FIRST DRAFT not to be quoted.

Price discovery in indicative and transaction FX prices:

A comparison of D2000-1 and EFX data

Kate Phylaktis* and Long Chen

Cass Business School

London

May 2006

Abstract

In this paper we compare four months of Reuters EFX high frequency indicative data with D2000-1 inter-dealer transaction data in DEM_USD and GBP_USD currency pairs. Contrary to previous studies, we find, using various information measurements, that the matched tick-by-tick indicative data bear no qualitative difference from the transaction data, and even have higher information content and react faster to market information than their corresponding transaction data.

*Corresponding author: Fax: 020 70408881; Email: K.Phylaktis@city.ac.uk

1. Introduction

With more and more firm quotes and transaction data available the value of the indicative data is gradually neglected or forgotten. Indicative data are assumed to be of low quality because of the lower importance attached to them by the traders, the excessive and irrelevant quotes from aggressive banks that need to build up a market presence, and the automatic quoting strategy of coping quotes from fellow banks. Therefore, any information extracted from indicative data should be treated with caution.

Although there is wide spread belief in that the indicative FX data associate with low information quality, formal and conclusive research on them is largely scanty. Even the few papers on the issue are baffled by a short sample period of either one day or one week to bear out some limited statistical results. For instance, Goodhart et al. (1996) compare one day of DEM_USD EFX data with those from D2000-2, but with no time stamp on the data. The two data sets are matched by maximising the correlation between the transaction and indicative data. After this procedure, they find at 10 minute frequency the statistical discrepancy between the two data sets is largely insignificant.

Danielsson and Payne (2002) extend the time window to five days (6^{th} to 10^{th} of October, 1997) for the comparison of the same two data sets, which is still too short for any meaningful conclusion. They process the data at a 20 seconds frequency as a compromise between the different frequencies of the two series. Basic statistics such as quote frequency, spread and return moments are investigated and compared for the two data sets. Using Hasbrouck's technique based on ECM, they find that D2000-2 has a dominant role in pricing the information except during midnight hours. However, all these differences disappear when they aggregate the data into 5 and 10 minute frequency.

Our paper’s indicative data are Reuters’ high-frequency EFX quotes on the DEM_USD and GBP_USD currency pairs. The corresponding transaction data are

D2000-1 inter-dealer transaction data on the same currency pairs.[1] With sample data spanning 82 trading days, we substantially increase the time window to make our empirical tests more reliable. Instead of only focusing on DEM_USD currency pair, we also introduce GBP_USD currency pair to make the comparison more conclusive.

In this paper we employ various methods to compare the indicative data and their corresponding transaction data. We conduct cross correlation tests to investigate the lead and lag relation between the indicative and transaction return series. We find that both indicative currency pair data sets lead transaction data by around 5 to 10 minutes. We utilise Hasbrouck (1995) information share technique to recover the information content of the two data sets, which further suggests that indicative data have the dominant role in mapping the fundamental information into the prices. We vary the data frequency by using real transaction time, 5 and 10-minute calendar time frequency to test the robustness of our results. To eliminate the trading zone effect on the 24-hour global FX trading, data are also tested under different trading sessions according to the openings and closings of major FX markets.

Then due to the growing importance of order flow in Exchange Rate theory, we introduce the order flow from the D2000-1 data to investigate whether our indicative data are relevant to the relationship between order flow and prices. Order flow is the difference between the volume of dollar buyers initiated trades and that of the seller initiated trades. Micro-based FX models treat order flow as an important channel to map the widely dispersed fundamental information into exchange rates. In Lyons and Evans' various models, the causality runs strictly from order flows to exchange rates. By applying generalized impulse response method to the both price series and order flow, we first find that the D2000-1 price takes longer (up to 5 to 10 minutes) to digest one standard error of shock from the EFX price than EFX price to digest one standard error of shock from the D2000-1 price, which confirms the results of the cross-correlation test. Adding to these findings is that the shock from EFX data imposes similar impact on order flow as the shock from D2000-1. However, we find

[1] The EFX and Reuters D2000-1 data sets are provided by Olsen & Associate and Martin Evans respectively.

that there is no significant impact from the shock of order flow to prices at both 5 and 10 minutes frequency. Though the prices and order flows display obvious co-movements during our sample period, it seems that order flow is more a latent response to fluctuations in prices, which is in contrast to Lyons's (2001) claim that order flow explains exchange rates at lower frequency (at 1 hour or daily frequency), instead of vice versa. To double check this result, we carry out Granger causality tests, which confirm our finding. This result casts some doubt on the order flow's value of being an important determinant of exchange rates.

The paper is structured as flows. In section 2, we give a description of our data sets and how we process them before we conduct our empirical tests. In section 3, we carry out a simple lead-lag return analysis and then the routine unit root and cointegration tests. In section 4, we conduct the Hasbrouck information share estimation of both prices. In section 5, we introduce the order flow and trivariate generalized impulse response on the two prices and Granger causality tests. In the final section we briefly summarize the results and the significant of our findings.

2. Data Sources and Sample Details

A. Sources of Data

EFX indicative high-frequency indicative data are collected from Reuters' EFX page. These tick-by-tick data are provided by different participating banks with each bid and ask pair stamped with time down to second and other information such as dealer's bank code and location. Indicative quotes are free from transaction obligations and are considered more as an advertising method to maintain banks' market presence. However, one should also note that due to reputation concern, the quote would not deviate too much from market price. Another important feature is that indicative data are not subjected to the consent from any other party like transaction data, hence are capable of instant update when news hit the market.

Reuters D2000-1 data are inter-bank transaction data. Because quotes and trades are executed electronically, an electronic record is produced. Each deal is time stamped to seconds with transaction size and transaction signs[2]. Unlike EFX data, transaction data usually include no detailed information on the involved counterparties; therefore no further heterogeneous information could be investigated.

B. Sample Details

There are eight fields included in EFX data: date, time, bid, ask, nation, city, bank and filter. Reuters D2000-1 contains nine fields: month, day, hour, minute, seconds, time index, transaction sign, price, and volume (see in Table 1). At first glance, EFX data provide more information on the quoting banks' identity, while D2000-1 offer unique record of the transaction signs and trading volume, though the later figures lacks accuracy[3].

The sample data span from 1st of May to 30th of August 1996, with total of 121 calendar days, including weekends and holidays. There are 313,845 and 612,260 quotes in GBP_USD and DEM_USD currency pairs respectively in EFX data set, which are much higher than those of D2000-1, with corresponding 52,318 and 257,398 ticks of data. This difference is more obvious in GBP_USD currency pair, with the size of EFX data being nearly six times of the D2000-1.

To check the basic statistics characteristics of both data sets, we first plot the average intraday quote frequency of the data sets in both currency pairs in Figure 1 and Figure 2. Each half-hour session's total quotes (trades) are divided by daily total quotes (trades). The peaks and lows of the two data sets generally coincide with each other throughout the day. However, the transaction data reveal more concentrated trading activities during London and New York trading hours. Indicative data instead display a relatively smooth quoting activity pattern during these sessions.

---

[2] If it is dollar buyer initiated trade, 1 is recorded. Otherwise 0 is filled in.
[3] As suggested in Evans (2002). He counts the sign of order flow as a proxy to the aggregate of real size of it.

Duration, which stands for the elapsed time (in seconds) in between two neighbouring quotes or transactions, is another method to study the intensity of the quoting or transaction activities. To eliminate the impact of automatic quoting, following Engel and Russell (1997), we exclude those quotes with price changes of less than 5 basis points in the indicative data. Both the EFX and D2000-1 data sets share the same highest clustering of duration under 10 seconds, nearly 17% of all quotes (see in Figures 3 and 4). From 40 seconds on, the duration patterns deviate from each other, with EFX experiencing a much smoother decline in density while D2000-1 swinging around EFX. One significant feature of D2000-1 is its relative lack of transaction duration of near 50-60 seconds. There is no explanation for this distinct feature as far as we know. However transaction frequency clusters again between 60 and 120 seconds, counting for over 30% of total transactions. The cumulative distribution lines confirm a much more stable pattern of EFX duration distribution.

Such differences indicate issue of unsynchronized data when comparing these two data sets. Indicative data can be updated without any transaction incurred while transaction data are the result of a mutual agreement of one pair of trading banks, therefore happen at a lower frequency. Since it is deemed that transaction data bear more information and take place at lower frequency compared with their indicative counterpart, we use transaction data as the benchmark to process our indicative data.

First, we filter indicative data using transaction data time stamp. For each transaction price, we locate its nearest indicative data in terms of time and form a matching transaction and indicative data pair. This procedure returns us with highly simultaneous price series in both currency pairs. The two data sets in GBP_USD currency pair have a time discrepancy of average 0.04 seconds and a standard deviation of 22 seconds, and the pairs in DEM_USD differ with average of 0.1 seconds and a standard deviation of 11 seconds. The differences between the processed time stamps of the pairs of indicative and transaction data sets in both exchange rates are insignificant under any strict criteria.

Second, to reflect the bid and ask shift of the transaction data, we choose the corresponding bid and ask price of the indicative data. This process is based on the transaction sign of each trade. For instance, for a dollar buyer initiated trade in D2000-1, the closest EFX bid is selected as the matching price.

After the above mentioned process, and excluding weekends and holidays, we have 82 trading days, 51,741 pairs of GBP_USD prices and 255,481 pairs of DEM_USD prices left. The 5 and 10-minute frequency data are obtained by choosing the last pair of prices in each time slot. Such a method sacrifices more available updated quotes in indicative data, but avoids comparing stale transaction data with indicative data. Therefore any subsequent empirical comparison of these two data sets may underestimate the information content of the indicative data.

The descriptive statistics of the both data sets are displayed in Table 2. The statistics on the first moments of the prices indicate that indicative data are generally a couple of basis points lower than transaction data. There is no documented explanation for such findings.

3. Preliminary Data Analysis

In this section, we first provide tests on the cross correlations of the return series in different data frequency, exploring the lead and lag pattern of the two data sets. Lead and lag analysis gives us a preliminary picture of the general relation between the returns of the prices by comparing the data at different lead and lags. Any significant lead or lag pattern may suggest one price's leadership over the other in mapping information into its returns.

We then perform unit root tests for each prices to investigate whether the price series are nonstationary and integrated of order one, i.e., a $I(1)$ process. Subsequently, we conduct the Johansen (1988) test to investigate whether the two sets of prices for each currency pair are cointegrated. By establishing the unit root and cointegration relation in the prices, we can investigate the information share between them, using a technique based on an Error Correction model.

A. Cross-correlations of Return Series

If two prices are based on the same fundamental asset, their return series should be correlated due to the shared determinants. The return series may present lead or lag patterns due to different information collecting and interpreting efficiency, transaction costs, execution (or quoting) speeds or market hierarchy, etc. For instance, ceteris paribus, if a market possesses private information which is unknown to other markets, its return series should lead other markets by revealing a positive lag value in a cross-correlation analysis.

Even though previous papers (see in Danielsson and Payne (1997), for example) associate indicative data with stale or lagged quotes compared to transaction data, we hold opposite opinion. One reason is that indicative quotes are in essence advertising signals to potential customers, hence should contain fresh price information to inform the market. Prolonged outdated quotes may put the effectiveness of this approach in doubt. Furthermore, theoretically indicative data could be updated at the absence of transaction, therefore should be more efficient in delivering information. Cross-correlation test of the return series offers a simple means to prove our hypothesis.

In Figures 5 and 6 we present the cross-correlation results of the return series of both data sets on the two currency pairs at 5-minute and 10-minute frequency respectively. The 95% confidence levels are formed by calculating $\pm 2/\sqrt{T}$, with T being usable observations, which are the dotted lines in both graphs. Positive lag at the X axis indicate that EFX quotes are at the lead and vice versa. At 5 minute frequency, we find that EFX leading quotes demonstrate a significant positive correlation with the lagged D2000-1 prices at lag 1 in GBP_USD and at both lag 1 and lag2 in DEM_USD. The 10-minute frequency results confirm that EFX quotes lead D2000-1 price around 10 minutes.

The lead-lag analysis demonstrates an asymmetric relation between the returns of the two data sets, with no significant correlation when D2000-1 is in the lag. Such findings are in contrast with Danielsson and Payne’s (1997). Based on 20 seconds frequency, they find that D2000-2 returns lead EFX returns by 2 and 3 minutes. However, positive correlation also exists when EFX is the one in the lag, which suggests that the cross-correlation is less asymmetric. In our case the predictive power runs only from EFX data to D2000-1, with D2000-1's return imposing no predictive power on EFX's return at all.

B. Unit Root Tests

To investigate the cointegration between the two prices, as a routine, we first test whether the time series in concern contain unit root. Specifically, if two time series are both nonstationary at their level, but become stationary after first difference, we denote them as $I(1)$ processes, or integrated of order one. Cointegration become relevant if the linear combination of both $I(1)$ series is stationary.

The augmented Dickey-Fuller (1981) test is used in our unit root tests, which extend basic Dickey-Fuller test by including a parametric correction for higher-order correlation by assuming that the time series follows an $AR(p)$ process and adding p lagged difference terms of the dependent variable to the right-hand side of the test regression. We select the lag length p by using Schwarz Bayesian criterion (Schwarz (1978)). Three types of unit root tests, including no intercept or trend, only intercept, and only trend, are conducted. Unless otherwise results are found, we only present the result with only intercept. Since our further empirical tests involve all three kinds of time frequency, i.e., transaction frequency, 5 and 10-minute frequency, unit root tests are conducted on each of them.

In Table 3 we present both t-statistics and p-value of the unit root tests. Overall, we conclude that all price series are $I(1)$ process.

C. Cointegration Tests

The object of the cointegration test is to determine whether two nonstationary series are cointegrated. As pointed out by Engle and Granger (1987), if a linear combination of two or more nonstationary series is stationary, then the series are said to be cointegrated. The stationary linear combination is called the cointegrating equation and may be interpreted as a long-run equilibrium relationship among the variables. Cointegration relation forms the basis of the VEC specification.

The cointegration tests used are based on the methodology developed by Johansen (1991, 1995a). We first consider a VAR of order p:

$$y_t = A_1 u_{t-1} + .. + A_p y_{t-p} + \varepsilon_t \tag{1}$$

where y_t is a k-vector of non-stationary $I(1)$ variables, and ε_t is a vector of innovations. We may rewrite this VAR as,

$$\Delta y_t = \Pi y_{t-1} + \sum_{t=1}^{p-1} \Gamma_i \Delta y_{t-i} + \varepsilon_t \tag{2}$$

where

$$\Pi = \sum_{i=1}^{p} A_t - I_t \,,\ \ \Gamma_t = -\sum_{j=i+1}^{p} A_j \ .$$

Granger's representation theorem asserts that if the coefficient matrix Π has reduced rank $r < k$, then there exist $k \times r$ matrix α and β each with rank r that $\Pi = \alpha\beta'$ and $\beta' y_t$ is $I(0)$. And subsequently r is the number of cointegrating relations and each column of β is the cointegrating vector. Johansen's method is to eliminate the Π matrix from an unrestricted VAR and to test whether the restrictions implied by the reduced rank of Π could be rejected. Cointegration tests establish the fact that there exists a long term equilibrium relation between two nonstationary series, which form the basis for VEC (Vector Error Correction). In Equation (2), the element of α are known as the adjustment parameters in the VEC model.

Though there is no possible arbitrage to keep the long run equilibrium relation between the indicative prices and transaction prices, as the former prices have no binding obligation of actual transaction, reputation and commercial concerns would drive dealers to quote on the fundamental market information. As a consequence, the indicative and transaction prices are both based on the same information on the currency pair, and these two prices are expected to be cointegrated.

To conduct the cointegration tests, we first choose the lag interval that minimizes the Schwarz information criterion. Based on the chosen lag length, we carry out the five standard types of cointegration tests with the option of including or excluding intercept or trend in the cointegration system. We use Schwarz information criterion to decide the lag structure of the model. We determine the number of cointegrating vectors by comparing the maximum eigenvalues with their corresponding critical values. The (nonstandard) critical values are taken from Osterwald-Lenum (1992). We present the cointegration tests results in Table 4. In all of the tests, we can reject that there is no cointegrating relationship but cannot reject there is one cointegrating vector at 5% significance level.

Based on the cointegration results, in the following section we explore the information share between the prices from the two data sets.

4. Static Information Share

A. Error Correction and Fundamental Value

In economics, fundamental value is essentially an abstract concept that cannot be observed directly. However, we can always assume that, in the long run, the fundamental value would manifest itself and transient information would disappear. On this specification, fundamental value could be identified as the permanent component of a price series. Price discovery therefore describes how one price series incorporates the permanent component into the price system, either in a static or dynamic sense.

Here we start with one pair of cointegrated price series. For most microstructure models, we assume that the efficient price follows a random walk, as stated in a structural model:

$$p_t = lm_t + s_t\,, \quad p_t = \begin{pmatrix} p_{1t} \\ p_{2t} \end{pmatrix}, \quad s_t = \begin{pmatrix} s_{1t} \\ s_{2t} \end{pmatrix} = \Upsilon(L)v_t\,, \quad l = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \tag{3}$$

where m_t is the underlying efficient price and s_t is the transient microstructure noise.

From (3) above, the first difference of p_t has the moving average representation

$$\Delta p_t = \Psi(L)\varepsilon_t = lu_t + \Delta s_t = lu_t + (1-L)\Upsilon(L)v_t\,, \quad \Psi(\mathrm{L}) = \sum_{\mathrm{j=0}}^{\infty}\Psi_j L^j \tag{5}$$

so that Δs_t is a covariance stationary but noninvertible moving average. Since

$$\Psi(L) \equiv \Psi(1) + (1-L)\Psi^*(L), \quad \Psi_{\mathrm{i}}^* = -\sum_{j=i+1}^{\infty}\Psi_j\,, \tag{6}$$

p_t and Δp_t can always be rewritten as

$$p_t = \Psi(1)\sum_{s=0}^{t}\varepsilon_s + \Psi^*(L)\varepsilon_t\,,$$

$$\Delta p_t = \Psi(1)\varepsilon_t + \Psi^*(L)\Delta\varepsilon_t = \Psi(1)\varepsilon_t + (1-L)\Psi^*(L)\varepsilon_t\,, \tag{7}$$

where $\Psi^*(L)\varepsilon_t$ is covariance stationary and its first difference $(1-L)\Psi^*(L)\varepsilon_t$ is a stationary, noninvertible moving average. Since $z_t = (1\ \ -1)\,p_t$ is stationary, it must omit the stochastic trend $\Psi(1)\sum_{s=0}^{t}\varepsilon_s$, implying that $(1\ \ \text{-}1)\Psi(1) = 0$ and thus $\Psi(1) = l(\psi_1\psi_2)$ which is intuitively obvious since the two prices share the same implicit efficient price.

This representation underlies the Hasbrouck (1995) information shares approach. The long run impact of ε_t, u_t, and v_t on Δp_t may be found by evaluating both

$\Psi(L)$ and $(1-L)\Upsilon(L)$ in (5) at $L=1$, yielding

$$\Psi(1)\varepsilon_t = lu_t \tag{8}$$

and the resulting perfect correlation arising from the relation $\psi'\varepsilon_t = u_t$ implies

$$E[ll'u_t^2] = E[\Psi(1)\varepsilon_t\varepsilon_t'\Psi(1)'] = E[l\psi'\varepsilon_t\varepsilon_t'\psi l'] \Rightarrow \sigma_u^2 = \psi'\Sigma_\varepsilon\psi \quad . \tag{9}$$

Hasbrouck information shares involve decomposing $\psi'\Sigma_\varepsilon\psi$ into components attributed to price innovations in the two markets, while in our case they are the two prices from the two data sets. This attribution is unique when said price innovations are uncorrelated, in which case the decomposition is given by:

$$1 = \frac{\psi_1^2}{\psi'\Sigma_\varepsilon\psi}\sigma_{\varepsilon_1}^2 + \frac{\psi_1^2}{\psi'\Sigma_\varepsilon\psi}\sigma_{\varepsilon_2}^2 \tag{10}$$

However, when the reduced form residuals are correlated, the decomposition is

$$1 = \frac{\psi_1^2}{\psi'\Sigma_\varepsilon\psi}\sigma_{\varepsilon_1}^2 + \frac{\psi_1^2}{\psi'\Sigma_\varepsilon\psi}\sigma_{\varepsilon_2}^2 + 2\frac{\psi_2^2}{\psi'\Sigma_\varepsilon\psi}\sigma_{\varepsilon_1}^2\sigma_{\varepsilon_2}^2 \tag{11}$$

and there is a range of possible attributions corresponding to different allocation of the covariance form to each market. Hasbrouck suggests change the order of the prices, hence the object of the Cholesky decomposition, till all possible orderings are realized, and then calculate the average result. In our case, since there are only two price series and hence only two possible rotations of the orders.

B. Information Shares Results

To circumvent the contemporaneous residual correlation problem and the ambiguity produced by the reordering procedure, Hasbrouck used ultimate high frequency price series to reduce this side effect on the information share formula. In our paper, due to the highly simultaneous data we processed, the correlation issue is less serious, which produces results with much tighter bands than most previous studies, which use the information share technique. In both currency pair, we test the information share of the two data sets at the transaction frequency. The results at 5 and 10-minute are also presented as robustness tests.

We first test the whole sample period in both currency pairs. In GBP_USD currency pair, after VEC test, the indicative and transaction data has a low residual correlation of 16.6% (see in last column of the third row in Table 5). The information share attributed to the indicative data EFX is as high as 83%, with relatively tight lower and upper bands of 77% and 89%, respectively. In DEM_USD, with the residual correlation as low as 4.2%, the information share attributes 85% of the total information to the EFX data. The lower and upper bands, as predicted, are only 1.5% away from the averaged result, indicating a much reliable decomposition. These figures suggest that EFX data take a dominant role in the price discovery process. Using 5 and 10-minute frequency, though the information shares are decreased for EFX data, the conclusion is not fundamentally changed.

To check the dynamic change of the information share during 24-hour trading days, we separate the trading hours into 7 sessions that corresponds to major FX markets' opening and closing times. The 7 sessions are 1) 21:00 to 8:00, New York closes till Tokyo closes, representing Asian trading hours; 2) 8:00-9:00, first hour of London opening with inventory effects; 3) 9:00-12:00, hours till New York opens; 4) 12:00-13:00, first hour of New York opening; 5) 13:00-15:00, overlapping trading hours of two major markets; 6) 17:00-18:00, London closing hour; 7) 18:00-21:00, hours till New York closes.

The results are displayed in the mid columns of Table 5[4]. We also plot them in Figures 7 and 8. The information share of EFX data peak during London trading (session 3) and overlapping (session 5) hours, with GBP_USD has a higher peak in London trading hours, and DEM_USD have a higher peak in overlapping hours. During closing and opening hours, EFX price experiences obvious drop of information share. These patterns suggest that during two of the peak trading hours, EFX data actually possess higher information content even though D2000-1 data substantially increase the transaction frequency.

[4] To calculate the information share of the each session, we use SUR (seemingly unrelated regression) and delete any lagged returns that belong to last day in each equation.

5. Order Flow and Generalized Impulse Response Analysis

Having established the information content of the two price series, in this section we focus on the dynamic interaction between them using impulse response analysis. Due to the importance of the order flow in recent developments in exchange rate theory, we include the order flow data from the D2000-1 data set in the analysis to investigate its relationship with the indicative and transaction prices.

A. Order Flow Analysis: An Introduction

Order flow is a measurement of the difference between buyer initiated trades and seller initiated trades. It is a well investigated factor in microstructure research in equity markets due to much earlier availability of the data. However, in FX market it is only first introduced by Goodhart and Flood in the late 80s and early 90s. With the seminal paper by Lyons (1995) and the gradually release of data from major trading banks and systems, order flow has become the utmost key word in exchange rate theories.

In FX microstructure theory, order flow is an important channel for heterogeneously dispersed liquidity information and asymmetric private information to exchange rates (see in Lyons (1995), Evans (2002), Breedon and Vitale (2005) and many others). In traditional canonical models (see e.g.Glosten and Milgrom, 1985) price and order flow have two-way causality with each other. In Evans and Lyons' (2002) model, the causality runs strictly from order flow to price. However, a further investigation reveals that they use hourly frequency and prices are taken as the last price while order flow is the interim aggregate. Therefore, if the actual causality happens at higher frequency, say down to 5 to 10 minute in our case, the information contained in the order flow can well lead the latent price at the end of an hour, even though the true lead and lag pattern is the other way around, i.e., price actually lead order flow. This mistake could be further amplified by using last daily price to compare intraday day accumulated order flow, as in Killeen et al. (2006) where they find the same conclusion of one way Granger causality from order flow to exchange rate. Our impulse response tests indicate that it might not be the case when using higher frequency data. Causality tests further confirm our findings.

B. Generalized Impulse Response Analysis

The order flow data in D2000-1 is the transaction sign counts of the inter-bank deal, which does not reveal the exact size of the deal involved. Though the sign of each trade itself could be random, the accumulated order flow could be non-stationary for a given time window. ADF tests indicate that order flow data in both exchange rates contain unit root and are $I(1)$ process at 5 and 10 minute frequency (see in Table 6).

We first look at the movements of the three series, i.e. EFX, D2000-1 and order flow, during our sample period (Figure 9 and 10). In order to demonstrate the positive correlation between order flow and the price levels graphically, order flows of the both currency pair are defined as net dollar seller initiated trades. Such a modification is applied in the following empirical tests. The general co-movement between the order flow and prices is apparent, especially in DEM_USD. This suggests that prices and order flow form a system and share the same fundamentals. We are more interested in the dynamic interaction among the three variables.

We introduce Hasbrouck's (1991) vector autoregression (VAR) model to investigate this issue,

$$\Delta p_{DL,t} = \sum_{i=1}^{\infty} a_i \Delta p_{DL,t-i} + \sum_{i=1}^{\infty} b_i \Delta p_{EF,t-i} + \sum_{i=1}^{\infty} c_i x_{t-i} + e_{1,t} \qquad 4.1$$

$$\Delta p_{EF,t} = \sum_{i=1}^{\infty} a_i \Delta p_{DL,t-i} + \sum_{i=1}^{\infty} b_i \Delta p_{EF,t-i} + \sum_{i=1}^{\infty} c_i x_{t-i} + e_{2,t} \qquad 4.2$$

$$\Delta x_t = \sum_{i=1}^{\infty} a_i \Delta p_{DL,t-i} + \sum_{i=1}^{\infty} b_i \Delta p_{EF,t-i} + \sum_{i=1}^{\infty} c_i x_{t-i} + e_{3,t} \qquad 4.3$$

where $\Delta p_{DL,t}$, $\Delta p_{EF,t}$ and Δx_t stand for D2000-1, EFX price change and order flow change.

As both prices p_t and the order flow x_t are $I(1)$ process, the differenced variables at the left hand sides are stationary. The changes of the order flow are divided by 1,000 to make them comparable to those of prices changes. The estimations are corrected by White's heteroskedasticity consistent standard errors. The optimum lag lengths are chosen by Schwarz criterion.

Dynamic analysis of VAR models is routinely carried out using the 'orthogonalized' impulse responses. However, the involved Cholesky decomposition is not invariant to the ordering of the variables in the VAR. Therefore we use Pesaran and Shin's (1998) generalized impulse response approach to analyze the interactions. The generalized impulse responses from an innovation to the *j*-th variable are derived by applying a variable specific Cholesky factor computed with the *j*-th variable at the top of the Cholesky ordering. It only coincides with orthogonalized approach when the investigated variable is put at the top of the ordering.

Figure 11 to Figure 16 display the impulse response of both prices and order flow in the two currency pairs. Due to space concern, we only present the results at the 5-minute frequency. There is no qualitative difference of the results at 10 minute frequency.

In GBP_USD currency pair, we find that the response of EFX price to one standard error of D2000-1 price impulse becomes insignificant around lag 2 or 10 minutes (Figure 11). However the effect of one standard error of EFX impulse in D2000-1 disappear after around lag 6 or 30 minutes (Figure 12). In both Figure 11 and 12, the responses of order flow to the shocks from the two prices build up from lag 1 till lag 3, and then slide to insignificance at lag four or 20 minutes. In stark contrast, we could not find any significant response from the two prices to the order flow shock at all (Figure 13).

In DEM_USD currency pair (Figure 14 to 16), the impulse response profiles are slightly different. Again, the response of EFX price to the D2000-1 impulse dies out at lag 2 to 3(Figure 14). However, the response of D2000-1 to the impulse of EFX takes 15 minutes (at lag 3) to reach insignificance. In both Figure 14 and 15, the responses of order flow to the price shocks quickly slide to near zero around 10 minutes, and then reach their peaks after another 5 minutes. The impacts of the prices' shocks on the order flow continue to exist even over 30 minutes, which is different from those profiles in GBP_USD currency pair. Finally, there is no significant response of prices to the order flow shock.

To summarize, the impulse response analysis in both currency pairs indicates that the shocks from EFX have much longer impact on D2000-1 data than vice versa. EFX data' impulses also have similar significant impact like those of D2000-1 on order flow. And order flow imposes no impact on prices, in contrast to the claims that order flow contains private information that is not revealed in prices.

In Table 7 and Table 8 we present the Granger causality tests on order flow with indicative and transaction data at both frequencies[5]. The causality is unambiguously one direction from prices to order flow, with literally no causality from order flow to prices. The results are more obvious at 10 minute frequency than 5 minute frequency. Combined with the impulse response analysis, order flow is a latent and passive response to prices at high frequency. Even if order flow does carry dispersed information among dealers, it can not happen at such high frequency, as opaque de-centralized market institution stops individuals from quickly aggregating information from disintegrated order flows.

6. Conclusion

By comparing various statistical features of the EFX and D2000-1 data sets, we find that, contrary to previous studies, the indicative data are not inferior in an

[5] Since there is no cointegration between prices and order flow as we have tested, the Granger causality test is not affect by complicate issues caused by ECM (see in Toda and Phillips (1993)).

information quality sense. More precisely, both lead-lag and impulse response analysis conclude that the indicative data lead the transaction data by 5 to 10 minutes. And information share technique indicates a dominant role of indicative data in mapping information. By adding order flow into the trivariate generalized impulse response analysis, we find that EFX price imposes similar impact on order flow like D2000-1 price.

These findings are supportive to studies using indicative data, as the quality of their data has never been formally tested. The different merits of indicative and transaction data to reflect market information suggest that we should combine both types of data to reveal the hidden picture of the heterogeneously distributed information in foreign exchange markets.

Reference

Breedon, F., and Vitale, P., 2005, An empirical study of liquidity and information effects of order flow on exchange rates, working paper.

Danielsson. J., Payne R., 2002, Real trading patterns and prices in spot foreign exchange markets. Journal of International Money and Finance, 21, 203-222.

de Jong. F., Mahieu R., and Schotman P., 2001, Price discovery on foreign exchange markets with differentially informed traders. Working paper.

Dickey, D. A., and W. A. Fuller, 1981, "Likelihood Ratio Statistics for Autoregressive Time Series with a Unit Root" Econometrica 49, 1057-1072.

Dufour, A., Engle, R. F., 2000. Time and the price impact of a trade. Journal of Finance 55, 2467-2498.

Engle, R.F., Granger C.W.J., 1987. Co-integration and error correction: Representation, estimation, and testing. Econometrica 55, 251-276.

Engle, R. F. and Patton. A. J., 2004. Impacts of trades in an error-correction model of quote prices. Journal of Financial Markets, 7, 1-25.

Engle R. F. and Russell J. R. 1997. Forecasting the frequency of changes in quoted foreign exchange prices with the autoregressive conditional duration model. Journal of Empirical Finance 4, 187-212.

Evans, M., D., 2002, FX trading and exchange rate dynamics, The Journal of Finance, VOL. LVII, NO. 6, p2405-2447.

Evans, M., D., and Lyons, R., K., 2002, Time-varying liquidity in foreign exchange, Journal of Monetary Economcs, 49, 1025-1051

Gonzalo, Jesus, and Serena Ng, 2001. A systematic framework for analyzing the dynamic effects of permanent and transitory shocks, Journal of Economic Dynamics and Control 25. 1527-1546.

Goodhart, C., Ito, T., Payne, R., 1996. One day in June, 1993: a study of the workings of Reuters D2000-2 electronic foreign exchange trading system. In: The Microstructure of Foreign Exchange Markets. University of Chicago Press, Chicago, IL, pp. 107-179.

Hasbrouck, J., 1991. Measuring the information content of stock trades. Journal of Finance 46 (1), 179-207.

____________ 1995. One security, many markets: Determining the contributions to price discovery, Journal of Finance 50, 329-339.

Johanson, S. (1988) ’’ Statistical Analysis of Co-integrating Vectors ’’ Journal of Economic Dynamic and Control, 12: 231-254.

Killeen., W., P., Lyons., R., K., and Moore., M., J., 2006, Fixed versus flexible: lesson from EMS order flow. Forthcoming, Journal of international Money and Finance.

Lyons, R. “Tests of Microstructural Hypotheses in the Foreign Exchange Market.” Journal of Financial Economics, 39 (1995), 321-351.

__________. The Microstructure Approach to Exchange Rates. Cambridge, MA: MIT press 2001.

Osterwald-Lenum, M., 1992. A note with quantiles of the asymptotic distribution of the maximum likelihood cointegration rank tests statistics. Oxford Bulletin of Economics and Statistics 54 (3), 461-472.

Pesaran M.H. and R.P. Smith, 1995, Estimation of long-run Relationships from Dynamic Heterogeneous Panels. Journal of Econometrics, 68, pp. 79-113.

Toda, H. Y. and Phillips P. C. B., 1993, Vector autoregressions and causality. Econometirca, Vol. 61, No. 6.

Table 1. Two ticks of sample data

EFX	Date	Time	Bid	Ask	Nation	City	Bank	Filter	
	1996-5-1	0:00:12	1.5	1.506	392	1	532	1	
D2000-1	Month	Day	Hour	Minute	Sec	T_index	B/S	Price	Vol
	5	1	1	15	35	501.05	0	1.5047	0

These two ticks of data are both in GBP_USD currency pair. The corresponding DEM_USD data are no different from these forms.

Table 2. Properties of processed data sets

GBP_USD	Mean	Median	Max	Min	Std. D.	Skew	Kurtosis
D2000-1	1.5397	1.5450	1.5650	1.4895	0.0167	-1.0137	2.9501
EFX	1.5393	1.5445	1.5682	1.4895	0.0167	-1.0136	2.9512
DEM_USD	Mean	Median	Max	Min	Std. D.	Skew	Kurtosis
D2000-1	1.5090	1.5182	1.5510	1.4638	0.0237	-0.1285	1.4205
EFX	1.5087	1.5180	1.5488	1.4635	0.0237	-0.1282	1.4198

The statistical results include all sample data of 82 trading days. There are total of 51,741 pairs of GBP_USD prices and 255,481 pairs of DEM_USD prices. We use D2000-1's time stamp as the benchmark time to locate the nearest EFX data. The EFX data' bid-ask selection is also based on D2000-1's order flow sign.

Table 3. Results of unit-root tests

	Transaction Time		5-Minute		10-Minute	
GBP_USD	EFX	D2000-1	EFX	D2000-1	EFX	D2000-1
t-Sta.	-2.01	-1.97	-2.39	-2.37	-1.94	-1.95
Prob.	0.28	0.30	0.14	0.15	0.31	0.31
DEM_USD	EFX	D2000-1	EFX	D2000-1	EFX	D2000-1
t-Sta.	-1.13	-1.05	-1.08	-1.09	-1.09	-1.08
Prob.	0.71	0.74	0.73	0.72	0.72	0.73

Table 4. Results of cointegration tests

Unrestricted Cointegration Rank Test (Maximum Eigenvalue)				
Hypothesized No. of CE(s)	Eigenvalue	Max-Eigen Sta.	5% Critical Value	1% Critical Value
GBP_USD - Transaction Frequency				
None *	0.07	3931.83	11.22	0.00
At most 1	0.00	1.18	4.13	0.32
GBP_USD - 5-minute Frequency				
None *	0.24	2264.14	15.89	1.00
At most 1	0.00	6.90	9.16	0.13
GBP_USD - 10-minute Frequency				
None *	0.24	1930.19	15.89	1.00
At most 1	0.00	4.89	9.16	0.30
DEM_USD - Transaction Frequency				
None *	0.03	6907.16	15.89	1.00
At most 1	0.00	2.09	9.16	0.76
DEM_USD - 5-minute Frequency				
None *	0.24	4877.84	15.89	1.00
At most 1	0.00	2.01	9.16	0.78
DEM_USD - 10-minute Frequency				
None *	0.24	2831.07	11.22	0.00
At most 1	0.00	0.82	4.13	0.42

The methodology is based on Johansen (1991 and 1995a). The optimum lag is chosen by Schwarz information criterion. The critical values are taken from Osterwald-Lenum (1992).

Table 5. Information share results of EFX data

GBP_USD EFX Information Share									
	1	2	3	4	5	6	7	ALL	Resid Corr
Trans.	60.5%	77.9%	89.5%	77.8%	93.0%	89.2%	73.9%	82.9%	16.6%
5-Min	48.8%	62.3%	82.7%	70.0%	80.5%	77.5%	64.7%	73.1%	49.6%
10-Min	42.0%	54.0%	74.0%	64.0%	67.0%	62.3%	60.2%	66.9%	60.0%
DEM_USD EFX Information Share									
	1	2	3	4	5	6	7	ALL	Resid Corr
Trans.	71.0%	78.6%	90.0%	82.0%	87.4%	82.3%	80.2%	85.0%	4.2%
5-Min	67.0%	76.0%	84.0%	69.1%	81.4%	76.6%	75.2%	81.2%	34.0%
10-Min	52.6%	71.0%	77.9%	60.8%	69.8%	66.2%	64.0%	76.4%	53.7%

The information share of the EFX data is calculated using Hasbrouck's (1995) technique based ECM. The columns numbered from 1 to 7 stand for the different trading zones of a complete trading day. The last two columns are the information share of the EFX data of the whole sample period and the residual correlation of the ECM, respectively.

Table 6. Unit-root test on order flow

Null Hypothesis: Order flow has a unit root Exogenous: Constant and trend Lag Length: Automatic based on SIC				
	GBP_USD		DEM_USD	
	5-Minute	10-Minute	5-Minute	10-Minute
t-Sta.	-1.49709	0.461431	-1.66038	-1.58001
*Prob.**	0.8309	0.9854	0.7687	0.8011

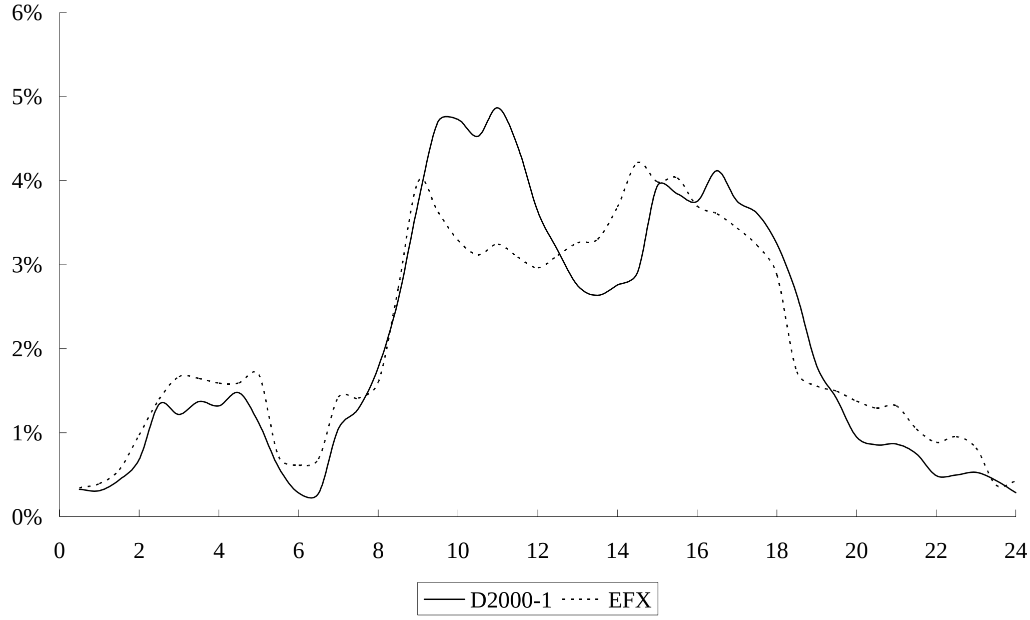
Table 7. Granger causality test of order flow and price series at 5-m frequency

5-Minute	Null Hypothesis:	F-Sta.	Prob.
GBP_USD Obs: 8034	Order flow does not Granger Cause EFX	1.18	0.10
	EFX does not Granger Cause Order Flow	4.63	2.40E-46
	Order flow does not Granger Cause D2000-1	1.16	0.13
	D2000-1 does not Granger Cause Order Flow	4.21	1.50E-39
DEM_USD Obs:17178	Order Flow does not Granger Cause EFX	1.08	0.22
	EFX does not Granger Cause Order Flow	10.71	2.00E-299
	Order flow does not Granger Cause D2000-1	1.12	0.13
	D2000-1 does not Granger Cause Order Flow	8.73	5.00E-230

Table 8. Granger causality test of order flow and price series at 10-m frequency

10-Minute	Null Hypothesis:	F-Sta.	Prob.
GBP_USD Obs: 7206	Order flow does not Granger Cause EFX	0.94	0.66
	EFX does not Granger Cause Order Flow	5.90	5.20E-67
	Order flow does not Granger Cause D2000-1	1.03	0.39
	D2000-1 does not Granger Cause Order Flow	4.41	1.60E-42
DEM_USD Obs: 10132	Order Flow does not Granger Cause EFX	0.93	0.69
	EFX does not Granger Cause Order Flow	17.00	1.00E-263
	Order flow does not Granger Cause D2000-1	1.16	0.13
	D2000-1 does not Granger Cause Order Flow	12.99	3.00E-193

Figure 1. GBP_USD intraday quote (trade) frequency



The intraday data frequency is calculated by counting each half hour's prices and then divided by total number of the sample average daily counts.

Figure 2. DEM_USD intraday quote (trade) frequency

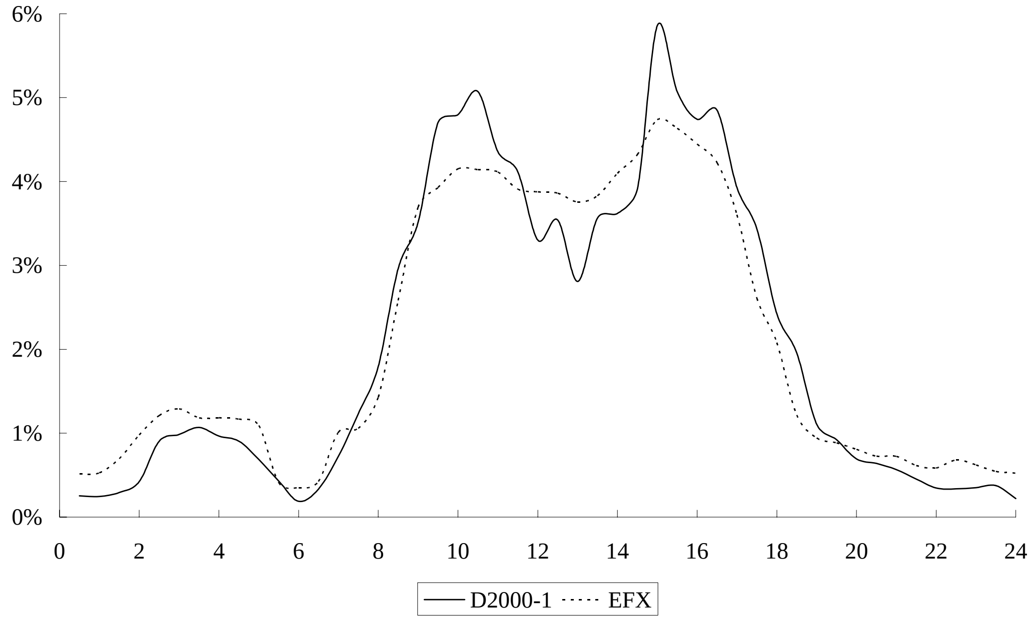
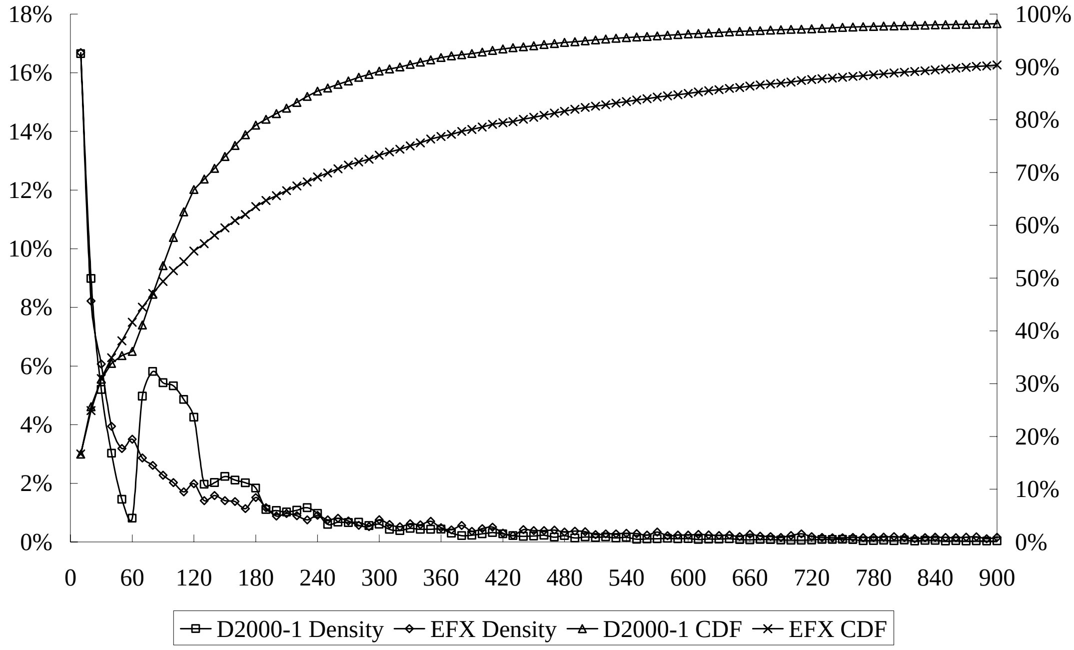


Figure 3. Duration distribution of the GBP_USD currency pair



The D2000-1 density is based on the duration distribution of all prices. Cumulative distribution function is the accumulated density for all duration. The duration density for EFX data is processed on quotes with changes larger than 5 basis points.

Figure 4. Duration distribution of DEM_USD currency pair

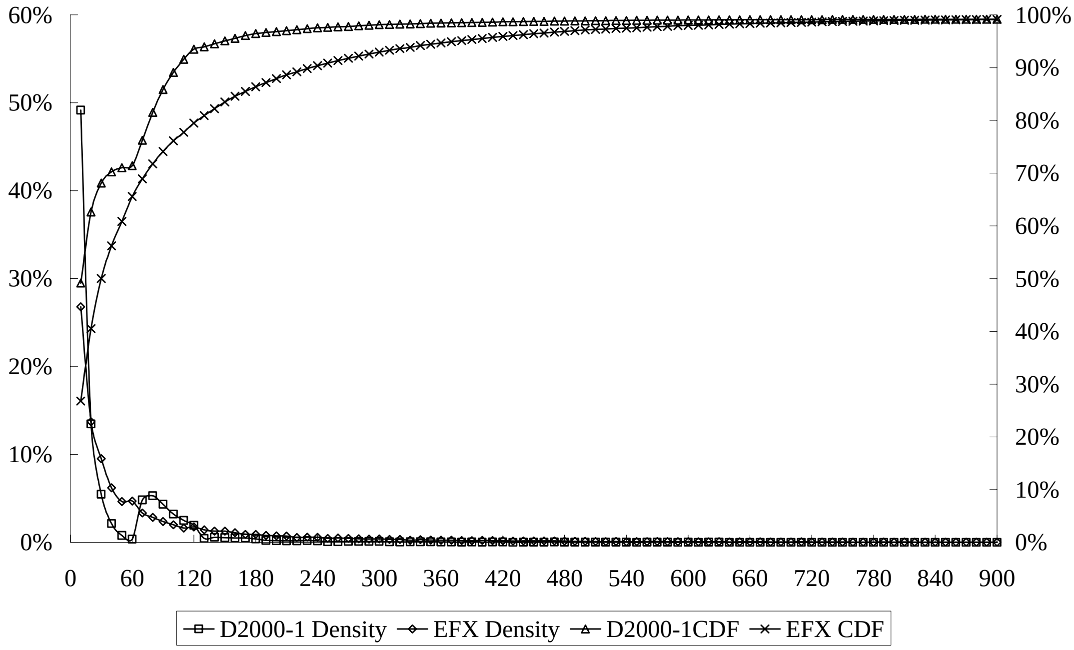
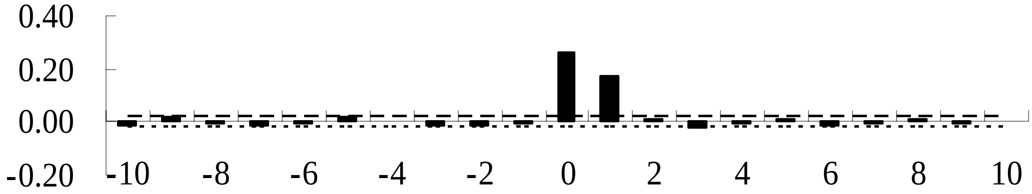
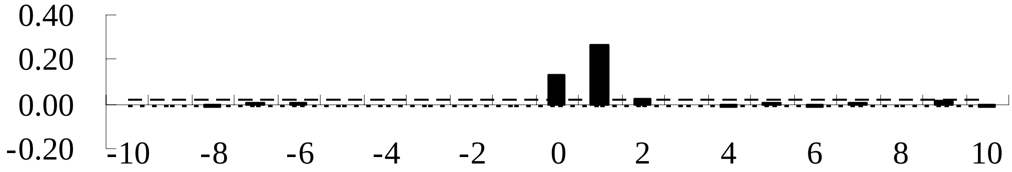


Figure 5. Cross-correlation of EFX and D2000-1 data at 5-m frequency

A. GBP_USD/5Min



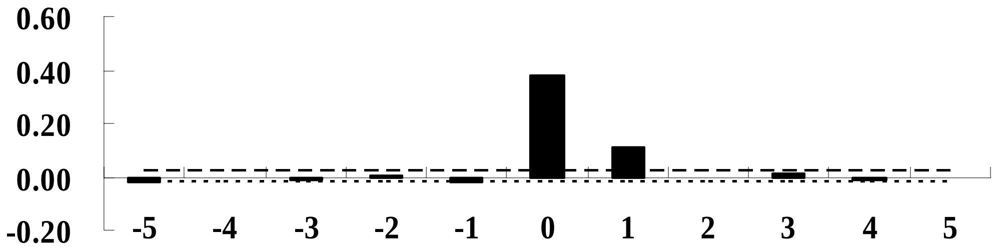
B. DEM_USD/5-Min



The 95% confidence levels are formed by calculating $\pm 2/\sqrt{T}$, with T being usable observations, which are the dotted lines in both graphs. Positive lag at the X axis indicate that EFX quotes are at the lead and vice versa.

Figure 6. Cross-correlation of EFX and D2000-1 data at 10-m frequency

C. GBP_USD/10Min



D. DEM_USD/10Min

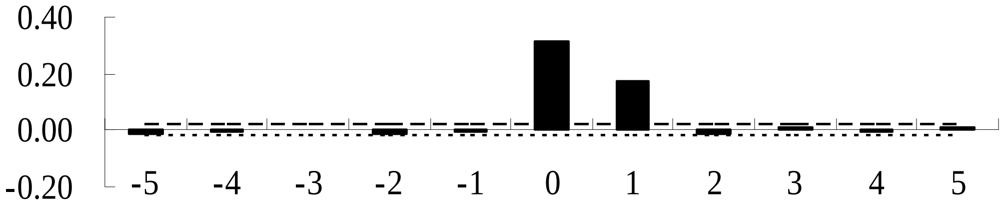
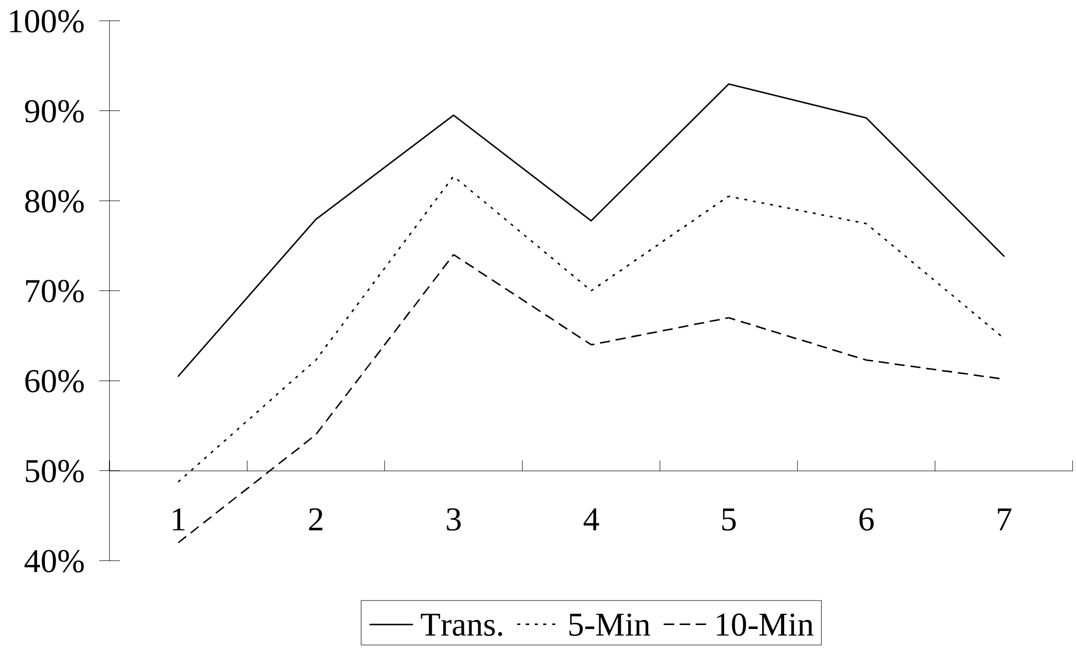


Figure 7. GBP_USD - information share of EFX during trading sessions



We separate the trading hours into 7 sessions that corresponds to major FX markets' openings and closings. The 7 sessions are 1) 21:00 to 8:00, New York closes till Tokyo closes, representing Asian trading hours; 2) 8:00-9:00, first hour of London opening with inventory effects; 3) 9:00-12:00, hours till New York opens; 4) 12:00-13:00, first hour of New York opening; 5) 13:00-15:00, overlapping trading hours of two major markets; 6) 17:00-18:00, London closing hour; 7) 18:00-21:00, hours till New York closes.

Figure 8. DEM_USD – information share of EFX during trading sessions

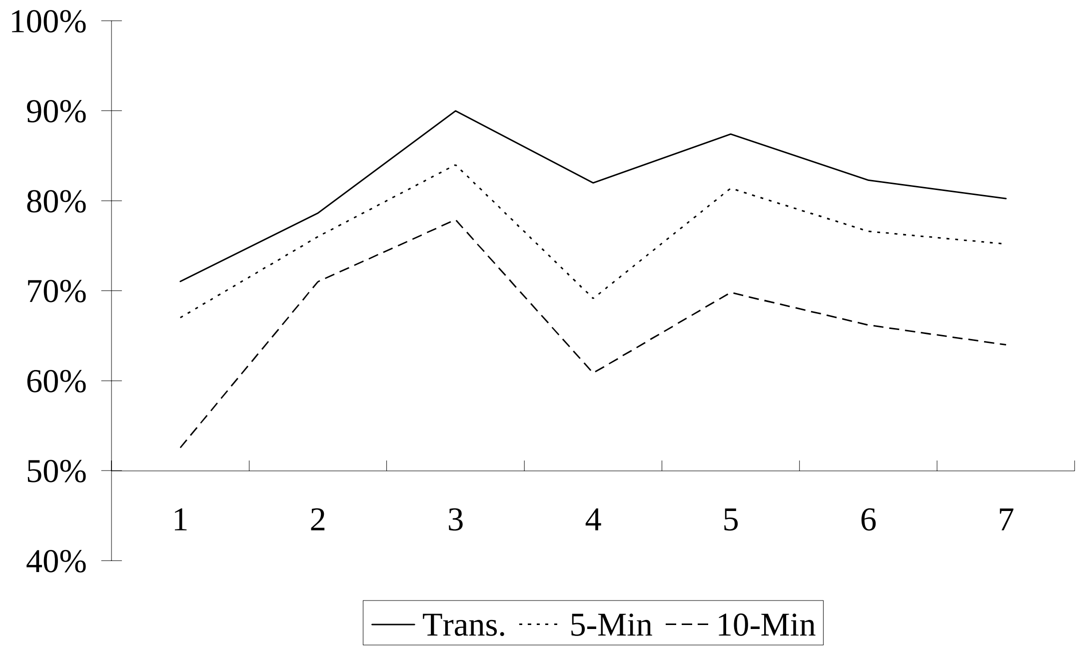
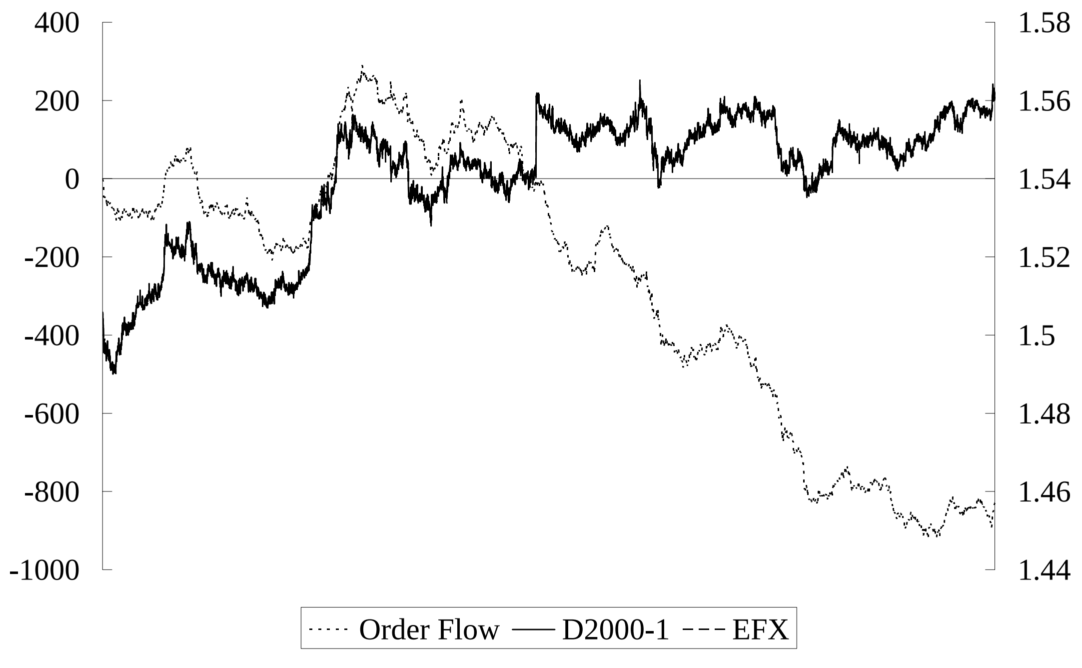


Figure 9. Order flow and prices in GBP_USD (10-M)



Order flow is the accumulated figure for each 10 minute session during our sample period. In order to show the strong co-movement of the order flow and exchange rate level, the order flow is the net dollar seller initiated trade instead of the usual definition. The 10-minute frequency prices of the two data sets literally overlap each other due to long time window.

Figure 10. Order flow and prices in DEM_USD (10-M)

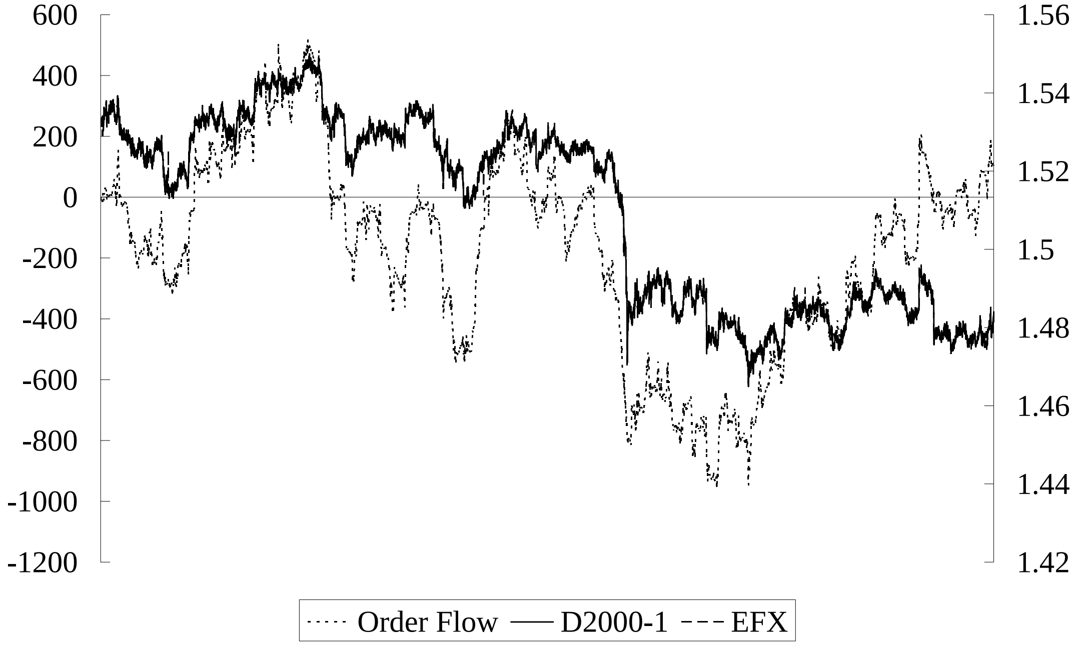
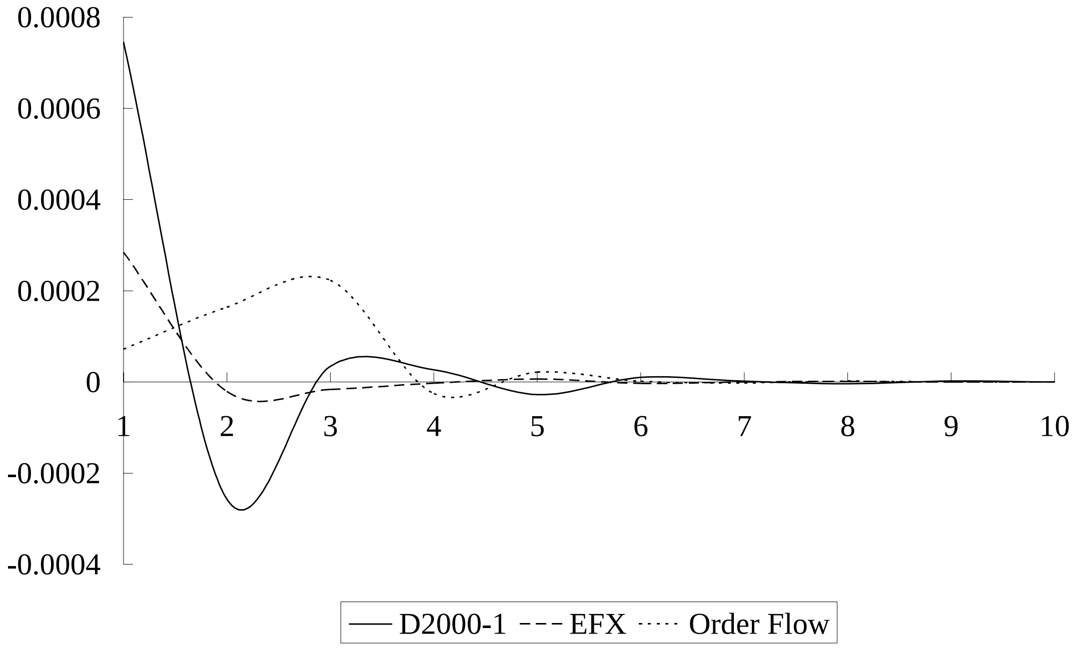


Figure 11. Generalized impulse response to one S.E. D2000-1 shock (GBP_USD)



These are the impulse responses to one S.E. of D2000-1 shock. Each lag stands for 5 minute.

Figure 12. Generalized impulse response to one S.E. EFX shock (GBP_USD)

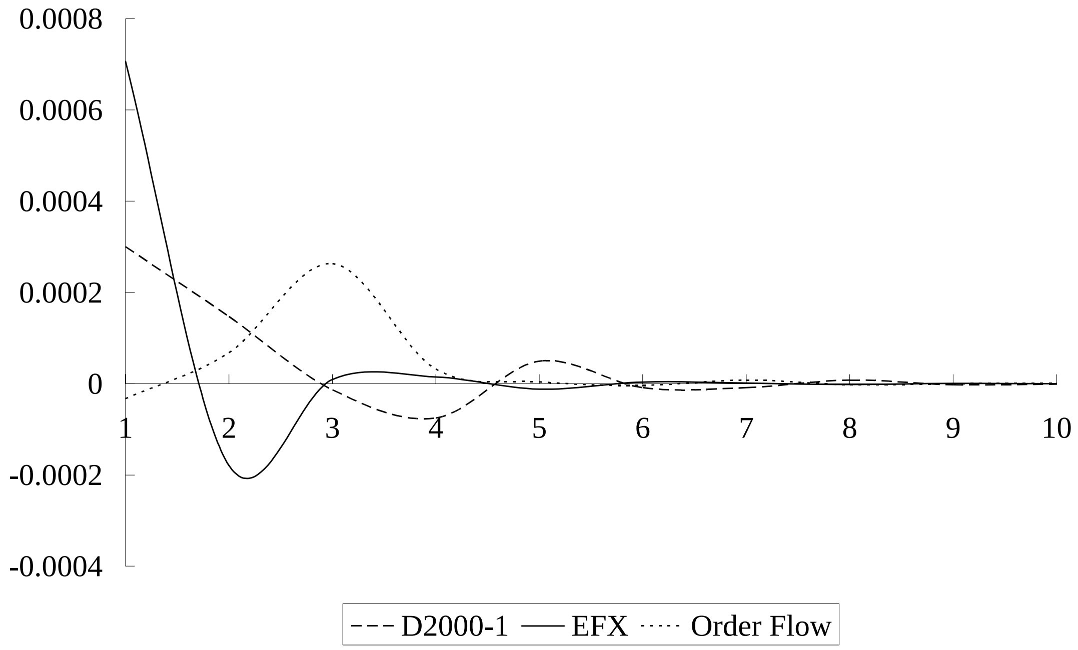


Figure 13. Generalized impulse response to one S.E. order flow shock (GBP_USD)

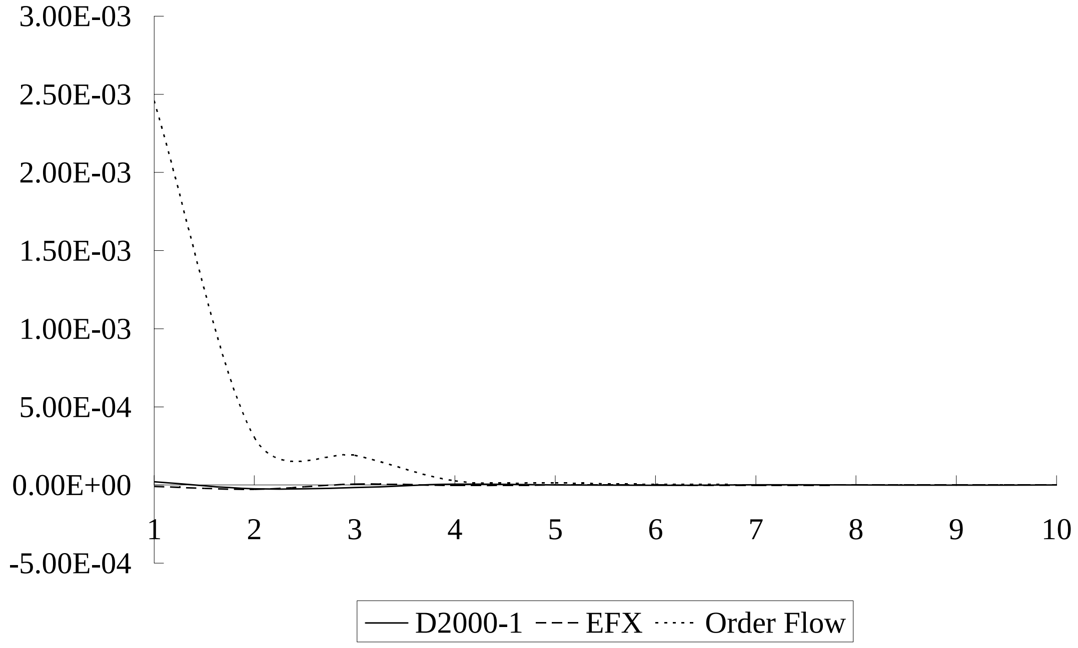


Figure 14. Generalized impulse response to one S.E. D2000-1 shock (DEM_USD)

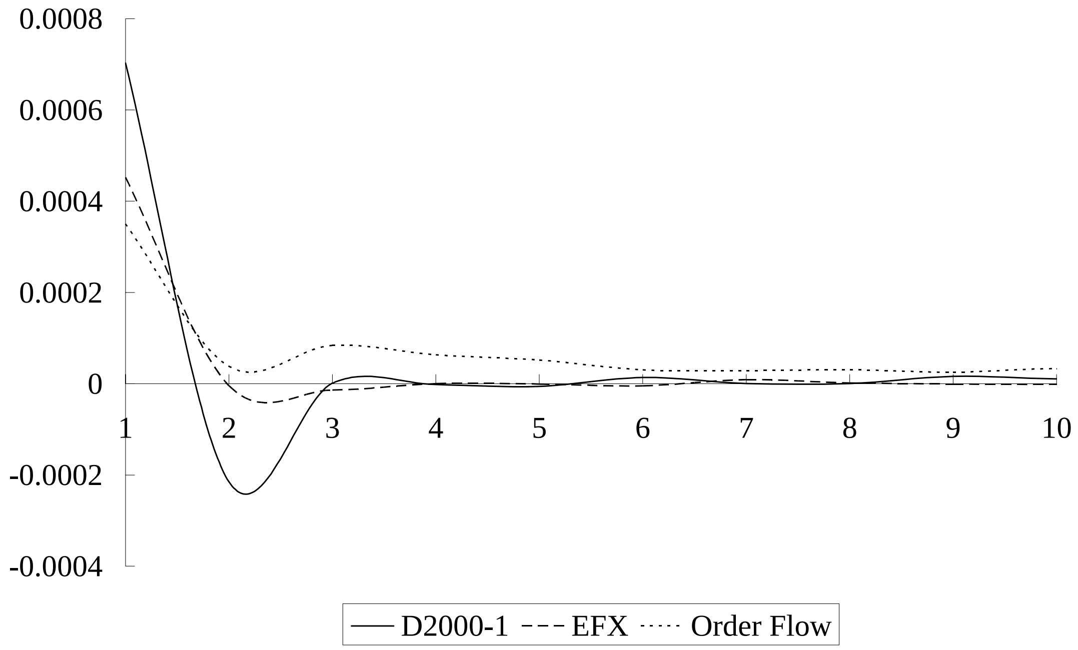


Figure 15. Generalized impulse response to one S.E. EFX shock (DEM_USD)

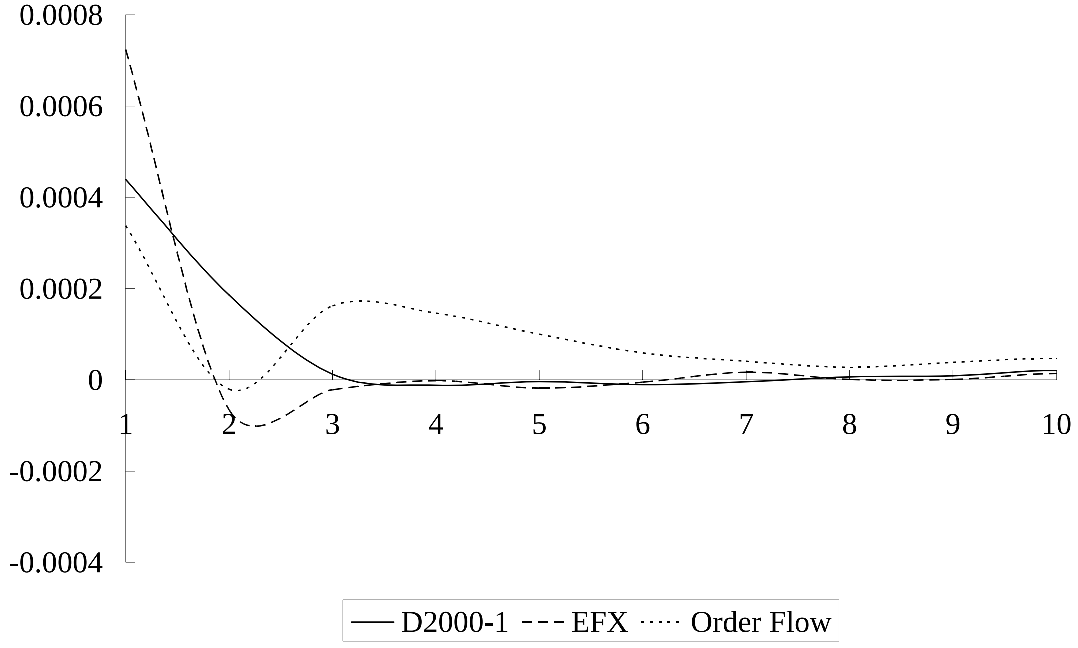
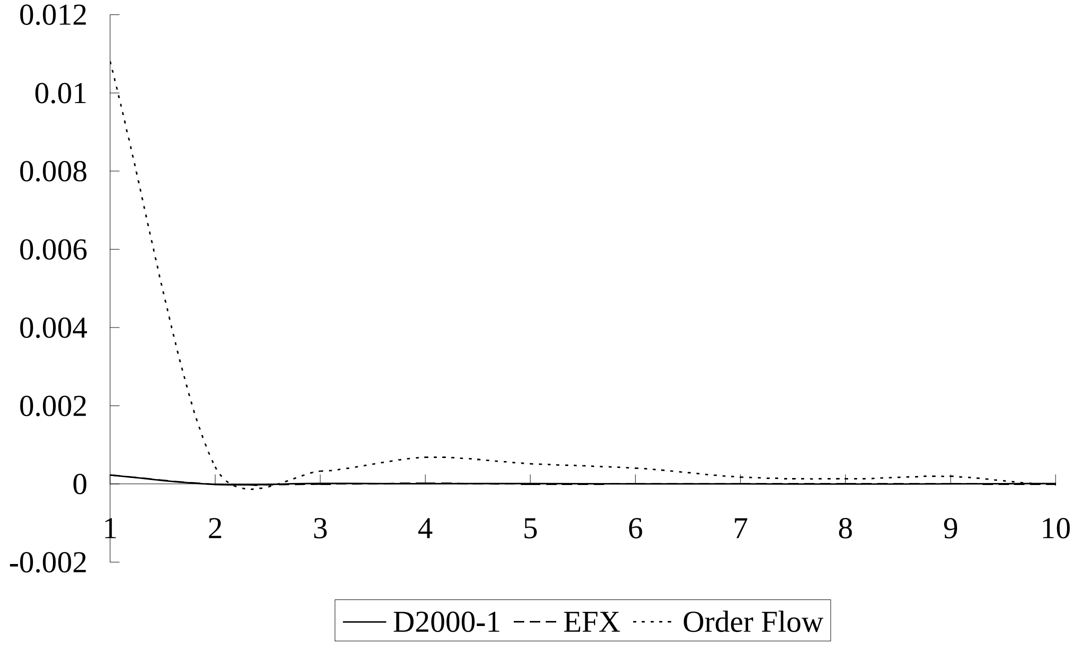


Figure 16. Generalized impulse response to one S.E. order flow shock (DEM_USD)



These are the three time series' impulse response to one unit of order flow shock. Each lag stands for 10 minute.