Asset Management with Price Impact and Fair Treatment of Clients

Michal Jezek and Stephen Satchell*

Version: May 2010

Abstract

In light of recent regulatory initiatives focusing on fair treatment of customers in financial markets, this paper examines the agency problem created by an asset manager with market impact, segregated accounts and preference-based contracts. It illustrates how aggregate client welfare and assets under management are affected by the order in which clients' accounts are sequentially traded and demonstrates that the manager is unlikely to have incentives for equal treatment of clients. Effectively, she may conduct limited invisible transfers of wealth among largely uninformed clients by granting preferential market access to some of them and this may be purely the result of her dollar-alpha maximization efforts rather than size/importance-based client discrimination. Increased transparency and/or effective regulation in this area seem socially desirable since the manager's incentives and client welfare generally appear to be misaligned.

JEL Classification: G11, G20, D18

Keywords: Multi-account portfolio optimization, market impact, treating customers fairly

*Jezek (e-mail: m.jezek.04@cantab.net) is from the Faculty of Economics and King's College, University of Cambridge, and gratefully acknowledges financial support from the Gates Cambridge Trust and the Economic and Social Research Council. Satchell is from Trinity College, University of Cambridge. We thank Oliver Williams, John Wylie, Bernd Scherer, John Knight, Kim Coppin and all workshop/seminar participants for helpful comments and suggestions. Any errors are our own.

1 Introduction

Recent regulatory initiatives have focused on fair treatment of customers in financial markets. For instance, the Financial Services Authority in the UK stipulates:[1]

> *A firm must pay due regard to the interests of its customers and treat them fairly. [It] must pay due regard to the information needs of its clients, and communicate information to them in a way which is clear, fair and not misleading. [It] must manage conflicts of interest fairly, both between itself and its customers and between a customer and another client.*

In many cases, successful implementation requires attention to details of which most clients may be unaware. This paper examines the agency problem created by an asset manager with market impact, segregated accounts and preference-based contracts. It shows that when it comes to execution quality, the manager has natural incentives not to act as envisaged by the regulation.

Practical asset management often involves portfolio rebalancing in markets with insufficient depth. Positions in small caps or illiquid emerging market stocks, for instance, are usually quite difficult to adjust without price effects and this regards, at least temporarily, a very broad range of securities in times of financial crisis such as that of the late 2000s. Large institutional investors are well aware of their market impact and are forced to incorporate it in their trading strategies.

In general, execution quality is an important issue for any investor.[2] Due to the effect of compounding, poor execution can result in substantial losses in a large frequently-rebalanced portfolio managed over a long period of time. This explains the recent growth in algorithmic trading and expansion of various dark pools of liquidity[3] as well as other mechanisms aimed at limiting price impact, such as iceberg orders with their full size hidden.[4] Nevertheless, market impact will not disappear entirely for large traders, even if the execution is spread over several days.[5] In fact, there are commercial analytical tools available to market participants specifically designed for portfolio optimization in the presence of price impact.[6]

As a consequence, a large asset manager with multiple clients[7] may strategically trade their accounts in such a way that some bear a disproportionately greater burden of the market impact costs than others. This is equivalent to invisible transfers between client accounts and even though they may seem relatively negligible, their impact on clients' final wealth can be quite noticeable. While for sophisticated market participants, e.g. at the interdealer trading level, it is simply their job to know whereabouts in the packet of orders they stand, there is a sizable segment of the market populated by much less informed participants. From less sophisticated institutional investors all the way down to the retail level, presumably a large number of clients are either genuinely unaware of the possibility of such arrangements or face prohibitive monitoring costs relative to their size. Such a client base allows for unobserved

[1] See Principles for Businesses No. 6-8 in the FSA Handbook at www.fsa.gov.uk. Similarly, rules on fair dealing with customers in the US can be found at www.finra.org.

[2] Bessembinder (2003) and Zhao and Chung (2007), for instance, show how institutional and regulatory changes in the past led to improvement in average execution quality.

[3] See e.g. Domowitz et al. (2008).

[4] De Winne and D'Hondt (2007) analyze their general strategic use in practice.

[5] In such a case, other costs enter into play as well. See Cai and Sofianos (2006).

[6] See e.g. Sofianos et al. (2007).

[7] See O'Cinneide et al. (2006) and Satchell and Scherer (2010) for pioneering work on multi-account portfolio optimization.

discretionary order flow prioritization by the manager, should there be incentives for that and should detection and enforcement of a potential explicit regulatory ban on such practices be weak.

There is more to transaction costs of portfolio management than market impact, such as brokerage commissions, fees charged by lenders of securities for short-selling or taxes. This paper, however, focuses specifically on the issue of market impact and the opacity with which the resulting loss might be distributed among clients, regardless of fairness or welfare considerations.[8] In particular, we show that even in the absence of pure favouritism, a rational manager interested primarily in her fee income will generally have incentives towards client discrimination. It is apparent that the problem is of substantial interest to clients, asset managers, regulators and supervisors alike. Our analysis seeks to provide relevant insights and stimulate further discussion by both academics and practitioners.

The paper is structured as follows. Section 2 presents a motivating example of single-account portfolio optimization with market impact, Section 3 outlines the main features of the multi-account optimization problem and Section 4 builds the microstructure framework in which the problem will be analyzed. Sections 5 and 6 then illustrate that distribution of market impact costs can be viewed as distribution of the manager's total alpha among her clients and that ordering of client trades affects not only the way in which the 'alpha pie' is sliced, with the corresponding welfare implications, but generally also its size. Section 7 demonstrates how a simple single-price rule takes such discretion away from the manager and enforces fair client treatment. Section 8 concludes.

2 Price Impact and Portfolio Choice

To motivate the problem, let us present a simple example of portfolio choice by a large investor with market impact. Suppose she has constant relative risk aversion (CRRA) preferences over her terminal wealth as popularly used in theoretical research. In particular, let the coefficient of relative risk aversion be unity. There are two types of securities, riskless and risky, and the investor allocates her wealth W between them so that her expected utility is maximized.

Formally, the investor's expected utility is

$$\mathrm{E}\left[u\left(\widetilde{W}_T\right)\right] = \mathrm{E}\left(\log \widetilde{W}_T\right) \tag{1}$$

where

$$\widetilde{W}_T = W \cdot [1 + r_f + \varphi \cdot (\widetilde{r} - r_f)] \tag{2}$$

is her terminal wealth, r_f is the rate of return on the riskless security and φ is the fraction of wealth allocated to the risky security which provides the rate of return

$$\widetilde{r} = \frac{\widetilde{P}_T}{P} - 1 \tag{3}$$

[8] Broadly speaking, there are many areas in which the principal-agent relationship in delegated portfolio management opens up a possibility for inefficient outcomes. These include, for instance, soft dollar arrangements (excess commissions to broker-dealers in exchange for additional services and favours, such as research or preferential IPO allocations) or payment for order flow (by stock exchanges or market makers to brokers for directing orders to them). However, unlike these previously addressed transparency issues, our analysis is largely confined to invisible in-house transfers of wealth with no external party directly benefiting from such undisclosed practices. In a certain sense, the manager's fiduciary duty is fully honoured in the aggregate and no Pareto improvement is possible.

with $\widetilde{P}_T$ being the (stochastic) terminal price of the security and $\overline{P}$ being its purchase price. If the risk premium is $\mathrm{E}(\widetilde{r}) - r_f > 0$, then the optimal allocation is $\varphi^* > 0$.[9] With an exogenously specified distribution of returns, it is a standard result that the fraction of wealth allocated to risky securities by CRRA investors is independent of their wealth.[10]

In practice, however, if the size of the investor's trade is large relative to the average trading volume of the security per the relevant time interval, this itself tends to result in a price movement that cannot be ignored. The larger the investor's wealth, the larger her demand for the risky security tends to be and hence the higher the (average) price $\overline{P}$ per unit of that security that she ends up paying.[11] As demonstrated in the Appendix, φ^* is strictly decreasing in $\overline{P}$ and so in effect, the investor's optimal allocation to risky securities increases less than proportionately with her wealth.

This example illustrates in the standard expected utility framework that ignoring price impact, calculating optimal demand at the prevailing market price and simply submitting a market order for that quantity is likely to lead to systematic errors resulting in suboptimal portfolio allocations and welfare loss.

3 Multi-Account Portfolio Optimization

We will examine a closely related practical problem faced by a large asset management company with market impact which has multiple clients and seeks to optimally allocate their individual portfolios. Not only does the manager have to incorporate price impact into individual clients' optimization but she also has the power to determine how the total loss due to this price impact is distributed among the clients. In general, a sophisticated manager will be maximizing her expected utility

$$v = v\left(\widetilde{f}, v_1^*, \ldots, v_N^*\right) \tag{4}$$

where $\widetilde{f}$ is the value of the fees she receives for her services and v_i^* is the maximized expected utility of client i conditional on the market prices at the time of trading on behalf of that client. All of the arguments depend on the manager's actions at present, particularly the order in which she sequentially trades clients' accounts and thus determines the distribution of the market impact costs. In general, she can split each client's order into a number of smaller orders and then trade these so-called 'clips' in any sequence she likes.

The key contribution of this paper is in pointing out that the perception of $\widetilde{f}$ by a sophisticated manager will generally vary with her actions in this respect. Even if she honours every single contract individually by maximizing each client's welfare given the market conditions, the fact that she co-creates those conditions provides her with some wiggle room for pursuing her own agenda. How she does that depends on the specification of (4). For instance, a benevolent manager may maximize what she views as aggregate client welfare $\sum_{i=1}^{N} \omega_i \cdot v_i^*$ irrespective of the impact on $\widetilde{f}$. Alternatively, cognizant of the problems with welfare aggregation, she may pursue fairness by distributing market impact costs in proportion to the size of client trades. At the other extreme, she may optimize her fee income regardless of fairness or aggregate client well-being. The latter may involve systematic favouring of some clients at the cost of others if that is what the manager deems instrumental to greater fees.

[9] This is true for any nonsatiated risk-averse investor. See Arrow (1971).

[10] *Ibid.*

[11] No price impact is assumed in the market for the riskless security.

Experience suggests that the case of a self-interested manager is likely to be the most relevant. In the following, we will examine the problem through a series of analytically tractable examples with contracts based on mean-variance preferences as utilized, rightly or wrongly, by a large number of practitioners.

4 Basic Microstructure

Let us consider an economy with two assets, a riskfree bond and a risky stock index. There is a large asset management company and a lot of small independent institutional and individual investors. Within the time frame available for trading, the residual supply and demand faced by the large asset manager are perfectly elastic in the case of bonds but imperfectly so in the case of stocks.[12] Technically, stocks are traded in an order-driven market and the large investor, demanding liquidity/immediacy, submits a market order which hits limit orders placed by the rest of the market. This moves the price away from its pre-trade level $\overline{p}$, which is the price that would prevail in the market in the absence of the large trader, and the degree of market resilience is such that the price will start reverting to that level only after the execution horizon of the large trader has elapsed.[13]

Pricing is discriminatory, i.e. if the quantity in the market order exceeds the quantity at the best available price, all orders at that price are matched and then the market order walks up/down the book, matching limit orders at less favourable prices. Specifically, we assume that the tick is infinitesimally small, the price grid in the limit order book is continuous and for the market depth parameter $\delta > 0$,

$$K(p) = \delta \cdot |p - \overline{p}| \tag{5}$$

describes the cumulative quantity on the bid and ask side of the book for $p < \overline{p}$ and $p > \overline{p}$, respectively.[14] The depth on both sides is limited, i.e. $K(p)$ is constant for p below or above a certain price threshold, but if the size of the market order exceeds the depth of the book, additional orders will arrive to extend the range in which (5) holds, reflecting partly the genuine reservation prices due to heterogenous beliefs of additional buyers or sellers and partly their exploitation of the presence of the large trader who needs to trade within a certain time frame.[15]

The large investor submits a market order with the quantity Q such that $Q > 0$ if it is a bid and $Q < 0$ if it is an ask. Then, given[16]

$$p(q) = \overline{p} + \frac{q}{\delta} \tag{6}$$

as implied by (5), the average share price for the investor is

$$\overline{P}(Q) = \frac{1}{Q} \cdot \int_0^Q p(q)\,\mathrm{d}q = \overline{p} + \frac{Q}{2 \cdot \delta} \tag{7}$$

[12] Thus, in the stock market, the large investor involuntarily plays the role of an 'elephant in the pool'.

[13] See e.g. Degryse et al. (2005) for an empirical analysis of market resilience.

[14] For simplicity of exposition, both sides of the book are linear and equally sloped. This is relaxed only later when it is relevant to the analysis.

[15] Thus, part of the market impact may be due to front-running traders who learn ahead that a large market order is due.

[16] In this framework, every infinitesimal share of stock in the market can trade at a different price and so for any bid (ask) order $q > 0$ ($q < 0$) between 0 and Q, $p(q)$ is the price of the *marginal* share bought (sold).

and the market impact cost is

$$C(Q) = \left[\overline{P}(Q) - \overline{p}\right] \cdot Q = \frac{Q^2}{2 \cdot \delta}. \tag{8}$$

Intuitively, an increase in the measure of market depth δ reduces the market impact cost.

Without loss of generality, let us assume that the asset management company has two clients with trading needs Q_1 and Q_2 such that $Q_1 > -Q_2 > 0$.[17] We will now illustrate some of the options the asset manager has to execute the trades. One option is to simply route the orders to the market,[18] which will result in the market impact costs $C(Q_1)$ for client 1 and $C(Q_2)$ for client 2. Alternatively, the quantity $N \in [0, -Q_2]$ could be matched in-house at some internal terms of trade and only the residual market orders for the quantities $Q_1 - N$ and $Q_2 + N$ would then be sent for execution to the market. Regarding the costs of price impact, we can state the following general result.

PROPOSITION 1: *For any given buy and sell orders from clients and a continuous nondecreasing $p(q)$ such that $p(0) = \overline{p}$, the aggregate market impact costs are minimized through maximal internalization of trading, i.e. netting orders in-house.*

PROOF: In the Appendix.

Thus, the total loss due to market impact is minimized if $N = -Q_2$ is traded in-house and only $Q_1 + Q_2$ is submitted to the market. In theory, however, the objective of the asset manager may not necessarily be to minimize the aggregate price impact costs, especially if those are borne by the clients themselves on top of explicit fees and charges. While fair customer treatment in terms of the latter may be required by regulation, the timing of internalization and the resulting terms of in-house trade might open an opportunity for client discrimination. In particular, internalization takes place at the price currently prevailing in the market and so it is at the discretion of the manager to determine at which stage of the price-moving transaction partial netting takes place.

To allow for differing depth of the ask and bid side of the limit order book in the following exposition, let $\delta = \delta_1$ on the ask side and $\delta = \delta_2$ on the bid side of the book with the substitutions made in (5)-(8) correspondingly. For each N, the manager can establish the price $p_N \in [p(Q_2 + N), p(Q_1 - N)]$ for a moment by routing only the relevant fraction of either the ask $Q_2 + N$ or the bid $Q_1 - N$ to the market, trade N shares in-house at p_N and only then complete the transaction by submitting the remainder of the total market order.[19]

If client 1 is the 'in-house favourite' and the manager minimizes the price of the block of shares purchased on behalf of this client, then obviously $p_N = p(Q_2 + N)$ and the formal

[17] For now, assume those are fixed irrespective of price fluctuations.

[18] We consider a single-arrival market in which each of the two market orders is matched with the existing limit orders as opposed to a batch-arrival market in which market orders are matched at the prevailing price and only the net order flow reaches the limit order book. In the latter framework, however, the same would be achieved by submitting either bids or asks with sufficient delay, assuming no relevant limit order book innovations in the meantime.

[19] Here, internal crossing of complementary orders of size N takes place at the most recent execution price p_N as the initial market order walks up/down the book. Alternatively, it could be the mid-price between the best bid and best ask following the execution of that order. Unlike our stylized model in which the bid-ask spread widens with no response from other market participants, newly arriving bids and asks in practice would tend to tighten it. More importantly, the manager in our model could submit a dummy order, namely a single-share limit order arbitrarily close to p_N, to attain the same price for internalization. For instance, Kazakov and Vasak (2006) report frequent use of single-share orders to establish one-tick bid-ask spreads at the Australian Stock Exchange as a result of specific rules for on-exchange crossing of orders below 1 million AUD. (Larger orders can be crossed off-exchange with no pricing rules imposed.)

problem becomes

$$\min_{N\in[0,-Q_2]} \left\{ \overline{P}(Q_1 - N) \cdot (Q_1 - N) + p(Q_2 + N) \cdot N \right\}, \tag{9}$$

which is solved by

$$N^* = \min \left\{ \frac{Q_1 - \frac{\delta_1}{\delta_2} \cdot Q_2}{1 + 2 \cdot \frac{\delta_1}{\delta_2}}, -Q_2 \right\}. \tag{10}$$

This implies that as long as

$$Q_1 < -Q_2 \cdot \left(1 + \frac{\delta_1}{\delta_2} \right), \tag{11}$$

the strategy of the asset manager that optimizes the overall terms of trade of client 1 results in incomplete internalization $N^* \in (0, -Q_2)$ and inefficiently high aggregate price impact costs, of which client 2 bears a disproportionately large part.[20]

It is straightforward to show that where (11) holds,

$$\frac{\mathrm{d}N^*}{\mathrm{d}\frac{\delta_1}{\delta_2}} < 0 \tag{12}$$

and so if there were an increase in the depth of the ask side of the book and/or a decrease in the depth of the bid side of the book, it would be optimal to dump a larger part of the order of client 2, namely $Q_2 + N^*$, in the market to depress the price temporarily and trade a smaller part of that order, namely N^*, in-house at that price because such manipulation would improve the overall terms of trade for client 1 at the expense of client 2.[21]

In practice, incomplete internalization would also give rise to extra transaction costs, but that might not necessarily invalidate the optimality of such a strategy if there were incentives to discriminate and no alternative means to do so. If we redefine client 2 as an aggregate of small clients with high relative costs[22] of switching to another asset management company, asset managers may indeed have incentives to compete with their rivals for large clients and grant them preferential treatment in equilibrium.

More relevantly, let us consider a situation in which the clients need to trade in the same direction, i.e. $Q_1 \cdot Q_2 > 0$. For notational simplicity, let us readopt a single δ for both sides of the order book as in (5)-(8). The aggregate price impact costs will be $C(Q_1 + Q_2)$ but their distribution across the clients is determined by the order in which the asset manager trades their accounts. In particular, if the account of client $i \in \{1, 2\}$ is traded first, she incurs $C(Q_i)$ and, consequently, the other client incurs $C(Q_1 + Q_2) - C(Q_i)$. Given

$$C(Q_1 + Q_2) = C(Q_1) + C(Q_2) + \xi(Q_1, Q_2) \tag{13}$$

where

$$\xi(Q_1, Q_2) = \frac{Q_1 \cdot Q_2}{\delta}, \tag{14}$$

[20] Client 2 bears $C(Q_2 + N|\delta = \delta_2) + \tau(N)$ while client 1 only $C(Q_1 - N|\delta = \delta_1) - \tau(N)$ of the aggregate price impact costs where $\tau(N) = (\overline{p} - p_N) \cdot N = -\delta_2^{-1} \cdot N \cdot (Q_2 + N) \geq 0$ is effectively an invisible transfer from the account of client 2 to the account of client 1.

[21] Since $\lim_{\delta_1/\delta_2 \to +\infty} N^* = -\frac{Q_2}{2}$, at least a half of the order of client 2 will always be internalized.

[22] The costs include, but are not limited to, time and effort associated with information acquisition and market analysis relative to the wealth at stake. Anecdotal evidence suggests that once a client has selected an asset manager, it takes a relatively strong impulse for her to switch, for which disproportionate sharing of hidden market impact costs seems unlikely to qualify.

the cost of being traded last is ξ and, effectively, it is at the discretion of the manager to decide from whose account this amount is taken.[23]

While fees and charges are relatively easily observable and verifiable by the clients, their priority queue in market access is not. Pure favouritism aside, a natural question of interest is whether there is some optimal way, from the perspective of the manager, of distributing ξ among uncoordinated clients who are largely unaware of its existence and observe only the total performance of their own portfolios. This is what we explore next.

5 Slicing of the Alpha Pie and Its Welfare Implications

Invariably, sophisticated investors can only achieve superior performance with a portfolio of limited size. By definition, no investment strategy can deliver excess profits with any amount of assets and in practice, the restriction on the volume of trades that can exploit existing mispricing and generate superior returns is likely to be relatively tight. This phenomenon is sometimes termed 'alpha decay', suggesting that the (marginal) alpha on the extra unit of money invested diminishes as the market moves against the investor.

For a given stock price p prevailing in the market, let alpha be defined as

$$\alpha = p_e - p \tag{15}$$

where p_e is the price level at which the expected return is justified by its risk as defined by the relevant equilibrium asset pricing model for the economy.[24] A sophisticated investor identifies mispricing through what she perceives to be a superior forecast of the future price of equity with the forecast error being incorporated in its risk.[25] Then, for p given by (6), the excess profit (dollar alpha) that the investor generates with an order of size Q is

$$\Pi(Q) = \int_0^Q \alpha(q)\,\mathrm{d}q = (p_e - \overline{p}) \cdot Q - C(Q), \tag{16}$$

which is maximized at $\widehat{Q} = \delta \cdot (p_e - \overline{p})$ such that $p\left(\widehat{Q}\right) = p_e$.

Thus, we can reframe the above problem of distributing market impact costs as the asset manager's problem of distributing her alpha among the clients. Π is referred to as the 'alpha pie' that is effectively sliced and distributed by the manager through her actions taken on behalf of the clients. In particular, different client ordering results in different slicing.[26] In general, the quantity traded by the manager will not be $\widehat{Q}$ but the investment company has incentives to expand by attracting new clients if it consistently underexploits the excess profit opportunities it is able to identify and, similarly, it has incentives not to overgrow.[27]

In the exposition above, clients' equity demand was fixed regardless of the execution price. Optimal demand derived from their preferences, however, generally varies with the price as

[23] For $Q_1 > -Q_2 > 0$, $-\xi$ is the amount saved in-house by internalization such that $C(Q_2)$ is saved by client 2 and $-\xi - C(Q_2) > C(Q_2)$ by client 1.

[24] For expositional purposes, alpha is defined in terms of price levels rather than rates of return. The concept may be less clear in this introductory setting with a single risky asset but it becomes more meaningful in a framework with multiple risky assets.

[25] Whether she is right or wrong is not pertinent to the analysis of her motives. Cf. Jezek (2009).

[26] Note that $\Pi(Q_1 + Q_2) = \Pi(Q_1) + \Pi(Q_2) - \xi(Q_1, Q_2)$.

[27] Unlike investment banks, the asset management company does not engage in proprietary trading, i.e. does not buy or sell securities for its own account. Perold and Salomon (1991) address the issue of the optimal amount of assets under management.

illustrated in Section 2 and so Q_1 and Q_2 also depend on the order in which client accounts are traded. The rational manager, trading on behalf of the clients based on their risk profiles and wealth, is aware of that and by the choice of their ordering, she may pursue her own objectives while still honouring her obligation to optimize each client's portfolio conditional on the current market price and the (expected) price impact of the transaction.

Let the preferences of client i over his terminal wealth $\widetilde{W}_{Ti} = \widetilde{W}_{Ti}(Q_i)$ be represented by the mean-variance utility function

$$V_i(Q_i) = \mathrm{E}\left(\widetilde{W}_{Ti}\right) - \frac{\rho_i}{2} \cdot \mathrm{var}\left(\widetilde{W}_{Ti}\right) \tag{17}$$

where $\rho_i > 0$ is a measure of aversion to absolute[28] portfolio volatility.[29] Initially, the client holds no stocks.[30] His wealth is W_i and the manager allocates it on his behalf between stocks and bonds based on her probabilistic assessment of the terminal stock price $\widetilde{P}_T$ and the client's preferences as in (17). Thus,

$$\widetilde{W}_{Ti} = [W_i - \overline{p} \cdot Q_i - C(Q_i) - I(i) \cdot \xi(Q_1, Q_2)] \cdot (1 + r_f) + \widetilde{P}_T \cdot Q_i \tag{18}$$

where

$$I(i) = \begin{cases} 0 & \text{if } i \text{ is traded first} \\ 1 & \text{if } i \text{ is traded last} \end{cases} \tag{19}$$

and r_f is the riskfree rate of return on bonds.

Hence, if the manager maximizes the utility of client i conditional on $I(i)$, the size of the market order submitted on his behalf is[31]

$$Q_i^* = \frac{\mathrm{E}\left(\widetilde{P}_T\right) - (1 + r_f) \cdot \left[\overline{p} + I(i) \cdot \delta^{-1} \cdot Q_j\right]}{(1 + r_f) \cdot \delta^{-1} + \rho_i \cdot \mathrm{var}\left(\widetilde{P}_T\right)}, \tag{20}$$

which depends on the quantity traded on behalf of client $j \neq i$ if $I(i) = 1$. As demonstrated in the Appendix, this implies

$$[Q_1^* + Q_2^*]_{12} = [Q_1^* + Q_2^*]_{21} \tag{21}$$

and

$$[V_1^* + V_2^*]_{12} \geq [V_1^* + V_2^*]_{21} \quad \Longleftrightarrow \quad \rho_1 \leq \rho_2 \tag{22}$$

where $V_i^* = V_i(Q_i^*)$ and the result inside the brackets $[\ ]_{ij}$ is obtained for $I(i) = 1 - I(j) = 0$.

[28] See the next section for an alternative mean-variance specification in which ρ_i measures aversion to relative portfolio volatility.

[29] The objective function can be viewed as the certainty equivalent of $\widetilde{W}_{Ti}$ under normality and constant absolute risk aversion (CARA) preferences with the coefficient of absolute risk aversion ρ_i. Since, however, neither normality nor CARA are realistic, we consider it an inferred index of client welfare over portfolio composition without necessarily embracing the expected utility framework in which variance is not an exact measure of risk [Rothschild and Stiglitz (1970)]. Thus, increases in expected wealth are desirable but there is a penalty for an *ad hoc* measure of portfolio volatility. Crude as this may seem in the light of economic theory, ρ_i is often the best that practitioners can establish about the client risk profile and mean-variance specifications are widespread in actual investment houses.

[30] In a dynamic context, we might interpret it as a situation in which the stock price forecast in the previous rebalancing period was such that it has been optimal to hold bonds only.

[31] No short-selling constraints are in place.

Thus, although individual contracts obligate the manager to maximize the welfare of each client conditional on current market prices by submitting a market order for the optimal quantity correctly derived from the client's risk profile, market impact introduces an additional dimension for optimization by the manager since she co-determines the market prices and the order in which she trades the clients' accounts matters. In this case, ordering of clients does not affect the aggregate quantity traded by the asset management company but it does affect aggregate welfare as measured by the simple summation of each client's welfare index (17). In particular, the manager will maximize aggregate welfare if she grants priority in market access to the client with less aversion to portfolio volatility. Effectively, as is apparent from (20), preferential market access is thus given to the client who wishes to trade more if allowed in the market first. The initial client wealth W_i is irrelevant for the results.

In sum, client ordering has no effect on the size of the alpha pie produced by the manager as measured by (16) but can well serve as an instrument for its slicing and distribution to the clients in a welfare-maximizing fashion. Before the application of this tool, however, the manager needs to establish what the relevant objective functions of the clients are and address the issue of their aggregation since that might affect the optimality of her actions.

For instance, let the clients have CARA preferences and $\widetilde{P}_T$ be normally distributed. Then, by the standard properties of the lognormal distribution,[32] the expected utility of client i is

$$\mathrm{E}\left[u_i\left(\widetilde{W}_{Ti}\right)\right] = \mathrm{E}\left[-\frac{1}{\rho_i}\cdot \mathrm{e}^{-\rho_i\cdot\widetilde{W}_{Ti}}\right] = -\frac{1}{\rho_i}\cdot \mathrm{e}^{-\rho_i\cdot V_i(Q_i)} = u_i\left[V_i(Q_i)\right] \tag{23}$$

with V_i given by (17). Since $u_i' > 0$, the optimal demand for equity is (20) and so (21) still holds. However, aggregate welfare defined as

$$\mathrm{E}\left[u_1\left(\widetilde{W}_{T1}\right)\right] + w\cdot \mathrm{E}\left[u_2\left(\widetilde{W}_{T2}\right)\right], \tag{24}$$

where $w > 0$ is a parameter capturing the clients' relative importance to the manager,[33] does not necessarily lead to the same optimal ordering as implied by (22).

In the absence of an analytical solution, numerical results are presented graphically in Figure 1. They show that clients' initial wealth W_i matters and as client 2 becomes wealthier, *ceteris paribus*, the region in the space of coefficients of absolute risk aversion in which aggregate welfare is maximized by trading that client first shrinks, suggesting that a benevolent asset manager would tend to distribute her alpha preferentially to less wealthy clients. Also, it shows that the weight w placed on client 2 would have to rise exponentially as W_2 increases in order to offset this effect.[34]

We have demonstrated that there are welfare consequences of the distribution of the manager's alpha among the clients through discriminatory market access. The construction of the aggregate welfare index by an asset manager seeking in-house welfare improvement should be based on careful inferences about clients' individual preferences and possibly the manager's broader objectives in terms of their relative importance. Clearly, if the manager pursues the objective of maximizing the aggregate certainty equivalent of true CARA clients as effectively done in (22), the action she takes might be suboptimal. On the other hand, it

[32] $\mathrm{E}(\mathrm{e}^z) = \mathrm{e}^{\mathrm{E}(z)+\frac{1}{2}\cdot\mathrm{var}(z)}$ for a normal z.

[33] Any positive affine transformation of the utility function u_i represents the same preferences and so w accounts for that as well.

[34] By relevant substitutions and rearrangement of (24), it can be shown easily that any deviation of W_2 from W_1, *ceteris paribus*, is neutralized if w is adjusted by a factor of $\mathrm{e}^{\rho_2\cdot(1+r_f)\cdot(W_2-W_1)}$.

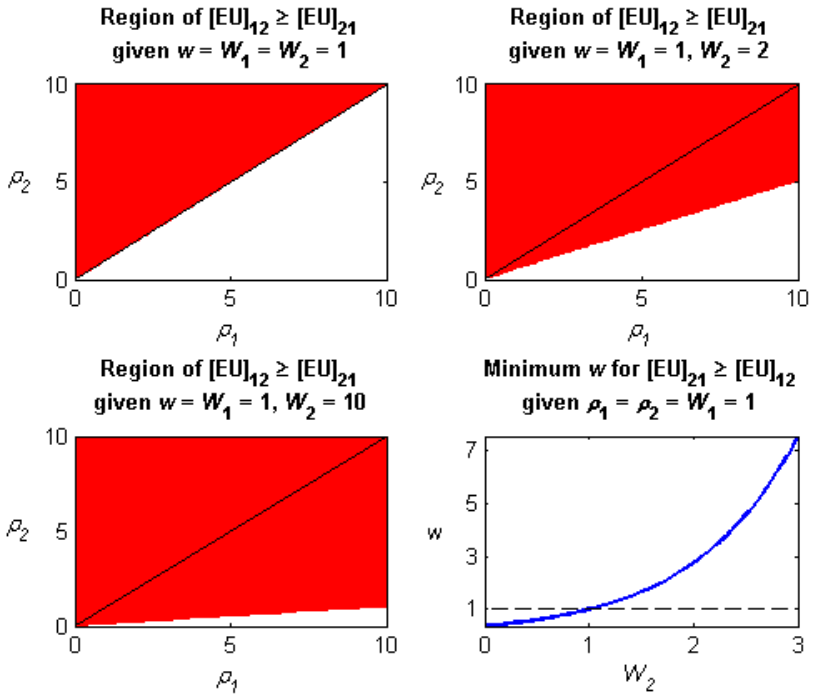

Figure 1: Numerical results for CARA preferences given $\delta = 10$, $r_f = 0.01$, $\overline{p} = 1$, $\mathrm{E}(\widetilde{P}_T) = 1.1$ and $\mathrm{var}(\widetilde{P}_T) = 0.15^2$. $\mathrm{EU} = \mathrm{E}(u_1) + w \cdot \mathrm{E}(u_2)$.

might be the preferred action if the risk-adjusted expected wealth in (17) approximates true welfare over investments better than (23).

6 Client Ordering and the Size of the Alpha Pie

Let us make a slight adjustment to the specification of clients' investment preferences. We remain in the mean-variance framework as it is widely utilized by practitioners and specify the welfare index of client i as[35]

$$V_i(\varphi_i) = W_i \cdot \left[1 + \mathrm{E}\left(\widetilde{r}_{pi}\right) - \frac{\rho_i}{2} \cdot \mathrm{var}\left(\widetilde{r}_{pi}\right)\right] \tag{25}$$

where

$$\widetilde{r}_{pi} = r_f + \varphi_i \cdot (\widetilde{r}_i - r_f) \tag{26}$$

is the portfolio rate of return, φ_i is a fraction of the wealth W_i allocated to stocks and

$$\widetilde{r}_i = \frac{\widetilde{P}_T}{\overline{P}_i} - 1 \tag{27}$$

is the rate of return on stocks with $\overline{P}_i$ being the average price in the trade on behalf of client i. Thus, the client has mean-variance preferences over the portfolio rate of return rather than final wealth as in (17). As a result, (25) exhibits the empirically appealing 'CRRA feature'

[35] Cf. (44) in the Appendix.

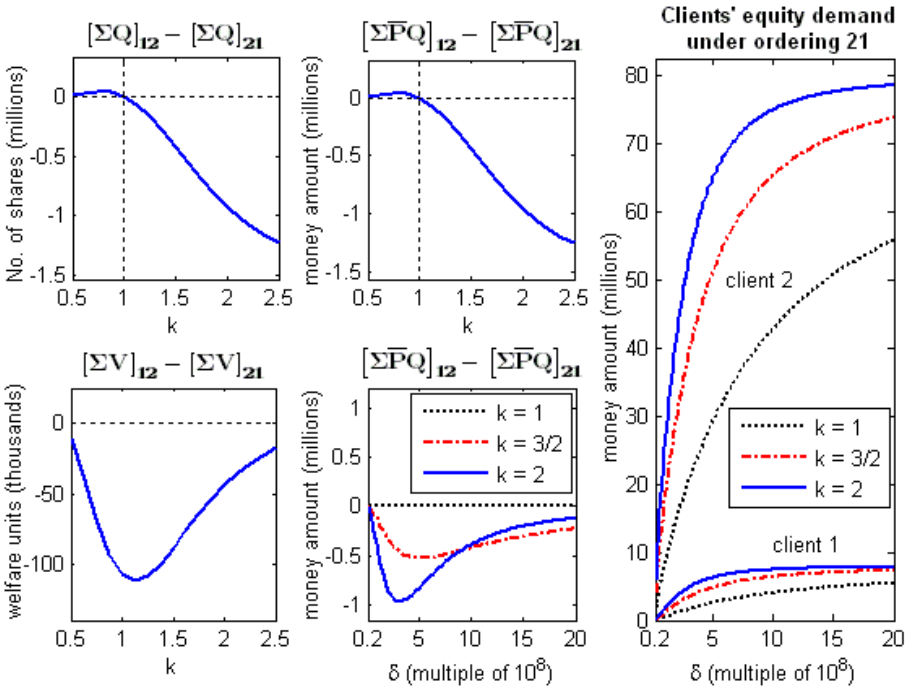

Figure 2: Numerical results for mean-variance preferences over the portfolio rate of return given $r_f = 0.01$, $\overline{p} = 1$, $\mathrm{E}(\widetilde{P}_T) = 1.1$, $\mathrm{var}(\widetilde{P}_T) = 0.15^2$, $W_1 = 10 \cdot 10^6$, $W_2 = 100 \cdot 10^6$, $\rho_1 = \rho_2 = 5$ and when fixed, $\delta = 3 \cdot 10^8$. $\Sigma Q = Q_1^* + Q_2^*$, $\Sigma V = V_1^* + V_2^*$ and $\Sigma \overline{P} Q = \overline{P}_1 \cdot Q_1^* + \overline{P}_2 \cdot Q_2^*$.

that for a fixed distribution of stock returns, i.e. with no market impact, the proportion of the portfolio allocated to stocks is independent of wealth.[36] Again, ρ_i is often the best that a practitioner is realistically able to establish about the client's risk profile with variance being merely a widely used *ad hoc* measure of portfolio volatility as a proxy for risk.

Let us also generalize the manager's market impact function. For now, let (5) become

$$K(p) = \delta \cdot |p - \overline{p}|^{\frac{1}{k}} \tag{28}$$

for $k > 0$ so that (6) becomes

$$p(q) = \overline{p} + (\operatorname{sgn} q) \cdot \left| \frac{q}{\delta} \right|^k, \tag{29}$$

allowing for nonlinear (power) price impact.

With two clients, the manager can sequentially trade their accounts as in the previous section and Figure 2 illustrates the outcomes for specific parameter values. It shows that with a strictly convex price impact ($k > 1$), given that the coefficient of aversion to portfolio volatility ρ_i is the same for both clients, the optimal quantity of stocks (and hence their monetary value) traded on behalf of the clients is larger if the larger client is traded first. That way, aggregate client welfare measured as a simple sum of the individual welfare indices $V_i^* = V_i(\varphi_i^*)$ is maximized as well. Note that for $k < 1$, although such a strictly concave price

[36] Whereas in (17), it is the *amount* invested in stocks that is independent of wealth.

	Ordering 12	Ordering 21
$p(Q_1^* + Q_2^*)$	1.03523	1.03640
$\overline{P}_1$	1.00023	1.03347
$\overline{P}_2$	1.01363	1.01021
$\overline{P}_1 \cdot Q_1^*$	$7.94 \cdot 10^6$	$4.89 \cdot 10^6$
$\overline{P}_2 \cdot Q_2^*$	$49.03 \cdot 10^6$	$53.04 \cdot 10^6$
$\sum_{i=1}^{2} \overline{P}_i \cdot Q_i^*$	$56.97 \cdot 10^6$	$57.93 \cdot 10^6$

Table 1: A numerical example under a quadratic price impact ($k = 2$) given $\delta = 3 \cdot 10^8$, $r_f = 0.01$, $\overline{p} = 1$, $\mathrm{E}(\widetilde{P}_T) = 1.1$, $\mathrm{var}(\widetilde{P}_T) = 0.15^2$, $W_1 = 10 \cdot 10^6$, $W_2 = 100 \cdot 10^6$ and $\rho_1 = \rho_2 = 5$.

impact is presumably less realistic, ordering 21 leads to larger aggregate welfare but smaller aggregate stock holdings than ordering 12 and it is only for $k = 1$ that either ordering results in the same size of the equity portfolio under the management of the investment company. The figure also illustrates how changes in the market depth parameter δ affect the outcomes. Intuitively, the clients wish to buy more stocks in deeper markets with smaller losses due to price impact.

As a particular numerical example, Table 1 provides the outcomes for each client ordering if the price impact is quadratic ($k = 2$), namely the price $p(Q_1^*+Q_2^*)$ of the last share purchased by the manager in the market, the average equity price $\overline{P}_i$ for each client and the monetary value $\overline{P}_i \cdot Q_i^*$ of the equity block purchased on behalf of each client. Given that the wealth of client 1 and client 2 is 10 million and 100 million, respectively, the results show that by changing the order in which clients' accounts are traded, the manager changes the total equity allocation on behalf of her clients by about a million in the portfolio of 110 million under management while she still honours her obligation to invest strictly according to each client's risk profile given the market conditions.

Relaxing the restriction on the functional form of $p(q)$ and allowing for more than two clients, we can state the following general result.[37]

PROPOSITION 2: *For a continuous nondecreasing $p(q)$ such that $p(0) = \overline{p}$ and $\mathrm{E}\left(\widetilde{P}_T\right) \neq (1 + r_f) \cdot \overline{p}$, let there be N clients with mean-variance investment preferences and zero initial stock holdings and let M be the set of all permutations of the client ordering $\{\, i \,\}_{i=1}^{N}$. Then*

$$\forall \rho_1, \ldots, \rho_N, W_1, \ldots, W_N > 0 \, \forall \boldsymbol{m}, \boldsymbol{m}' \in M : \left[\sum_{n=1}^{N} Q_n^*\right]_{\boldsymbol{m}} = \left[\sum_{n=1}^{N} Q_n^*\right]_{\boldsymbol{m}'} \tag{30}$$

if and only if the price impact $p(q) - \overline{p}$ is linear in its economically relevant domain given by the range of $\sum_{n=1}^{N} Q_n^*$.

PROOF: In the Appendix.

Due to welfare aggregation problems as illustrated in the previous section, it may be difficult for the manager to correctly determine the true welfare-maximizing order in which

[37] While price impact might alternatively be specified in terms of the average price, note that by (6) and (7), if the impact of trade on p is linear, it is so for $\overline{P}$ as well.

clients' accounts should be sequentially traded. Even if it were unambiguously possible, however, the manager's incentives may not be aligned with aggregate welfare maximization. While she is in the position of a social planner, assuming that the effects under consideration are small enough so that no client defects to a rival investment house, she may not necessarily act in a benevolent fashion. In particular, having estimated a nonlinear price impact of her trades, she is aware that client ordering will generally result in a different structure of the total portfolio under her management and, therefore, she may pursue her own agenda subject to the contractual constraint of optimizing individual portfolios conditional on current market prices.

In some situations, the manager may wish to simply maximize the amount of risky assets under her management without much strategic thought. In general, however, she will pursue a longer-term strategy that takes into account the impact of her performance on her future client base and fees. Most importantly, the structure of performance fees tends to be such that it is personally more profitable for the manager to achieve above average returns with a smaller total portfolio than average returns with a larger one.[38] In our stylized setting with a single risky asset, the performance fee for the manager might be a pre-specified fraction of the excess profit (alpha pie) generated by the investment house, defined as in (16) but now generalized for a nonlinear price impact $p(q) - \overline{p}$. Changes in client ordering result not only in a different slicing and distribution of the alpha pie among the clients but generally also in its different size. Thus, the manager will select the client ordering $\boldsymbol{m}$ that results in $\sum_{n=1}^{N} Q_n^*$ generating dollar alpha closest to that generated with $\widehat{Q}$, which may or may not coincide with maximization of the trading volume and/or client welfare.[39]

In sum, the contractual obligation to invest strictly according to clients' risk profiles tends to leave some, however narrow, room for maneuvering on the part of a large asset management house with market impact and sophisticated managers can be expected to take advantage of that in an optimal fashion, generally irrespective of client welfare.

7 Alternative Trading Arrangements

The scope of trading arrangements available to the manager is broader than trading one client after another. In theory, she could split each client's order into infinitesimal clips and trade those sequentially across clients in such a way that at the end of the day, their portfolios are optimized. That would enable her to vary the size of the alpha pie continuously.

More generally, with nonzero initial holdings, sequential trading might become rather inefficient when clients' portfolios are rebalanced. Firstly, some clients may wish to sell while others buy and so by Proposition 1, the manager could reduce the aggregate market impact costs through internalization, i.e. simultaneous in-house trading at the prevailing market price, taking only the excess demand/supply out to the market. Secondly, there may be situations in which a client demands extra equity on top of her current holding at the price she faces if she is placed at the front of the queue by the manager, while the same client would be willing to sell some of her current holding at the price she faces if she is placed at the back of the queue.[40] In the latter case, she would generally be willing to do so at a slightly

[38] See Perold and Salomon (1991).

[39] Implicitly, the manager is neutral to the risk of her own alpha forecasts. That, however, is not central to the conclusions.

[40] With zero initial stock holdings, no shuffling of clients can result in a switch between buying and short-

lower price as well and since the previous client has been purchasing stocks at such a price, it might be efficient to additionally internalize some of their trading. Thus, even after the initial internalization at $\overline{p}$, purely sequential trading of the residual clients might still be an inefficient trading arrangement.

Order splitting and internalization add further flexibility to the manager's conduct, which is nearly impossible to monitor for unsophisticated clients. However, to attract such clients, some managers might advertise their commitment to provide equal market access to all their clients, the credibility of which would come from the existence of supervisory authorities that would fine and make public any deliberate client misinformation. Alternatively, regulatory authorities might define equal market access in unambiguous terms and enact it as a default standard with appropriate enforcement, unless stated explicitly otherwise in the prospectus of the asset management company.

Fairness in the sense of pro-rata sharing of price impact costs can always be achieved in this framework. In particular, let us assume that the manager has a self-imposed and/or regulatory obligation to make a notional market in which all her clients obtain equal treatment in terms of market access and, at the same time, she must exert her best effort to accomplish the highest overall execution quality, i.e. minimum loss due to market impact. Let $P = \overline{P}_i$, $i \in \{1, \ldots, N\}$, be a fixed price per share at which the manager offers to trade on behalf of all her clients. Client i initially holds Q_{0i} shares and riskfree bonds worth B_i, thus valuing his portfolio at[41]

$$W_i = B_i + \overline{p} \cdot Q_{0i} > 0. \tag{31}$$

The portfolio structure is the result of decisions made in the previous investment/rebalancing periods and subsequent realization of uncertainty. In general, it will be optimal for the client in the current period if the manager rebalances his portfolio by selling some shares to buy more bonds or by selling some bonds to buy additional shares. Q_i denotes demand for shares net of the initial position Q_{0i} and so wealth of client i at the end of the current rebalancing period is[42]

$$\widetilde{W}_{Ti} = (B_i - P \cdot Q_i) \cdot (1 + r_f) + \widetilde{P}_T \cdot (Q_{0i} + Q_i). \tag{32}$$

Given the above, the optimal rebalancing of the portfolio is determined by maximizing (44), or equivalently (25), which results in the net aggregate order

$$\sum_{i=1}^{N} Q_i^*(P) = A + B \cdot P \tag{33}$$

where

$$\begin{aligned} A &= \frac{\mathrm{E}\left(\widetilde{P}_T\right)}{\mathrm{var}\left(\widetilde{P}_T\right)} \cdot \sum_{i=1}^{N} \frac{W_i}{\rho_i} - \sum_{i=1}^{N} Q_{0i}, \\ B &= -\frac{1 + r_f}{\mathrm{var}\left(\widetilde{P}_T\right)} \cdot \sum_{i=1}^{N} \frac{W_i}{\rho_i} < 0. \end{aligned}$$

selling of any client as shown in the Appendix.

[41] Shares in the initial portfolio are valued at $\overline{p}$ rather than P. The difference between those two prices is merely due to a temporary impact of the manager's current trading activity in the market and is considered to be a transaction cost.

[42] If $P \cdot Q_i > B_i$, the rebalanced position involves bonds sold short, which is equivalent to borrowing and so shares are traded on margin. If $Q_{0i} = 0$, then $W_i = B_i$ and $\widetilde{W}_{Ti}$ is specified as in the previous sections.

The following result then holds.

PROPOSITION 3: *Let $p(q)$ be continuously nondecreasing such that $p(0) = \overline{p}$ and let there be a rule that internalization be conducted at the current market price or at the same terms as the manager's overall residual trading in the market. Then there exists a unique fixed share price $P^* > 0$ at which the asset manager can optimally rebalance the portfolios of all clients with mean-variance investment preferences while keeping market impact at a minimum.*

PROOF: In the Appendix.

This means that the manager is always able to set a price P^* at which internal crossing of complementary orders takes place such that the residual order $\sum_{i=1}^{N} Q_i^*(P^*)$ can be traded in the market at exactly that price per share, given the price impact. Such a single-price rule eliminates discretion from the manager and reduces clients' uncertainty about the management of their assets. Should there be pure off-market internalization, it must take place at the market price $\overline{p}$.[43]

In practice, the single-price rule might stipulate that trading at different prices can only take place after a minimum time interval has elapsed between the completion of trading on behalf of one client and the start of trading on behalf of another client in order to make sure that the manager is not strategically riding on the 'price wave' she made herself. In sum, a possible arrangement for prevention of unconsented client discrimination can be made, although its practical implementation might involve some challenges in terms of compliance and enforcement.

8 Conclusion

We have pointed out an extra dimension for optimization by an asset manager with market impact and multiple clients. While our model is stylized, it clearly demonstrates that formal analysis of the manager's problem, using estimated market impact, is likely to yield practical benefits. This is so for asset managers seeking optimal strategies subject to existing regulatory rules as well as for regulators and supervisors who create and enforce those rules.

We have built a framework in which a single asset manager is in charge of a given number of client portfolios and we have shown how her incentives may be diverted from equitable distribution of market impact costs and/or maximization of the overall client welfare if she pursues the most favourable fee income prospects by maximizing her dollar alpha. Of course, while sophisticated managers might take advantage of such opportunities, their unsophisticated counterparts might simply ignore them and at the same time fail to actively ensure fairness. Given the structure of the asset management industry, it might be equally plausible if such subtle considerations were in many cases overshadowed by asset managers' competition for large clients due to economies of scale and lower relative costs of defection of those clients to a rival asset manager. Then, there would be incentives to simply grant priority in market access to large clients at the expense of their small counterparts.

Anecdotal and off-the-record evidence suggests that provision of preferential market access to selected clients has indeed been quietly practiced by a number of market participants,

[43] Technically, whenever $A > 0$, the manager could set $P = -\frac{A}{B} > 0$ which by (33) would lead to a single-price equilibrium with no trading in the market. This rule eliminates such a possibility unless $\overline{p} = -\frac{A}{B}$. If the manager trades in the market, the single price is anchored beyond her control, thus eliminating any pricing discretion on her part as long as she honours her obligation to optimize clients' portfolios according to their risk profiles.

regulatory rules regarding fair client treatment notwithstanding. The contribution of this paper is in explicit analysis of the problem and in demonstrating that such practices might in some situations be motivated by little more than the managers' pursuit of an optimal structure of their total portfolio with neither themselves nor any third party directly profiting from the client discrimination. In any case, irrespective of the motives for such discriminatory actions, we have illustrated that without adequate monitoring, there is little reason to expect provision of equal market access to all clients or maximization of their total welfare, if the latter could be unambiguously specified, other than by coincidence.

It is common for commercial banks to offer higher interest rates and preferential treatment to clients with larger deposits and, similarly, it is common for asset managers to provide more favourable terms to clients with larger investment portfolios, not necessarily excluding preferential market access. Nevertheless, since preferential treatment of one group of clients, legitimate as it may be, is equivalent to discrimination of the remaining (presumably uninformed) clients, transparency about any systematic in-house distribution of market access is in principle socially desirable.[44] Unfortunately, practical enforcement of fairness/transparency rules may be quite challenging. Potentially, trading records of randomly selected asset managers could be analyzed by the supervisory authorities and properly designed computer algorithms could detect any such tacit practice over an extended period of time, which should reduce the frequency of failures to disclose such arrangements to all clients. In fact, mere public awareness of this issue might prompt competitive forces towards that.

A Appendix

A.1 Logarithmic Preferences and Portfolio Choice

The first-order condition for the maximization of (1) is

$$\frac{\mathrm{dE}\left(\log \widetilde{W}_T\right)}{\mathrm{d}\varphi} = \Phi\left(\varphi^*, \overline{P}\right) = 0 \tag{34}$$

where

$$\Phi(\varphi, P) = \mathrm{E}\left\{\left(\frac{\widetilde{P}_T}{P} - 1 - r_f\right) \cdot \left[1 + r_f + \varphi \cdot \left(\frac{\widetilde{P}_T}{P} - 1 - r_f\right)\right]^{-1}\right\} \tag{35}$$

is a continuously differentiable function in the neighbourhood of $\left[\varphi^*, \overline{P}\right]$ with $1 + r_f > 0$, $P > 0$ and stochastic (nondegenerate) $\widetilde{P}_T > 0$, which implies

$$\frac{\partial \Phi}{\partial \varphi} = -\mathrm{E}\left\{\left(\frac{\widetilde{P}_T}{P} - 1 - r_f\right)^2 \cdot \left[1 + r_f + \varphi \cdot \left(\frac{\widetilde{P}_T}{P} - 1 - r_f\right)\right]^{-2}\right\} < 0 \tag{36}$$

[44] For instance, see Treating Customers Fairly at www.fsa.gov.uk:
Q: "Is it unfair that some customers have to pay more for certain financial products or services?"
A: "We view a firm's decision to provide products (or not) and at what price (whether over the phone, internet or by email) as a commercial decision by that firm. Furthermore, to treat customers fairly does not mean that a firm is required to offer the same products or levels of service to all customers, as long as it delivers the product or level of service promised, and that customers are protected from unpleasant surprises from the products they buy."

and

$$\frac{\partial \Phi}{\partial P} = -\frac{1+r_f}{P^2} \cdot \mathrm{E}\left\{ \widetilde{P}_T \cdot \left[1 + r_f + \varphi \cdot \left(\frac{\widetilde{P}_T}{P} - 1 - r_f \right) \right]^{-2} \right\} < 0. \tag{37}$$

Hence, by the Implicit Function Theorem, there exists a continuously differentiable function $\varphi = \varphi(P)$ in the neighbourhood of $\overline{P} > 0$ such that $\Phi[\varphi(P), P] \equiv 0$, $\varphi\left(\overline{P}\right) = \varphi^*$ and

$$\frac{\mathrm{d}\varphi\left(\overline{P}\right)}{\mathrm{d}P} = -\frac{\frac{\partial \Phi\left(\varphi^*, \overline{P}\right)}{\partial P}}{\frac{\partial \Phi\left(\varphi^*, \overline{P}\right)}{\partial \varphi}} < 0. \tag{38}$$

A.2 Proof of Proposition 1

Proof of Proposition 1: Without loss of generality, let $Q_1 > -Q_2 > 0$ where Q_1 and Q_2 are sums of all buy and sell orders, respectively. Given

$$C(Q) = \int_0^Q p(q)\,\mathrm{d}q - \overline{p} \cdot Q, \tag{39}$$

the aggregate price impact costs are

$$L(N) = C(Q_1 - N) + C(Q_2 + N) \tag{40}$$

where the first and second term are the costs of price impact on the ask and bid side of the order book, respectively. By the properties of p,

$$L'(N) = p(Q_2 + N) - p(Q_1 - N) \leq 0 \tag{41}$$

for $N \in [0, -Q_2]$ and so in that interval, $L(N)$ is minimized at $N = -Q_2$. □

Note that the minimum is not necessarily unique. A sufficient condition for uniqueness would, for instance, be strict monotonicity of p which would turn the weak inequality in (41) into a strict one.

A.3 Aggregate Values under Mean-Variance Preferences (17)

Using (20), we obtain

$$[Q_1^* + Q_2^*]_{ij} = \frac{\left[\mathrm{E}\left(\widetilde{P}_T\right) - (1+r_f)\cdot\overline{p}\right] \cdot \left[(1+r_f)\cdot\delta^{-1} + (\rho_i + \rho_j)\cdot\mathrm{var}\left(\widetilde{P}_T\right)\right]}{\left[(1+r_f)\cdot\delta^{-1} + \rho_i\cdot\mathrm{var}\left(\widetilde{P}_T\right)\right] \cdot \left[(1+r_f)\cdot\delta^{-1} + \rho_j\cdot\mathrm{var}\left(\widetilde{P}_T\right)\right]} \tag{42}$$

and (21) follows immediately. Plugging (20) into (18) and substituting the resulting expression into (17) leads to

$$[V_1^* + V_2^*]_{12} - [V_1^* + V_2^*]_{21} = \frac{(\rho_2 - \rho_1)\cdot\mathrm{var}\left(\widetilde{P}_T\right)\cdot(1+r_f)^2\cdot\left[\mathrm{E}\left(\widetilde{P}_T\right) - (1+r_f)\cdot\overline{p}\right]^2}{2\cdot\delta^2\cdot\left[(1+r_f)\cdot\delta^{-1} + \rho_1\cdot\mathrm{var}\left(\widetilde{P}_T\right)\right]^2\cdot\left[(1+r_f)\cdot\delta^{-1} + \rho_2\cdot\mathrm{var}\left(\widetilde{P}_T\right)\right]^2} \tag{43}$$

and hence (22).

A.4 Q_i as Control under Mean-Variance Preferences (25)

Given $\widetilde{W}_{Ti} = W_i \cdot (1 + \widetilde{r}_{pi})$, (25) can be rewritten as[45]

$$V_i(Q_i) = \mathrm{E}\left(\widetilde{W}_{Ti}\right) - \frac{\rho_i}{2} \cdot \frac{\mathrm{var}\left(\widetilde{W}_{Ti}\right)}{W_i} \tag{44}$$

where

$$\widetilde{W}_{Ti} = \left(W_i - \overline{P}_i \cdot Q_i\right) \cdot (1 + r_f) + \widetilde{P}_T \cdot Q_i \tag{45}$$

and if $i = m_n$ in the client ordering $\boldsymbol{m} = [m_1, \ldots, m_N] \in M$, then[46]

$$\overline{P}_i = \frac{1}{Q_i} \cdot \int_{\sum_{j=1}^{n-1} Q_{m_j}}^{\sum_{j=1}^{n-1} Q_{m_j} + Q_i} p(q)\,\mathrm{d}q. \tag{46}$$

By the Fundamental Theorem of Calculus,

$$\frac{\mathrm{d}\left(\overline{P}_i \cdot Q_i\right)}{\mathrm{d}Q_i} = p\left(\sum_{j=1}^{n-1} Q_{m_j} + Q_i\right) \tag{47}$$

and hence the first-order condition for a maximum of (44) gives[47]

$$p\left(\sum_{j=1}^{n-1} Q_{m_j} + Q_i^*\right) = \frac{\mathrm{E}\left(\widetilde{P}_T\right)}{1 + r_f} - \frac{\rho_i \cdot \mathrm{var}\left(\widetilde{P}_T\right)}{W_i \cdot (1 + r_f)} \cdot Q_i^*. \tag{48}$$

A.5 Proof of Proposition 2

Given $1 + r_f$, $\overline{p}$, $\mathrm{E}\left(\widetilde{P}_T\right)$, $\mathrm{var}\left(\widetilde{P}_T\right) > 0$, let

$$K = \left\{ q \in \mathbb{R} : \operatorname{sgn} q = \operatorname{sgn}\left[\frac{\mathrm{E}\left(\widetilde{P}_T\right)}{1 + r_f} - p(q)\right] = \operatorname{sgn}\left[\frac{\mathrm{E}\left(\widetilde{P}_T\right)}{1 + r_f} - \overline{p}\right] \right\}. \tag{49}$$

LEMMA 1: *Let $N = 2$ and, without loss of generality, $\boldsymbol{m} = [m_1, m_2] = [1, 2]$, i.e. $I(1) = 1 - I(2) = 0$.*

i) $\forall \rho_1, \rho_2, W_1, W_2 > 0$ $\exists! Q_1^*$ $\exists! Q_2^*$ and Q_1^*, Q_2^*, $Q_1^* + Q_2^* \in K$.

ii) $\forall q, Q \in K$, $|Q| > |q| > 0$, $\exists! \frac{\rho_1}{W_1} > 0 : Q_1^* = q$, $\exists! \frac{\rho_2}{W_2} > 0 : Q_1^* + Q_2^* = Q$ and varying q from 0 to Q implicitly defines $\frac{\rho_2}{W_2}$ as a strictly decreasing continuous function of $\frac{\rho_1}{W_1}$ with the domain and range $(l_Q, +\infty)$ such that $\frac{\rho_1}{W_1} = l_Q$ results in $Q_1^* = Q$.

[45] Note that (25) is thus obtained from (17) by replacing ρ_i with $\frac{\rho_i}{W_i}$ and, therefore, such a replacement of parameters in the relevant results obtained in Section 5 under the latter specification immediately leads to the results that would be obtained under the former.

[46] Assuming $Q_i \neq 0$. For $n = 1$, $\sum_{j=1}^{n-1} Q_{m_j}$ is replaced with 0. Note that $\varphi_i = \frac{\overline{P}_i \cdot Q_i}{W_i}$.

[47] The second-order condition is satisfied given a continuously nondecreasing $p(q)$.

Proof: Let us rewrite (48) as

$$Q_i^* = \frac{W_i \cdot (1+r_f)}{\rho_i \cdot \mathrm{var}\left(\widetilde{P}_T\right)} \cdot \left[\frac{\mathrm{E}\left(\widetilde{P}_T\right)}{1+r_f} - p\left[I(i) \cdot Q_j + Q_i^*\right]\right]. \tag{50}$$

i) The existence of a unique optimal portfolio for client 1 such that $Q_1^* \in K$ springs immediately from (50) for $[i,j] = [1,2]$, the property of $p(q)$ being continuously nondecreasing and the Intermediate Value Theorem. Similarly, the latter two combined with (50) for $[j,i] = [1,2]$, given $Q_1^* \in K$, result in a unique optimal portfolio for client 2 such that $Q_2^* \in K$ and $Q_1^* + Q_2^* \in K$.

ii) (50) for $[i,j] = [1,2]$ and the properties of K immediately imply a unique $\frac{\rho_1}{W_1} > 0$ for a given $q \in K$, $q \neq 0$, such that $\lim_{q \to 0} \frac{\rho_1}{W_1} = +\infty$, $\frac{\rho_1}{W_1}$ strictly decreases with $|q|$ and its infimum in the relevant interval is l_Q. Given $Q_2^* = Q - q$, the properties of K combined with (50) for $[j,i] = [1,2]$ result in a unique $\frac{\rho_2}{W_2} > 0$. Given Q, $|Q_2^*|$ strictly decreases in $|q|$ and it is also straightforward to see from (50) that $\frac{\rho_2}{W_2}$ strictly decreases in $|Q_2^*|$. Thus, $\frac{\rho_2}{W_2}$ is implicitly defined as a continuous function of $\frac{\rho_1}{W_1}$ strictly decreasing in its domain $(l_Q, +\infty)$ which equals its range.

□

Lemma 2: *For* $\mathrm{E}\left(\widetilde{P}_T\right) \neq (1+r_f) \cdot \overline{p}$, *if*

$$\forall \rho_1, \rho_2, W_1, W_2 > 0 : [Q_1^* + Q_2^*]_{12} = [Q_1^* + Q_2^*]_{21}, \tag{51}$$

then the price impact $p(q) - \overline{p}$ *is linear in its economically relevant domain given by the range of* $Q_1^* + Q_2^*$.

Proof: Assume that (51) holds and $p(q) - \overline{p}$ is not linear in $K \neq \{0\}$, which is the range of $Q_1^* + Q_2^*$ by Lemma 1. Then (50) for $[j,i] = [1,2]$, which is a notation implying $I(1) = 1 - I(2) = 0$, and (50) for $[j,i] = [2,1]$ result, together with (51), in

$$\frac{[Q_1^* + Q_2^*]_{21} - [Q_1^*]_{12}}{[Q_1^* + Q_2^*]_{12} - [Q_2^*]_{21}} = \frac{\rho_1 \cdot W_2}{\rho_2 \cdot W_1}. \tag{52}$$

Similarly, the combination of (50) for $[i,j] = [1,2]$ and for $[j,i] = [2,1]$ gives

$$p([Q_1^* + Q_2^*]_{21}) - p([Q_1^*]_{12}) = \frac{\rho_1 \cdot \mathrm{var}\left(\widetilde{P}_T\right)}{W_1 \cdot (1+r_f)} \cdot ([Q_1^*]_{12} - [Q_1^*]_{21}) \tag{53}$$

and the combination of (50) for $[j,i] = [1,2]$ and for $[i,j] = [2,1]$ gives

$$p([Q_1^* + Q_2^*]_{12}) - p([Q_2^*]_{21}) = \frac{\rho_2 \cdot \mathrm{var}\left(\widetilde{P}_T\right)}{W_2 \cdot (1+r_f)} \cdot ([Q_2^*]_{21} - [Q_2^*]_{12}). \tag{54}$$

By (51), $[Q_1^*]_{12} - [Q_1^*]_{21} = [Q_2^*]_{21} - [Q_2^*]_{12}$ and so (52), (53) and (54) yield

$$\frac{p([Q_1^* + Q_2^*]_{21}) - p([Q_1^*]_{12})}{[Q_1^* + Q_2^*]_{21} - [Q_1^*]_{12}} = \frac{p([Q_1^* + Q_2^*]_{12}) - p([Q_2^*]_{21})}{[Q_1^* + Q_2^*]_{12} - [Q_2^*]_{21}}. \tag{55}$$

It follows from Lemma 1 that for all $Q \in K \neq \{0\}$ such that

$$[Q_1^* + Q_2^*]_{12} = [Q_1^* + Q_2^*]_{21} = Q, \tag{56}$$

varying $\frac{\rho_1}{W_1}$ continuously from l_Q to $+\infty$ implies $\frac{\rho_2}{W_2}$ being continuously varied over the same interval in the opposite direction and hence $[Q_1^*]_{12}$ varies continuously from Q to 0 as $[Q_2^*]_{21}$ varies continuously from 0 to Q. But then (55) immediately implies that $p(q)$ has a constant slope in $K \neq \{0\}$, i.e. it is affine, and so $p(q) - \overline{p}$ is linear, which is a contradiction. □

With the above auxiliary results, we can proceed to the main proof.

Proof of Proposition 2: Using (48), a trivial extension of the relevant parts of the proof of Lemma 1 reveals that K is the range of $\sum_{n=1}^{N} Q_n^*$ for $N \in \mathbb{N}$.[48] Lemma 2 proves necessity for the special case of $N = 2$. Let us assume more generally that for some $N \geq 2$, if (30) holds, then $p(q)$ is affine in K. Now we demonstrate that such an implication must hold for $N+1$ as well. In particular, if the premise of the implication holds for all $\boldsymbol{m}, \boldsymbol{m}' \in M$ and for all $\frac{\rho_1}{W_1}, \ldots, \frac{\rho_{N+1}}{W_{N+1}} > 0$, then it holds in a subset of the set of permutations in which the newly added client $N+1$ is traded first and at the same time in a subset of the parameter space in which $\frac{\rho_{N+1}}{W_{N+1}}$ is held fixed so that $Q_{N+1}^* = q_1$. Since the implication holds for N, it follows immediately that $p(q)$ is affine in K_1 defined by replacing $\overline{p}$ with $p(q_1)$ in the definition of K in (49). Letting $\frac{\rho_{N+1}}{W_{N+1}} \to +\infty$, we obtain $q_1 \to 0$, $p(q_1) \to \overline{p}$ and hence $K_1 \to K$. By induction, this proves necessity for $N \in \mathbb{N}\backslash\{1\}$.

To prove sufficiency, let $p(q)$ be affine such that $p(q) = \overline{p} + \beta \cdot q$, $\beta \geq 0$. For $\beta = 0$, individual demands are independent of $\boldsymbol{m}$ and sufficiency follows immediately. For $\beta > 0$,[49] given $\boldsymbol{m} = [m_1, \ldots, m_N] \in M$, (48) leads to the recursive formula

$$\sum_{i=1}^{n} Q_{m_i}^* = \frac{\mathrm{E}\left(\widetilde{P}_T\right) - (1+r_f)\cdot\overline{p} + \frac{\rho_{m_n}}{W_{m_n}} \cdot \mathrm{var}\left(\widetilde{P}_T\right) \cdot \sum_{i=1}^{n-1} Q_{m_i}^*}{(1+r_f)\cdot\beta + \frac{\rho_{m_n}}{W_{m_n}} \cdot \mathrm{var}\left(\widetilde{P}_T\right)}. \tag{57}$$

By repeated substitution, we obtain[50]

$$\begin{aligned}\sum_{i=1}^{N} Q_{m_i}^* = {} & \frac{\mathrm{E}\left(\widetilde{P}_T\right) - (1+r_f)\cdot\overline{p}}{\prod_{i=1}^{N}\left[(1+r_f)\cdot\beta + \frac{\rho_{m_i}}{W_{m_i}} \cdot \mathrm{var}\left(\widetilde{P}_T\right)\right]} \cdot \left\{\prod_{i=2}^{N}\left[\frac{\rho_{m_i}}{W_{m_i}} \cdot \mathrm{var}\left(\widetilde{P}_T\right)\right]\right. \\ & + \sum_{n=1}^{N-2}\prod_{i=n+2}^{N}\left[\frac{\rho_{m_i}}{W_{m_i}} \cdot \mathrm{var}\left(\widetilde{P}_T\right)\right] \cdot \prod_{i=1}^{n}\left[(1+r_f)\cdot\beta + \frac{\rho_{m_i}}{W_{m_i}} \cdot \mathrm{var}\left(\widetilde{P}_T\right)\right] \\ & \left. + \prod_{i=1}^{N-1}\left[(1+r_f)\cdot\beta + \frac{\rho_{m_i}}{W_{m_i}} \cdot \mathrm{var}\left(\widetilde{P}_T\right)\right]\right\}. \end{aligned} \tag{58}$$

For all $n \in \mathbb{N}$, $n \leq N$, let $\Gamma_{\boldsymbol{m}}^n = \left\{C_j^n\right\}_{j \in \left\{1,\ldots,\binom{N}{n}\right\}}$ be the family of all n-subsets (n-combinations)

[48] Note that the economic restriction $p(q) > 0$ never binds.

[49] As in (6).

[50] For $N = 2$, the middle term (summation) inside the braces vanishes.

C^n of the elements of $\boldsymbol{m} \in M$. Then (58) can be rewritten as

$$\left[\sum_{n=1}^{N} Q_n^*\right]_{\boldsymbol{m}} = \frac{\left[\mathrm{E}\left(\widetilde{P}_T\right) - (1+r_f)\cdot\overline{p}\right]\cdot[(1+r_f)\cdot\beta]^{N-1}}{\prod_{n=1}^{N}\left[(1+r_f)\cdot\beta + \frac{\rho_n}{W_n}\cdot\mathrm{var}\left(\widetilde{P}_T\right)\right]} \cdot\left\{1+\sum_{n=1}^{N-1}\left([(1+r_f)\cdot\beta]^{-n}\cdot\sum_{C^n\in\Gamma_{\boldsymbol{m}}^n}\prod_{i\in C^n}\left[\frac{\rho_i}{W_i}\cdot\mathrm{var}\left(\widetilde{P}_T\right)\right]\right)\right\}. \tag{59}$$

Trivially, $\forall n \in \mathbb{N},\ n \leq N,\ \forall \boldsymbol{m}, \boldsymbol{m}' \in M : \Gamma_{\boldsymbol{m}}^n = \Gamma_{\boldsymbol{m}'}^n$, which completes the proof. □

A.6 Proof of Proposition 3

LEMMA 3: *Let $p(q) > 0$ be continuous and nondecreasing on $\left(\underline{q}, \overline{q}\right)$ for any $\underline{q} < 0$ and $\overline{q} > 0$. Then*

$$\overline{P}(Q) = \frac{1}{Q}\cdot\int_0^Q p(q)\,\mathrm{d}q > 0 \tag{60}$$

is continuous and nondecreasing on $\left(\underline{q}, 0\right)$ and $(0, \overline{q})$.

Proof: The sign and continuity follow immediately from the integral properties. Assume that $\exists Q \in \left(\underline{q}, 0\right) \cup (0, \overline{q})$:

$$\frac{\mathrm{d}\overline{P}(Q)}{\mathrm{d}Q} = \frac{1}{Q}\cdot\left[p(Q) - \overline{P}(Q)\right] < 0. \tag{61}$$

By the Mean Value Theorem, there exists q strictly between 0 and Q such that $p(q) = \overline{P}(Q)$. It follows from (61) that $p(q) > p(Q)$ if $Q > 0$ and $p(q) < p(Q)$ if $Q < 0$, which contradicts the monotonicity of p. □

Now we can proceed to the main proof.

Proof of Proposition 3: By (33), which gives the residual market order on behalf of the clients after maximal internalization has been conducted so that market impact costs are minimized by Proposition 1, we have $\sum_{i=1}^{N} Q_i^*(P) = 0$ if and only if $P = -\frac{A}{B}$. Let us define an auxiliary function

$$\Psi(P) = \begin{cases} \overline{P}\left[\sum_{i=1}^{N} Q_i^*(P)\right] & \text{for} \quad P \neq -\frac{A}{B} \\ \overline{p} & \text{for} \quad P = -\frac{A}{B} \end{cases}. \tag{62}$$

By l'Hospital's Rule,

$$\lim_{Q\to 0}\overline{P}(Q) = p(0) = \overline{p} \tag{63}$$

and hence (33) with Lemma 3 imply that $\Psi(P) > 0$ is continuously nonincreasing in the economically relevant interval $[0, +\infty)$. By the Brouwer Fixed Point Theorem, $\exists P^* \in [0, \Psi(0)] : P^* = \Psi(P^*)$. Furthermore, the monotonicity of Ψ guarantees a unique $P^* > 0$. □

References

Arrow, K. J. (1971): *Essays in the Theory of Risk-Bearing*, Amsterdam, Netherlands, and London, UK: North-Holland Publishing Company.

Bessembinder, H. (2003): "Trade Execution Costs and Market Quality after Decimalization," *Journal of Financial and Quantitative Analysis*, 38, 747–777.

Cai, T. and G. Sofianos (2006): "Multi-day Executions," *Journal of Trading*, 1, 25–33.

De Winne, R. and C. D'Hondt (2007): "Hide-and-Seek in the Market: Placing and Detecting Hidden Orders," *Review of Finance*, 11, 663–692.

Degryse, H., F. De Jong, M. Van Ravenswaaij, and G. Wuyts (2005): "Aggressive Orders and the Resiliency of a Limit Order Market," *Review of Finance*, 9, 201–242.

Domowitz, I., I. Finkelshteyn, and H. Yegerman (2008): "Cul de Sacs and Highways: An Optical Tour of Dark Pool Trading Performance," *Investment Technology Group* research paper, August, available at www.itg.com.

Jezek, M. (2009): "Passive Investors, Active Traders and Strategic Delegation of Price Discovery," *Cambridge Working Papers in Economics* 0951, University of Cambridge.

Kazakov, V. and T. Vasak (2006): "DMA Trading and Crossings on the Australian Stock Exchange," *Quantitative Finance Research Centre* Research Paper 180, University of Technology Sydney.

O'Cinneide, C., B. Scherer, and X. Xu (2006): "Pooling Trades in a Quantitative Investment Process," *Journal of Portfolio Management*, 32, 33–43.

Perold, A. F. and R. S. Salomon (1991): "The Right Amount of Assets Under Management," *Financial Analysts Journal*, 47, 31–39.

Rothschild, M. and J. E. Stiglitz (1970): "Increasing Risk I: A Definition," *Journal of Economic Theory*, 2, 225–243.

Satchell, S. E. and B. Scherer (2010): "Fairness in Trading: A Microeconomic Interpretation," *Journal of Trading*, 5, 40–47.

Sofianos, G., S. Takriti, and I. Tierens (2007): "Including Trading Costs in Portfolio Optimization," *Goldman Sachs Street Smart* report, No. 30, December 5.

Zhao, X. and K. H. Chung (2007): "Information Disclosure and Market Quality: The Effect of SEC Rule 605 on Trading Costs," *Journal of Financial and Quantitative Analysis*, 42, 657–682.