of industry participants. For the latter, we have information on the models used by Citi (BECS), Investment Technology Group (ITG ACE®), J.P.Morgan, and Northfield.

- **Trade size** – This key determinant of \( MI \) has been defined in a variety of ways, including in absolute and relative terms. For example, Jones and Lipson (2001) analyze the number of shares in an order. Both Almgren et al. (2005) and Northfield initially define their model with trade size defined as the number of shares traded, but then estimate their models in relative trade size terms. Relative trade size has been specified in a number of ways:
  - *Trade / Average or median daily volume (ADV)* – This appears to be the most common definition, and has been used by: Lillo et al. (2003); Chiyachantana et al. (2004); Obizheava (2012); Borkovec and Heidle (2010) and the ITG ACE® model (see ITG, 2008); Citi BECS; J.P.Morgan; and Pritamani and Williams (2016). Nevertheless, the period over which ADV is estimated varies. For example, Berkowitz et al. (1988) scale trades by total volume over the trade day; Chiyachantana et al. (2004) use a 5-day average; ITG ACE® uses a 21-day average, whereas Lillo et al. (2003) scale by volume over 1-year.
  - *Trade / Volume over Duration of Trading* – This approach is used by Berkowitz et al. (1988), who examine daily data; as well as Almgren et al. (2005).
Trade / Shares on Issue – The number of shares traded relative to shares on issue is examined by Keim and Madhavan (1997), Breen et al. (2002) and Johanning et al. (2015); as well as Northfield, Loeb (1991) and Christoffersen et al. (2007) use trade value relative to market capitalization. Frino et al. (2006) use dummy variables based around percentiles for number of shares traded in a package, relative to shares on issue.

Trade size by percentile ranking – Chan and Lakonishok (1995) use seven dummy variables that represent percentile rankings based on the value of the trade parcels.

- **Functional form** – The functional form of the relation between trade size and MI varies from convex through to concave. The assumption of a concave function appears to have broad support, e.g. Hasbrouck (1991); Almgren et al. (2005); Lillo et al. (2003); Citi BECS; ITG ACE®. A power function with an exponent of 0.5 is a common choice, but is not universally accepted. For instance, Lillo et al. (2003) estimate exponents that vary between 0.1 and 0.5 with market capitalization as the input variable; while Almgren et al. (2005) argue for a 0.6 power law. Barclay and Warner (1993) provide a theoretical argument in support of a concave function based on stealth trading, under which orders containing information are broken up and have proportionately greatest impact in the medium-trade size segment. Nevertheless, a linear relation is estimated by many authors, including Berkowitz et al. (1988), Keim and Madhavan (1997), Jones and Lipson (2001), Breen et al. (2002), Chiyachantana et al. (2004) and Johanning et al. (2015). Almgren et al. (2005) apply a linear function for the permanent MI component, while using a power function for the temporary component. Northfield also uses a combination of linear and power terms. Meanwhile, Loeb (1991) estimates a convex quadratic function defined over block size relative to market capitalization.

- **Trading strategy** – A variety of strategies are available when undertaking trades, distinguished in part by the degree to which an investor pursues fast execution, versus distributing orders over time. Some strategies include: immediate execution; spreading orders with a view to secure a share of volume through time (e.g. uniform, VWAP by horizon, VWAP by participation); adopting a contingent strategy such as a limit order or trading opportunistically; or following some optimal strategy (see Leinweber, 2002; ITG 2008; Borkovec and Heidle, 2010). It is expected that institutions would execute trades strategically, and that this would influence realized MI. Evidence exists that this is indeed the case. Keim and Madhavan (1995) find that larger orders are spread over longer periods, and that the choice of order type is related to fund style. Over half the trade packages identified by Chan and Lakonishok (1995) are executed over four days or more. Chiyachantana and Jain (2009) find that information quality and trading conditions influence trading strategy, and that just under 50% of orders in their database are not completely filled. Research confirms that trading strategy is a determinant of MI. Breen et al. (2002) find that forecast errors for their MI model are negatively related to indicators reflecting more patient trading strategies. Jones and Lipson (2001), Almgren et al. (2005) and Chiyachantana and Jain (2009) find that MI is related to the extent to which orders are ‘worked’, including the degree to which orders are split and trade duration. Anand et al. (2013) provide evidence that institutions follow differing trading styles with differing MI. Commercial

---

39 For instance, Hasbrouck (2009) and Edelen et al. (2013) both estimate MI as a function of the square root of trade volume.

40 One industry practitioner noted during feedback that, in their experience, market impact does not always adhere to the smooth functional forms assumed under modeling. They suggested that market impact can be small or even negative for very small orders; then slope up to around 10% of ADV; and can be flat up to about 40% of ADV.

41 VWAP is Volume-Weighted Average Price.

42 Researchers also find that MI varies with fund style. Some authors examine the Plexus database which designates funds into index, fundamental value and technical traders, and find evidence that MI is highest for index funds and lowest for value funds (Keim and Madhavan, 1997; Jones and Lipson, 2001; Breen et al., 2002). Chan and Lakonishok (1995) analyze trade data provided by SEI Corp, and find that MI is higher for funds following strategies that are growth-oriented versus value-based; or involve higher turnover. Leinweber (1993,
predictive models, such as Citi BECS, ITG ACE® and J.P.Morgan’s, typically estimate transaction costs conditional on the trading strategy specified by the user.43

• **Security characteristics** – *MI* models often include various security characteristics to control for determinants of liquidity, other than those set by the investor such as order size and trading strategy. Again there is no consistency across models in terms of what variables are included. Some of the more common characteristics are listed below.

  o **Market capitalization** – This aims to control for the fact that liquidity tends to be higher, and hence *MI* lower, for larger stocks. Raw market capitalization appears in the models of Loeb (1991), Chan and Lakonishok (1995), Keim and Madhavan (1997), Domowitz et al. (2001), Lillo et al. (2003), Chiyachantana et al. (2004), Christoffersen et al., 2007, and Northfield. Breen et al. (2002) include relative market capitalization; while Frino et al. (2006) use dummy variables based around market capitalization percentiles.

  o **Average or median daily trading volume (ADV)** – This liquidity proxy appears in the ITG ACE® model, and the analysis of Breen et al. (2002). The former use *ADV* estimated over 21-days; while the latter include *ADV* relative to the market average over 3-months.

  o **Turnover** – This proxy for liquidity appears in the form of the ratio of *ADV* to shares outstanding in Christoffersen et al. (2007) based on 20-day *ADV*; as well as Northfield. Almgren et al. (2005) use the inverse measure based on 10-day *ADV*.

  o **Volatility** – A number of authors include security volatility as it is also known to be correlated with illiquidity. Again, there is no consistency in the measure employed. Almgren et al. (2005) and Citi BECS use an intraday estimator; ITG ACE® predict volatility using a proprietary model; J.P.Morgan employ volatility as a scaling factor; while Northfield state that they use the inverse of volatility, but do not mention how it is estimated. Domowitz et al. (2001) and Chiyachantana et al. (2004) find volatility is positively related to trading costs across global equity markets, but only insignificantly so for the latter.

  o **Bid-ask spread** – Some models include the bid-ask spread, including Frino et al. (2006), Christoffersen et al. (2007), Obizhaeva (2012), Johanning et al. (2015), J.P.Morgan and ITG ACE® (5-day average in the latter).

  o **Price level** – This is sometimes included to capture bid-ask and tick size effects, e.g. see Keim and Madhavan (1997); Breen et al. (2002); Jones and Lipson (2001); Chiyachantana et al. (2004).

  o **Other** – A range of other security characteristics appear in various models, even though the relation with liquidity may not be clearly established. For instance, Breen et al. (2002) include 13 variables in total. An example of a variable used in a few models is price momentum.

• **Side** – Whether trades are a buy or sell is often included in the model, on the basis that there can be *MI* asymmetries by market ‘side’. There was an initial view that buys have larger *MI* than sells (e.g. see Chan and Lakonishok, 1995; Keim and Madhavan, 1997); with Saar (2001), suggesting that this reflects permanent *MI* for buys as they are more likely to be information-driven. However, some researchers have found the cost of transacting buys to be lower than sells (e.g. Jones and Lipson, 2002) reports higher trading costs for a value management style, and the lowest for ‘earnings surprise’. Christoffersen et al. (2007) find that active managers incur lower *MI* costs relative to passive managers, which they attribute to greater discretion over trading. Frino et al. (2006) find that passive funds incur greater *MI* costs due to demand for immediate liquidity; whereas active funds are more exposed to permanent *MI* related to information leakage.

43 Pritamani and Williams (2016) use the ITG ACE® model to demonstrate how execution costs vary with trading strategy for an Australian superannuation fund investing in Australian and international equities.
Chiyachantana et al. (2004) provide evidence that the effect of side depends on market state, i.e. whether the investor is trading with or against market direction.

- **Time of day** – MI may vary with the time of day, and this may be taken into account in more granular transaction cost models. The ITG ACE<sup>®</sup> model, for instance, takes time of day into account.

- **Market conditions** – Chiyachantana et al. (2004) and Anand et al. (2013) provide evidence that MI impact varies with market conditions. J.P.Morgan’s algorithms have a flag for ‘special’ days that might impact the modeling of execution cost, such as index rebalancing or option expiry.

- **Market structure** – MI can vary across exchanges, including whether securities are traded on a listed market, over-the-counter, and perhaps in dark pools. For example, Keim and Madhavan (1997) find that trading costs are lower on NASDAQ than the NYSE; while the prediction errors in Breen et al. (2002) imply the converse. Chan and Lakonishok (1997) suggest that exchange on which stocks are traded impacts on execution costs via interacting with trade size, or time period. Both Domowitz et al. (2001) and Chiyachantana et al. (2004) find that MI is higher in emerging markets after controlling for other determinants; while the latter provide evidence that MI is lower for markets that are liberalized and have stronger shareholder protections. Most commercial providers estimate their models separately for different markets.

- **Temporary and permanent components** – Some, but not all, models explicitly allow for separate estimation of temporary and permanent MI components, including: Breen et al. (2002); Almgren et al. (2005); Frino et al. (2006); ITG ACE<sup>®</sup>; J.P.Morgan; and Northfield.

### Definition of Market Impact

Transaction cost models are also applied to a range of definitions for MI, and hence are often trying to explain differing dependent variables. Any measurement of MI requires specifying a fair price benchmark against which the realized trade price can be compared. Collins and Fabozzi (1991) outline four fair price measures: pre-trade, post-trade,<sup>44</sup> intraday (e.g. versus VWAP), and factor-adjusted.  

Our primary interest is in the estimation of $E[IS]$ with reference to the decision point, of which $E[M_I]$ is a key component. This implies that any model should ideally be able to generate expected trade prices for an entire planned change in portfolio position, evaluated with respect to the decision price. Unfortunately, this does not line up with the empirical estimation for many transaction cost models that are reported in the literature. Some researchers use fair price benchmarks that are not consistent with measuring IS, such as VWAP (e.g. Berkowitz et al., 1988; Domowitz et al. 2001; Christoffersen et al., 2007); or price at the commencement of a trade period or ‘bin’ (e.g. Breen et al., 2002). Many authors base their analysis around the commencement of the trading of an order or ‘arrival time’, thus using arrival price as a pre-trade benchmark.<sup>45</sup> This approach is used by Chan and Lakonishok (1995), Jones and Lipson (2001), Lillo et al. (2003), Almgren et al. (2005), Frino et al. (2006), Chiyachantana and Jain (2009) and Johanning et al. (2015). Order level data poses challenges for measuring IS, as the time that an order is recorded could occur after the decision is made.<sup>46</sup> The closest to observing decision price is where the database contains a record of the decision to trade (e.g. Plexus data). Such data has been used by Keim and Madhavan (1997) and Chiyachantana et al. (2004); and commercial providers

---

<sup>44</sup> Post-trade benchmarks can be used to extract the permanent component of MI by comparing realized trade price with the price at some point after the order is completed, e.g. see Almgren et al. (2005).

<sup>45</sup> Some researchers only have access to trade data, although attempts have been made to aggregate trades into parcels that reflect entire orders following Chan and Lakonishok (1995), with some success.

<sup>46</sup> For example, Keim and Madhavan (1996) find signs of significant price movement prior to the date of block trades, consistent with information leakage as the block is ‘shopped around’.
such as Citi BECS and ITG also have access to order data. In any event, observed orders may only reflect part of the position change that a fund manager ultimately intends to undertake, and hence may provide a partial view of the potential MI associated with rebalancing a portfolio.

The fact is that most trade databases and hence MI models appear to be built around orders, rather than intended changes in portfolio positions. This dictates the manner in which these models might be employed for capacity analysis under predictive models. A transaction cost model could be reliably used to estimate the expected MI for individual orders submitted into the market. For example, it might provide MI estimates for completing an order of (say) X units over a day, or Y units per week, assuming optimal execution of a specific order conditioned on a specified trading strategy. These orders would comprise the schedule of trades under which a change in portfolio position is given effect. The predictive model is used to specify the optimal schedule of trades (orders), by trading off E[EC] derived from the transaction cost model, against E[OC] as well as risk. The trick would be to extract any permanent MI from the transaction cost model and accumulate it, so that it can be incorporated into the predictive model and hence the specification of optimal trade schedule and/or portfolio.

Estimation Issues

The difficulty of estimating transaction cost models is compounded by data issues. For example, Breen et al. (2002) conduct their analysis on net turnover, defined as buyer less seller initiated volume, meaning that their analysis is not focused on MI associated with individual orders. The time span analyzed can vary, with some researchers examining trades occurring over the course of a day (e.g. Almgren et al., 2005; Christoffersen et al., 2007), others analyzing trading over consecutive trading intervals or ‘bins’ (e.g. Breen et al., 2002; ITG ACE®), and others focusing on multi-day trading patterns (e.g. Chan and Lakonishok, 1995). Many trade datasets contain limited data on larger trades, which reduces confidence when applying transaction cost models to block trades. Sometimes the order side needs to be inferred from the direction of trades, which creates scope for error. Data often reflects a mixture of discretionary and non-discretionary trades (see Borkovec and Heidle, 2010). Obizhaeva (2012) argues that estimating execution costs based on observed trades may lead to underestimating MI due to selection bias related to the propensity for only more attractive trades to be undertaken. Korajczyk and Sadka (2004) suggest that the reporting of only more beneficial larger trades may give rise to the appearance of concavity, when the underlying trade-off being faced might even be convex. In summary, a variety of data issues, combined with the fact that the underlying trading strategy is typically unknown, creates problems in estimating transaction cost models for predictive purposes. This is especially the case when the model is to be used to estimate expected execution costs where discretion exists, and there is considerable scope to pursue optimal trading strategies.

5.5. Expected Alpha and Opportunity Costs

There is much research on the magnitude of excess returns associated with various anomalies, which we will not review here. However, there appears to be limited research on the accrual profile for these excess returns. Often researchers appear to simply impose a rebalancing cycle (e.g. yearly), without

---

47 Trades are classified as buyer-initiated or seller-initiated on the basis of whether the trade price is greater than or less than the midpoint of the prevailing best bid and ask quotes.
48 We received feedback from one market participant that market impact in Australian small-cap stocks is often less than implied by transaction cost models due to the ability to secure large blocks on many occasions.
49 Obizhaeva (2012) examines portfolio transition trades, and finds estimates of trade costs that are comparable to other authors, implicitly suggesting that any selection bias may not be substantial.
directly addressing the period over which returns to the anomaly actually occur.\textsuperscript{50} The main exception relates to the momentum anomaly. In their seminal paper, Jegadeesh and Titman (1993) consider a range of periods over which past returns are measured, and over which the associated mimicking portfolios are held (the ‘J’ and ‘K’ categorization). They also report cumulative returns on a month-by-month basis. The tradition of focusing on timing issues continued with subsequent momentum research, e.g. see Galariotis (2010). Lakonishok et al. (1994) profile the returns arising from decile portfolios sorted by value measures, reporting payoffs on a yearly basis out to five years. While this analysis lacks granularity, it does suggest that returns from value investing may take 3–4 years to accrue. Schwartz and Steil (2002) survey Chief Investment Officers on the expected timing of price correction for mis-priced stocks. The most popular answer was ‘one month to one year’, followed by ‘over one year’. However, these survey results only represent perceptions, albeit informed.

Qian et al. (2007) introduce the concept of ‘information horizon’ as a method for characterizing the profile over which a signal decays. They define the ‘lagged information coefficient (IC)’ as the correlation between a (factor-based) signal and returns over each future period; and ‘horizon IC’ as the correlation between the signal and the accumulation of returns over various horizons. They report lagged IC and horizon IC, as well as related information ratios (IR), for price momentum and value (trailing earnings yield) signals over three quarters. Their illustrative analysis finds that the momentum signal has decayed by the third quarter, whereas the value signal continues to generate a high IR.

While there may be papers that we have missed, it nevertheless seems that there is scant publically available research on the rate at which expected returns accrue for various investment signals. This appears to be a fruitful area for future research, especially given that the period over which returns accrue may be an important determinant of the capacity associated with a signal.


This report discusses the issues surrounding capacity – including its definition and drivers – and outlines various methods by which it might be for evaluated. In doing so, it focuses on capacity in an equity market context, largely reflecting how much of the existing knowledge and applications have been framed. The report constitutes a ‘foundation document’ for further work. So, what is next?

Looking forward, two additional outputs are being planned as part of this CIFR project:

- **Capacity from an asset owner perspective** – We will expand the focus in an opinion piece that considers capacity from the perspective of institutional asset owners, such as a pension funds. One intention is to discuss capacity in the context of multi-asset portfolios. This will include observing how capacity and its drivers vary across asset classes and strategies; and how capacity might be managed by funds of various sizes. Another is to address some of the related agency issues, such as the relation between capacity and the incentives faced by investment management organizations and their fund managers.

- **Illustrating and contrasting the methods** – Another report will apply some of the methods for measuring capacity as identified above to data for an actual portfolio, plus selected factor-based signals. The analysis will be conducted with two aims. First is to illustrate how capacity analysis might be undertaken in practice. Second is to gauge the sensitivity of the findings to how capacity is modeled. We hope to provide some insight into whether moving beyond simulations to predictive

\textsuperscript{50} Fama and French (2007) examine how migration between portfolios contributes to returns for the size and value anomalies; although this only vaguely relates to the timing of payoffs, and they impose a yearly rebalancing assumption.
models is worth the additional effort; and gauge the extent to which results vary depending on the specification or parameterizations.

The current report also points to the gaps in the capacity literature, and hence potential areas for future research. We flag three such areas. The first (and most obvious) relates to the lack of much publically available research into capacity for asset classes other than equities. There are good reasons to expect that the issues and drivers will vary across asset classes. In fixed income markets, trading structures and liquidity conditions differ to equity markets. In bond markets, there is a regular process of refinancing existing debt as it rolls over; and illiquidity can become problematic after the primary placement is completed and an issue moves ‘off the run’. Strategies that rely on derivatives or arbitrage opportunities – such as certain hedge fund activities – face their own particular capacity issues. And, as mentioned earlier, for some unlisted assets the ability to source and manage investments may be more relevant to capacity than the magnitude of FUM. All these aspects provide fertile grounds for future research.

The second research area relates to methods for analysis of capacity under predictive models, as described in Section 5 of this report. Predictive models where both the portfolio and trade schedules are jointly optimized to account for transaction costs and alpha decay have been developed only recently. However, existing models may be viewed as too complex and opaque by many investment industry participants, given their highly technical nature (notwithstanding the resort to simplifying assumptions). The ideal would be to develop methods for portfolio construction and trade scheduling that might be readily understood and applied by the average portfolio manager. Perhaps the challenge is too great given the complexity of the task – but the idea is worth throwing out there.

A third area where there is ample room for additional research relates to the profiling of expected returns arising from a signal. This can be a key driver of capacity, which is likely to be greater for signals where excess return is larger in magnitude, accrues over a longer period, and is sourced from a broader range of securities. As indicated in Section 5.5, research into the profile of return generation is quite limited.
References


Citi. BECS 2.0. Product overview, Citi Equities, Best Execution Consulting Services.


Eggins, John (2008), Boutique investment managers: Adding a touch of spice, with a good dose of care, Forum, Russell Investments (May).


Leinweber, David J. (2002). Using information from trading in trading and portfolio management: Ten years later. Social science working paper No. 1135, California Institute of Technology.


