

Federal Reserve Bank of New York  
Staff Reports

# **Nonlinear Pricing with Competition: The Market for Settling Payments**

Adam Copeland  
Rodney Garratt

Staff Report No. 737  
August 2015



This paper presents preliminary findings and is being distributed to economists and other interested readers solely to stimulate discussion and elicit comments. The views expressed in this paper are those of the authors and do not necessarily reflect the position of the Federal Reserve Bank of New York or the Federal Reserve System. Any errors or omissions are the responsibility of the authors.

## **Nonlinear Pricing with Competition: The Market for Settling Payments**

Adam Copeland and Rodney Garratt

*Federal Reserve Bank of New York Staff Reports*, no. 737

August 2015

JEL classification: L11, L51, L97, E42

### **Abstract**

The multiple payments settlement systems available in the United States differ on several dimensions. The Fedwire Funds Service, a utility that operates a U.S. large-value payments-settlement service, offers the fastest speed of settlement. Recognizing that payments differ in the urgency with which they need to be settled, Fedwire offers banks a decreasing block-price schedule. This approach allows Fedwire to price discriminate, charging high fees for urgent payments and low fees for less urgent ones. We analyze banks' demand for Fedwire Funds given this nonlinear scheme, taking into account competing settlement systems. We show that how banks respond to Fedwire's pricing depends crucially on the need to settle payments quickly. If the urgency for immediate settlement is great enough, banks will respond to marginal price; otherwise, they will respond to average price. We test whether banks respond to marginal or to average price. Our identification comes from exogenous variation in Fedwire's pricing, which results in differential changes in marginal and average price for comparable banks. We find that banks respond to average price.

Key words: nonlinear pricing, marginal versus average pricing

---

Copeland, Garratt: Federal Reserve Bank of New York (e-mail: adam.copeland@ny.frb.org, rodney.garratt@ny.frb.org). The authors thank Emmanuel Dechenaux for comments and Rosa Hayes for providing excellent research assistance. They also thank Ken Isaacson for discussions about the market for settling payments. The views expressed in this paper are those of the authors and do not necessarily reflect the position of the Federal Reserve Bank of New York or the Federal Reserve System.

# 1 Introduction

In the United States, there are a variety of payment-settlement systems. One of these, the Fedwire<sup>®</sup> Funds Service (Fedwire), offers a real-time gross settlement service to financial institutions that hold an account at a Federal Reserve Bank. A main feature that distinguishes Fedwire from its competition is that it provides immediate and final settlement. To raise revenues more effectively and thus facilitate its regulatory mandate to recover costs, Fedwire uses a nonlinear pricing scheme. In particular, Fedwire offers its customers, which are mostly banks, a decreasing block-price schedule. This nonlinear scheme allows Fedwire to price discriminate according to the urgency of the payments. For those payments that require immediate settlement and thus have little alternative to being settled over Fedwire, there is a high fee. For those payments that could be settled with delay and thus could be routed over a competing settlement service, there is a low fee.

We analyze banks' demand for Fedwire given this nonlinear pricing scheme, taking into account the existence of competing settlement services. We present a model of a bank's demand for Fedwire relative to an alternative service that does not offer immediate settlement. Because the total number of payments a bank needs to settle is independent of the pricing policies of settlement services, we focus on how a bank allocates its given stock of payments across services, in particular, Fedwire, which offers immediate settlement, and an alternative, which does not. Payments differ in the immediacy with which they need to be settled, and this characteristic is not observed by the payment-settlement systems. Anecdotally, large-value payments related to financial market activities are often said to be urgent, whereas small-value payments between two households can typically be settled with a delay. We assume that a bank minimizes the cost of settling its payments, taking as given the pricing schedules it faces as well as the varying costs of delaying the settlement of different payments.

The interesting aspect of the decreasing block schedule is that, when the alternative settlement system sets a constant marginal price, the resulting isocost curve is convex and piecewise linear. Given that payments can be perfectly substituted across processing systems, the isoquant is linear. This means that in the absence of costs to delaying urgent payments, a bank will minimize costs by processing all its payments through whichever system has the lowest average cost. However, if a bank faces a schedule of increasingly urgent payments, then the model permits an interior solution whereby a bank may utilize more than one settlement sys-

tem. The implication is that some banks will respond to price changes on the extensive margin and some will respond on the intensive margin, with the allocation of payments between the two processing services, in the latter case, being determined by marginal prices.

Our model, then, presents two fairly different predictions about banks' demand for Fedwire. In our empirical work, we measure which prediction prevails in this market for payment settlement. We draw on two data sources. The first is a history of Fedwire's price schedules, which specify the fees charged to banks based on their current and historical payment volumes. The second is transaction-level data on payments sent over Fedwire. The second data source allows us to observe the payment volumes each bank sent and received on Fedwire. Using both data sources, we can compute the marginal and the average price Fedwire charged each bank.

We cannot use the data on payment volumes to observe whether banks are at corner or interior solutions to their cost-minimization problem. This is because the theoretical model focuses on the set of payments for which a bank has discretion over settlement, whereas in the data, we observe all payments, which includes a mixture of payments over which the bank can and cannot decide how to settle. For example, some clients may demand that a bank settle payments over Fedwire. The theoretical model, though, ties together a bank's response to marginal and average price to whether a bank is at a corner solution. Hence, by estimating whether banks respond to marginal or to average price, we can determine whether banks are responding to Fedwire's price changes on the intensive or on the extensive margin.

We test whether banks respond to marginal or to average price using two approaches. First, our theoretical model predicts that a bank should not select quantities of Fedwire payments near the thresholds that define the three tiers in Fedwire Funds' block-pricing schedule.<sup>1</sup> We look at the empirical distribution of banks' monthly payment volumes and find little evidence that banks avoid the thresholds. Banks, however, may not be responding to marginal price but rather are using an alternative price when deciding how many payments to settle over Fedwire. Alternatively, banks could just be unresponsive to price. Our second test uses a regression approach. Following Ito (2014), we use an encompassing test (Davidson and MacKinnon 1993) to determine whether banks are responding to marginal or to average price. We employ a year-over-year approach and regress changes in a bank's monthly payment volume on changes

---

<sup>1</sup>This is the opposite prediction obtained when agents face an increasing block schedule for pricing. In that case, consumers should bunch at the threshold quantities (see Ito (2014)).

in that bank's marginal and average price, controlling for bank-type and month fixed effects.

Because the marginal and average prices a bank faces in a particular month depend on the bank's concurrent volume of payments sent over Fedwire Funds, there is an endogeneity problem. We therefore take the usual approach of constructing instruments for the changes in price, where the instruments are equal to the year-over-year changes in the marginal and average price, given a fixed volume of payments. This fixed volume is equal to the bank's activity six months prior.

We are able to identify whether a bank responds to marginal or to average price because of a new, bank-specific, pricing scheme that Fedwire introduced in 2011. Under this "incentive-pricing" scheme, banks become eligible for heavily discounted prices after their monthly payment volume crosses a benchmark threshold. This threshold is bank specific, being equal to half the average monthly payment volume a bank achieved over the past five years. The combination of incentive pricing and Fedwire's usual changes to the parameters defining its decreasing block-price schedule results in comparable banks' experiencing increases in average price and decreases in marginal price, and vice versa. Consequently, there is enough variation in marginal and average prices to meaningfully estimate the regression coefficients on both price measures.

We find that banks respond to average rather than to marginal price. In the regression specification that includes both price measures, we find that the coefficient on average price is both economically and statistically significant. In contrast, the coefficient on marginal price is statistically significant but economically tiny. The interpretation is that after conditioning on average price, the inclusion of marginal price has little to no explanatory power over decisions on volume of payments. The estimated coefficients imply that the average elasticity of demand for Fedwire services with respect to average price is -0.491. This result is robust to reestimating the benchmark regression for different subsets of the data or the banks.

The result that banks respond to average price, through the lens of the model, implies that costs of delay for those payments for which the average bank has discretion over settlement do not appear to be large. Thus, for this set of payments Fedwire's advantage over competing services of immediate and final settlement is diminished. In response to Fedwire's decreasing block schedule, then, banks solve their cost-minimization problem by simply comparing the total costs of using Fedwire to those of a competing service.

According to our elasticity estimate, on average the demand for Fedwire's services is inelastic, a result consistent with elasticity estimates for other services provided by utilities: for example, Reiss and White (2005) estimate that households' mean elasticity for electricity is  $-0.39$ . At the bank level, our model predicts a discontinuous response to average price because each bank is calculating at which corner of its minimization problem its costs will be minimized. At the aggregate level, though, we find that a 1 percent increase in average price will decrease the volume of payments volume by 0.491 percent.

Finally, we conclude our analysis by discussing how a two-part tariff would allow Fedwire to better price discriminate across payments of different urgencies. Using the estimated price elasticity, we predict that using a two-part tariff, where the second price is set to Fedwire's marginal cost, would increase the volume of payments over Fedwire by 156 percent. Although this prediction should be considered with care, it highlights the large potential gains of setting each bank's average and marginal price to Fedwire's marginal cost. Furthermore, we argue that implementing a two-part tariff generates welfare gains because it shifts payments that are settled on competing settlement systems with delay to Fedwire, where they are settled immediately.

This paper builds on the literature focused on price discrimination in markets of imperfect competition. Fedwire uses third-degree price discrimination because each bank's price schedule is a function of its historical usage. In addition, Fedwire's decreasing block schedule involves indirect price discrimination over payments' unobserved urgency.<sup>2</sup> Specifically, Fedwire is effectively charging high fees to urgent (and thus price-inelastic) payments and low fees to nonurgent (and thus price-elastic) payments. A main innovation of this paper is linking a bank's response to Fedwire's decreasing block schedule to the unobserved urgency of that bank's payments. In particular, a bank with sufficiently urgent payments will respond to the marginal price set by Fedwire; otherwise, the bank will respond to the average price. Given the rich payment data to which we have access and because of exogenous variation in Fedwire's nonlinear pricing, we are then able to infer whether banks are responding to average or to marginal price and thus learn about the (unobserved) urgency with which payments need to be settled.

Within the price discrimination literature, a number of papers are focused on nonlinear pricing and utilities. In our paper, we study a utility that provides services to financial institutions,

---

<sup>2</sup>Stole (2007) and Lambrecht, Seim, Vilcassim, Cheema, Chen, Crawford, Hosanagar, Iyengar, Koenigsberg, Lee, Miravete, and Sahin (2012) are two reviews of price discrimination.

whereas most papers in this literature focus on utilities that provide services to households, such as water (Olmstead, Hanemann, and Stavins 2007) and energy (Reiss and White 2005). Our econometric approach is similar to Ito (2014), a paper that explores whether households in California respond to the average or to the marginal price for electricity. A main difference between our paper and Ito's is that we develop a theoretical model that details how responding to average price is a fully rational cost-minimizing response by banks. In contrast, Ito (2014) frames a household's response to average price within the behavioral economics literature.

A last major difference is our welfare analysis. Typically, the introduction of nonlinear pricing affects welfare by changing overall output. In contrast, given the low fees associated with settling payments, we assume that the number of payments that need to be settled is independent of pricing.<sup>3</sup> Our welfare analysis, then, is not based on the usual output metric of total payment volumes but rather on the speed of settlement.

In the next section, we provide background information on Fedwire and alternative competing payment systems. We then introduce the data in section 3 and describe our theoretical model in section 4. We present our empirical results in section 5. Finally, we discuss the empirical results and their policy implications in section 6 and then conclude in section 7.

## **2 Institutional details**

As part of the normal course of business, financial institutions are required to make payments to one another to settle a variety of obligations. These obligations result from both the bank's own financial activity and those of its clients. U.S. financial institutions can use a number of payment systems to settle their U.S. dollar obligations. Our focus is on the Fedwire<sup>®</sup> Funds Service (Fedwire), a real-time gross settlement payment system operated by the Federal Reserve Banks, which processes more than \$3 trillion worth of payments each day.

To be eligible to send and receive payments on Fedwire, a financial institution must hold an account at a Federal Reserve Bank. When using Fedwire, institutions are thus using their reserves held at the Federal Reserve, or central bank money, to settle their obligations. Fedwire is a credit transfer service, whereby the participant sending the payment originates the trans-

---

<sup>3</sup>We assume that the volume of derivative or foreign exchange trades between financial institutions is independent of the fees charged by Fedwire or its competitors.

fer by requesting Fedwire to debit its own account and credit another participant's account. Payments over Fedwire are immediate, final, and irrevocable.

Fedwire's price schedule is detailed in the following section, but prices range from 2 to 65 cents per payment sent or received. In a consideration of these prices and what information they might contain about banks' willingness to pay, it is important to note that regulation demands that Fedwire price its service so as to recover costs, not to maximize profits.<sup>4</sup> Finally, Fedwire does not offer a liquidity-savings mechanism. When a bank initiates a transfer over Fedwire, the bank needs to have reserves in its account or be eligible to receive the necessary amount of intraday credit from the Federal Reserve.<sup>5</sup>

Financial institutions have a number of alternative payment systems they can use to settle obligations. However, just two are relevant as likely substitutes for Fedwire. These are the Clearing House Interbank Payments system and the Automated Clearing House system. Although these payment systems differ along many dimensions, for the purposes of this paper it is useful to focus on three characteristics as points of comparison: the speed of settlement, the price charged to send a payment, and the availability of liquidity-savings mechanisms.

The speed of settlement is the time it takes between the initiation of a payment and its settlement. For example, debit card transactions typically take at least a day to settle, whereas a cash payment provides immediate settlement. The price characteristic is simply the price the payment-settlement system charges banks for using its service. Finally, the last characteristic provides a measure of how much liquidity is needed to settle obligations on a specific payment system. As detailed below, some payment systems offer tools for reducing the amount of liquidity banks are required to provide.

The Clearing House Interbank Payments system (CHIPS) is the closest competing service to Fedwire Funds. CHIPS is a bank-owned, privately operated electronic payment system.<sup>6</sup> A main difference between CHIPS and Fedwire is that payments sent over CHIPS are netted. As a result, an institution participating in CHIPS needs to settle its net obligation only against other institutions participating in CHIPS. The netting service that CHIPS provides is a liquidity-

---

<sup>4</sup>For a detailed analysis of how these costs are determined, see Green, Lopez, and Wang (2003).

<sup>5</sup>As part of its role as a central bank, the Federal Reserve provides intraday credit to ensure the smooth functioning of payment and settlement systems. For more detail, see the Board of Governors of the Federal Reserve System website on intraday credit policies, [http://www.federalreserve.gov/paymentsystems/psr\\_policy.htm](http://www.federalreserve.gov/paymentsystems/psr_policy.htm).

<sup>6</sup>For more information, see <https://www.chips.org/home.php> and the Fedpoint article at <http://www.newyorkfed.org/aboutthefed/fedpoint/fed36.html>

savings mechanism. This tool allows an institution to settle a potentially large gross amount of obligations with a small net amount of cash. Compared to Fedwire, then, CHIPS offers settlement with lower liquidity demands on banks. This advantage in settling a net amount comes with the cost that payments are not guaranteed to settle immediately. CHIPS runs its multilateral netting algorithm at regular intervals throughout the day, and most payments settle very quickly. However, there is a chance that a payment will not be netted and so settled quickly, or even at all. At 5 p.m., when CHIPS closes, a bank's remaining payments are either settled individually, in gross terms, or these payments are released back to the bank to be settled outside of CHIPS.<sup>7</sup>

CHIPS's pricing schedule is not publicly available. Because CHIPS is owned by its members, the price a member pays for using CHIPS is roughly tied to the costs of running it. Our understanding is that CHIPS and Fedwire have somewhat similar operating costs, implying that banks face roughly similar average prices on CHIPS and Fedwire.

CHIPS is a substitute for Fedwire only for payments being sent among CHIPS members. As of October 21, 2013, 50 institutions were participating in CHIPS, ranging from large U.S. banks such as JPMorgan Chase Bank and Citibank, to foreign banks with a small U.S. presence, such as Bangkok Bank Public Company and the State Bank of India. Although 50 is a tiny fraction of the roughly 7,000 Fedwire participants, this group of institutions accounts for a substantial number of payments. In 2012, the average daily number of payments settled over Fedwire was 524,452 compared to 386,972 over CHIPS.<sup>8</sup>

The most likely competitor to Fedwire for non-CHIPS members is the Automated Clearing House system (ACH). However, ACH offers a substantially different settlement service from Fedwire. The ACH system is a network through which financial institutions send each other batches of credit or debit transfers. ACH acts as a central clearing facility, accumulating payments from its members. Once a day, ACH runs a batch process to settle the stock of accumulated payments. Because this approach has significant economies of scale, the price of sending a payment over the ACH service is at least an order of magnitude smaller than

---

<sup>7</sup>In this latter case, a bank is likely to use Fedwire to send the payment and settle its obligation. Fedwire closes at 6:30pm, although the deadline for initiating transfers over Fedwire for the benefit of a third party is 6pm.

<sup>8</sup>Aggregate Fedwire payments volume data can be found on the Board of Governors website, at [http://www.federalreserve.gov/paymentsystems/fedfunds\\_data.htm](http://www.federalreserve.gov/paymentsystems/fedfunds_data.htm). Aggregate CHIPS payments volume are published by CHIPS, and can be found at <https://www.theclearinghouse.org/payments/chips/helpful-info>.

Table 1: Comparison of Fedwire, CHIPS, and ACH

Payments System	Timing of settlement	Pricing of services	Liquidity-savings mechanism
Fedwire Funds	Immediate	Between 2 and 65 cents	No
CHIPS	Small to large delay	Not publicly available	Yes
ACH	Large delay	Fraction of a cent	No

Note: CHIPS is Clearing House Interbank Payments System; ACH is Automated Clearing House.

Fedwire’s. A payment sent over ACH, however, is substantially delayed, typically settling the next day.<sup>9</sup> Similar to Fedwire, ACH does not offer a liquidity-savings mechanism.

A summary of the features of the three competing payment systems is provided in Table 1. Across the three payment systems, the speed of settlement declines as we move from Fedwire to CHIPS to ACH. Our understanding is that the average price across Fedwire and CHIPS is of the same magnitude, whereas ACH offers a much lower price for its service. Finally, only CHIPS offers banks a liquidity-savings mechanism to settle payments. Overall, then, when deciding how to settle an obligation, banks can choose to pay a relatively high price for immediate settlement over Fedwire, or they can choose to delay the payment and pay a lower price. CHIPS offers a small delay in settlement and a small decrease in price (reflecting the lower liquidity demands on banks), and ACH offers a substantial delay in settlement with a substantial decrease in price.

A last important feature of this part of the payment landscape is the ability of banks to access Fedwire through other banks. Rather than establish its own account on Fedwire, a bank has the option of establishing a correspondent account at another bank that has direct access to Fedwire and using that bank to send and receive payments over Fedwire. Given the fixed costs of setting up an account and Fedwire’s price schedule, the details of which are laid out in the next section, it can be economical for a bank to avoid setting up a Fedwire account and simply establish a correspondent account at another bank. A banker’s bank is an example of

<sup>9</sup>ACH has begun to offer a new service in which certain types of payments can be settled the same day. Nevertheless, there is still a large delay in the settlement of the payment relative to Fedwire. Furthermore, there are other significant product differences between Fedwire and ACH; for example, unlike in Fedwire, ACH payments can be reversed. For more information on ACH’s faster settlement service, see [http://www.frbservices.org/serviceofferings/fedach/sameday\\_service.html](http://www.frbservices.org/serviceofferings/fedach/sameday_service.html).

a financial institution that offers a variety of wholesale banking services to banks, including access to Fedwire.

### 3 Data

We have two data sources. The first is information on the pricing of Fedwire’s services over time, which is publicly available. The second source is transaction-level data on payments sent over Fedwire.

#### 3.1 Pricing data

Banks are charged monthly for using Fedwire based on their payment volume (both sent and received). Over the past two decades, Fedwire has dramatically changed its pricing schedule. Until 1999, Fedwire used a flat pricing scheme whereby banks were charged a constant price to send or receive a payment. Starting in 1999, however, Fedwire embraced nonlinear pricing and began to use a three-tier decreasing-block schedule. Let  $x^i$  denote the quantity of payments a bank  $i$  has sent or received over Fedwire in a given month. Under Fedwire’s three-tier pricing scheme, the cost of using Fedwire to settle these payments,  $F$ , is given by

$$F(x^i) = \begin{cases} Ax^i & \text{if } x^i < t_1, \\ At_1 + B(x^i - t_1) & \text{if } t_1 \leq x^i < t_2, \\ At_1 + B(t_2 - t_1) + C(x^i - t_2) & \text{if } x^i \geq t_2, \end{cases} \quad (1)$$

where  $A > B > C > 0$  and  $(t_1, t_2)$  are thresholds defining the three payment tiers. Notice that  $F$  is a decreasing step function in  $x^i$ .

From 1999 to 2010, Fedwire maintained its three-tier, decreasing block schedule structure but changed the parameters of its pricing scheme. Each time, Fedwire has adjusted its pricing schedule on an annual basis, with the new pricing schedule being publicly announced in the fall and implemented at the start of the calendar year. For the most part, only the tiered-pricing parameters,  $(A, B, C)$ , were changed, although in 2008 and 2010 the thresholds defining the tiers,  $(t_1, t_2)$ , were increased (see Figure 1 for a visualization of these parameter changes).

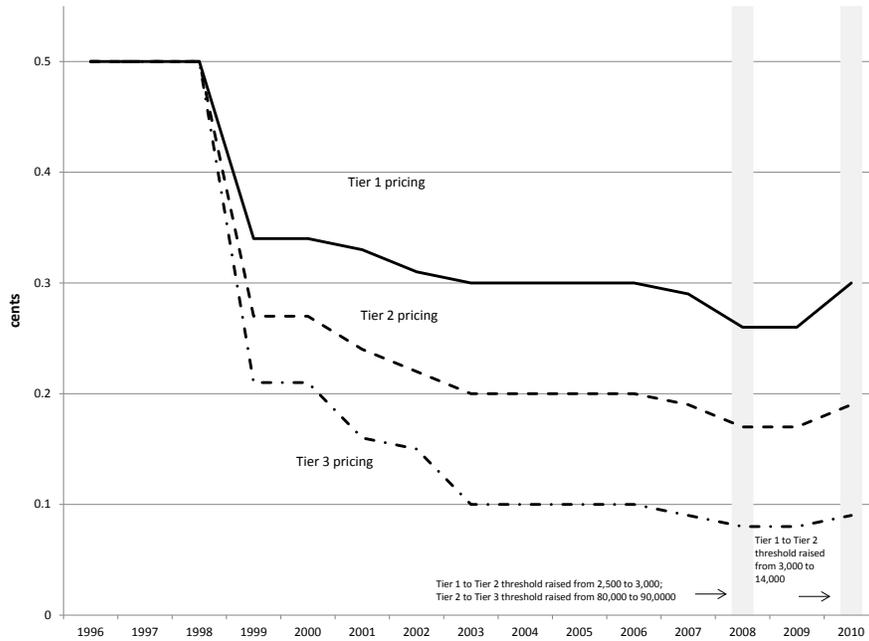


Figure 1: Fedwire Funds Service Three-Tier Pricing Schedule from 2000 to 2010

In 2011, Fedwire introduced a new feature to its pricing scheme called incentive pricing. Fedwire offered two sets of three-tier pricing, a “regular” and an “incentive-pricing” set. Fedwire also assigned *each* financial institution a benchmark threshold,  $T^i$ . Until an institution’s accumulated payment volume (payments sent plus payments received) within a month reached its benchmark threshold, a financial institution faced the regular set of prices. After reaching its benchmark threshold, the institution faced the lower, incentive-pricing set of prices for all subsequent payments sent and received. The benchmark threshold was set to half an institution’s

five-year average monthly volume.<sup>10</sup> The incentive-pricing scheme is

$$I(x^i, T^i) = \begin{cases} F(x^i) & \text{if } x^i \leq T^i, \\ J(x^i) + H(T^i) & \text{if } x^i > T^i, \end{cases} \quad (2)$$

where  $F$  is defined in equation (1), and

$$J(x^i) = \begin{cases} \hat{A}x^i & \text{if } x^i < t_1, \\ \hat{A}t_1 + \hat{B}(x^i - t_1) & \text{if } t_1 \leq x^i < t_2, \\ \hat{A}t_1 + \hat{B}(t_2 - t_1) + \hat{C}(x^i - t_2) & \text{if } x^i \geq t_2, \end{cases}$$

$$H(T^i) = \begin{cases} (A - \hat{A})T^i & \text{if } T^i < t_1, \\ (A - \hat{A})t_1 + (B - \hat{B})(T^i - t_1) & \text{if } t_1 \leq T^i < t_2, \\ (A - \hat{A})t_1 + (B - \hat{B})(t_2 - t_1) + (C - \hat{C})T^i - t_2 & \text{if } T^i \geq t_2. \end{cases}$$

If a bank's payment volume in a month does not exceed its benchmark volume,  $x^i \leq T^i$ , then the bank faces the regular set of prices as defined by  $F$ . If a bank's payment volume exceeds its benchmark volume, then the new set of discounted prices apply, as defined by  $J$  and  $H$ .  $J$  has the same decreasing-block structure as  $F$ , except that the marginal prices  $\{A, B, C\}$  are replaced with the incentive price discounts of  $\{\hat{A}, \hat{B}, \hat{C}\}$ , where  $\hat{A} > \hat{B} > \hat{C}$  and  $C > \hat{A}$  (see Figure 2 for an illustration of these differences in prices for 2011, 2012, and 2013).  $H$  captures the fact that banks face the "regular" set of prices of  $\{A, B, C\}$  up until their total accumulated payments on Fedwire in a month exceed their threshold level,  $T^i$ .

To illustrate the incentive-pricing scheme, consider the case of a bank assigned a benchmark threshold of 25,000, sending and receiving a total of 50,000 payments for a month in 2011 (refer to the black solid and dashed lines in Figure 2 for 2011 pricing). The price schedule dictates that the bank pays  $A = 52$  cents on each of the first 14,000 payments (this is the regular tier one price). The bank then pays  $B = 23$  cents on each of the next 11,000 payments

---

<sup>10</sup>Fedwire's methodology is to compute each bank's daily average over the past five years. A bank's benchmark threshold is then equal to this daily average multiplied by the number of business days for that particular month divided by two. If fewer than five years of data are available, Fedwire computes the daily average based on the data that are available.

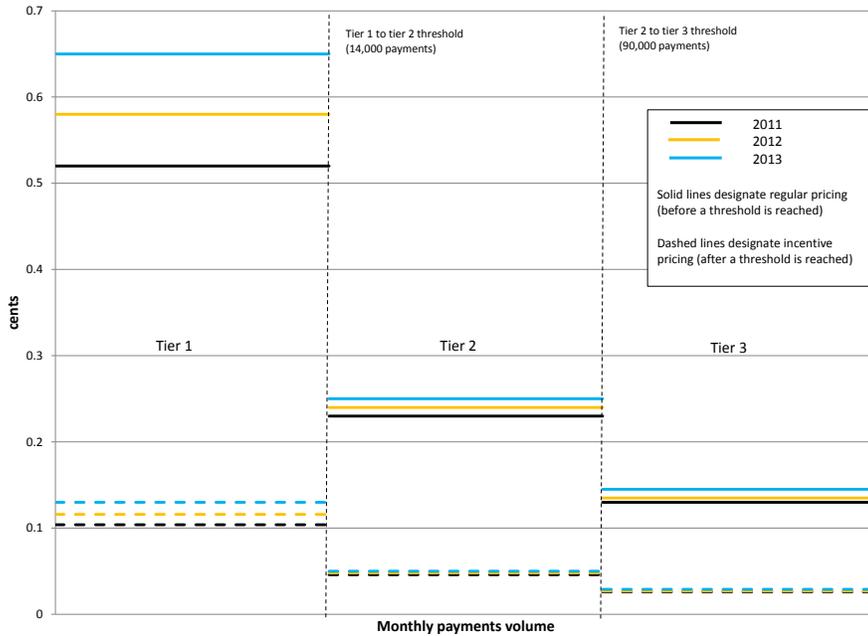


Figure 2: Fedwire Incentive-Pricing Schedule, 2011, 2012, and 2013

(the regular tier two price). The incentive-pricing discount then applies to all the remaining payments made, so that the bank pays  $\hat{B} = 4.6$  cents on the final 25,000 payments made over the remainder of the month (the incentive-pricing tier two price).

An important feature of the new scheme is its individualized nature. Because the benchmark threshold each bank faces is a function of that bank's five-year history of Fedwire usage, each bank faces its own nonlinear pricing scheme. As explained in more detail in section 5, this feature of institution-specific pricing is crucial to our identification strategy in our empirical work.

The motivation for the new pricing scheme is to encourage institutions to route more payments over Fedwire. This is accomplished by dropping the marginal price that banks face. In the example above, a bank's marginal price fell from 23 cents to 4.6 cents, an 80 percent de-

Table 2: Types of Institutions Using Fedwire Services

Institution type	Volume	Percent of total	Number
Domestic commercial bank	679,146,501	86.4	4,435
Foreign banking organization	44,510,686	5.7	150
Savings bank	24,147,438	3.1	521
Credit union	22,039,690	2.8	1,124
Savings and loan association	3,449,174	0.4	248
All others	12,514,403	1.6	430
Total	785,807,892	100	6,908

Note: Institution types are listed from largest to smallest by payments volume. Volume is the number of payments sent and received over Fedwire from 2011 to 2013. Number is the number of institutions of each type that were active on Fedwire in the first quarter of 2011. Foreign banking organizations includes both U.S. agencies of a foreign bank and U.S. branches of a foreign bank.

crease. Furthermore, the design of the new pricing scheme reduces the volatility in revenues, allowing Fedwire to better forecast its annual revenues and appropriately set prices to recover its costs.

### 3.2 Payment data

We combine the information on pricing with detailed confidential data on payments sent over Fedwire. We focus on the time period from January 1, 2011, to December 31, 2013, because starting in 2011 Fedwire introduced its new incentive-pricing scheme. For each payment, we observe which institution sent the payment, which institution received the payment, the amount of the payment, and the time when the payment was sent. With these data, we calculate each institution's total monthly volume and incentive-pricing threshold.<sup>11</sup> With these volume statistics, we can use Fedwire's pricing schedule to compute each institution's average and marginal price for each month in our sample.

Using bank identifiers, we are able to merge information on bank characteristics into our volume and price statistics. This is particularly helpful because a variety of types of institutions use Fedwire (see Table 2). Our theoretical model is focused on institutions that need to

<sup>11</sup>Because this threshold is based on a bank's five year history, we actually draw on data from as far back as 2005.

Table 3: The Distribution of Monthly Payment Volumes

	Percentile							
	1	10	25	50	75	90	99	99.95
Average monthly volume	1	5	39	148	455	1,515	27,460	1,483,387

Note: The data cover January 2011 to December 2013. The mean average volume is 3,356.

send and receive payments, where these payments differ in the cost of delay. Depository institutions best fit this characterization, given that a main benefit of establishing a deposit is its use as a means of payment. Consequently, for our analysis we focus on domestic commercial banks, foreign banking organizations, savings and loan associations, savings banks, and credit unions—a group of financial institutions we label “banks” and that account for 98.4 percent of all payments sent and received over Fedwire from January 2011 to December 2013.<sup>12</sup>

There is a tremendous range in the monthly payment volume across banks. As depicted in Table 3, half of all banks sent and received fewer than 148 payments a month, and nine-tenths sent and received fewer than 1,515 payments a month. Although the number of small banks far outweighs that of large banks, the large banks are Fedwire’s main revenue source. Indeed, banks above the 90th percentile in volume account for almost 80 percent of total revenue (see Table 4).

## 4 Model

We now present a theoretical model that seeks to explain how banks route payments over different settlement systems. We take as given that a bank is obligated to send and receive a certain number of payments in a particular month. In particular, we do not believe that the cost of payment processing affects the decisions of customers to undergo economic activity that generates the need for a payment. This makes sense given that, regardless of the choice of payment system, the cost of individual payments is measured in cents. However, we do not need to consider all payments made by a bank. For most banks, a certain fraction of payments will be sufficiently urgent in nature that Fedwire is the only option. This is true for time-critical

<sup>12</sup>Foreign banking organizations includes both U.S. agencies of foreign banks and U.S. branches of foreign banks.

Table 4: Contribution to Total Revenue by Monthly Payment Volume

Monthly payment volume (percentile)	Contribution to total revenue (percent)	Cumulative contribution to total revenue (percent)
1	0.47	0.47
10	0.88	1.34
25	2.12	3.47
50	4.51	7.98
75	6.80	14.77
90	8.14	22.92
99	25.75	48.67
99.95	35.40	84.06

Note: Revenues are the authors' calculations and reflect fees from banks' use of Fedwire as detailed in the paper, as well as monthly participation, large-value and late-in-the-day fees. These additional fees are discussed in appendix A. Banks are grouped by their monthly payment volume, where the percentiles mark the upper bound of a bin. A bank is in the 90th percentile bin if its monthly volume is less than the 90th percentile, but greater than the 75th percentile. The cumulative contribution does not sum to 100 percent because banks with monthly volumes above the 99.95th percentile are not included.

payments such as those sent to Continuous Linked Settlement (CLS) to settle foreign exchange obligations or those that need to be made before certain markets close. Likewise, regularly scheduled payments, such as direct deposits, are almost exclusively settled over ACH, and the economics of settling such payments make it improbable that a bank would ever use Fedwire to settle them. Furthermore, a client may demand that a bank use a specific payment system when settling that client's obligation, and debit transfers must go through ACH. Our focus is therefore on the bank's decision problem of determining how to settle the remaining "flexible" payments. We denote the quantity of flexible payments over which bank  $i$  has discretion by  $x^i$ .

We simplify the analysis by assuming that banks consider only Fedwire and one alternative for their flexible payments. For CHIPS members, it is natural to assume the alternatives are Fedwire and CHIPS. For non-CHIPS members, the alternatives must be Fedwire and ACH. Let  $x_F^i \leq x^i$  denote the number of payments that bank  $i$  settles on Fedwire. Abstracting from the new incentive-pricing scheme for now, the settlement cost from such a decision is  $F(x_F^i)$ , where  $F$  is defined in equation (1).

For a given bank, the cost of settling payments through its alternative payment system

depends on that system's pricing schedule and the delay cost imposed on the bank by selecting that option. The delay cost is a function of the urgency of the individual payment and the delay time. We assume that all payments settled through an alternative payment system will be delayed for the same amount of time but that each bank  $i$ 's profile of payments  $x^i$  will have different urgency characteristics. We therefore denote bank-specific cost functions for the alternative processor, which we denote by  $G^i$ . We assume that  $G^i(x_A^i) = ax_A^i + g^i(x_A^i)$ , where  $g^i$  is strictly convex and  $x_A^i$  denotes the quantity of payments processed by bank  $i$  using the alternative payment system. Our interpretation of this function is that banks that use the alternative payment system are charged a constant marginal price,  $a$ , and incur an urgency cost that increases at an increasing rate. The former assumption is valid for ACH, and we believe it is reasonable for CHIPS, given what we know about its operating structure. The latter assumption is justified on the grounds that the bank will optimally allocate the least urgent payments to the alternative system first. Hence, each additional payment added to the alternative system will be more urgent than the last and hence will add to the total cost at an increasing rate.

Given our assumption that bank  $i$  faces a fixed quantity  $x^i$  of flexible payments that it seeks to settle, the cost-minimization problem of bank  $i$  is

$$\begin{aligned} \min_{x_F^i, x_A^i} F(x_F^i) + G^i(x_A^i) \\ \text{subject to } x_F^i + x_A^i = x^i. \end{aligned}$$

If all flexible payments are nonurgent, the cost-minimization problem yields only corner solutions. This can be seen graphically in Figure 3, where we plot the bank's isoquant and isocost curves. The isoquant is linear with a slope of -1, because each payment is settled individually regardless of the settlement system. The isocost curve is convex and piecewise linear, reflecting the decreasing-block schedule of Fedwire as well as the linear pricing of the alternative settlement scheme. As a result, the isocost curve which minimizes the bank's settlement costs will necessarily result in a bank processing all its payments through the alternative processor if  $F(x^i) > ax^i$  or through Fedwire if  $F(x^i) < ax^i$ . In other words, a bank will process all its payments using the processor that has the lowest average cost. This result, which is depicted in the two panels of Figure 3, is in stark contrast to the utility-maximization problems studied

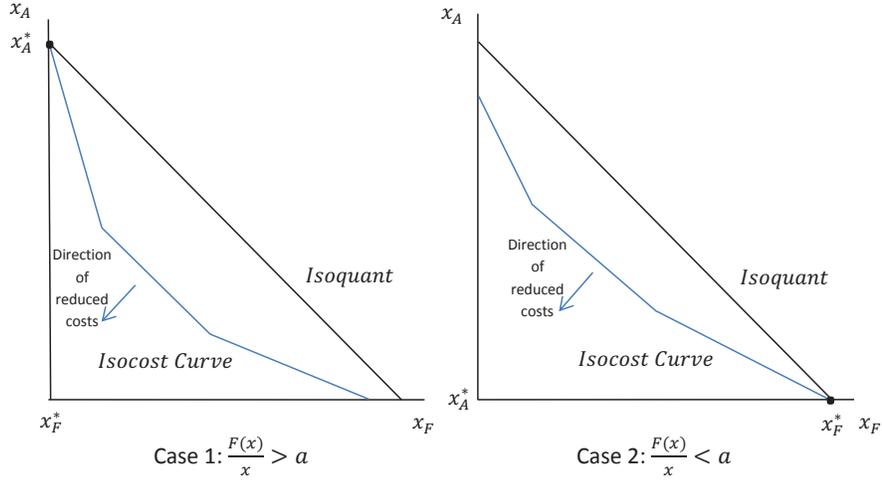


Figure 3: Isocost Curve with No Urgent Payments

by Ito (2014) and others in the context of utilities pricing, which employ an increasing-block schedule. An increasing-block schedule preserves the convexity of the budget set and permits interior solutions.

When flexible payments have differing degrees of urgency (keeping in mind that none are sufficiently urgent to require settlement over Fedwire), interior solutions to the cost-minimization problem are possible, and in such cases the optimal allocation of payments across the two payment systems will be obtained at a point where the ratio of marginal costs equals 1. This situation is depicted in Figure 4 (we drop the superscripts  $i$  in the figure). At any point  $(x_A^i, x_F^i)$  the slope of the isocost curve is given by  $A/(a + g'(x_A^i))$  if  $0 < x_F^i < t_1$ ,  $B/(a + g'(x_A^i))$  if  $t_1 < x_F^i < t_2$  and  $C/(a + g'(x_A^i))$  if  $x_F^i > t_2$ . In Figure 4, the solution  $(x_A^{i*}, x_F^{i*})$  is defined by the equations

$$\frac{C}{a + g'(x_A^{i*})} = 1,$$

$$x_A^{i*} + x_F^{i*} = x^i.$$

The implication of the analysis of urgent payments is that banks may minimize costs by dividing their flexible payments between two systems and the allocation of payments to Fed-

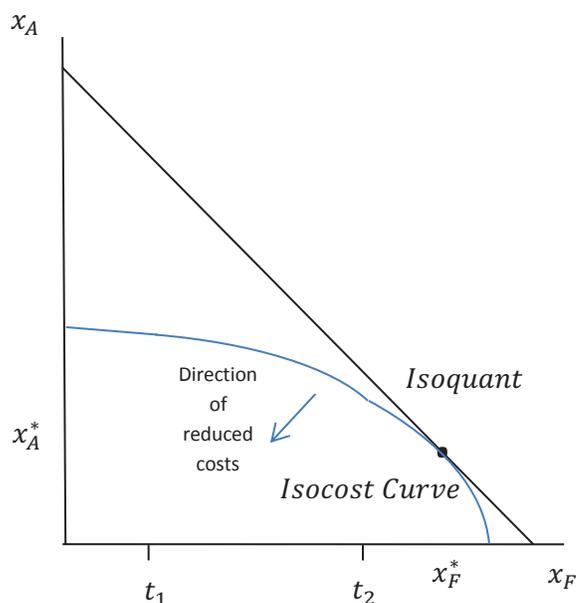


Figure 4: Isocost Curve with Urgent Payments.

wire should vary with the marginal price it charges. Banks with few or no urgent payments will base their processing decision on a comparison of average costs. Because the banks are at a corner solution, the demand for Fedwire’s processing services will not vary continuously in either the marginal or the average cost of Fedwire payments.

Interestingly, the cost-minimization problem has a solution characteristic that is the opposite of the utility-maximization problem with increasing block prices. In the latter, demand should spike around kinks. In the former, small changes in price should lead to discontinuous jumps in usage near the price thresholds. No bank should operate right at a quantity threshold.

The above results continue to hold if we replace Fedwire’s three-tier pricing scheme,  $F$ , with the new incentive-pricing scheme,  $I$ . The incentive-pricing scheme generates more kinks in the isocost curve, but preserves its convex and piecewise linear shape.

## 5 Empirical analysis

In our data, we do not observe which payments are flexible. Hence, when we observe a specific bank sending and receiving payments over Fedwire, we do not know if that bank has chosen to settle all, some, or none of its flexible payments on Fedwire. The data on bank volumes over Fedwire, then, are not enough to tell us whether banks are facing an interior solution to their cost-of-settlement minimization problem.

The theoretical model, however, ties together banks' responses to marginal and average prices with their settlement decisions on flexible payments. In particular, if we estimate that a bank is responding to marginal price, then we can infer that the bank faces an interior solution to its settlement problem (as depicted in Figure 4). If the bank is responding to average price, then we can infer that the bank is at a corner solution, meaning that all or none of its flexible payments are being settled on Fedwire (as depicted in Figure 3).

In this section, we estimate whether banks that use Fedwire are responding to marginal or to average price, with the ultimate goal of understanding their allocations of flexible payments. We present two results. First, we test the implication of the theoretical model that banks responding to marginal price should not cluster around the thresholds that define the payment tiers. Second, we present the main empirical result of the paper—an encompassing test to determine whether banks respond to marginal or to average price.

### 5.1 Nonbunching at kink points of price schedules

The theoretical model predicts that banks that have an interior solution to their cost-minimization problem, and so respond to marginal price, will not choose to make and receive a total number of payments near a tier threshold. In our sample period, these tier thresholds are set at 14,000 and 90,000 payments. We plotted banks' monthly volumes around these thresholds over 2010 to 2013. For the lower threshold of 14,000, we find little evidence that banks are not bunching at the threshold. Graphical evidence is provided in Figure 5, which illustrates that for 2012 banks' monthly volumes are relatively smooth over the range from 13,000 to 15,000. (This figure looks quite similar for the other years in our sample.) For the higher threshold of 90,000, it is difficult to determine whether banks are not bunching because there are so few observations. For example, in 2012, over the range of 87,000 to 92,000 payments, there are

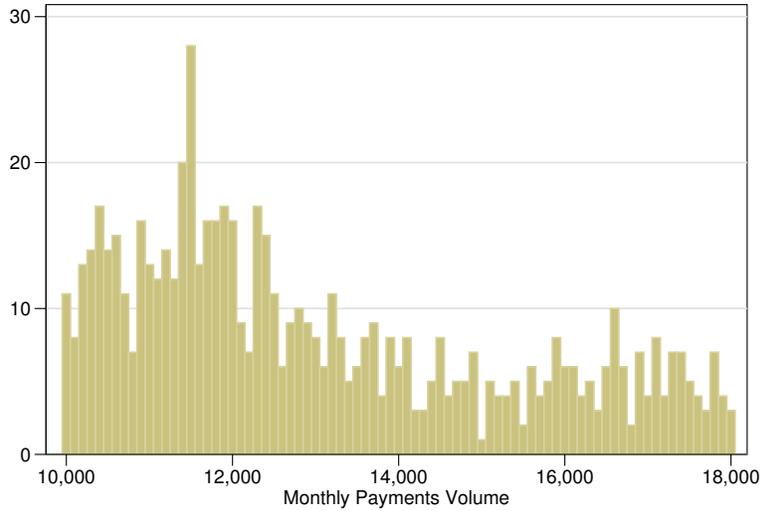


Figure 5: Banks' Monthly Payments Volume around the 14,000 Threshold

Note: An observation is a bank's monthly payments volume on Fedwire for every month, 2012.

only eight observations.

The lack of evidence of nonbunching around these payment-tier thresholds is a weak test of the model, if nothing else because theory does not tell us the size of the neighborhood around the kink for which we should observe the nonbunching. Nevertheless, the graphical evidence around the lower tier of 14,000 payments suggests that banks with payment volumes near these thresholds are not responding to marginal price. Alternatively, these banks may not be price sensitive. In the next section, we further explore this issue using regression analysis.

## 5.2 Encompassing test of marginal and average price

A more rigorous approach to test whether banks respond to marginal or to average price uses the encompassing test (Davidson and MacKinnon 1993). Let  $x_t^i$  denote the total number, or volume, of payments that bank  $i$  sends and receives in month  $t$ , and let  $MP_{it}(x_t^i; T_t^i)$  and  $AP_{it}(x_t^i; T_t^i)$  denote the marginal and the average price, respectively, that bank  $i$  faced in month  $t$ , given a total monthly volume of  $x_t^i$  and a benchmark threshold  $T_t^i$ . Let  $d_i \in \{1, 2, 3, 4, 5\}$  denote whether bank  $i$  is a domestic commercial bank, foreign bank, savings and loan association,

savings bank, or credit union. Following Ito (2014), we take a year-over-year approach and so eliminate bank-by-month fixed effects. This approach is particularly useful because there are seasonal variations in monthly payment volumes. We define the year-over-year change in log volume as  $\Delta \log(x_t^i) = \log(x_t^i) - \log(x_{t-12}^i)$  and the year-over-year change in log marginal price is  $\Delta \log(\text{MP}_{it}) = \log(\text{MP}_{it}(x_t^i; T_t^i)) - \log(\text{MP}_{i,t-12}(x_{t-12}^i; T_{t-12}^i))$ . The change in log average price is similarly defined. Using this notation, we specify the following regression,

$$\Delta \log(x_t^i) = \alpha \Delta \log(\text{MP}_{it}) + \beta \Delta \log(\text{AP}_{it}) + \sum_{k=1}^5 \sum_{s=1}^T 1_{k=d_i} 1_{s=t} \gamma_{ks} + \varepsilon_{it}, \quad (3)$$

where  $\gamma_{ks}$  are bank type-by-time fixed effects and  $\varepsilon_{it}$  is an error term. This empirical strategy tests whether banks respond to marginal price or to average price. According to our model, if banks have interior solutions to their cost-minimization problem and thus are responding to marginal price, we should find that  $\alpha$  is negative and  $\beta$  is equal to 0.

This specification suffers from an endogeneity problem in that both marginal and average price depend on concurrent volume and thus are correlated with  $\varepsilon_{it}$ . Following past studies, we construct an instrument for price change using past volume. Given the year-over-year structure of the regression, we know that  $\varepsilon_{it}$  is a function of volume in periods  $t$  and  $t - 12$ . To construct our instruments for price change, we chose the payments volume at  $t - 6$ .<sup>13</sup> Specifically, the instruments for marginal price are  $\Delta \log(\tilde{\text{MP}}_{it}) = \log(\text{MP}_{it}(x_{t-6}^i; T_t^i)) - \log(\text{MP}_{i,t-12}(x_{t-6}^i; T_{t-12}^i))$ , and the instruments for average price are constructed similarly. Accordingly, the instruments take a fixed payment volume,  $x_{t-6}^i$ , and construct the change in price faced by bank  $i$ , given the price schedule in place in period  $t$  as opposed to the price schedule in place in period  $t - 12$ .<sup>14</sup>

These instruments should not be correlated with the error term. They do not depend on payment volumes in periods  $t$  and  $t - 12$ . Furthermore, even if payment volumes are serially correlated, these instruments are uncorrelated with the error term as long as the serial correlation depends only on the time difference between the error terms (Blomquist and Selin 2010).<sup>15</sup>

<sup>13</sup>The choice of an instrument that is at the midpoint between  $t$  and  $t - 12$  is suggested by Saez, Slemrod, and Giertz (2012) and follows Ito (2014).

<sup>14</sup>As detailed in section 2, over our sample period the Fedwire price schedule was changed every calendar year.

<sup>15</sup>This result rests on the fact that the instrument is constructed using payment volumes that are equally spaced between  $t$  and  $t - 12$ .

In addition, these instruments are highly correlated with actual price change; we calculate that the instruments and changes in marginal price (average price) have a correlation of 0.75 (0.62).

The identification of  $(\alpha, \beta)$  hinges on the incentive-pricing feature of the nonlinear pricing schedule. By assigning each bank its own threshold (based on volumes over the past five years), Fedwire created a different pricing schedule for each bank. As a result, we observe different changes in marginal and average price across banks. We illustrate this fact in Figure 6 where we plot the change in instrumented marginal price versus the change in instrumented average price for each bank in each month of the sample. (In appendix B, we present the same figure, but using actual marginal and average prices.) This figure highlights that banks experienced the full set of possibilities with respect to changes in instrumented marginal price and instrumented average price: both prices increasing, both prices decreasing, marginal price increasing and average price decreasing, and marginal price decreasing and average price increasing. It is this wide range in the changes of instrumented marginal and average price that allows us to identify the coefficients of interest.

We run an instrumental-variables estimation for four versions of equation (3): using only marginal price; using only average price; using both marginal and average price; and using one month lagged marginal and average price. When including only marginal price, we estimate that banks respond to marginal price, and the price elasticity is  $-0.279$  (see column (1) in Table 5). A larger price elasticity,  $-0.547$ , is estimated when using only average price (see column (2)). The encompassing test, where both marginal and average price are included in the regression, delivers the main result: changes in average price, not marginal price, are driving banks' decisions on payment volume over Fedwire. Formally, we find that the estimated impact of average price on volumes is hardly affected by the addition of marginal price to the instrumental-variables regression. Including marginal price changes the elasticity of volume with respect to average price slightly from  $-0.547$  to  $-0.491$  (compare columns (2) and (3)), where  $-0.491$  is statistically significant. Furthermore, we find that the elasticity of volume with respect to marginal price is  $-0.064$ , a tiny economic effect.

These estimates are particularly striking because our fairly standard model predicts that these banks are at corner solutions with respect to their cost-minimization problem. In response to Fedwire's decreasing block-price schedule, banks in general are either routing all or none of their flexible payments through Fedwire.

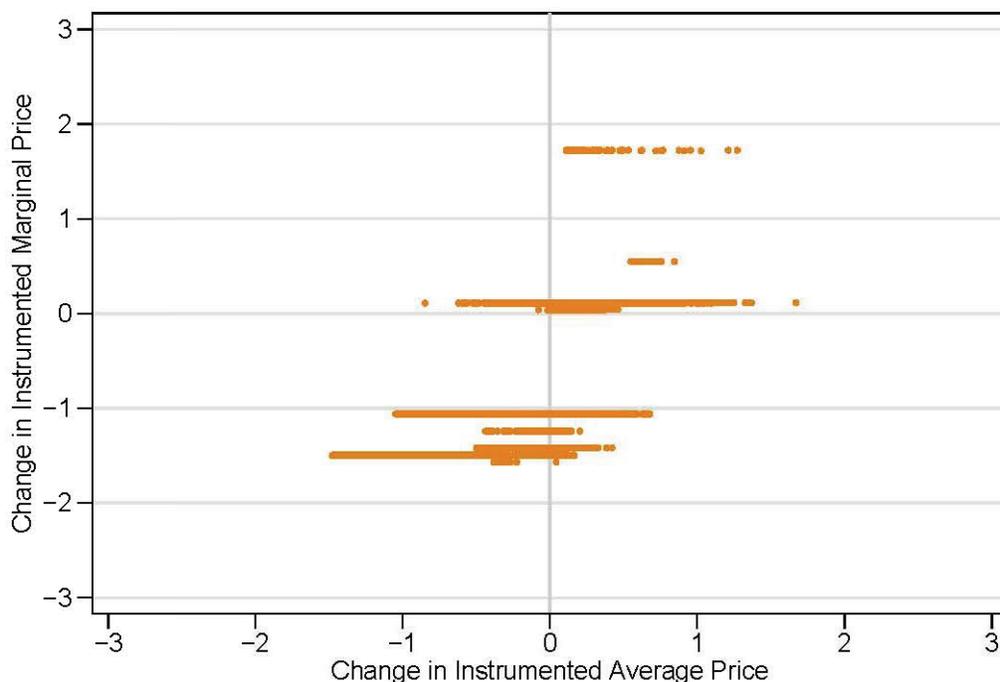


Figure 6: Changes in Instrumented Marginal and Average prices

A second feature of our result is that the estimated price elasticity is less than 1 in absolute value, implying that Fedwire is operating on the inelastic portion of the payment services demand curve. Our result implies that Fedwire could increase its average price and earn more revenues. Fedwire does not likely operate on the elastic portion of the demand curve because it is mandated to cover costs and not necessarily to maximize profits. Indeed, estimating an elasticity less than 1 in absolute value is in line with other empirical studies that analyze pricing of services provided by utilities.<sup>16</sup>

Although Fedwire advertises in advance the change in its pricing schedule, it might be the case that banks respond to these price changes with a lag. We check how our results vary when accounting for lags in adjustment by using two approaches. First, we consider the case that banks are reactive, adjusting the volume of payments settled on Fedwire only after they receive

<sup>16</sup>For example, Reiss and White (2005) and Ito (2014) estimate that households' mean elasticity of electricity demand is -0.39 and -0.05, respectively.

Table 5: Results from the Encompassing Test

Price variables	(1)	(2)	(3)	(4)	(5)
$\Delta \log(\text{MP}_{it})$	-0.279*** (0.020)		-0.064*** (0.024)		0.034 (0.027)
$\Delta \log(\text{AP}_{it})$		-0.547*** (0.023)	-0.491*** (0.029)		-0.659*** (0.035)
$\Delta \log(\text{MP}_{i,t-1})$				-0.021 (0.023)	
$\Delta \log(\text{AP}_{i,t-1})$				-0.374*** (0.028)	
Observations	205,270	205,174	205,174	196,169	102,331

Note: This table presents the estimated coefficients of 5 instrumental-variables regressions. Not shown are the estimated fixed effects. The standard errors are shown in parenthesis and are clustered by bank. \*\*\* denotes that the associated p-value is less than 0.01.

a monthly bill from Fedwire. To accommodate this possibility, we reestimate the instrumental-variables regression with one-month lagged values of the change in marginal price and average price (see column (4) in Table 5). We continue to find that banks respond to average price; the coefficient on marginal price is now both statistically and economically insignificant. The estimated coefficient on average price remains large and significant, although it does increase slightly from  $-0.491$  to  $-0.374$ .<sup>17</sup>

Our second approach accounts for banks being able to transfer payments from one settlement system to another only with a substantial lag. For example, it may take banks a few months to change their back office procedures so as to reroute payments off of or onto Fedwire. To accommodate this possibility, we reestimate the benchmark regression using only data from July to December of each year. By focusing on the latter half of the calendar year, banks have had plenty of time to react to changes in the Fedwire price schedule (which goes into effect at the start of the calendar year). These regression results reinforce our main result that banks respond to average, and not marginal, price. In particular, we estimate that the coefficient on marginal price is insignificant, whereas the coefficient on average price is significant and equal

<sup>17</sup>Using two- or three-month lags produces the same qualitative results. Marginal price remains small and statistically insignificant. Average price remains statistically significant, although the estimated elasticity declines with each longer lag.

Table 6: The Encompassing Test for Banks of Different Sizes

Price variables	First quartile	Second quartile	Third quartile	Fourth quartile
$\Delta \log(\text{MP}_{it})$	-0.090 (0.066)	-0.083** (0.042)	0.027 (0.085)	-0.012 (0.064)
$\Delta \log(\text{AP}_{it})$	-0.432*** (0.111)	-0.514*** (0.046)	-0.553*** (0.054)	-0.518*** (0.064)
Observations	23,226	55,516	62,499	63,933

Note: This table presents the estimated coefficients of an instrumental-variables regression for mutually exclusive sets of banks. Not shown are the estimated fixed effects. The standard errors are shown in parenthesis and are clustered by bank. Banks are divided into four groups, where each group is a quartile in the distribution of average monthly volume. The number of observations differs across specifications because some banks (especially low-volume ones) have zero monthly volumes.\*\*\* and \*\* denote that the associated p-value is less than 0.01 and 0.05, respectively.

to  $-0.659$  (see column (5) in Table 5).

### 5.2.1 Robustness results

The approach above estimates an average local effect across all banks. Here we explore whether we continue to find that banks respond to average price when considering subsets of banks. Because there are large differences in payment volumes across banks, we first redo our analysis for banks grouped by size. We divide banks into four groups based on the quartiles of the distribution of payment volume described in Table 3. For each of these groups, we then reestimate the regression detailed in the previous section (equation (3)), using the same instrumental-variables approach.

For all four groups, we continue to find strong evidence that banks respond to average price. Furthermore, the estimated coefficients on average price are quite similar across banks grouped in the second, third, and fourth quartiles of the distribution of payment volume, with the coefficients ranging from  $-0.514$  to  $-0.553$  (see Table 6). Although the estimated coefficient on average price for banks in the first quartile is slightly higher at  $-0.432$ , this coefficient still provides convincing evidence that this set of the smallest banks is responding to average price.

Our second robustness exercise is to allow separate coefficients on marginal and average price for banks that are CHIPS members to differentiate them from banks that are not. Because CHIPS offers settlement with only a slight delay relative to Fedwire and is similar to Fedwire

Table 7: The Encompassing Test for CHIPS Member Banks

Price variables	Estimated coefficients
$\Delta \log(MP_{it})$	-0.065*** (0.024)
$\Delta \log(MP_{it}) * CHIPS$	0.030 (0.048)
$\Delta \log(AP_{it})$	-0.492*** (0.029)
$\Delta \log(AP_{it}) * CHIPS$	0.174 (0.270)

Note: This table presents the estimated coefficients of an instrumental-variables regression. Not shown are the estimated fixed effects. The standard errors are shown in parenthesis and are clustered by bank. CHIPS is an indicator variable equal to 1 when a bank is a CHIPS member. There are 205,174 observations. \*\*\* denotes that the associated p-value is less than 0.01.

in other respects, CHIPS members may be more likely to face a corner solution to their cost-minimization problem and so react to average price. We explore this hypothesis using the benchmark regression, but include a dummy variable equal to 1 for banks that are CHIPS members. This dummy variable is interacted with both marginal and average price. We find that the estimated coefficients for the CHIPS interaction terms are not statistically significant (see Table 7). The evidence suggests that CHIPS member banks react to average price, and thus are also at a corner solution with respect to their cost-minimization problem.

## 6 Discussion of results and policy implications

One main motivation behind Fedwire’s nonlinear pricing scheme is to price discriminate across (unobserved) types of payments. The goal is to charge high prices for those payments that need to settle on Fedwire, and thus are price inelastic, and to charge low prices for those payments that can be settled elsewhere, and thus are sensitive to price. Fedwire’s decreasing-block schedule implements this price-discrimination strategy, where the setting of payment tiers (i.e., tiers one, two, and three) as well as the bank-specific thresholds reflects Fedwire’s understanding of the share of banks’ monthly payments that are price inelastic.

Our model predicts that banks will respond to Fedwire’s decreasing-block schedule in one

of two ways. First, a bank with a set of flexible payments with insufficiently high urgency for immediate settlement will find itself at a corner solution. This bank will settle all or none of its flexible payments on Fedwire, depending on whether the average cost of settlement on Fedwire is greater than or less than the average cost of settlement on a competing settlement system. Second, a bank with flexible payments with high enough urgency for immediate settlement will find itself at an interior solution. The most urgent payments will be settled on Fedwire and the remainder on a competing system.

Our empirical work demonstrates that the average bank is at a corner solution and thus is responding to Fedwire's average price. An implication of this result is that changes in banks' volume of payments in response to changes in average price should not be continuous (at the individual bank level). Rather, small changes in average price may induce large changes in volume of payments, because a bank may decide to reroute all its flexible payments onto, or off of, Fedwire. In future work, we aim to further analyze this (predicted) nonlinear response by banks to average price.

In addition, our elasticity estimate implies that banks' demand for Fedwire services is inelastic. Accordingly, Fedwire could increase revenues (to cover rising costs) by increasing the average price banks must pay. The downside of raising average price is that banks may exit and no longer use Fedwire.<sup>18</sup> Using a probit, we explore the degree to which changes in average price affect a bank's probability of exit (see the details of this analysis in appendix C). We find that average price has an economically and statistically significant impact on a bank's decision to leave Fedwire. Indeed, we compute that the marginal effect of a 1 percent increase in average price is a 9.3 percent increase in the probability of exit.

We conclude our analysis by presenting predictions of aggregate payment volumes under the counterfactual where Fedwire implements a two-part tariff. Furthermore, we argue there are social benefits to routing more payments over Fedwire. A two-part tariff provides a clean way for Fedwire to price discriminate across price-elastic and price-inelastic payments. The first price of this tariff is a fixed cost of accessing Fedwire, and the second price is a constant fee for sending or receiving payments over Fedwire.<sup>19</sup> Because the bank must pay the first part

---

<sup>18</sup> As discussed in section 2, banks have the option of accessing Fedwire through a correspondent banking account. In these instances, a bank can exit Fedwire and set up an account with another bank that has access to Fedwire. In this case, the exiting bank is still able to access Fedwire, albeit as a client to another bank.

<sup>19</sup>The first price in this two-part tariff is equivalent to the monthly participation fee that Fedwire currently

of the tariff to settle its set of non-price-sensitive payments, the cost-minimization problem of the bank focuses on the additional costs of settling its set of flexible payments. Under a two-part tariff, the bank's average price and its marginal price for using Fedwire to settle the set of flexible payments are both equal to the second price.<sup>20</sup>

To illustrate by how much Fedwire's volumes of payments might increase under a two-part tariff, we use our empirical estimates. We consider the demand for Fedwire services at the market level and assume that the volume of aggregate payments is given by

$$x = \gamma p^\beta, \tag{4}$$

where  $x$  is the volume of payments sent over Fedwire,  $p$  is average price, and  $\beta$  is the elasticity of Fedwire volumes to price. From our empirical work, we set  $\beta = -0.491$  and note that  $\gamma$  is a fixed effect that is pinned down in the data using the observed aggregate volume and average price.

To predict the impact of a two-part tariff, we set the second price equal to Fedwire's marginal cost. We approximate for this unobserved cost by using the lowest observed price in the data, 5.2 cents.<sup>21</sup> The first price is not a concern for this comparative static exercise, because banks must pay this fixed fee for Fedwire access to settle their set of nonflexible (price insensitive) payments.

We use data from March 2011 to back out  $\gamma$  and then rely on the aggregate demand function to predict Fedwire's volume, given a price of 5.2 cents. The predicted result is an increase in payment volumes of 156 percent, up from roughly 11 million to 28 million (see Table 8). The transactional revenue earned from charging banks fees to send and receive payments falls by 61 percent. Accordingly, Fedwire would need to set a high fee for accessing its settlement

---

charges (see appendix A for more details), except that in this case the fee would differ across banks.

<sup>20</sup>For the unusual case in which all of a bank's payments are flexible, then the bank would include the first fixed fee for accessing Fedwire in its cost-minimization problem.

A difficulty with implementing a two-part tariff is setting the first price, the fixed fee to access Fedwire, correctly. Fortunately, Fedwire is able to implement perfect price discriminate, because of its ability to charge bank-specific prices based on past Fedwire volumes or other characteristics. A key constraint facing Fedwire, though, is a bank's option to exit Fedwire but retain access through other banks or institutions.

<sup>21</sup>This is the smallest fee charged under incentive pricing in 2011 and is equal to the price charged for sending a payment plus the price charged for receiving a payment, whereby both the sending and receiving banks are in tier three and have crossed their bank-specific thresholds.

Table 8: Effects of Introducing a Two-Part Tariff

Pricing scheme	Fedwire volume	Transaction revenue (dollars)
Actual	10,974,556	3,760,759
Two- part tariff	28,119,701	1,462,224

Note: Fedwire Volume is the total number of payments sent over Fedwire. Transaction revenue is the revenue raised from sending payments over Fedwire, excluding the high value, late-in-the-day, and monthly participation fees. For the two-part tariff, transaction revenue is computed using only the second price of the two-part tariff, and is equal to volume  $\cdot$  0.052 .

system under a two-part tariff to raise enough revenue to cover its costs.

The estimated 156 percent increase in volume should be considered with care. The elasticity estimate driving the large volume increase is a local average effect and thus may perform poorly under large changes in price. Furthermore, this counterfactual holds the actions of competing payment systems fixed, when in fact they are likely to respond. Nevertheless, our counterfactual exercise illustrates the potential for large increases in payment volume over Fedwire with the adoption of a two-part tariff-pricing scheme, while still allowing Fedwire to fully recover its costs.<sup>22</sup>

Although the above discussion focuses on the specific problem Fedwire faces in raising revenues, the pricing of Fedwire also has implications for social welfare along two specific dimensions. First, costs of delay are introduced when payments are moved off of Fedwire. These delay costs are deadweight losses and thus are inefficient. Using a two-part tariff to bolster payment volumes by 156 percent suggests that these welfare gains are more than negligible.

Second, using a two-part tariff may have a larger effect on the payment settlement landscape by causing a shift of all payment volumes from CHIPS to Fedwire. Because there is a duplication of (some) fixed costs in running two settlement systems, migrating all payments from CHIPS to Fedwire would free up resources to be used elsewhere, thus increasing social welfare.<sup>23</sup>

<sup>22</sup>There are enough payments on CHIPS and ACH such that a 156 percent increase in the volume on Fedwire is feasible.

<sup>23</sup>In the literature, a main efficiency concern of nonlinear pricing by a utility is that output or volume of services demanded is less than that in the case of perfect competition. This because in a variety of nonlinear

The migration of all payment volumes from CHIPS to Fedwire is feasible, because any payment settled on CHIPS can also be settled on Fedwire.<sup>24</sup> In contrast, there are many types of payments that settle on Fedwire that cannot be settled on CHIPS (e.g., payments to or from a bank that is not a CHIPS member). Furthermore, this nonlinear response to price is consistent with our theoretical model, showing banks' moving to a corner solution (that is, settling all flexible payments over Fedwire) in their cost-of-settlement problem.

Banks will move all their payments from CHIPS to Fedwire only if two basic assumptions hold. The first is that the total cost of settling payments on CHIPS is higher than the incremental cost to the bank of settling that volume on Fedwire. With Fedwire implementing a two-part tariff, banks will use CHIPS only if the average price of using CHIPS is below 5.2 cents.

Second, it must be the case that CHIPS provides the same service as Fedwire (i.e., CHIPS and Fedwire are not differentiated products). *Prima facie*, this assumption is incorrect because CHIPS provides a netting service. Other than netting and the associated costs of delay, however, CHIPS provides a very similar settlement service to that of Fedwire. Recall that the CHIPS netting mechanism is a liquidity-savings tool. We argue, however, that in the current environment the costs of Fedwire's liquidity demands are minimal and thus that the value of the CHIPS netting service is negligible. In particular, banks currently have high reserve holdings and payments in excess of reserve holdings that can typically be made at zero cost in the form of collateralized overdrafts. Consequently, CHIPS and Fedwire essentially offer similar services.

## 7 Concluding remarks

We analyze banks' demand for Fedwire's payment-processing services. We find that when facing Fedwire's decreasing block-price schedule and given the existence of competing services, banks respond to average price. Through the lens of our model, this behavior is driven

---

pricing schemes, such as the increasing block schedules analyzed in Olmstead, Hanemann, and Stavins (2007) and Reiss and White (2005), marginal price is above marginal cost. In our environment, there are not concerns that changing Fedwire's prices will affect output (the total number of payments ( $x^i$ ) is fixed). Instead, Fedwire's price schedule determines which payment system is used to settle obligations.

<sup>24</sup>This is evidenced by CHIPS's rules and procedures that state in an emergency a CHIPS bank may send its payment order over Fedwire (see Emergency Procedures in the CHIPS Rules and Administrative Procedures (March 2013)).

by banks' finding themselves at corner solutions. For those payments over which they have discretion, banks will route all or none of these payments over Fedwire depending on the comparison of the average cost of settlement on Fedwire with that of a competing service. Our elasticity estimate also implies that on average there is inelastic demand for Fedwire's services. In the aggregate, we find that a 1 percent increase in average price will decrease payment volume by 0.491 percent. In future work, we aim to deepen our understanding of banks' demand for Fedwire's settlement services by finding and incorporating more information on banks' characteristics as well as payment characteristics into our analysis. Using such data should yield insights into differences in banks' response to changes in fees and price schedules.

## References

- BLOMQUIST, S., AND H. SELIN (2010): “Hourly wage rate and taxable labor income responsiveness to changes in marginal tax rate,” *Journal of Public Economics*, 94(11), 878–889.
- DAVIDSON, R., AND J. MACKINNON (1993): *Estimation and Inference in Econometrics*. Oxford University Press, USA.
- GREEN, E., J. LOPEZ, AND Z. WANG (2003): “Formulating the Imputed Cost of Equity Capital for Priced Services at Federal Reserve Banks,” *Economic Policy Review*, pp. 55–81.
- ITO, K. (2014): “Do Consumers Respond to Marginal or Average Price? Evidence from Nonlinear Electricity Pricing,” *American Economic Review*, 104(2), 537–63.
- LAMBRECHT, A., K. SEIM, N. VILCASSIM, A. CHEEMA, Y. CHEN, G. CRAWFORD, K. HOSANAGAR, R. IYENGAR, O. KOENIGSBERG, R. LEE, E. MIRAVETE, AND O. SAHIN (2012): “Price Discrimination in Service Industries,” *Marketing Letters*, 23(2), 423–438.
- OLMSTEAD, S., M. HANEMANN, AND R. STAVINS (2007): “Water Demand Under Alternative Pricing Structures,” *Journal of Environmental Economics and Management*, 54(2), 181–198.
- REISS, P., AND M. WHITE (2005): “Household electricity demand, revisited,” *Review of Economic Studies*, 72(3), 853–883.
- SAEZ, E., J. SLEMROD, AND S. GIERTZ (2012): “The elasticity of taxable income with respect to marginal tax rates: A critical review,” *Journal of Economic Literature*, 50(1), 3–50.
- STOLE, L. (2007): *Handbook of Industrial Organization* chap. Price Discrimination and Competition in Competitive Environments. NorthHolland.

## A Fedwire Funds' other fees

Starting in 2009, Fedwire charged banks a monthly participation fee of \$60. This fee is essential a monthly fixed cost, which is incurred after a bank sends or receives a payment on Fedwire. Further, Fedwire imposed additional fees on high value payments and late-in-the-day payments. For those payments with a value greater than or equal to \$10 million, Fedwire imposed a fee of 12 cents on both the sending and receiving bank starting in 2012. An additional fee of 30 cents is imposed on payments equal to or greater than \$100 million, starting in 2013. Starting in 2011, Fedwire charged a fee of 18 cents on banks originating payments over Fedwire after 5 p.m. (eastern time). Table AI shows how these fees have changed over time.

Table AI: Fedwire's non-volume fees

Fee	2009	2010	2011	2012	2013
Monthly participation (dollars)	60	75	75	85	85
High value ( $\geq$ \$10 million) (cents)	—	—	—	12	12
High value ( $\geq$ \$100 million) (cents)	—	—	—	—	30
Late in the day (cents)	—	—	18	20	21

The High value fee applies to payments both sent and received, whereas the Late in the day fee only applies to the bank sending the payment.

## B Changes in marginal and average price

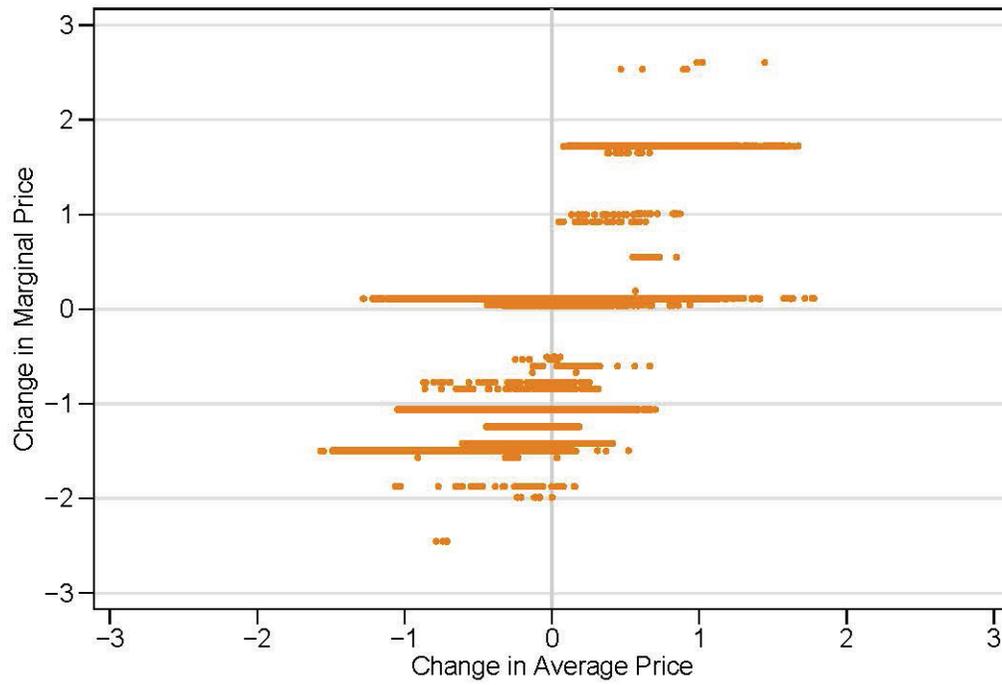


Figure B1: Changes in marginal and average prices

## C Changes in average price impact the probability of exit

The rise in average price may also increase the probability that a bank leaves Fedwire. As discussed in Section 2, banks have the option of accessing Fedwire through a correspondent banking account. In these instances, a bank can exit Fedwire and set up an account with another bank which has access to Fedwire. In this case, the exiting bank is still able to access Fedwire, albeit as a client to another bank. In addition, the increasing differences between the average price paid by small and large banks further encourages small banks to exit Fedwire. This is because a larger difference implies there are larger gains to trade for a smaller bank set up a correspondent banking account with a larger bank.

Because we can observe banks exiting Fedwire, we are able to explore to what extent changes in Fedwire pricing are driving banks off of Fedwire.<sup>25</sup> We deduce a bank has exited Fedwire by observing that its monthly payments volume goes to and remains at zero. Given the back office complexities of re-routing payments off of Fedwire, we anticipate that exit will take place over several months. Our empirical strategy then, is to identify the set of banks that left Fedwire in a given calendar year. Banks, however, may leave Fedwire for reasons unrelated to pricing. To screen out exits unrelated to price, we exclude from our analysis exiting banks that either merged with another bank or became insolvent. Using data that tracks the organizational structure of banks we can identify the set of banks that both exited Fedwire Funds and were involved in a merge with another bank. Using data from the Federal Deposit Insurance Corporation, we know which banks became insolvent.

Over the sample period of 2011 to 2013, we find 820 banks stopped using the Fedwire yet remained viable banks and were not involved in a merger.<sup>26</sup> These banks were small in that their average monthly volume is 217 payments per month, with 90 percent of them sending less than 421 payments a month. We use a probit to determine if changes in average price impacted these banks' decision to leave Fedwire. We first construct a data set of comparable banks, by selecting banks which sent on average less than 500 payments per month over Fedwire. We then construct an exit dummy variable,  $X_{iy}$  which is equal to 1 if the bank  $i$  left Fedwire in year

---

<sup>25</sup>In principle, we could also measure whether entry onto Fedwire is impacted by changes in average price. Our perception however, is that a bank's entry onto Fedwire will most likely be driven by factors other than Fedwire's pricing schedule.

<sup>26</sup>444 banks exited Fedwire over this same period but were insolvent or involved in a merger. The number of exiting banks is roughly the same across all three years.

Table CII: Exit Analysis

Variable	Estimated Coefficient
Change in average price	1.053*** (0.118)
2011 year dummy	0.061 (0.048)
2012 year dummy	0.104** (0.047)
Constant	-1.940*** (0.044)

Note: This table presents the estimated coefficients of a probit. The standard errors are shown in parenthesis and are computed using a sandwich estimator. There were 15,281 observations and the pseudo r-squared is 0.0195. \*\*\* and \*\* denote that the associated p-value is less than 0.01 and 0.05, respectively.

$y + 1$  (i.e., neither sent nor received any payments in  $y + 1$ ), remained solvent in  $y + 1$ , and did not merge with another bank in  $y + 1$ .

We construct average price change over calendar years using mean monthly volume. Specifically, denoting the mean monthly volume for bank  $i$  in  $y$  as  $\bar{x}_y^i$ , our measure of average price change is  $\Delta \log(\tilde{A}P_{iy}) = \log(AP_{i,y+1}(\bar{x}_y^i)) - \log(AP_{i,y}(\bar{x}_y^i))$ . This variable forecasts how much more bank  $i$  would have paid in  $y + 1$  compared to  $y$ , using the mean monthly payments volume in year  $y$ . We estimate the following probit

$$Pr(X_{iy} = 1) = \Phi(\zeta_0 + v\Delta \log(\tilde{A}P_{iy}) + \sum_{k=2011}^{2012} 1_{y=k}\zeta_k), \quad (C1)$$

where  $\zeta_k$  are calendar year dummies.

We estimate that  $v$  is equal to 1.053 and is statistically significant (see Table CII). For these small banks over this sample period, the mean change in average price is 12 percent. Using this average change in price, we compute that the marginal effect of a one percent increase in average price is a 8.9 percent increase in the probability of exit.

Overall then, we find that a sizeable number of banks are exiting Fedwire over the 2011 to 2013 period. For these banks, an increase in average price has a large impact on exit.