

Evaluating Executive Compensation Packages

Arantxa Jarque and John Muth

Executive compensation is a topic that has received attention both in the media and the academic literature. This article discusses issues relevant to the construction and interpretation of compensation figures typically reported in both sources. First, it is not clear what precisely should be included within a measure of the chief executive officer's (CEO's) income tied to his firm. Second, the study of executive compensation remains constrained by the availability of data. We discuss the main source of data used in most studies on the topic: Execucomp. We highlight where the lack of data requires a deviation between a theoretical “ideal” measure of compensation and that which the researcher must use as an approximation. In this way, we hope our article will be a useful first introduction for those looking to do further research on the topic.

We propose a measure of realized annual pay, compare it to other measures used in the literature, and illustrate the difficulties in calculating it. Using data in Execucomp, we provide our pay measure for CEOs of large U.S. firms in the period 1993–2012 and use it to estimate sensitivity of pay to firm performance. The main difficulties in this exercise lie in the fact that compensation packages of most executives include stock and option grants on their own firm's shares, which typically come with requirements that they be held by the executive for at least three or four years.¹ This implies two important

■ We thank the editor, Ned Prescott, and the referees, Kartik Athreya, Zhu Wang, and Peter Debbaut, as well as Huberto Ennis and Todd Keister, for helpful comments. The views expressed in this article are those of the authors and do not necessarily represent the views of the Federal Reserve Bank of Richmond or those of the Federal Reserve System. E-mail: arantxa.jarque@rich.frb.org.

¹ Moreover, it is a fact that most CEOs hold on to stock for which selling restrictions have expired, or to options that are exercisable and in the money. The reasons for these “voluntary” holdings are not entirely clear, since CEOs are risk averse and

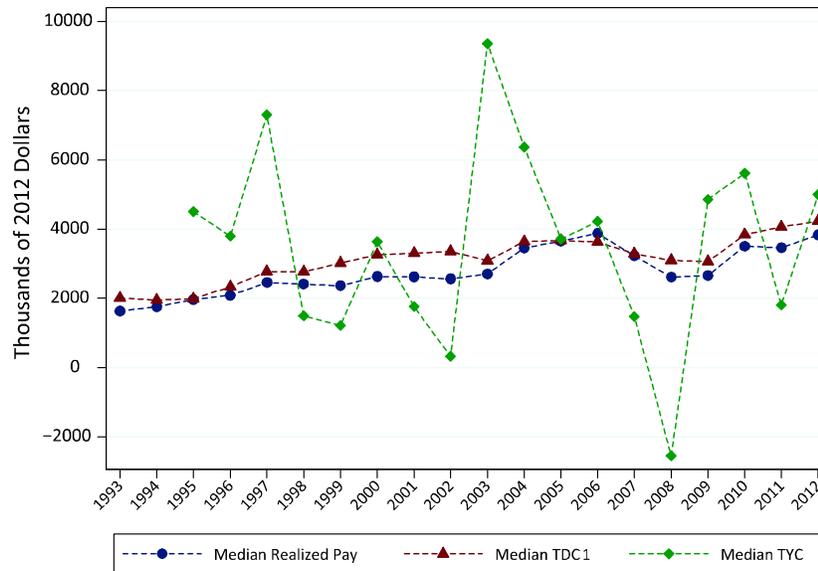
things. First, the compensation figures that are reported by firms (and are readily available to the press and researchers) are a combination of both *expected* value of compensation (for deferred compensation in the form of restricted stock and option grants that are not convertible into cash right away) and *realized* value (salaries, bonus payments, and perks). Second, a given year's compensation package provides income for several years to follow, since the CEO will be able to realize gains from selling and exercising stock and option grants once their vesting restrictions expire. That is, an important part of the annual realized pay of a CEO in any given year comes from his net gains from trading stock that he received in a past grant. Due to the fact that stock price realizations may differ from ex-ante expectations of those prices, the ex-post realized gains from those trades will typically differ from the valuation made at the time of the grant.

A measure of what is sometimes called *direct compensation* (the sum of salary, bonus, other compensation such as pension plans or perks, and the value of new stock and option grants during the year) is readily available in Execucomp (variable TDC1).² As we just discussed, grants included in this measure are valued in expectation. Our objective in this article is to provide a measure of *realized pay* instead. We define realized pay as the sum of salaries, bonuses, and other compensation, plus the gains from trades that the CEO realizes in a given year. We will argue that this measure is close to the one first proposed by Antle and Smith (1985) and used later by important contributions such as Hall and Liebman (1998) and Gayle and Miller (2009). *Total yearly compensation* is defined in these studies as the change in the wealth of the CEO that is tied to his employment in the firm, and it is calculated in practice as direct compensation plus the year-on-year change in the market value of stock and option holdings of the CEO from past grants. This measure is, hence, still a measure of expected pay, although more sophisticated than TDC1. The main departure of our measure of realized pay with respect to this total yearly compensation is that it does not attribute changes in the value of grants that are not yet exercised to the realized pay in the year when they occur; rather, the final realized value is captured in gains from trades and attributed to the period of exercise of the grants. This simplification is useful in terms of the calculation of the measure—we need to rely less heavily on assumptions about the unavailable details of grants.

standard economic theory would suggest that they would value a diversified portfolio of assets more. Overconfidence, privileged information, or personal tax considerations have been proposed in the literature as potential explanations (Jin and Kothari 2008).

²This measure has been studied, for example, in Gabaix and Landier (2008) and Frydman and Saks (2010).

Figure 1 Median Realized Pay, and Mean Expected Pay as Measured by Execucomp in the Variable TDC1, as well as Measured in Total Yearly Compensation (TYC)



Still, only part of the information that we need for our measure (about trades or vesting restrictions and exercise prices of past grants) is available in Execucomp. When approximating the gains from trades, in particular, we follow closely the algorithm used in Clementi and Cooley (2009) to recover the executive's holdings of stocks and options of his firm.³ In the Appendix, we walk the reader through the step-by-step construction of the portfolio, discussing the shortcomings of the available data in Execucomp and how different assumptions about the unknowns may affect the compensation numbers.

We use our measure of realized pay to provide an updated account of CEO compensation through the year 2012. Figure 1 presents a comparison of our measure of realized pay versus two measures of expected pay used in the literature: “direct compensation,” the variable TDC1 in Execucomp, and “total yearly compensation,” as calculated by us following the implementation in Clementi and Cooley (2009) of the

³ For another recent application of the algorithm first developed in Antle and Smith (1985), see Gayle and Miller (2009).

concept introduced by Antle and Smith (1985). Median realized pay is mostly below median direct compensation. The main difference observable with total yearly compensation is that it is a lot more variable than either of the other two measures. This figure suggests that different measures of pay present different pictures of CEO compensation, and it is important to understand what is behind the measurements before using them to evaluate pay practices.

We use our realized pay measure to perform a sensitivity analysis of annual realized pay to performance, with a special focus on the finance sector throughout the recent crisis in 2008. We simplify some of the difficulties of the analysis by assuming that the choice of selling and buying stock is invariant to the stock price movements in our counterfactual exercises; i.e., only the profits from the trades change, not the quantities. We find that in the aftermath of the crisis the realized pay of CEOs of finance firms has decreased in level relative to other industries. Moreover, the sensitivity exercise suggests that, during the whole sample period, mean realized pay for CEOs in finance firms changes with the performance of the firm in similar magnitudes than that of the average CEO.

We proceed as follows. In Section 1, we introduce compensation instruments included in most CEO pay packages and discuss data availability and measurement challenges. In Section 2, we present a simplified model of compensation accounting to illustrate the differences between three different measurement alternatives: the measure of realized pay that we construct in this article, and two measures of expected pay—the simple measure of expected pay readily available in Execu-comp, direct compensation, and the one based on the concept of total yearly compensation introduced by Antle and Smith (1985). Section 3 presents the results on the implied measure of realized pay over time, with a special focus on pay sensitivity, as well as a detailed look at the financial sector before and after the recent financial crises. Section 4 concludes. The Appendix provides the technical details on how we construct our realized pay measure from the data available.

1. UNDERSTANDING COMPENSATION PACKAGES

Nowadays, companies pay their top executives mainly through different combinations of the following instruments: a salary, a bonus program, a signing bonus, stock grants (also referred to as “restricted stock,” since they are usually granted with restrictions on the ability to sell them), grants of options on the stock of the firm, and perks and long-term incentive plans that specify severance payments, as well as pension plans.

Table 1 Summary of Annual Compensation Information Available in Execucomp

Instrument (Average % of TDC1)	Information in Execucomp
Salary (32%)	Value
Bonus and Incentive Compensation (23%)	Value, some details on targets (after 2006)
Perks and Other Compensation (6%)	Value
Restricted Stock Grants (11%)	Value (stock price times number of shares)
Stock Option Grants (28%)	Value (Black and Scholes), number of shares underlying options

Notes: Information available in Execucomp about the components of CEO compensation packages. For the percent calculations, the sample includes the CEOs of the largest 1,500 public firms in the United States in the period 1993–2010.

The publicly available information on CEO compensation comes from the compensation tables included by firms in their annual reports, as mandated by the Securities and Exchange Commission (SEC). This is the same data that Execucomp has compiled since 1992 and has been used in numerous empirical studies of CEO compensation, including this article. When the press publicizes information on CEO pay, it usually reports a summary measure of total or “direct compensation,” which is also readily available in Execucomp as the variable TDC1. Direct compensation is the sum of cash compensation (wage, bonus, and incentive compensation), pension contribution and other perks, plus the expected value of new stock and option grants given to the CEO within a given year. Execucomp also reports separately the different components of total compensation, and it includes some limited information on stock ownership and the portfolio of unvested restricted stock and option grants of the executives. A brief description of each of the instruments and further details on the information available about them in Execucomp follows. Table 1 presents statistics for their relative importance as a share of total pay using data from 1993 to 2010 and summarizes the information on availability.

Salaries are the simplest compensation instrument: They are not contingent on performance and information on their level is readily available on the proxy statements of firms.⁴ Bonus plans and incentive pay typically depend on yearly accounting results. Information is available mainly on payouts and more recently on some limited details of

⁴ The source for the shares of compensation that are reported come from Jarque and Gaines (2012). See the article for details on sample selection.

Table 2 Classification of Compensation Instruments

	Current (within year)	Deferred
Non-Contingent	Salary, perks, signing bonus	Pension plan
Contingent	Bonus plan	Options, stock, severance, future pay

the bonus plans. Information on perks and other compensation is also available, although not to a great level of detail. Grants of restricted stock of the firm make pay depend on the results of the firm over a longer time horizon, since the CEO is restricted from selling them until their vesting period expires. Execucomp compiles information on their expected value at the time of the grant (number of shares times market price of stock), but it does not have separate information on the number of shares granted. Grants of stock options allow the executive to purchase stock of the firm at a pre-established price (the “exercise price”) and are also typically granted with restrictions as to how soon they can be exercised. These also provide incentives for longer-term performance, but they only pay off if the stock price of the firm is above the exercise price. For option grants, Execucomp has information on both the number and the Black and Scholes value of the total grants during the year. Typically, both stock and option grants come with a clause that forces the executive to forfeit them in the event of employment termination. Information on the vesting periods is not generally available in Execucomp for either stock or option grants.⁵

It should be apparent that compensation instruments can be classified according to two criteria: whether or not they are contingent on the performance of the firm, and whether or not they are deferred.⁶ Table 2 summarizes this classification of the main compensation instruments.

Given that executives are risk averse, paying them with contingent instruments, such as bonuses, stocks, and options, comes at a cost, since they will demand higher expected payments to compensate them for the risk. The most accepted explanation for the inclusion of compensation instruments that are contingent on the performance of the firm is the existence of a moral hazard problem: The separation of ownership and control of the firm implies the need to provide incentives to the

⁵ A commonly cited length of this restriction period is four years, with vesting taking place proportionally over this period—see Hall and Liebman (1998).

⁶ Firm performance is typically proxied by accounting measures such as return on equity, sales, and profit, or on market-based measures such as the stock price.

CEO that align his interests with those of the firm owners.^{7,8} Within this context of incentive provision, it is also commonly accepted that expectations over future wages or jobs (career concerns), as well as the threat of dismissal, are also important compensation instruments—although less easy to study due to the lack of hard information on them.⁹

Deferral of pay also comes at a cost if CEOs are more impatient (i.e., they discount the future more) than the shareholders of the firms they manage. Several reasons may explain the use of deferred instruments. Perhaps the most accepted one is that, despite the cost of waiting, deferral is valuable—in combination with commitment to long-term contracts—because it allows to smooth incentives over time, making (costly) exposure to risk less necessary.¹⁰ Other reasons include retention purposes in the face of lack of commitment to long-term contracts or provision of incentives for hidden actions with long-term effects.¹¹

In most cases, instruments that are “cashed” within the year (labeled “current” in the table) are straightforward to value. In contrast, for contingent deferred instruments an expected value needs to be calculated, which presents some challenges. For example, the actual amount of compensation that the CEO will receive from stock and options granted to him in a given fiscal year will depend on the stock price of the firm at the moment he sells or exercises them. Similarly, the value of future compensation will depend on the performance of the firm during the tenure of the CEO. The value of pension payments will be contingent on the firm being solvent once the CEO retires. The value of severance payments is typically pre-set at the time of contracting, but a full list of the contingencies that may lead to termination is not written in the employment contract of the CEO. Hence, in order to calculate the expected value of compensation, one needs to know both the set of contingencies that trigger each payment (for example, the circumstances that trigger firing of the CEO or the performance targets for granting salary increases), as well as the probability attached to each of these performance contingencies (for example, the probability

⁷ See Prescott (1999) and Jarque (2010) for an introduction to static and dynamic moral hazard problems, respectively. Classical references in the literature include Grossman and Hart (1983), as well as Spear and Srivastava (1987).

⁸ Bebchuck and Fried (2004) argue that captive boards may use stock and option grants as a less obvious instrument to transfer excessive amounts of pay to their CEOs.

⁹ See Jensen and Murphy (1990); Gibbons and Murphy (1992); and Jenter and Kanaan (forthcoming).

¹⁰ Wang (1997) fleshes out this explanation using a repeated moral hazard model.

¹¹ See Bolton, Sheinkman, and Xiong (2006); Clementi, Cooley, and Wang (2006); and Edmans and Liu (2011).

distribution over future stock prices of the firm). These difficulties are important when choosing a measure of CEO pay.

Measurement of Pay: Expected versus Realized Value

There are two main approaches to measuring CEO pay:

1. *Expected* value of pay: The expected value of compensation granted in a given year, which includes the cash (realized value) he receives in salary and bonus, plus the expected value of the deferred contingent instruments such as stock and options;
2. *Realized* pay: The actual amount of money received in a given year, which includes the cash he receives in salary and bonus, plus the proceeds from selling past stock and option grants for which selling restrictions have expired (all realized).

Any attempt at valuing contingent deferred compensation, either in expectation or its realized value, will be constrained by the availability of data. Table 3 summarizes the data available in proxy statements and compiled by Execucomp about CEO holdings of stock and options of his own firm, the evolution of which is key to measurements in both categories. For stock holdings, we have the number of shares held by the CEO at the end of the fiscal year, as well as the number and value of both stock that remains restricted and of stock that vested during the year. For option holdings, we know the number of options exercised during the year, as well as their value. We also know the number and value of options exercisable (but still unexercised) and those whose vesting restrictions did not yet expire. These values, however, are calculated using the “intrinsic” valuation (stock price at the end of the year minus exercise price, times number of options, if positive), hence ignoring the options that are currently out of the money, and provide a simplistic evaluation (Black and Scholes would be a more accurate choice).

We choose our measure of realized pay (presented in the next section) in light of these data availability issues. Our choice tries to minimize the sensitivity of our measurements to assumptions about the unknown details of compensation packages, while still exploiting the information we have available on the portfolio of stock and options of the CEO.

Before we present our measure, it is important to note that we view expected and realized measures of pay as complements rather than substitutes when trying to understand incentives for CEOs. Expected

Table 3 Summary of Information Available in Execucomp about Stock and Option Holdings

	Information in Execucomp
Stock Holdings	Number of unrestricted Number of restricted Value of restricted Number vested during the year Value of vested during the year
Option Holdings	Number exercised during the year Value of exercised during the year Number of all unexercised vested Value of in-the-money unexercised vested (intrinsic) Number of all restricted Value of restricted in-the-money (intrinsic)

pay is a forward-looking measure, which gives important information about the value of the current compensation package given to the CEO. However, it is a difficult task to get a realistic valuation of stock or options for the CEO, especially because of selling restrictions and risk aversion considerations. In practice, the data in Execucomp reflects the firm's estimate of that value for CEOs. For options, usually a pricing model based on arbitrage conditions, such as Black and Scholes' option valuation model, is used to provide a value in the company's report with the SEC. Ad hoc modifications are often used to accommodate the fact that CEOs are risk averse and there are selling restrictions on the option grants.¹²

Realized pay, instead, is a backward-looking measure: Given past performance, we can calculate how much payoff the CEO actually got in the given period. In contract theory terms, we can view this measure as a description of the contract payoffs on the equilibrium path. That is, we observe what the CEO gets for the actual performance that materialized, but we do not have information on what the payoffs would have been for better or worse performances. For an estimate of these off-the-equilibrium-path payoffs, in Section 3 we perform sensitivity analyses that exploit the fact that we have some information on the number of stocks and options the executive sold or exercised.

¹² See Hall and Murphy (2002) for a quantitative evaluation of the difference between the executive's value of options and the cost to the firm in providing them.

One advantage of our realized pay measure is that we do not need to take expectations over the value of deferred contingent pay. Hence, we will be able to use the publicly available information on compensation packages without resorting to assumptions about the future value of contingent compensation. Still, even for the purposes of measuring realized pay, we are missing some important information on these deferred contingent instruments. As reflected in Table 3, Execucomp records the *value* of stock and the *value* and *number* of stock underlying options at the time when they are granted to the CEO. The values are approximations to the expected income that the CEO will realize in the future, when their restrictions expire. However, we do not have explicit information on the vesting schedules of these grants, or the exact date when the vested stocks are sold or the options exercised, or the market price of the stock at those times. This information is key to compute the actual cash the CEO receives as a result of the original grant. Our construction of a realized pay measure will necessarily involve assumptions on these unknown characteristics of the compensation, which we discuss in detail in the Appendix.

Larcker, McCall, and Tayan (2011) have a short and interesting essay in which they also point out the differences in measuring expected and realized pay.¹³ The authors include illustrative examples of the difference between expected and realized compensation based on data for a handful of firms in the year 2010. In this article we will use a larger number of firms and a longer period of time to illustrate quantitatively the difference between the two measures.

2. CONSTRUCTING A MEASURE OF REALIZED PAY

In this section, we provide a framework for comparing different measures of compensation. For this, we describe the types and timing of the different components in a typical compensation package. Using this framework, we introduce our proposed measure of realized contingent pay, denoted I_t , which is defined as the sum of salary, bonus, and gains from selling stock and exercising options in the current year. To construct it, we use information on the several components of pay packages that is publicly available, along with some assumptions. We refer to the model to illustrate the need for these assumptions and to justify

¹³ Larcker, McCall, and Tayan (2011) also present a third measure that they call *earned pay* (the value of pay at the moment when all selling restrictions are lifted, which does not necessarily coincide with the value at the time the CEO decides to sell). We do not have enough information in Execucomp to calculate this measure.

our choices. Then we illustrate in the context of the model what the differences are between our measure and two alternative ones: (1) direct compensation, which is defined as the sum of salary, bonus, perks, and other compensation, and the value of stock and options at the time of grant, and (2) total yearly compensation, which is defined as direct compensation plus dividends, plus the change in the value of stock and options in the portfolio of the CEO.

Consider a CEO who lives for T years. He starts his tenure with a firm at year $t = 1$. He receives compensation for all the years he is working, and after he retires he consumes out of his accumulated wealth and pension payments. We assume he has no sources of income other than what he receives as payments for his job as CEO, which we denote as I_t . The value he attaches to his employment at the beginning of period 1, denoted V_0 , is equal to the *expected* stream of income that he expects to receive in exchange for his work in each of the periods of his life:¹⁴

$$V_0(\mathbf{e}^*) = E \left[\sum_{t=1}^T \frac{I_t(p_1, \dots, p_t)}{(1+r)^{t-1}} \mid \mathbf{e}^* \right], \tag{1}$$

where the expectation is with respect to stock price realizations (which summarize the performance of the firm in this simple model), conditional on the sequence of effort choices by the CEO (denoted \mathbf{e}^*) given the optimal contract. We denote the market interest as r .

In this article, we want to measure the *realized* value of I_t . A more ambitious objective, which would relate more directly to theoretical models of CEO compensation based on repeated moral hazard models (Wang 1997), would be to try to measure $V_t(\mathbf{e}^*)$. We discuss some of the added difficulties of this measurement at the end of this section.

Realized pay I_t will not all be delivered directly in cash. Rather, the executive will receive an annual compensation, C_t , that will consist of two elements: a cash-based portion, or current liquid payment, denoted L_t , and a grant-based portion, denoted G_t . We assume compensation is received only once per year, at the end of the fiscal year. We have that

$$C_t = L_t + G_t \quad \forall t, \tag{2}$$

where

$$L_t = W_t + B_t + D_t + K_t \quad \forall t.$$

¹⁴ Note that the utility the CEO may get from a given value of employment will also depend on his wealth from sources other than the executive's employment. There is typically no information on this outside wealth to be used in empirical studies of CEO compensation.

That is, L_t is the sum of annual salary W_t , bonus payment B_t , which usually will depend on the annual results of the firm, dividends D_t , and perks and contributions to pension plans K_t .¹⁵ Grants consist of both restricted stock of the firm, s_t^r , and options to buy stock, o_t^r , and are valued at any $t' \geq t$ as¹⁶

$$\begin{aligned} G_t^{t'} &= EV(s_t^r; p_{t'}) + EV(o_t^r; x_t, p_{t'}) \\ &= s_t^r p_{t'} + EV(o_t^r; x_t, p_{t'}). \end{aligned}$$

In this expression, $EV(s_t^r; p_{t'})$ is the estimated value of restricted stock, i.e., the amount of stock, s_t^r , valued at the stock price at the time of valuation, $p_{t'}$. The estimated value of options, $EV(o_t^r; x_t, p_{t'})$, stands for some version of the Black and Scholes (1973) option valuation formula and depends both on the market price at the time of valuation, $p_{t'}$, and the exercise price, x_t .

Our Measure of Realized Pay

The stream of realized pay I_t that the CEO will receive from the firm while working will be equal to the cash part of his compensation, L_t , plus whatever net gains from trade he gets from buying and selling unrestricted stock (or vested exercising options). To compute these gains from trade, it will be important to keep track of the accumulated number of stock and option grants that have vested, what we will refer to as the “portfolio” of the CEO.¹⁷ Let S_{t-1} denote his holdings of unrestricted stock at the beginning of period t , and O_{t-1} denote his holdings of vested options. Let $T_t(S_{t-1}, O_{t-1})$ denote the gains from the sales of stock and exercises of options at period t . Then, we can write realized pay as

$$I_t = L_t + T_t(S_{t-1}, O_{t-1}).$$

Tracking the holdings S_t and O_t involves understanding the law of motion of the quantities of vested stock and options available to the CEO. Under the assumption that the CEO did not own any stock or options of the firm before his employment as CEO started, we have

¹⁵ Note that dividends are not included in Execucomp’s TDC1 (which we will compare later to our own proposed measure of income). We include them because they are attached to the grants given to the CEO, and hence they are income that he receives because of his association with the firm.

¹⁶ Here and in the rest of the model description, we use capital letters to denote values and lowercase letters to denote quantities.

¹⁷ Note that option grants also come with expiration dates; we are abstracting from those in this discussion, since the information we have on expirations is limited.

that his holdings in the beginning of year 1 are equal to zero:

$$\begin{aligned} S_0 &= 0, \\ O_0 &= 0. \end{aligned}$$

Any subsequent year, the quantities available to trade will change for two main reasons:

1. some of the past grants will have vested, or the CEO may choose to buy unrestricted stock; these actions will increase his holdings;
2. some of the past grants in his holdings will be sold or exercised, decreasing his holdings.

It is worth noting here that accurately evaluating the evolution of the holdings of the CEO would necessitate a large amount of information. For example, the CEO may choose to buy or sell stock, or exercise options, at different times during the year—with different market prices for each transaction. Also, he may choose to exercise options and hold on to the stock that he obtains with this transaction. Moreover, he may inherit or donate stock at any time. Unfortunately, the only data we have for the holdings of stock and options is their quantities and value at the end of each fiscal year (see Table 3), and we are lacking the details on the specific transactions that determine their evolution. Hence, we make the following important simplifying assumptions. First, we assume each of the possible trades happens only once in the fiscal year. Note that this still accommodates for a given sale of options to include options from different past grants, which implies different exercise prices. Second, we assume that the executive never purchases options, and that he exercises options only if he plans to sell the stock immediately. Third, we ignore any inheritances or donations.

We can summarize the above discussion in a formal law of motion for the holdings of stock and options by introducing some notation. The vesting restrictions on the stock and option grants determine the available S_t and O_t in each period. Typically, only a portion of the previous years' restricted stock vests every t . Denoting the vested shares in year t by s_t^v and vested options in year t by o_t^v , the accumulated number of shares and options available for selling in year t is

$$\begin{aligned} S_t &= S_{t-1} - (s_t^s - s_t^b) + s_t^v, \\ O_t &= \sum_{o_g \in O_{t-1}} o_g - o_{g,t}^e + o_t^v, \end{aligned} \quad (3)$$

where we are denoting the three types of trades that can happen at time t as follows:

1. selling stock s_t^s of the unrestricted stock available at period t , S_{t-1} , at price p_{s_t} ,
2. buying an amount s_t^b of stock from the market, at price p_{b_t} ,
3. buying stock through the exercise of $o_{g,t}^e$ of any vested option grant g (with corresponding exercise x_g) at price p_{e_t} .

With this notation, we can write an expression for the gains from trade:

$$T_t(S_{t-1}, O_{t-1}) = s_t^s p_{s_t} - s_t^b p_{b_t} + \sum_{o_g \in O_{t-1}} \max \{0, o_{g,t}^e (p_{e_t} - x_g)\}. \quad (4)$$

This completes the description of our measure of realized pay, I_t . Next, before moving on to the estimates of I_t using data, we use the model in this section to compare our measure of realized pay with alternative measures used in the literature.

Alternative Measures: Expected Pay

As we discussed in Section 1, the literature has used compensation measures based on the expected value of pay. The theoretical measure of expected pay is described by (1). The employment value, V_t , is the sum of the expected stream of realized pay. For the measurement of V_t (\mathbf{e}^*) in the data, however, one would have to make assumptions about the terms of the contract offered to the CEO regarding compensation in future periods (i.e., what would trigger a wage increase, or what is the schedule of future grants contingent on realized performance). One would also need to understand the CEO's expectations about stock prices in the future, which will determine his future realized gains from trade. One would also need to understand his expectations regarding his transitions to other firms and their consequences for his realized pay. Moreover, one would need to model how performance during the CEO's working life will affect his pension payments. To the best of our knowledge, no study has provided a reliable measure of V_t . Instead, two different approximations to V_t have been widely used: "direct compensation" (TDC1) and "total yearly compensation" (TYC). We define each of these using our notation, in turn, and compare them to our measure of realized pay.

The Execucomp variable TDC1 can be written in terms of our notation as

$$TDC1_t = W_t + B_t + K_t + G_t^t.$$

This measure of expected pay does not closely correspond to the theoretical V_t , since it does not include any estimation of future wages,

bonuses, and new grants. It includes an estimate of the expected future value of the grants given to the CEO *in the current year*, $G_t^t = s_t^r p_t + EV(o_t^r; x_t, p_t)$, but it ignores the changes in the value of past grants, or the realized gains from exercising them once they are vested, as well as the dividends that correspond to the CEO from holding stock. The main difference between our I measure and TDC1 is that we do not include the value of grants, G_t , but rather the realized net gains from trade, T_t . Also, dividends are included in I_t but not in $TDC1_t$.

A second alternative measure of expected pay, TYC, has been used in the literature since Antle and Smith (1985) proposed it. The idea behind it is to calculate the expected value that the CEO attaches to working in his firm, every period, as the current expected value of stock and option holdings plus the expected future compensation; then one can interpret the annual change in this expected value from one period to the next as the TYC of the executive.¹⁸ Because the expected value of grants is updated every year, this measure presents a more accurate picture of the incentive value of the CEO's contract. However, the measure is not without problems. For example, a common simplifying assumption when computing this measure is to assume that salary and bonus payments remain constant in future years and that the expected value of future grants is zero.¹⁹

We follow the description in the Appendix of Clementi and Cooley (2009) to replicate their measure of TYC, assuming wages, bonuses, and perks remain constant throughout the work life of the CEO, and no turnover. We graph it for comparison purposes in Figures 1, 2, and 5. In terms of our notation, TYC can be written as

$$TYC_t = W_t + B_t + K_t + D_t + \sum_{\tau=1}^t (G_\tau^t - G_\tau^{t-1}),$$

where G_τ^t in this case denotes the updated expected value during period t of stock and (unexpired) option grants that were given at period $\tau \leq t$ and are still unexercised.²⁰

The measure TYC attributes initial grants as compensation in the year when they are granted, and then subsequent appreciations and

¹⁸ Examples of different implementations of this concept of expected pay include Jensen and Murphy (1990); Garen (1994); Haubrich (1994); Hall and Liebman (1998); Haubrich and Popova (1998); Schaefer (1998); Aggarwal and Samwick (1999); Baker and Hall (2004); Clementi and Cooley (2009); Edmans, Gabaix, and Landier (2009); and Gayle and Miller (2009).

¹⁹ See, for example, Clementi and Cooley (2009; 2, 29).

²⁰ Note that $G_\tau^{t-1} = 0$ whenever $\tau > t$. Also, note that this re-evaluation of grants coincides conceptually with our measure of gains from trade, for the portion of the vested portfolio that is converted to cash in period t . That is, if, for example, only grants given at $t - 4$ are exercised at t , then $T_t(S_{t-1}, O_{t-1}) = G_{t-4}^t$.

depreciations of the grants to the periods when they happen—even if they do not translate into realized pay in that particular period. In comparison, our measure I of realized pay records only the realized value of grants when they get exercised, and it attributes the gains from trade to the particular period when they happen. It is easy to see that the simple sum of $\sum_{t=1}^T I_t = \sum_{t=1}^T TYC_t$; however, the individual year entries will differ, and hence the properly discounted sum will differ as well.

3. MEASUREMENTS

In this section, we present the empirical measurement of pay according to the methodology described above. In the Appendix, we provide the details on how to map the elements of pay described in the previous section to the data available in Execucomp.

In this article, we work with the August 2013 release of Execucomp, which includes annual observations through the fiscal year 2012. We drop CEOs who own 50 percent or more of the shares of their company, since we want to focus on measuring incentives in relationships for which there is an agency problem. Our final sample includes 3,345 different firms, for a total 34,497 firm-CEO-year observations.^{21,22}

Figure 1 presented the median of our measure of realized pay from 1993 to 2012. We compare it to the two measures of expected pay discussed earlier in this article: “total compensation” reported in Execucomp as the variable TDC1 and our own calculation of TYC following Clementi and Cooley (2009).²³

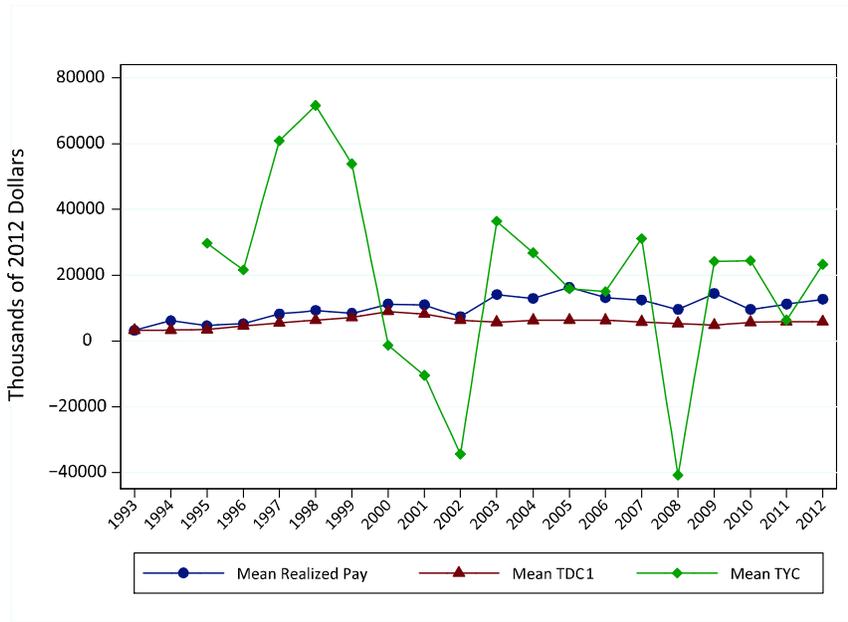
Two features emerge from Figure 2. First, averages are much larger than medians. This is well known for the measure TDC1, and it is confirmed for our measure of realized pay, I . Second, average realized pay is more volatile over time than average total compensation, and it is typically above $TDC1_t$, while it was typically below it when we looked at the medians in Figure 1. However, TYC_t is more volatile than either of the other two measures. This is true both when looking at medians, in Figure 1, or when looking at means, here. Our analysis of the

²¹ The database includes up to five executives of a firm per year, but we restrict our sample to those designated as the CEO by the Execucomp variable CEOANN.

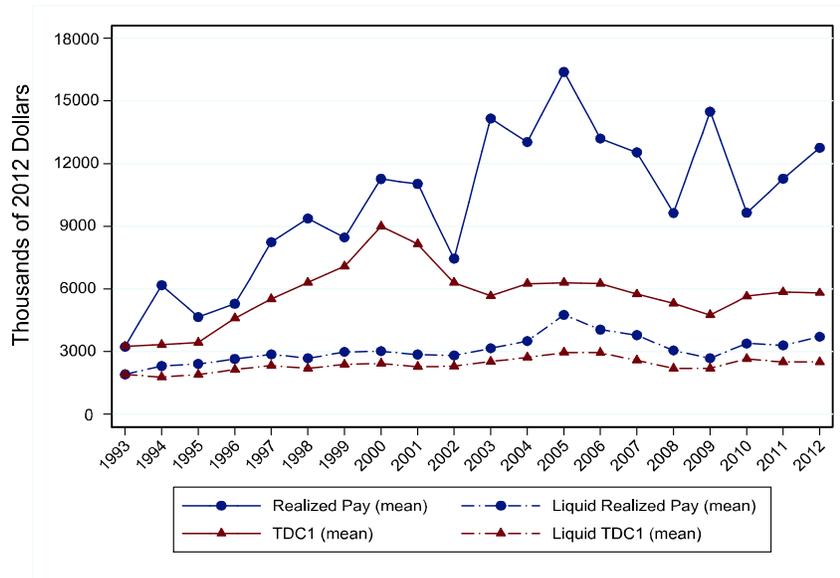
²² We also exclude from our analysis Warren Buffett, the CEO of Berkshire Hathaway, and Larry Ellison, the CEO of Oracle Corporation, because their values of trades are extreme outliers.

²³ We replicate Clementi and Cooley’s simpler calculation of TYC, which uses intrinsic valuations for options when their value is updated with new stock prices at the end of the fiscal year. Clementi and Cooley report in their manuscript that their results do not change substantially when they use Black and Scholes to produce those revaluations.

Figure 2 Mean Realized Pay and Mean Expected Pay as Measured by TDC1 and TYC



different components of pay shows that the estimated gains from trading stock are causing the volatility in realized pay. Also, every year there are a few CEOs who realize very large gains from trading stock, making the averages of the two measures of compensation differ more than the medians. Moreover, the large revaluations of the portfolio of the CEOs with changes in the stock price do not seem to translate into gains from trades, causing the large deviation of the measure TYC from the measure I . One potential explanation would be that CEOs have in their portfolios a large fraction of restricted stock and options, so even if their value increases they are not able to realize those gains. However, the information available in Execucomp about restricted stock and options does not seem to support this hypothesis (the restricted grants are a small part of the portfolio of the CEO at any point in time). However, it is still plausible that implicit selling restrictions are in place even after the explicit vesting period expires, presumably with the objective of strengthening the market perception about the confidence of the CEO in the performance of his own firm.

Figure 3 Liquid Portion of Compensation

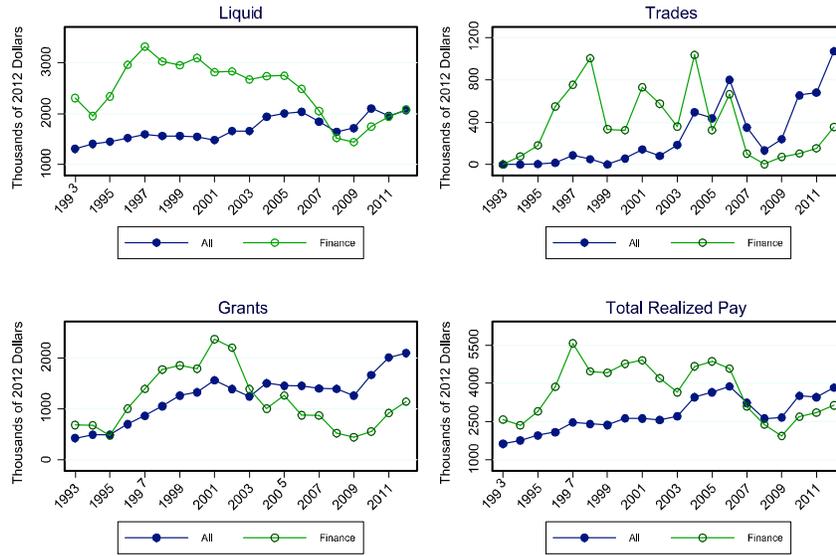
Notes: The blue line presents mean total realized pay, I_t , and its liquid component, L_t (wage, bonus, perks, and dividends). The difference equals mean trades, T_t . The red line presents mean total expected pay as measured in $TDC1_t$ and its liquid component (wage, bonus, and perks). The difference equals grants, G_t .

In Figure 3, we display the liquid portion of compensation for mean realized pay, I_t , and for mean total expected pay as measured in $TDC1_t$. We see that the higher volatility of mean I_t compared to that of mean $TDC1_t$ is mainly driven by the volatility of trades. Figure 4 plots separately the medians of the different components of realized pay, L_t and T_t , and the median of I_t . (Figure 4 plots also these statistics for finance firms, which we will discuss in the next subsection.) Both components, as well as the total I_t , are increasing over time. For comparison, the median value of grants, G_t , is included as well. The value of grants is also increasing over time.

As a robustness check, we replicate Figure 2 in Figure 5 for a subsample of the firms including only the CEOs that own less than 1 percent of the shares of their company.²⁴ The level of TYC_t is much

²⁴ This subsample includes 2,169 out of our 3,345 firms, and 16,302 out of our 34,497 observations.

Figure 4 All CEOs versus Finance CEOs



Notes: A comparison of the medians of liquid compensation, L_t , net gains from trading and stock options, T_t , the expected liquid value of stock and option grants, G_t , and total realized pay, I_t . Note that although $I_t = L_t + T_t$, the sum of the median of L_t and T_t is not equal to the median of I_t .

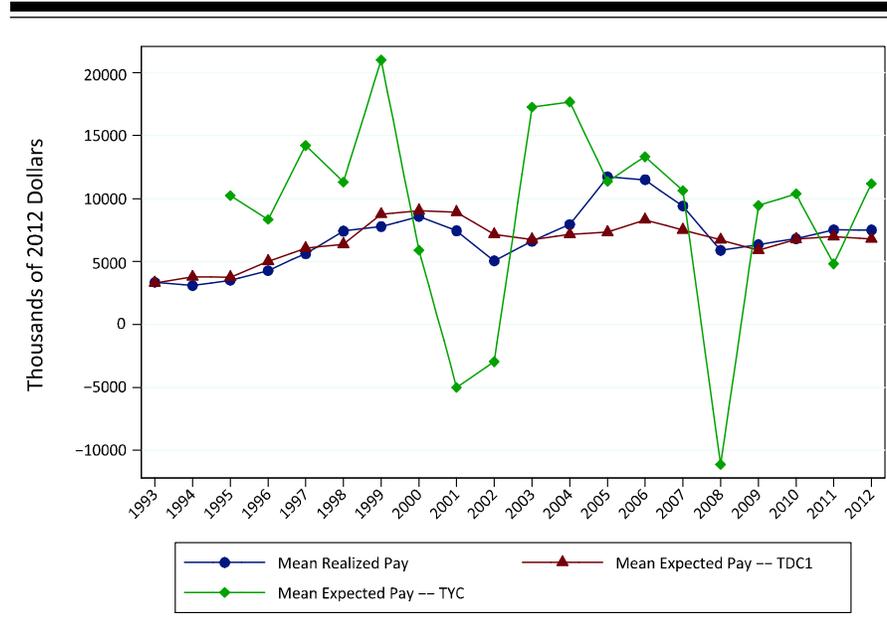
lower, and mean realized pay is sometimes above $TDC1_t$. The main difference for this sample continues to be the higher volatility of TYC_t .

Finance Firms

In Figure 4, we include statistics for firms in the finance sector with the statistics for firms in all sectors.²⁵ Note that firms in the finance sector are, on average, larger (in the sample, the average size in finance is between five and six times larger than the average size for all firms, year by year, with a decreasing trend between 2004 and 2009). Because the level of total compensation (TDC1) has been shown to be positively

²⁵ Firms in the finance sector are those with SIC classification in the 6,000–6,300 range. There are 144 firms per year, on average, in our subsample of finance. We performed the same analysis with a broader category including real estate firms as well as insurance, and the plots looked qualitatively similar.

Figure 5 Mean Realized Pay and Expected Pay, as Measured both by TDC1 and TYC, for CEOs Who Own Less Than 1 Percent of the Stock of Their Firm

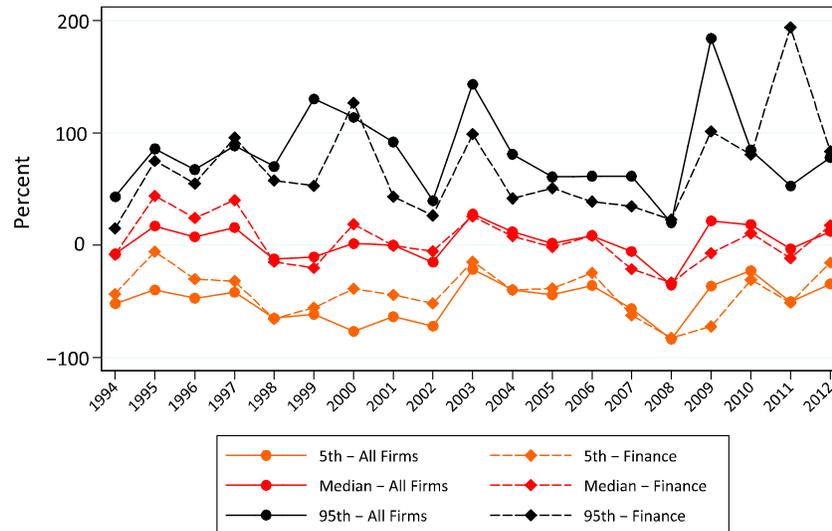


correlated with size, we expect a higher realized pay for CEOs in finance. This is confirmed in the data up to the financial crisis of 2008. Figure 4 shows that the composition of realized pay is slightly different among finance firms, with higher liquid compensation and higher value of trades (which are also more volatile, although this could be due to the smaller number of firms).

When looking in detail at the period since the 2008 financial crisis, it is apparent in the graphs that there has been a steeper decline in median realized pay—both for liquid compensation and trades—for firms in finance than for the full sample of firms. It is worth noting that the median value of grants is, for both groups of firms, well above the median value of trades. The adjustment pattern of median grants during the crisis is similar to that of realized pay, i.e., we see a steeper decline for firms in finance.

Sensitivity of Realized Pay to Performance

Hall and Liebman (1998) provide a measure of sensitivity of pay to performance by using information on stock holdings to construct

Figure 6 Stock Returns by Percentile

Notes: Evolution of the 5th percentile, median, and 95th percentile stock returns for the largest 1,500 firms in our sample. For comparison, the same percentiles of returns for all firms in finance are included as well.

counterfactuals.²⁶ First, they construct a measure of the portfolio of the CEOs, similar to our S_t and O_t holdings of stock and options. Then, using the realized distribution of performances (stock returns), they evaluate the holdings of each CEO in the data for different performance scenarios corresponding to different percentiles of the distribution of returns. We follow this methodology and provide a similar counterfactual for our measure of annual realized pay. An important caveat of this measure is that the quantities of stock traded and of options exercised are assumed to remain constant when stock prices vary in the counterfactual. A model of how these trades would vary in a more realistic setup is beyond the scope of this article.

For our performance counterfactuals, we need to propose the support and distribution of stock returns. For this, we use the observed distribution of stock returns in each given year. We denote the annual

²⁶ Given the limited quantitative importance of bonuses in total compensation, we will ignore changes in bonus payments in our sensitivity analyses.

stock price return as

$$r_t = \frac{p_t - p_{t-1}}{p_{t-1}}. \quad (5)$$

This measure has the advantage of being comparable across firms, as opposed to the stock price itself. In Figure 6, we summarize the evolution of these distributions of returns r_t of the 1,500 largest firms in our sample over time by plotting the return value for the median, and the 5th and 95th percentiles.

Each realization of returns in the support of the distribution can be translated into a stock price for each individual firm using (5). That is, when calculating the counterfactual value of T_t for an individual executive working for firm j , we will construct a counterfactual stock price for various percentiles of the return distribution. We use a hat to denote a variable's counterfactual value, and a superscript nth to indicate the percentile to which we are setting the performance of the firm. For the nth percentile, the counterfactual price for firm j at time t is

$$\hat{p}_{j,t}^{nth} = (1 + r_t^{nth}) p_{j,t-1}.$$

With this price $\hat{p}_{j,t}^{nth}$, a new valuation of $T_{j,t}$ can be produced, assuming the return of the firm was equal to the nth percentile return, r_t^{nth} . Recall that we approximate the gains from trade coming from stock purchases and sales as $\max[0, \bar{p}_t q_t]$, where \bar{p}_t is the average price within the year. We will set the counterfactual for this average price to

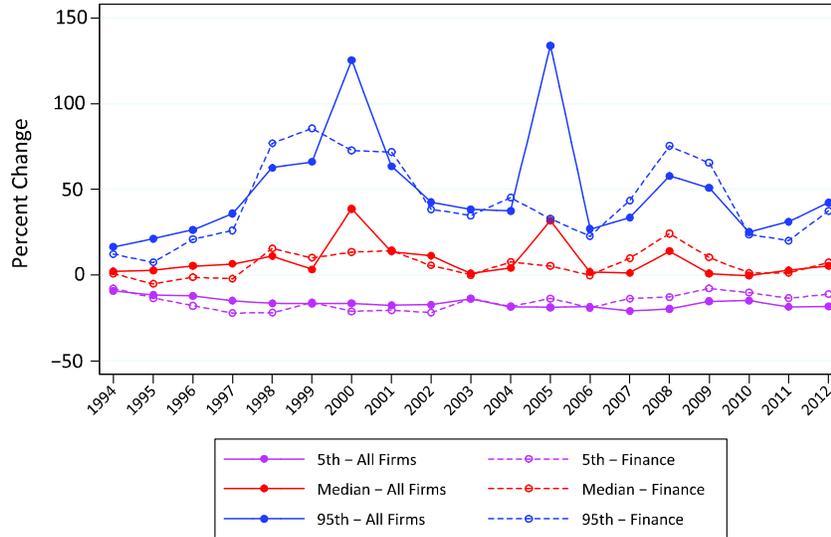
$$\widehat{\bar{p}}_t^{nth} = \frac{\bar{p}_t}{p_t} \hat{p}_t^{nth}, \quad (6)$$

that is, we assume that the proportionality between the average price and the end-of-the-year price is maintained in the counterfactual.

For the portion of the gains from trade that comes from exercising options, we will need several pieces of information. First, in order to compute the net benefit per option exercised, $(\hat{p}_{e_t} - x_g)$, we would need to construct the counterfactual for the stock price at the time of exercise, \hat{p}_{e_t} , possibly using p_{e_t} , and we would need to know the exercise price, x_g , corresponding to each option exercised. Unfortunately, as discussed earlier, we do not know p_{e_t} or x_g (we do not know which particular past grant g was used to purchase the shares). The value of exercised options is recorded in Execucomp:

$$\sum_{o_g \in O_0} o_{g,t}^e (p_{e_t} - x_g) \equiv OPT_EXER_VAL_t \quad \forall t.$$

Figure 7 Mean Counterfactual Income



Notes: Average percentage change in income for three different performance counterfactuals, for all firms and for firms in finance only.

We also have the number of options exercised within the year:

$$o_t^e = \sum_{o_g \in O_0} o_{g,t}^e \equiv OPT_EXER_NUM_t \quad \forall t.$$

To produce an estimate for the counterfactual value of exercising options, we assume $p_{e_t} = \bar{p}$, the average price during the year, and we solve for an “effective” exercise price \tilde{x} using

$$o_t^e (\bar{p} - \tilde{x}) = \sum_{o_g \in O_0} o_{g,t}^e (p_{e_t} - x_g).$$

Finally, we also assume that CEOs do not exercise options in the counterfactual if they are “out of the money” (that is, if $\widehat{p}_t^{nth} < \tilde{x}$). With these assumptions, we have that our counterfactual for gains of trade is

$$\hat{T}_{j,t}^{nth} (S_{j,t}, O_{j,t}) = \max[0, \widehat{p}_t^{nth} q_t] + \max \left[0, o_t^e \left(\widehat{p}_t^{nth} - \tilde{x} \right) \right].$$

This, together with the actual liquid compensation for the executive in the data, L_t , which is not contingent on stock price realizations, amounts to a calculation of a counterfactual $\hat{I}_{j,t}^{nth}$.

The numerical results are listed in Tables 4 (levels) and 5 (percentage changes). We display the percentage changes for the 5th, median, and 95th percentile counterfactuals graphically in Figure 7. Keeping in mind that percentage changes are bounded below by -100 percent, we see that there is an obvious asymmetry in changes when the firm performs better rather than worse. This responds to the uncontingent nature of the wage and the bonus in our calculations. Also, we see in Figure 7 that the gains for the 95th percentile (i.e., outstanding stock return performance) is very extreme in particular years. Two things can lead to high net gains from trade: particularly good stock returns in the given year (i.e., the 95th percentile stock return is an outlier when compared to the other 95th percentile returns in other years) or particularly generous past grants that imply a large number of stock or options are available for trade. We can use the distribution of stock returns, plotted in Figure 5, to track which of the two explanations seems more plausible. The years 2000, 2003, and 2009 represent examples of outlier stock return performance in the 95th percentile; however, only in the year 2000 does this translate into a very large counterfactual mean realized pay in the 95th percentile. The spikes in income for the years 2005 and, to a lesser extent, 2008–09 may correspond instead to particularly large net quantities traded, as computed by us from the portfolios of the CEOs.

Sensitivity for Finance Firms

We observed a sharper decrease in median realized pay for firms in finance during the recent financial crisis (see Figure 4). However, this does not seem to correspond to a very different sensitivity of realized pay to performance for financial firms during the crisis. Tables 5 and 7 replicate the sensitivity analysis of Tables 4 and 6 for firms in finance. That is, using the stock and option holdings of financial firms, we feed in the same percentile stock returns used in Tables 4 and 6 (i.e., those from the distribution of stock for the overall population of firms) to calculate their counterfactual realized pays. We find that the sensitivity estimates align with those of the general sample for the whole sample period.²⁷ It is worth referring back to Figure 4 and noting that the median liquid (uncontingent) compensation of CEOs in finance is

²⁷ Given the way we construct the counterfactuals, any differences in level between Tables 1 and 3 is due to the original differences in the level of actual compensation between the average finance firm and the average firm in the sample.

**Table 4 Counterfactual Income: Mean Level of Income if
Certain Percentile Stock Return Had Been
Achieved—All Firms**

Year	5th	25th	Median	75th	95th	Actual
1994	4,301	5,630	6,378	7,203	8,821	6,182
1995	3,438	4,296	4,926	5,562	6,832	4,659
1996	3,694	4,923	5,734	6,554	8,157	5,286
1997	5,291	7,231	9,083	10,809	14,289	8,243
1998	5,246	9,150	12,397	15,911	24,021	9,364
1999	4,775	7,060	8,668	11,217	20,479	8,460
2000	10,618	27,698	41,383	55,636	87,588	11,268
2001	6,859	11,909	14,968	18,089	27,074	11,022
2002	4,439	6,956	8,910	10,472	13,455	7,448
2003	9,840	12,994	14,718	17,314	26,321	14,156
2004	7,917	11,463	13,211	15,242	20,491	13,023
2005	10,266	14,414	17,072	20,330	26,911	16,382
2006	8,883	11,849	13,802	15,503	19,754	13,200
2007	6,991	10,090	12,326	14,769	19,817	12,531
2008	5,076	8,934	12,177	15,324	20,812	9,636
2009	8,078	11,230	13,450	17,288	28,866	14,474
2010	6,899	8,447	9,459	10,749	13,825	9,640
2011	6,443	8,809	10,325	11,913	15,215	11,270
2012	7,996	10,718	12,080	13,623	18,149	12,747

**Table 5 Counterfactual Income: Mean Level of Income if
Certain Percentile Stock Return Had Been
Achieved—Finance Firms Only**

Year	5th	25th	Median	75th	95th	Actual
1994	3,687	3,987	4,190	4,414	4,851	4,228
1995	5,256	5,700	6,072	6,488	7,679	7,047
1996	6,515	9,468	11,628	13,792	18,017	9,881
1997	6,656	8,323	11,956	15,389	22,304	12,473
1998	5,664	8,173	10,442	12,902	18,538	9,452
1999	6,341	8,800	10,453	12,996	22,127	9,485
2000	6,189	10,105	16,523	24,242	41,568	12,604
2001	8,507	13,528	16,610	19,757	28,852	14,333
2002	6,345	9,544	12,237	14,405	18,536	10,947
2003	8,898	11,051	12,293	14,159	20,626	12,467
2004	8,933	13,099	15,118	17,452	23,474	13,744
2005	11,591	15,828	18,226	21,159	27,077	17,473
2006	9,153	12,596	14,916	16,925	21,937	13,284
2007	10,455	14,958	18,105	21,476	28,427	14,184
2008	4,957	8,523	11,570	14,561	19,803	7,918
2009	5,600	7,378	8,645	10,847	17,502	8,275
2010	5,291	6,322	7,025	7,910	10,125	6,621
2011	4,314	5,350	6,024	6,719	8,152	8,980
2012	5,535	7,174	8,023	8,983	11,791	7,594

Table 6 Counterfactual Income: Mean Percent Change in Income if Certain Percentile Stock Return Had Been Achieved—All Firms

Year	5th	25th	Median	75th	95th
1994	-9.2	-2.2	2.1	6.8	16.4
1995	-11.5	-3.5	2.7	9.0	21.4
1996	-12.1	-1.6	5.4	12.5	26.5
1997	-15.0	-3.7	6.6	16.3	35.9
1998	-16.4	-1.7	11.2	25.5	62.6
1999	-16.6	-5.0	3.4	16.8	66.1
2000	-16.5	13.7	38.8	65.3	125.3
2001	-17.6	1.4	13.7	26.5	63.6
2002	-17.2	-1.5	11.5	22.1	42.4
2003	-13.8	-4.3	1.1	9.4	38.3
2004	-18.6	-3.7	4.2	13.4	37.5
2005	-18.7	4.3	31.8	65.6	133.9
2006	-18.4	-6.4	1.7	8.9	26.9
2007	-20.8	-8.2	1.3	11.8	33.5
2008	-19.7	-1.9	14.0	29.9	57.8
2009	-15.3	-6.1	0.8	13.1	50.8
2010	-14.8	-6.2	-0.4	7.1	25.2
2011	-18.6	-5.8	2.8	12.0	31.3
2012	-18.2	-2.8	5.5	14.8	42.1

particularly large compared to the entire sample, up until the recent crisis. This, together with the fact that sensitivity estimates are similar to those of the overall sample, suggests that the quantities of stock and options held by finance CEOs are larger than those in other industries, hence implementing a similar risk in their realized pay in spite of larger uncontingent compensation levels.

4. CONCLUSION

Information on CEO pay is typically obtained from the mandatory disclosure of compensation required by the SEC for large public firms. A good measure of *realized* pay for CEOs, which includes the actual gains from trading stock rather than their expected value at the time when the firm awards them to the CEO, is not readily in this source. This article discusses how to construct an approximation to the value of realized pay using the partial information compiled in the database Execucomp on the stock owned, bought, and sold by CEOs each year. We present our estimates for the period 1993–2012 and compare them to two alternative measures of *expected* annual total compensation that are frequently used in the media and the academic literature: direct compensation (the sum of salary, bonus, other compensation, and the

Table 7 Counterfactual Income: Mean Percent Change in Income if Certain Percentile Stock Return Had Been Achieved—Finance Firms Only

Year	5th	25th	Median	75th	95th
1994	-7.7	-2.6	0.9	4.8	12.5
1995	-13.3	-8.9	-5.2	-1.1	7.5
1996	-17.8	-8.2	-1.2	6.3	21.0
1997	-22.1	-11.9	-2.2	7.2	26.1
1998	-21.9	-1.6	15.6	34.2	76.9
1999	-16.0	-0.5	10.1	26.6	85.6
2000	-21.0	-3.2	13.5	31.6	72.8
2001	-20.4	0.6	14.4	29.0	71.9
2002	-21.8	-8.1	5.7	17.0	38.5
2003	-13.7	-5.0	0.1	7.9	34.8
2004	-18.3	-1.1	7.9	18.3	45.1
2005	-13.6	-1.9	5.4	14.5	32.9
2006	-18.9	-7.8	-0.2	6.4	22.8
2007	-13.7	-0.2	10.0	20.9	43.4
2008	-12.7	6.8	24.3	42.5	75.4
2009	-7.8	2.6	10.4	24.1	65.5
2010	-10.2	-3.4	1.2	7.1	23.8
2011	-13.5	-4.9	1.1	7.3	20.2
2012	-11.1	0.9	7.6	15.2	37.5

market value of new grants) and total yearly compensation (which includes the year-on-year change in the value of the stock holdings of the CEO). Our measure of realized pay tends to be more volatile over time than direct compensation, mainly due to the volatility of the gains that CEOs realize from trading stock. However, total yearly compensation is markedly more volatile than the other two measures. We find that, while the average realized pay level has historically been at or above that of direct compensation, its median has consistently been lower. We provide descriptive statistics of realized pay for firms in the finance sector. In the aftermath of the crisis the realized pay of CEOs of finance firms seems to have decreased in level relative to the realized pay of CEOs in all industries. Our calculations suggest, however, that realized pay of finance CEOs changes with the performance of their firm in similar magnitudes to that of the average CEO for the whole 1993–2012 period.

APPENDIX

In this Appendix, we show how to map the variables defined in Section 2 to the Execucomp database. We discuss the elements of our ideal measure of compensation that are missing in the data, and what assumptions we make to go around these difficulties.

As we list the objects needed to calculate I_t , we will note how the change in reporting requirements of the SEC in 2006 changes the availability of data (or, sometimes, simply the name of the Execucomp variable that corresponds to a given concept). For this purpose, we will refer to the reporting period before 2006 as P_1 , and the one after as P_2 .

Measuring Liquid Compensation, L_t

Our measure of liquid or cash-based compensation, L_t , is the sum of the executives' annual salary, bonus, dividends, and any perks received within the year, such as contributions to pension plans. Data on annual salary W_t is directly available in Execucomp:

$$W_t \equiv SALARY_t, \forall t.$$

Our measure of bonus, B_t , is the sum of the Execucomp variable BONUS and two variables that capture payments received from hitting "objective" performance targets such as sales growth or stock price performance.²⁸

$$B_t \equiv \begin{cases} BONUS_t + LTIP_t & \text{if } t \in P_1 \\ BONUS_t + NONEQ_INCENT_t & \text{if } t \in P_2. \end{cases}$$

We also have information in the data about the dividend yield (dividends per share, divided by p_t , times 100) that the executive receives from his stock ownership of the company. We back out the total dividend payments as follows:

$$D_t \equiv \frac{DIV_YIELD_t}{100} \times PRCCF_t \times SHROWN_EXCL_OPTS_t \quad \forall t,$$

²⁸ Specifically, after 2005 Execucomp's BONUS variable was modified to only include discretionary or guaranteed bonuses. So to include payments from objective targets, we sum BONUS with NONEQ_INCENT, the amount of income received in the year pursuant to non-equity incentive plans being satisfied. Whenever NONEQ_INCENT is missing (i.e., prior to 2006), we add BONUS with LTIP, the amount of income received in the year pursuant to long-term incentive plans that measure performance over more than one year.

where $PRCCF_t$ is Execucomp's record of the stock price at the closing of the fiscal year:

$$p_t = PRCCF_t \quad \forall t.$$

Finally, our measure of perks and pension payments K_t is the sum of Execucomp variables related to "other compensation":

$$K_t \equiv \begin{cases} ALLOTHTOT_t + OTHANN_t & \text{if } t \in P_1 \\ DEFER_RPT_AS_COMP_t + OTHCOMP_t & \text{if } t \in P_2. \end{cases}$$

Tracking Grants, G_t

Our measure of grant-based compensation G_t is the sum of the value of restricted stock grants and options in the period. We have data on the value of the stock component of that sum, $EV(s_t^r; p_t)$, with the following variables:²⁹

$$EV(s_t^r; p_t) \equiv \begin{cases} RSTKGRNT_t & \text{if } t \in P_1 \\ STOCK_AWARDS_FV_t & \text{if } t \in P_2. \end{cases}$$

In reality, there may be N grants within the year, each with a quantity $s_{t,n}$ and a market price at the time of granting of $p_{t,n}$, for $n = 1 : N$. The variables above that we observe in Execucomp will not have the disaggregated information grant by grant, but rather they correspond to

$$EV(s_t^r; p_t) = \sum_{n=1}^N s_{t,n} p_{t,n}.$$

The value of options awarded in the period is recorded in the data as follows:³⁰

$$EV(o_t^r; x_t, p_t) \equiv \begin{cases} OPTION_AWARDS_BLK_VALUE_t & \text{if } t \in P_1 \\ OPTION_AWARDS_FV_t & \text{if } t \in P_2. \end{cases}$$

²⁹ Both variables measure the value of stock awards as of the grant date. RSTKGRNT was reported by the companies themselves in the Summary Compensation Table, while STOCK_AWARDS_FV is calculated by Execucomp. Strictly speaking, each also contains restricted stock units and phantom stocks.

³⁰ OPTION_AWARDS_BLK_VALUE is calculated by Compustat, during that period of time when—prior to FAS 123R—companies typically expensed options using the "intrinsic value" method, i.e., the difference between grant date stock price and exercise price of the option, which nearly always led to no expensing of options. OPTION_AWARDS_FV is the grant date fair value of option awards in the year, reported by the company per FAS 123R using some version of Black and Scholes (1973) or a similarly accepted calculation.

Again, these variables aggregate all grants within a year, so effectively we will set

$$EV(o_t^r; x_t, p_t) = \sum_{g=1}^M EV(o_{g,t}^r; x_{t_g}, p_{g,t}),$$

where M is the total number of option grants in the year. There is some partial information in Execucomp about the date and exercise price of the different grants for an executive in a given year. However, we do not have their vesting schedule or the date of their exercise (that is, we do not know what the stock market price was at the time when the executive exercised the options). See the related discussion in the realized pay sensitivity analysis in Section 3.

Computing Net Gains from Trading Stock, T_t

We will now define the components of our net gains from trade measure, T_t . To begin, recall that we assume each of these trades happens only once in the fiscal year, and if the executive exercises options, he sells the acquired shares immediately.

The portion of T_t that comes from exercising options is captured by the Execucomp variable `OPT_EXER_VAL`:³¹

$$\sum_{o_{g,t}^e \in O_{t-1}} o_{g,t}^e (p_{e,t} - x_g) \equiv OPT_EXER_VAL_t, \forall t.$$

The portion of T_t that comes instead from buying and selling stock on the open market, $s_1^s p_{s_1} - s_1^b p_{b_1}$, must be estimated, because we cannot observe in the data the quantities s_t^s or s_t^b (and, correspondingly, the prices p_t^s or p_t^b). We use an algorithm similar to Clementi and Cooley (2009) to estimate this difference, with slightly different assumptions that we discuss later in this section. From the law of motion for vested stock in (3), we have that the difference between last year's unrestricted stock holdings and this year's is either coming from the newly vested stock this year, s_t^v , or net purchases. We denote the net quantity of shares sold in t as $q_t \equiv s_t^s - s_t^b$. Rearranging (3) and substituting q_t , we have

$$q_t = S_{t-1} - S_t + s_t^v, \forall t. \quad (7)$$

Typically, q_t will be positive in the data, i.e., the CEO will sell more shares than he buys in a given year. Occasionally, however, q_t

³¹ `OPT_EXER_VAL` is the total value realized from option exercises in the year, and is measured (for each g award, in our notation) as the difference between the exercise price and stock price on the date of exercise.

calculated as in (7) will be negative. This could be due to violations of our assumption that the CEO immediately sells stock acquired through the exercise of options.³² Because we would rather bias our measure of realized pay upward, we set q_t in our calculations equal to the maximum of q_t from (7) and 0.

To calculate q_t using (7) we need S_{t-1} and S_t , which correspond to the CEO's holdings of unrestricted stock. We observe this variable directly in Execucomp.³³

$$S_t \equiv SHROWN_EXCL_OPTS_t, \forall t.$$

We also need the variable s_t^v , the stock vested within the year. This variable maps directly into Execucomp's SHRS_VEST_NUM in the reporting period P_2 . For observations in P_1 , when it is missing, we estimate it by examining annual changes in aggregate restricted stock holdings and annual grants. Specifically:

$$s_t^v \equiv \begin{cases} \left[\begin{array}{l} STOCK_UNVEST_NUM_{t-1} \\ -STOCK_UNVEST_NUM_t + s_t^r \end{array} \right] & \text{if } t \in P_1, \\ SHRS_VEST_NUM_t & \text{if } t \in P_2. \end{cases},$$

where our measurement of the number of stocks granted within the year, s_t^r , is an approximation to the real total number of stock (unavailable in the data) that we recover from $EV(s_t^r)$ by assuming all grants are valued at the average price within the year, denoted \bar{p}_t .³⁴

$$s_t^r = \frac{EV(s_t^r)}{\bar{p}_t}.$$

Note that \bar{p}_t is not in Execucomp. We match the firms in Execucomp to a different database from the Center for Research in Security Prices (CRSP) containing daily stock prices, and we construct the average price ourselves. For this, we take the 12-month window of each firm's fiscal year. To summarize, in our notation, our estimate for the amount of stock vested within t is

$$s_t^v = S_{t-1}^r - S_t^r + s_t^r.$$

Once we get q_t from (7), we estimate the value $s_t^s p_{s_1} - s_t^b p_{b_1}$ by assuming the q_t shares were traded at the average market price over

³² In addition to what we have described, there are two other types of transactions that will change CEO holdings: stock inheritances and stock donations. We abstract from them, as these transactions will typically be small, if non-zero. However, these could also be behind some of the negative q_t in the data.

³³ SHROWN_EXCL_OPTS reports shares of the firm owned by the CEO, excluding options that are exercisable or will become so within 60 days. This amount is reported as of some date between the fiscal year-end and proxy publication.

³⁴ Clementi and Cooley (2009) use the end-of-the-fiscal-year price for this calculation. We choose average price hoping to avoid some of the idiosyncrasy of p_t due to volatility of stocks.

the year, i.e., $p_{s_1} = p_{b_1} = \bar{p}_t$. Given our assumption of non-negative net quantities traded, this amounts to stating

$$s_t^s p_{s,t} - s_t^b p_{b,t} \equiv \max[0, \bar{p}_t q_t].$$

Thus, adding the stock and option portions of T_t , we get

$$T_t(S_{t-1}, O_{t-1}) \equiv \max[0, \bar{p}_t q_t] + OPT_EXER_VAL_t, \forall t.$$

Note that there are two differences between our estimation of net revenue from trade and the calculations in Clementi and Cooley (2009). First, we use average instead of end-of-year prices to recover the quantity of shares granted in a given year, s_t^x , from the value of the grants; this influences our estimate of the net quantities traded, q_t . Second, we use `OPT_EXER_VAL` directly to account for the proceeds of options sales during the year: This variable is the true value of option exercises collected in `Execucomp` and hence uses actual exercise prices and actual stock prices on date of exercise. Clementi and Cooley (2009) instead choose to lump the stock purchases resulting from option exercises in with other stock sales, and they assume that they are acquired at the average price.

REFERENCES

- Aggarwal, Rajesh K., and Andrew A. Samwick. 1999. "The Other Side of the Trade-off: The Impact of Risk on Executive Compensation." *Journal of Political Economy* 107 (February): 65–105.
- Antle, Rick, and Abbie Smith. 1985. "Measuring Executive Compensation: Methods and an Application." *Journal of Accounting Research* 23 (Spring): 296–325.
- Baker, George P., and Brian J. Hall. 2004. "CEO Incentives and Firm Size." *Journal of Labor Economics* 22 (October): 767–98.
- Black, Fischer, and Myron S. Scholes. 1973. "The Pricing of Options and Corporate Liabilities." *Journal of Political Economy* 81 (May/June): 637–54.
- Bebchuk, Lucian, and Jesse Fried. 2004. *Pay without Performance: The Unfulfilled Promise of Executive Compensation*. Cambridge, Mass.: Harvard University Press.

- Bolton, Patrick, Jose Sheinkman, and Wei Xiong. 2006. "Executive Compensation and Short-termist Behaviour in Speculative Markets." *Review of Economic Studies* 73 (July): 577–610.
- Clementi, Gian Luca, and Thomas F. Cooley. 2010. "Executive Compensation: Facts." Working Paper 15426. Cambridge, Mass.: National Bureau of Economic Research (October).
- Clementi, Gian Luca, Thomas F. Cooley, and Cheng Wang. 2006. "Stock Grants as a Commitment Device." *Journal of Economic Dynamics and Control* 30 (November): 2,191–216.
- Edmans, Alex, and Qi Liu. 2011. "Inside Debt." *Review of Finance* 15 (1): 75–102.
- Edmans, Alex, Xavier Gabaix, and Augustin Landier. 2009. "A Multiplicative Model of Optimal CEO Incentives in Market Equilibrium." *Review of Financial Studies* 22 (December): 4,881–917.
- Frydman, Carola, and Raven E. Saks. 2010. "Executive Compensation: A New View from a Long-Term Perspective, 1936–2005." *Review of Financial Studies* 23 (May): 2,099–138.
- Gabaix, Xavier, and Augustin Landier. 2008. "Why Has CEO Pay Increased So Much?" *The Quarterly Journal of Economics* 123 (1): 49–100.
- Garen, John E. 1994. "Executive Compensation and Principal-Agent Theory." *Journal of Political Economy* 102 (December): 1,175–99.
- Gayle, George-Levi, and Robert A. Miller. 2009. "Has Moral Hazard Become a More Important Factor in Managerial Compensation?" *American Economic Review* 99 (December): 1,740–69.
- Gibbons, Robert, and Kevin J. Murphy. 1992. "Optimal Incentive Contracts in the Presence of Career Concerns: Theory and Evidence." *Journal of Political Economy* 100 (June): 468–505.
- Grossman, Sanford J., and Oliver D. Hart. 1983. "An Analysis of the Principal-Agent Problem." *Econometrica* 51 (January): 7–45.
- Hall, Brian J., and Jeffrey B. Liebman. 1998. "Are CEOs Really Paid Like Bureaucrats?" *The Quarterly Journal of Economics* 113 (August): 653–91.
- Hall, Brian J., and Kevin J. Murphy. "Stock Options for Undiversified Executives." *Journal of Accounting & Economics* 33 (February): 3–42.

- Haubrich, Joseph G. 1994. "Risk Aversion, Performance Pay, and the Principal-Agent Problem." *Journal of Political Economy* 102 (April): 258–76.
- Haubrich, Joseph G., and Ivilina Popova. 1998. "Executive Compensation: A Calibration Approach." *Economic Theory* 12 (3): 561–81.
- Jarque, Arantxa. 2010. "Hidden Effort, Learning by Doing, and Wage Dynamics." Federal Reserve Bank of Richmond *Economic Quarterly* 96 (4): 339–72.
- Jarque, Arantxa, and Brian Gaines. 2012. "Regulation and the Composition of CEO Pay." Federal Reserve Bank of Richmond *Economic Quarterly* 98 (4): 309–48.
- Jensen, Michael C., and Kevin J. Murphy. 1990. "Performance Pay and Top-Management Incentives." *Journal of Political Economy* 98 (April): 225–64.
- Jenter, Dirk, and Fadi Kanaan. Forthcoming. "CEO Turnover and Relative Performance Evaluation." *Journal of Finance*.
- Jin, Li, and S. P. Kothari. 2008. "Effect of Personal Taxes on Managers' Decisions to Sell Their Stock." *Journal of Accounting and Economics* 46 (September): 23–46.
- Larcker, David F., Allan L. McCall, and Brian Tayan. 2011. "What Does it Mean for an Executive to 'Make' \$1 Million?" Stanford Closer Look Series No. CGRP-22 (December 14).
- Prescott, Edward S. 1999. "A Primer on Moral-Hazard Models." Federal Reserve Bank of Richmond *Economic Quarterly* 85 (Winter): 47–77.
- Schaefer, Scott. 1998. "The Dependence of Pay-Performance Sensitivity on the Size of the Firm." *The Review of Economics and Statistics* 80 (August): 436–43.
- Spear, Stephen E., and Sanjay Srivastava. 1987. "On Repeated Moral Hazard with Discounting." *Review of Economic Studies* 54 (October): 599–617.
- Wang, Cheng. 1997. "Incentives, CEO Compensation, and Shareholder Wealth in a Dynamic Agency Model." *Journal of Economic Theory* 76 (September): 72–105.

The Business Cycle Behavior of Working Capital

Felipe Schwartzman

Firms require short-term assets or liabilities in order to facilitate production and sales. Those “working capital” requirements are often incorporated in macroeconomic models designed to study the impact of monetary or financial shocks.¹ They are important for the propagation of those shocks since they affect the marginal cost of funds faced by some set of agents in the economy. If firms require working capital in order to acquire variable inputs, a change in the cost of funds faced by firms translates into immediate changes in macroeconomic activity.² This article investigates the cyclical properties of the three main components of working capital—inventories (raw materials, work-in-process, and finished goods), cash and short-term investments, and trade credit—aggregated across all firms and with special attention to their correlations across time with output. The key objective is to obtain stylized facts. While theory informs what kind of facts are worth examining, the uncovering of stylized facts also serves as an input for the development of new theories. The discussion above provides a couple of examples of existing theoretical models that motivate the exploration that follows, but the results stand on their own as useful

■ The views expressed in this article are those of the author and do not necessarily represent the views of the Federal Reserve Bank of Richmond or those of the Federal Reserve System. E-mail: felipe.schwartzman@rich.frb.org.

¹ Technically, the accounting definition of working capital is the difference between the sum of short-term assets and the sum of short-term liabilities. In the article, as in the literature, I use the term more broadly to refer to the collection of short-term assets and short-term liabilities rather than the aggregate accounting concept.

² Examples of articles that model working capital requirements explicitly are Christiano and Eichenbaum (1992) and Fuerst (1992), who develop the canonical model of working capital in monetary economics, and Jermann and Quadrini (2012), who advance working capital as a key part of the transmission mechanism for financial shocks. Working capital also plays a prominent role in the emerging markets business cycles literature, much of which emphasizes the aggregate impact of shocks affecting the supply of foreign funds. Neumeyer and Perri (2004) is a primary example of the latter.

for potentially any theory in which working capital plays a significant role.

In the simplest models, working capital is needed in advance of production. This requirement implies that, so long as data is available at a high enough frequency, the relevant components of working capital ought to be more strongly correlated with future values of cash flows than with current values. This, however, need not be generally the case. In an environment with credit frictions, working capital could also lag production. Credit frictions commonly imply that firms have a borrowing capacity that is increasing in the size of their balance sheet. In particular, interest rates can increase with leverage, as in Bernanke and Gertler (1989), or there might be outright leverage limits, as in Kiyotaki and Moore (1997).³ Models with credit frictions generate endogenous propagation, since profits retained in a given period increase the size of firms' balance sheets, which in turn allow firms to subsequently expand their borrowing and their acquisition of working capital.

To evaluate the lead-lag relationships, I use data from the Financial Accounts of the United States.⁴ The data set is put together by the Federal Reserve Board and distributed online four times per year. The accounts are constructed based on a variety of data sources to provide a comprehensive view of how different sectors of the economy (households and different types of corporations) interact with one another, as well as providing a breakdown of the assets and liabilities held in each one of those sectors. The time series span most of the post-WWII period, from 1952 onward, and I use all of the data in my analysis. The advantage of using this data set over firm-level data, such as COMPUSTAT, is that it provides a comprehensive view of the economy, including noncorporate businesses, whereas COMPUSTAT data only include the largest firms. For all the time series, I compare correlations before and after 1984. This marks the end of the 1981 recession and the beginning of the "Great Moderation." The motivation for splitting the sample follows Lubik, Sarte, and Schwartzman (2014), who find that around the same time as the onset of the Great Moderation there was a marked change in key business cycle properties of the U.S. economy. Strikingly, these changes in correlations survive the end of the Great Moderation after 2008. Since the focus of the article is on correlations

³ These two articles also correspond to the two most widely used microfoundations for credit frictions, which are costly state verification and imperfect commitment, respectively

⁴ These data were previously called the "Flow of Funds Accounts of the United States."

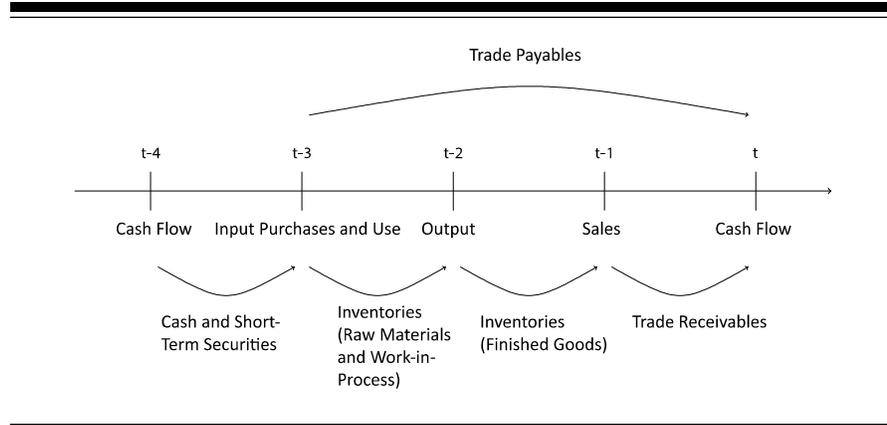
and not on volatilities, I treat the whole period from 1984 onward as a single one.

The findings are as follows: First, inventories lag business cycles in the years before 1984 by about three quarters but by only one or two quarters in the more recent period. This is consistent with the view that before 1984 inventory accumulation was determined by previous cash flow accumulation by firms but less so afterward. The second finding is that cash holdings broadly defined to include short-term investments commonly lead the business cycle, consistent with the cash-in-advance model for short-term production decisions. This echoes classic results by Sims (1972) and updated in Stock and Watson (1999) showing that monetary aggregates are a good leading indicator of output. However, and in contrast to monetary aggregates, the lead-lag relationship between cash holdings and output is considerably more robust, remaining in place in the past 30 years, a period in which the relationship between conventional monetary aggregates and output has broken down. Finally, I find that trade credit lags output, although less markedly than inventories.

This article has a very simple structure. I first discuss in more detail how decisions made by a firm over time can give rise to the various components of working capital. The following three sections examine in turn each of the three major components of working capital (inventories, cash and short-term investments, and trade credit). I provide for each component additional background information about existing theories explaining why firms are willing to hold them, as well as some broad descriptive statistics about how relevant those components are on firms' balance sheets, the long-run trends in those holdings, if any, and the cyclical properties of those different components. The last section concludes.

1. WORKING CAPITAL DEMAND

In models, working capital requirements often arise out of timing restrictions. As an example of such restrictions, consider a firm whose production and sales process follows a seasonal flow, so that cash flows are only realized every four periods $\{\dots, t-4, t, t+4, \dots\}$. As an example of a real activity, one could think of this as a Christmas decorations producer that only sells its products in the last quarter of the year. However, in order to receive a cash flow at t , the firm needs to perform several activities throughout the year that result in accumulating working capital between $t-3$ and t . If one were to look at the balance sheet of this firm, one would see working capital peaking in the quarters

Figure 1 Timeline

between cash flow accumulation periods and the cash flows peaking in periods $\{\dots, t-4, t, t+4, \dots\}$.

Figure 1 shows a detailed breakdown of the production cycle, depicting the different components of working capital. The flows are depicted by the vertical lines and stocks are described by the arrows. In the example, the firm starts the year with some cash flow that it receives in $t-4$. It may choose to distribute some of this cash flow to shareholders as dividends, to use it to pay outstanding debts or to dedicate it to long-term investments. It may also choose to retain some of the cash for future use, an option that is attractive if external funds are costly to acquire.

The production cycle starts in the spring, in $t-3$, with the acquisition and use of inputs, including materials and labor. These can be paid for using the cash that the firm has on its balance sheet or with credit. The typical “cash-in-advance” assumption is that a subset of the inputs that firms acquire in $t-3$ require it to have cash available from the previous period, $t-4$, onward. The required cash may be a leftover of period $t-4$ cash flows that were not put to alternative uses, raised through financial intermediaries, or acquired by issuing new shares. Alternatively, the firm might choose to defer payment for inputs to which the cash-in-advance constraint does not apply, acquiring an account payable. In the example, those accounts payable remain on the firm’s balance sheet until it receives new cash flows in t and uses those to pay the accounts payable out.

The raw materials that the firm purchases in the spring, in $t-3$, are incorporated into raw materials inventories. Some part of it is processed right away, and the combination of the cost of those materials with

labor and overhead costs involved in the processing are incorporated into work-in-process inventories. Raw materials and work-in-process inventories remain on the firm's balance sheet until production is finalized in the summer, in $t - 2$. At that point, all the inventories become finished goods inventories, which remain on the balance sheet until the fall in $t - 1$, when the Christmas decorations producer sells the goods to wholesalers. However, since wholesalers will only sell those goods to final customers in the last quarter of the year, the producer may agree to let them delay the payment, acquiring an account receivable, which is canceled at t . Firms can then use the associated cash flows to cancel outstanding accounts payable and restart the production cycle.

The assumption of a seasonal pattern may be appropriate for certain firms and industries but not for others. Some models of working capital requirements such as in Christiano and Eichenbaum (1992) incorporate a seasonal-like pattern. However, instead of taking place over the year, the seasonality takes place within each period, with working capital being required in the beginning of the period so that cash flows can be realized in the end of the period. Since model periods are chosen to correspond to periods in the data, the seasonality is not observable to an econometrician. A perhaps more natural case (although not usually explicitly modeled in the literature) is for firms to run multiple production processes simultaneously, with working capital being accumulated in any point in time for the sake of production in the following period.

The different forms of working capital assets require the firm to commit funds ahead of cash flows. The marginal cost of those funds can be determined in different ways depending on the details of the environment in which the firms find themselves. In the simplest case in which there are no credit market frictions, the marginal cost of funds dedicated to working capital assets is given simply by the interest rate on financial assets of similar maturity. If, however, credit frictions impose a wedge between the interest rate on borrowing and the return on financial assets, the marginal cost of funds will depend on whether the firm is a borrower. More generally, if the firm faces credit rationing, the marginal cost of funds is given by the return on alternative uses of those funds, for example in illiquid, long-term investment projects.

Finally, note that the demand for different components of working capital emerges for very different reasons. The demand for inventories arises because of a discrepancy between the timing of purchase and use of inputs, production, and sales that is likely to arise largely for technological reasons. However, the demand for cash and trade credit is largely a function of the type of access that the firm and its trading

partners have to payment and credit institutions. We will examine each component of working capital in the following sections.

2. INVENTORIES

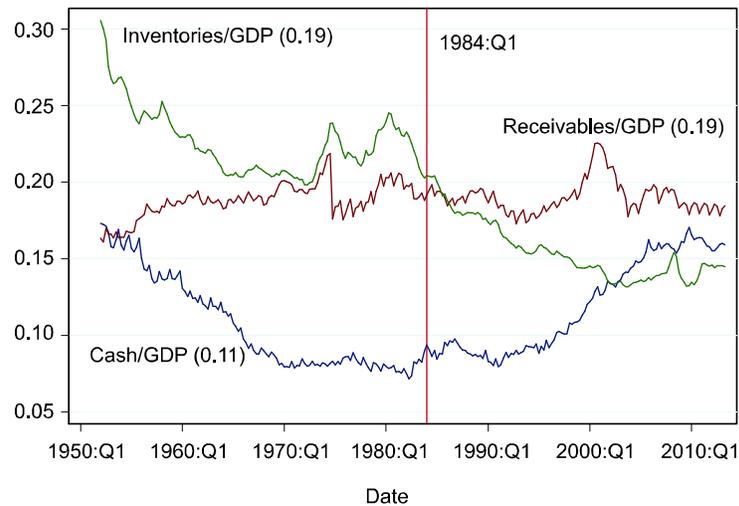
There is a large literature on inventories, some of it summarized in Ramey and West (1999), but it is still evolving. Hornstein (1998) also provides a detailed overview of stylized facts associated with inventory investment. Holding inventories is inherently costly, because by dedicating funds to the purchase of inputs that will only result in cash flows in the future, firms forgo the return on financial investments. Furthermore, they might have to incur storage costs. Given those costs, there are two dominant views of why firms hold inventories. One emphasizes firms' desire to avoid stockouts, i.e., situations in which customers desire to purchase some good or the firm desires to use some input but cannot because it is not available at that moment.^{5,6} The second view points to fixed costs of moving goods between locations, which leads firms to purchase inputs or deliver output to retailers in batches.⁷

In both views, inventories are a pre-condition for sales and, to the extent that these theories also explain the holding of raw materials inventories, they are a pre-condition for production. Given either stockout avoidance or fixed delivery costs, firms choose the inventory/sales ratio to balance out the costs associated with very low inventories against the opportunity cost of funds and storage costs associated with holding those inventories. For a given target inventory/sales ratio, changes in the economic environment that lead firms to increase their prospective sales are, therefore, likely to be accompanied by a prior buildup of inventories. Likewise, changes in the opportunity cost of holding inventories due to less expensive bank credit or lower return on financial investments might also lead firms to build up inventories and, subsequently, increase their cash flow. In both cases, a buildup in inventories precedes increases in cash flows. Alternatively, to the extent that reduced cash holdings are associated with a higher

⁵ For a recent article analyzing the implications of this view for the macroeconomy, see Wen (2011).

⁶ A closely related view is that firms hold inventories in order to smooth production in the face of erratic demand shocks. While still an important building block of inventory models, production smoothing is, by itself, at odds with the fact that production is generally more volatile than sales (Ramey and West 1999).

⁷ See Khan and Thomas (2007) for an analysis of the implications of this view for macroeconomic dynamics.

Figure 2 Components of Working Capital/GDP

Notes: Share of GDP averages are in parentheses.

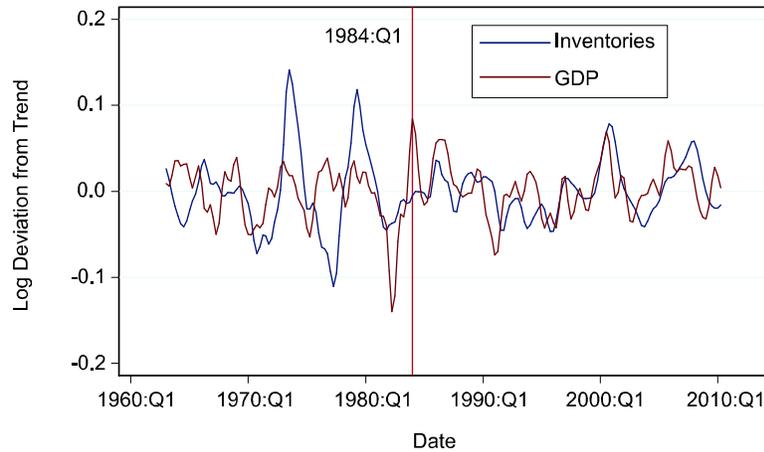
opportunity cost of funds for the firm, a reduction in output or sales may precede reductions in inventories holdings.⁸

Figure 2 shows the evolution of the different components of working capital, as calculated using the Financial Accounts of the United States all normalized by gross domestic product (GDP). The normalization is chosen to control for underlying trends, and to give a sense of the importance of inventories in production. In the specific case of inventories, we can see that between 1952 and 2013 nonfinancial businesses have held an amount of inventories equal to around 19 percent of GDP. Furthermore, from the early 1980s onward there is a well-documented secular decline in the inventories/GDP ratio (Ramey and West 1999).⁹

Figure 3 shows the cyclical component of inventories together with the cyclical component of GDP, where both GDP and inventories were

⁸ More complicated dynamics are certainly possible. For example, if demand for products increases unexpectedly and firms need time to ramp up production, final goods inventories might decline momentarily with an increase in output and sales following that decline.

⁹ When calculating ratios, I use nominal values in both the numerator and the denominator.

Figure 3 Cyclical Components of GDP and Inventories

deflated using the GDP deflator. The cyclical component of the deflated series is extracted using the band-pass filter to isolate variation in the data corresponding to cycles with amplitude between four and 32 quarters. Thus, it excludes seasonal variation (which have an amplitude of four quarters) and fluctuations at lower than what is typically considered business cycle frequencies (which have amplitudes of eight years or fewer), including long-run trends. From the figure, it is almost immediate that inventories have lagged business cycles before the mid-1980s, but that the lead-lag relationship becomes less salient afterward.

Table 1 confirms the visual impression. For each column, the first line of the table shows the correlation of the cyclical component of GDP at t with the cyclical component of inventories in some $t + k$, with each column corresponding to a different value of k . We say that inventories lead output if the peak correlation occurs for $k < 0$ and that it lags output if it occurs for $k > 0$. The table omits standard errors for simplicity, but as a rule of thumb correlations above 0.2 in absolute value are statistically significant. The table shows that before 1984 GDP correlated most with inventories three quarters in the future. After 1984, the peak of the lead-lag difference shortens from three quarters to one quarter, and the difference between the peak and the contemporaneous correlation becomes less salient. The result provides a different perspective on the stylized facts pointed out by Lubik, Sarte, and Schwartzman (2014), who show that inventory/sales

Table 1 Correlations Between Inventories and Measures of Economic Activity

	$t-4$	$t-3$	$t-2$	$t-1$	t	$t+1$	$t+2$	$t+3$	$t+4$
1952–1983									
GDP	-0.48	-0.31	-0.11	0.13	0.39	0.61	0.76	0.81	0.77
Final Sales	-0.50	-0.32	-0.11	0.13	0.38	0.61	0.78	0.84	0.82
Cash Flow 1	-0.56	-0.53	-0.44	-0.29	-0.09	0.15	0.37	0.53	0.60
Cash Flow 2	-0.52	-0.43	-0.29	-0.11	0.11	0.35	0.55	0.66	0.66
1984–2013									
GDP	0.06	0.22	0.40	0.57	0.70	0.78	0.77	0.63	0.40
Final Sales	0.23	0.40	0.55	0.67	0.73	0.77	0.75	0.60	0.37
Cash Flow 1	-0.36	-0.23	-0.09	0.03	0.11	0.21	0.36	0.50	0.58
Cash Flow 2	-0.11	0.05	0.23	0.38	0.47	0.51	0.54	0.52	0.48

ratios were strongly countercyclical prior to 1984 but became acyclical or even somewhat pro-cyclical afterward.

As Figure 1 suggests, production begets inventories, thus implying mechanically the possibility of a lead-lag relationship. The bottom rows of each of the panels in Table 1 examine this possibility by investigating whether the lead-lag relationship uncovered for GDP is also present for final sales and cash flows. Final sales are defined as being equal to GDP with inventory investment excluded from it. For cash flow, I use two alternative definitions. The first one defines cash flows to be equal to net income plus the consumption of capital of both corporate and noncorporate firms. Adding the consumption of capital back to net income is necessary in order to obtain a sensible measure of cash flow since the consumption of capital (which is closely related to depreciation) does not reduce firm cash flows even if it reduces the economic income. The second one adds interest payments, thus separating the ability of the firm to generate cash flow from the financial position of the firm and the timing of interest payments. These definitions of cash flow are imperfect in that net income is recognized at the time of sale, not at the time in which trade receivables are paid out. Thus, in terms of the diagram in Figure 1, the measured cash flow might be recognized closer to time $t-1$ than to t . In all cases, inventories lag the particular flows considered, demonstrating that the lead-lag relationship with output is not an artifact of timing restrictions.

3. CASH AND SHORT-TERM INVESTMENTS

Cash and short-term investments represent cash and all securities readily transferable to cash. This includes, apart from cash on hand,

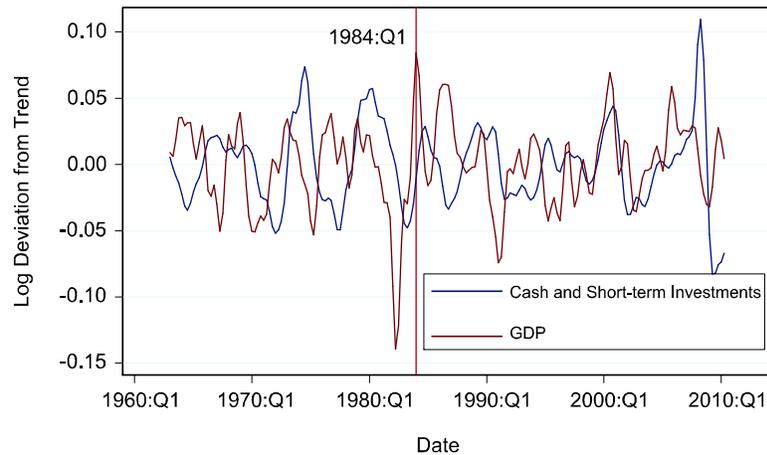
certificates of deposits, commercial paper, government and other marketable securities, demand deposits, etc. Firms hold cash and short-term investments for many reasons, including to facilitate day-to-day payments of variable inputs (Christiano and Eichenbaum 1992), to serve as cushions to allow firms to insure against negative cash flow shocks (Bates, Kahle, and Stulz 2009), to help firms take advantage of fleeting investment opportunities (Kiyotaki and Moore 2012), or to help them with their tax management (Foley et al. 2007). Of those motives, business cycle models in which firms demand cash typically focus on the first, which is the payments for variable inputs. These models are normally posited as “cash-in-advance” models, in which firms need to have cash at hand for a nontrivial period of time before the time in which they use the cash.

For cash-in-advance constraints to play a meaningful economic role, it must be the case that cash pays a rate of return below the opportunity cost of funds for firms. This is trivially the case if cash is understood to include only currency, which pays no interest rate and the value of which declines with inflation. In that case, the opportunity cost of holding cash is given by the nominal rate of interest on bonds. However, firms also hold a variety of assets that are “as good as cash,” in the sense that they either mature very quickly or can be converted into cash at very short notice. The opportunity cost of holding these “short-term” investments is given by their liquidity premia, that is, by the difference between the rate of return on those securities and the rate of return on alternative, illiquid investments.

Using the Financial Accounts of the United States data, I calculate cash and short-term investments for both corporate and noncorporate nonfinancial businesses. For noncorporate businesses, these are the sum of checkable deposits and currency, time and savings deposits, money market fund shares, Treasury securities, and municipal securities. For nonfinancial corporate businesses, cash includes, in addition to those just listed, foreign deposits and agency and GSE-backed securities. From Figure 2, we can see that between 1952 and 2013 corporate businesses have held on average 11 percent worth of GDP in cash. Furthermore, in the last few decades there has been a secular increase in the shares of cash and short-term investments, a fact pointed out in articles by Foley et al. (2007) and Bates, Kahle, and Stulz (2009), among others, who have found firms holding increasing amounts of cash in the last three decades.

Figure 4 shows the cyclical component of cash and short-term investments held by corporate businesses together with the cyclical component of GDP, with both series deflated by the GDP deflator, and

Figure 4 Cyclical Components of GDP and Cash and Short-Term Investments



filtered using the band-pass filter for variations at cycles with amplitudes between four and 32 quarters. As Table 2 makes clear, cash leads business cycles throughout the period under analysis, although the relationship weakens after 1984. The relationship is only hard to discern when cash flow 1 (incoming profits plus depreciation, net of interest expenses) is used as a measure of economic activity, but it is again apparent with cash flow 2 (incoming profits plus depreciation, gross of interest expenses). Such a lead-lag relationship echoes the old monetarist view that money is a good leading indicator for business conditions, as well as formal analysis by Sims (1972), updated by Stock and Watson (1999). Table 3 revisits these results by showing the lead-lag relationship between M2 (which includes currency, demand deposits, money market mutual funds, and other time deposits) and GDP, both deflated by the GDP deflator and band-pass filtered, for the whole sample and broken down before and after 1984. The lead-lag relationship of M2 with GDP is very strong before 1984, but disappears afterward. Given the comparison with the behavior of M2, it is remarkable that the lead-lag relationship between cash and short-term investments held by firms with output is as robust as it is.

The finding goes along with the assertion by Lucas and Nicolini (2013) and Belongia and Ireland (2014) that traditional monetary aggregates do not measure adequately the amount of liquidity in the economy, and that more carefully constructed measures of aggregate

Table 2 Correlations Between Cash and Short-Term Investments and Measures of Economic Activity

	$t-4$	$t-3$	$t-2$	$t-1$	t	$t+1$	$t+2$	$t+3$	$t+4$
	1952–1983								
GDP	0.46	0.60	0.69	0.70	0.61	0.42	0.18	-0.09	-0.32
Final Sales	0.47	0.60	0.68	0.70	0.62	0.44	0.21	-0.04	-0.28
Cash Flow 1	0.17	0.34	0.48	0.57	0.61	0.56	0.40	0.19	-0.01
Cash Flow 2	0.35	0.48	0.55	0.59	0.56	0.44	0.24	0.02	-0.17
	1984–2013								
GDP	0.25	0.41	0.54	0.57	0.51	0.43	0.31	0.19	0.09
Final Sales	0.33	0.42	0.48	0.49	0.46	0.41	0.33	0.22	0.11
Cash Flow 1	0.08	0.08	0.07	0.08	0.12	0.20	0.28	0.33	0.34
Cash Flow 2	0.20	0.26	0.28	0.24	0.20	0.18	0.17	0.16	0.17

liquidity have retained the ability to forecast output. Of course, a measure of liquidity based on cash and short-term investments held by firms is distinct from measures such as M2 or others in that it does not include cash held by households. A closer investigation of whether liquid assets held by firms are specially correlated with future output as compared to those held by households is an interesting avenue for future work.

4. TRADE CREDIT

The third major component of working capital is trade credit, with trade receivables as part of the assets and trade payables as part of the liabilities. Trade receivables represent amounts owed by customers for goods and services sold in the ordinary course of business. Conversely, trade payables represent trade obligations due within one year, or the normal operating cycle of the company.

Trade credit is an active area of research in corporate finance, with an abundant theoretical and empirical literature. To a large degree, theories of trade credit emphasize the fact that, relative to financial institutions, suppliers often have advantages in securing repayment from their customers. Among other reasons for that advantage, the literature mentions information advantages for suppliers (Mian and Smith 1992), incentives for customers to preserve their relationship with suppliers (Cuñat 2007), and the fact that, since goods are harder to divert than cash, borrowers have less incentive to default (Burkart and Ellingsen 2004).

The opportunity cost of holding trade receivables is given by the difference between the rate of return on alternative investments and

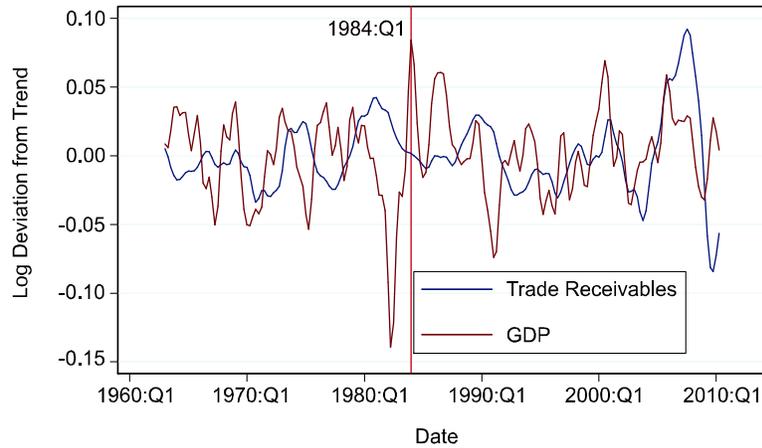
Table 3 Correlations Between M2 and Short-Term Investments and Measures of Economic Activity

	$t-4$	$t-3$	$t-2$	$t-1$	t	$t+1$	$t+2$	$t+3$	$t+4$
1952–1983									
GDP	0.66	0.77	0.82	0.78	0.65	0.46	0.23	-0.01	-0.22
Final Sales	0.66	0.76	0.80	0.78	0.67	0.47	0.22	-0.02	-0.23
Cash Flow 1	0.24	0.44	0.62	0.75	0.80	0.74	0.61	0.43	0.23
Cash Flow 2	0.45	0.60	0.71	0.77	0.74	0.62	0.43	0.22	0.02
1984–2013									
GDP	-0.02	-0.08	-0.14	-0.22	-0.29	-0.32	-0.27	-0.15	-0.01
Final Sales	-0.14	-0.20	-0.23	-0.26	-0.29	-0.30	-0.25	-0.14	0.01
Cash Flow 1	0.06	0.01	-0.08	-0.19	-0.31	-0.42	-0.49	-0.52	-0.50
Cash Flow 2	-0.07	-0.13	-0.21	-0.31	-0.38	-0.41	-0.38	-0.30	-0.20

the interest rate paid by customers. If the latter is smaller than the former, it will be costly for firms to hold trade receivables. Conversely, there is a cost associated with issuing trade payables if the interest rate on trade payables is higher than the rate of return on real or financial investments.

When analyzing trade credit, I focus on trade receivables, which I define to include consumer credit held by corporate and noncorporate nonfinancial firms. Including consumer credit follows the spirit of including in trade receivables all short-term credit conceded by the firm to other parties in order to facilitate production and sales. I focus only on receivables rather than payables since, in a closed economy, whenever a firm issues a trade payable, the counterpart acquires a trade receivable. Because the U.S. economy is not closed, the two numbers do not exactly coincide. Furthermore, even after accounting for foreign holdings and issuance of trade credit, the difficulties in collecting accurate data are significant enough that there exists a nontrivial discrepancy between aggregate trade payables and aggregate trade receivables. Finally, trade payables do not include consumer credit. In spite of those differences, both measures of trade credit behave very similarly, so that for brevity I will only discuss trade receivables.

From Figure 2 we can see that between 1952 and 2013 corporate businesses hold a value of trade receivables equal to 19 percent of GDP. Furthermore, unlike inventories and cash, there is no clear trend in the ratio of trade receivables to GDP. Figure 5 shows the cyclical component of receivables together with the cyclical component of GDP, both deflated using the GDP deflator and extracted using a band-pass filter for frequencies between four and 32 quarters. Table 4 presents the cross-time correlation. Trade receivables lag output by a quarter both

Figure 5 Cyclical Components of GDP and Trade Receivables

before and after 1984. This is in line with the diagram depicted in Figure 1, which predicts that firms accumulate trade receivables after production and sales have taken place. A comparison with final sales and the different measures of cash flow shows a similar pattern. This is still in line with the diagram, since net income is recognized at the time of sale, not at the time in which final payment is received. Thus, to the extent that firms tend to provide financing for their customers, one would expect trade receivables to lag cash flows defined using data from income.

5. CONCLUSION

Working capital is an important part of many macroeconomic models that emphasize the impact of fluctuations in the cost of capital on firm decisions. I find that the cyclical properties of the different components are quite different. In particular, cash holdings consistently lead the business cycle, whereas inventories and trade receivables are lagging. Interestingly, the lead-lag relationships for inventories appear to weaken after 1984. To the extent that those relationships are indicators of payment and financial frictions, the reductions in the lead-lag relationships between inventories and economic activity are consistent with the view, argued by Jermann and Quadrini (2006), that financial markets became more efficient after the early 1980s. A second set of interesting facts concerns cash holdings, which are particularly

Table 4 Correlation of Trade Receivables with Different Measures of Economic Activity

	$t-4$	$t-3$	$t-2$	$t-1$	t	$t+1$	$t+2$	$t+3$	$t+4$
	1952–1983								
GDP	−0.25	−0.03	0.25	0.53	0.73	0.81	0.78	0.68	0.52
Final Sales	−0.22	−0.03	0.23	0.50	0.73	0.84	0.83	0.74	0.58
Cash Flow 1	−0.55	−0.49	−0.32	−0.04	0.25	0.47	0.59	0.63	0.60
Cash Flow 2	−0.40	−0.29	−0.08	0.18	0.45	0.64	0.70	0.62	0.48
	1984–2013								
GDP	−0.04	0.13	0.32	0.50	0.64	0.71	0.71	0.64	0.54
Final Sales	0.08	0.25	0.41	0.56	0.67	0.72	0.71	0.65	0.53
Cash Flow 1	−0.47	−0.35	−0.22	−0.07	0.09	0.25	0.38	0.47	0.50
Cash Flow 2	−0.14	0.04	0.19	0.31	0.40	0.45	0.47	0.47	0.45

noteworthy because the facts are robust over time. This is in contrast to the lead-lag relationship between M2 and GDP, which broke down after the 1980s. The results suggest that availability of cash is an important precursor of economic activity, giving some credence to models that emphasize cash-in-advance type constraints.

REFERENCES

- Bates, Thomas W., Kathleen M. Kahle, and René M. Stulz. 2009. “Why Do U.S. Firms Hold So Much More Cash Than They Used To?” *The Journal of Finance* 64 (October): 1,985–2,021.
- Belongia, Michael T., and Peter N. Ireland. 2014. “Interest Rates and Money in the Measurement of Monetary Policy.” Working Paper 20134. Cambridge, Mass.: National Bureau of Economic Research (May).
- Bernanke, Ben, and Mark Gertler. 1989. “Agency Costs, Net Worth, and Business Fluctuations.” *The American Economic Review* 79 (March): 14–31.
- Burkart, Mike, and Tore Ellingsen. 2004. “In-Kind Finance: A Theory of Trade Credit.” *The American Economic Review* 94 (June): 569–90.

- Christiano, Lawrence J., and Martin Eichenbaum. 1992. "Liquidity Effects and the Monetary Transmission Mechanism." Working Paper 3974. Cambridge, Mass.: National Bureau of Economic Research (January).
- Cuñat, Vicente. 2007. "Trade Credit: Suppliers as Debt Collectors and Insurance Providers." *Review of Financial Studies* 20 (March): 491–527.
- Foley, C. Fritz, Jay C. Hartzell, Sheridan Titman, and Garry Twite. 2007. "Why Do Firms Hold So Much Cash? A Tax-Based Explanation." *Journal of Financial Economics* 86 (December): 579–607.
- Fuerst, Timothy S. 1992. "Liquidity, Loanable Funds, and Real Activity." *Journal of Monetary Economics* 29 (February): 3–24.
- Hornstein, Andreas. 1998. "Inventory Investment and the Business Cycle." Federal Reserve Bank of Richmond *Economic Quarterly* 84 (Spring): 49–72.
- Jermann, Urban, and Vincenzo Quadrini. 2006. "Financial Innovations and Macroeconomic Volatility." Working Paper 12308. Cambridge, Mass.: National Bureau of Economic Research (June).
- Jermann, Urban, and Vincenzo Quadrini. 2012. "Macroeconomic Effects of Financial Shocks." *The American Economic Review* 102 (February): 238–71.
- Khan, Aubhik, and Julia K. Thomas. 2007. "Inventories and the Business Cycle: An Equilibrium Analysis of (S, s) Policies." *The American Economic Review* 97 (September): 1,165–88.
- Kiyotaki, Nobuhiro, and John Moore. 1997. "Credit Cycles." *The Journal of Political Economy* 105 (April): 211–48.
- Kiyotaki, Nobuhiro, and John Moore. 2012. "Liquidity, Business Cycles, and Monetary Policy." Working Paper 17934. Cambridge, Mass.: National Bureau of Economic Research (March).
- Lubik, Thomas, Pierre-Daniel Sarte, and Felipe Schwartzman. 2014. "What Inventories Tell Us About How Business Cycles Have Changed." Manuscript.
- Lucas, Jr., Robert E., and Juan Pablo Nicolini. 2013. "On the Stability of Money Demand." Manuscript, University of Chicago.
- Mian, Shehzad L., and Clifford W. Smith, Jr. 1992. "Accounts Receivable Management Policy: Theory and Evidence." *The Journal of Finance* 47 (March): 169–200.

- Neumeyer, Pablo A., and Fabrizio Perri. 2005. "Business Cycles in Emerging Economies: The Role of Interest Rates." *Journal of Monetary Economics* 52 (March): 345–80.
- Ramey, Valerie A., and Kenneth D. West. 1999. "Inventories." In *Handbook of Macroeconomics, Vol. 1*, edited by J. B. Taylor and M. Woodford. Philadelphia: Elsevier, 863–923.
- Sims, Christopher A. 1972. "Money, Income, and Causality." *The American Economic Review* 62 (September): 540–52.
- Stock, James H., and Mark W. Watson. 1999. "Business Cycle Fluctuations in U.S. Macroeconomic Time Series." In *Handbook of Macroeconomics, Vol. 1*, edited by J. B. Taylor and M. Woodford. Philadelphia: Elsevier, 3–64.
- Wen, Yi. 2011. "Input and Output Inventory Dynamics." *American Economic Journal: Macroeconomics* 3 (October): 181–212.

Pecuniary Externalities, Segregated Exchanges, and Market Liquidity in a Diamond-Dybvig Economy with Retrade

Borys Grochulski

Price changes affect economic agents primarily by altering their budget constraints. In many economic environments, however, price changes additionally impact the agents by altering other constraints agents face. Those additional ways in which prices affect agents, other than through budget constraints, are known as pecuniary externalities.¹ Examples of the additional constraints that can be affected by prices include incentive compatibility, participation, and collateral constraints.

Numerous recent macroeconomic studies have shown that pecuniary externalities can lead to market failure.² The intuition behind this failure is as follows. In standard Arrow-Debreu economies, where

■ The author would like to thank Tee Kilenthong, Sam Marshall, Wendy Morrison, Pierre Sarte, Felipe Schwartzman, and Ned Prescott for their helpful comments. The views expressed in this article are those of the author and not necessarily those of the Federal Reserve Bank of Richmond or the Federal Reserve System. E-mail: borys.grochulski@rich.frb.org.

¹ The term pecuniary externality has been used more broadly than this definition. Viner (1932) uses it to describe the impact of a change in the price of an input on the production cost curve of a firm. Greenwald and Stiglitz (1986) and, more recently, Bianchi (2011) and Dávila et al. (2012) among others, use it in reference to the generic constrained-inefficiency of competitive equilibria in economies with exogenously imposed market incompleteness (studied in, e.g., Stiglitz [1982] and Geanakoplos and Polemarchakis [1986]). In the language of Prescott and Townsend (1984), the definition we use corresponds to prices having a direct impact on the agents' consumption possibility set, in addition to the budget constraint.

² See, e.g., Kehoe and Levine (1993), Golosov and Tsyvinski (2007), Lorenzoni (2008), and Di Tella (2014).

prices only affect budget constraints, equilibrium allocations are efficient. It is therefore impossible to alter equilibrium prices (perhaps by imposing taxes) and obtain a Pareto improvement (i.e., make an agent better off without making someone else worse off). An increase in the price of good x , for example, will relax budget constraints of some agents, loosely speaking the sellers of x , making them better off, but it will tighten budget constraints of others, the buyers of x , making this group worse off. The equilibrium price of good x cannot therefore be improved upon in Pareto sense.

The same may no longer hold true when prices affect not only budget but also some other constraints that can be tightened or relaxed for all agents simultaneously. If an increase in the price of x relaxes everyone's incentive compatibility constraint, for example, then not only the sellers of x but also the buyers of x can benefit from a higher price of x , as long as the relaxed incentive constraint helps them more than the tightened budget constraint hurts them. The benevolent social planner—a stand-in concept we use to calculate optimal allocations—will take this effect into account. In a market economy, however, agents take prices as independent of their individual actions. By ignoring the general equilibrium impact of their actions on prices, agents also ignore the indirect effect they have on how tight their own incentive constraints are. The planner's and the agents' costs-benefit calculus are thus different, which leads to suboptimal equilibrium outcomes.

By relaxing a constraint that all agents face, a high price of good x has in the preceding example a positive “external” effect similar to, e.g., a clean environment or a good public highway system. Agents' inability to coordinate on a sufficiently high price for good x in equilibrium is therefore similar to the failure to internalize an external effect, which has led to the name pecuniary externality.

In this article, we discuss the pecuniary externality that leads to underprovision of liquidity in the banking model of Diamond and Dybvig (1983) (hereafter, DD). We introduce the DD economy in Section 1. In this economy, agents have access to two assets: a short-term, liquid asset with net return normalized to zero and a long-term, illiquid asset with positive net return $\hat{R} - 1 > 0$. Agents face random liquidity shocks: They may become impatient, i.e., find themselves having to consume before the illiquid asset matures, or remain patient, in which case they can postpone consumption until the illiquid asset pays off. By investing a part of their initial endowment/wealth in the low-yielding liquid asset, agents purchase insurance against the liquidity shock.

In Section 2, we derive the efficient allocation of liquidity in this economy, i.e., the optimal levels of investment in the two assets along with the resulting amounts of consumption for the agents who do and

do not experience the need for liquidity. At the optimum, the liquidity shock is partially insured: The impatient agents are able to capture a part of the return on the long-term asset despite the fact that they have to consume before this asset matures.

There are several variants of the DD model in the literature. The variant we consider follows closely Jacklin (1987) and Farhi, Golosov, and Tsyvinski (2009). It has been designed to focus on market provision of liquidity and not on the possibility of bank runs.³ In particular, we assume that liquidity shocks are agents' private information, but we do not assume a sequential service constraint: Trade can be organized after all agents have received their realizations of the liquidity shock. To study pecuniary externalities, we follow Farhi, Golosov, and Tsyvinski (2009) in giving the agents access to an anonymous, hidden market in which they can borrow and lend at the market-determined gross rate of return R . As this rate of return (the price of credit) affects incentive compatibility constraints, it gives rise to a pecuniary externality. This pecuniary externality makes competitive equilibria inefficient.

To show this inefficiency, we analyze in Section 3 a simple model of trade with incomplete markets. In this model, agents invest directly in the two assets *ex ante* and trade the long-term asset for cash *ex post*, i.e., after they find out their liquidity needs. Diamond and Dybvig (1983) showed that competitive equilibrium in this simple, incomplete-markets model is inefficient. In this model, a no-arbitrage condition determines how the return on the long-term asset is allocated in equilibrium: The whole net return $\hat{R} - 1$ is captured by the patient agents, leaving the impatient agents with zero net return on their investment, which is too low relative to the optimal allocation. In this incomplete-markets equilibrium, thus, agents do not obtain sufficient liquidity insurance.

This inefficiency prevails even when markets for state-contingent contracts are introduced. Jacklin (1987) and Farhi, Golosov, and Tsyvinski (2009) show that when agents can borrow and lend privately in a hidden retrade market, liquidity is underprovided in competitive equilibrium with complete markets and fully state-contingent contracts (or banks). The inefficiency is caused by a pecuniary externality that, as we mentioned, enters the model through the agents' incentive compatibility constraints that depend on the retrade interest rate R . In equilibrium, this interest rate is too high, which, by arbitrage, forces the secondary-market price for the long-term asset to be too low. The impatient agents, thus, re-sell their holdings of the long-term asset

³ See Ennis and Keister (2010) for a review of the literature on bank runs in the DD model.

in the secondary market for too little. As in the incomplete-markets model, they are unable to capture any part of the long-run net return $\hat{R} - 1$, which again is inefficient. We review this result in detail in Section 4.

As is the case with standard externalities like pollution, the market failure caused by the pecuniary externality creates a role for government intervention. Farhi, Golosov, and Tsyvinski (2009) consider direct government intervention imposing a minimum requirement on the level of liquid investment. They show that this intervention decreases the retrade interest rate R and increases the return on the initial investment in the liquid asset. This allows the impatient agents to capture a part of \hat{R} and eliminates the effect of the pecuniary externality.⁴

If the extent of an externality can be costlessly and verifiably quantified, the problem of excessive externality can also be addressed with a more decentralized approach that can be implemented through the so-called cap-and-trade mechanism. An explicit assignment of property rights over the extent of the externality lets markets for these rights emerge. In these markets, agents face prices for generating the externality, which makes them take into account the full impact of the externality and thus restores the efficiency of the equilibrium outcome.⁵ Pollution is a textbook example of a negative external effect. Currently, emission of greenhouse gasses is regulated through the cap-and-trade mechanism in many countries.⁶

In a recent article, Kilenthong and Townsend (2011) (hereafter, KT) study a market solution to the pecuniary externality problem analogous to cap-and-trade.⁷ In addition to a class of moral hazard environments, they consider a DD economy with retrade.⁸ In their model, the impact of one's liquidity demand on the retrade interest rate is priced, which results in efficient ex ante investment, sufficient liquidity, and an optimal amount of retrade in competitive equilibrium. Clearly, this approach is interesting because it implies no need for direct government intervention into markets. Similar to the cap-and-trade

⁴ In this article, we do not present details of the implementation of this intervention. The interested reader is referred directly to Farhi, Golosov, and Tsyvinski (2009).

⁵ See Chapter 11 of Mas-Colell, Whinston, and Green (1995).

⁶ The first and to-date largest implementation of this mechanism is the European Union Emission Trading Scheme; see Ellerman and Buchner (2007).

⁷ Bisin and Gottardi (2006) use a similar approach in the Rothschild-Stiglitz adverse selection economy.

⁸ Kilenthong and Townsend (2014a) study the model with segregated exchanges in a class of environments with collateral constraints. Kilenthong and Townsend (2014b) extend the analysis of segregated exchanges to a generalized framework nesting collateral and liquidity constraints, incentive constraints with retrade, and exogenously incomplete markets.

mechanism, this approach requires that agents' activities generating the externality—in this case retrade—be observable. We discuss the KT model in Section 5.

In the KT market model, retrade is allowed but only within access-controlled ex-post markets called segregated exchanges. Agents are admitted to membership in an exchange upon payment of an entry fee. The size of the entry fee depends on the composition of the agent's investment portfolio. The defining characteristic of a segregated exchange is the price at which agents expect to be able to (re)trade the long-term asset ex post. In equilibrium, these expectations must be correct. This market structure is free of pecuniary externalities because agents can no longer take retrade prices as independent of their actions. The portfolio-contingent exchange entry fee, similar to the price for greenhouse gas emissions in the cap-and-trade mechanism, creates an explicit connection between the investment decisions an agent makes ex ante and the price at which he is able to trade ex post. Consequently, equilibrium with segregated exchanges does not suffer from the problem of underprovision of liquidity, and the market outcome is efficient.

Our exposition of the KT mechanism in Section 5 extends the exposition in Kilenthong and Townsend (2011). We explicitly solve for equilibrium entry fees associated with each segregated exchange and show how with these prices the agent's ex ante utility maximization problem becomes aligned with the planner's problem of maximization of ex ante welfare.

In Section 6, we conclude the article with a discussion of the question of whether the possibility of retrade in the DD model implies the need for government intervention. The literature we review makes it clear that the answer depends on the agents' ability to commit themselves to restrict retrade to access-controlled venues with priced entry. This means that retrade itself does not imply the existence of a pecuniary externality requiring government intervention, only hidden retrade without commitment does. Which of these two kinds of retrade possibilities financial firms face in reality is an important empirical question.

The Appendix contains proofs of two auxiliary results and a precise definition of the incomplete-markets equilibrium studied in Section 3. Table 1 summarizes the frictions and outcomes associated with all allocation mechanisms we discuss in this article.

1. A DIAMOND-DYBVIIG ECONOMY WITH RETRADE

The version of the Diamond-Dybvig economy that we consider here is close to those studied in Jacklin (1987); Allen and Gale (2004); Farhi, Golosov, and Tsyvinski (2009); and Kilenthong and Townsend (2011). There is a continuum of ex ante identical agents. There are three dates: $t = 0, 1, 2$. There is a single consumption good at each date. Each agent is endowed with resources e at date 0. These resources can be invested in two available technologies/assets. The short-term asset pays the return of 1 unit of the consumption good at date 1 per unit of resources invested at date 0. We will often refer to this asset as the cash asset. The long-term asset pays nothing at date 1 and $\hat{R} > 1$ at date 2 per unit invested at date 0. Note that the long-term asset is technologically illiquid at date 1, i.e., it cannot be physically turned into the consumption good.

Agents do not consume at date 0. Their preferences over consumption at dates 1 and 2 are represented by a DD utility function

$$u(c_1 + \theta c_2),$$

where $\theta \in \{0, 1\}$ is an idiosyncratic shock with $\Pr\{\theta = 0\} = \pi > 0$. Note that if $\theta = 0$, the agent is extremely impatient: He only values consumption at date 1. The standard interpretation of this shock is that with $\theta = 0$ the agent experiences at date 1 a critical need for liquidity. If $\theta = 1$, however, the agent is extremely patient: He is in fact indifferent to the timing of consumption between dates 1 and 2.⁹ We follow DD in assuming that relative risk aversion is larger than 1, i.e., $-cu''(c)/u'(c) > 1$ for all c . As we will see, this assumption implies that the impatient agents will be allocated consumption with present value larger than the value of their initial endowment e .

A consumption allocation c consists of $\{c_1(0), c_2(0), c_1(1), c_2(1)\}$, where $c_t(\theta) \geq 0$ denotes date- t consumption for an agent with shock θ . Associated with allocation c are initial asset investment $s \geq 0$ in the liquid asset and $x \geq 0$ in the illiquid asset. To ensure that resources at date 1 and 2 are sufficient to provide consumption as specified in c , initial investment (s, x) associated with allocation c must satisfy

$$s \geq \pi c_1(0) + (1 - \pi)c_1(1), \tag{1}$$

⁹ Note that with these preferences the DD economy violates standard smoothness and convexity assumptions. In particular, the shadow interest rate (i.e., the rate at which an agent is willing to refrain from borrowing or saving) is plus infinity for the impatient type and one for the patient type regardless of the allocation of consumption.

and

$$\hat{R}x \geq \pi c_2(0) + (1 - \pi)c_2(1). \tag{2}$$

The amounts s and x that can be invested in the two technologies are constrained by the amount e of resources available at date 0:

$$s + x \leq e. \tag{3}$$

Substituting (1) and (2) into (3), we can express the economy’s aggregate resource constraint in terms of just the consumption allocation c :

$$\pi \left(c_1(0) + \frac{c_2(0)}{\hat{R}} \right) + (1 - \pi) \left(c_1(1) + \frac{c_2(1)}{\hat{R}} \right) \leq e. \tag{4}$$

Allocation c gives an agent an expected utility value of

$$\mathbb{E}[u(c_1 + \theta c_2)] = \pi u(c_1(0)) + (1 - \pi)u(c_1(1) + c_2(1)). \tag{5}$$

Since all agents are ex ante identical, the expected utility of the representative agent measures total utility, or social welfare, attained in this economy.

We follow DD in assuming that realizations of θ are private information. That is, given an allocation $c = \{c_1(0), c_2(0), c_1(1), c_2(1)\}$, an agent can obtain either $\{c_1(0), c_2(0)\}$ or $\{c_1(1), c_2(1)\}$ depending on what realization of θ he reports.

In addition, we follow Farhi, Golosov, and Tsyvinski (2009) and Kilenhong and Townsend (2011) in assuming that individual final consumption is also private and that agents have access to a hidden retrade market where they can lend and borrow from one another “behind the back” of the planner, i.e., with all trades in this market being hidden from everyone but the parties directly involved. More precisely, at date 1 agents have access to a perfectly competitive market for one-period IOUs. Given an allocation $c = \{c_1(0), c_2(0), c_1(1), c_2(1)\}$, an agent reporting shock realization $\tilde{\theta}$ obtains the bundle $(c_1(\tilde{\theta}), c_2(\tilde{\theta}))$. But this bundle does not have to be his actual consumption. Rather, this bundle becomes his endowment of goods in the hidden retrade market. The agent’s final consumption is determined by his retrade activity. At the hidden-market interest rate R , the agent can either save some of his $c_1(\tilde{\theta})$ for consumption at date 2, or borrow against $c_2(\tilde{\theta})$ for consumption at date 1. Specifically, given an allocation c and a gross interest rate R in the hidden retrade market, an agent of type θ selects a report $\tilde{\theta} \in \{0, 1\}$, IOU purchases b , and a final consumption bundle

$(\tilde{c}_1, \tilde{c}_2) \geq (0, 0)$ that solve

$$\begin{aligned} \tilde{V}(c, R; \theta) &= \max_{\tilde{\theta}, \tilde{c}_1, \tilde{c}_2, b} u(\tilde{c}_1 + \theta \tilde{c}_2) \\ &\text{s.t.} \\ &\tilde{c}_1 + b \leq c_1(\tilde{\theta}), \\ &\tilde{c}_2 \leq Rb + c_2(\tilde{\theta}). \end{aligned} \quad (6)$$

The value $\tilde{V}(c, R; \theta)$, thus, is determined by the agent's best strategy with respect to reporting his realization of the shock θ as well as saving/borrowing in the hidden market.

Allocation c is incentive compatible (IC) if agents prefer to reveal their type truthfully and not use the retrade market. That is, c is IC if it satisfies

$$u(c_1(\theta) + \theta c_2(\theta)) \geq \tilde{V}(c, R; \theta) \quad (7)$$

for both θ , with R being an equilibrium gross interest rate in the hidden retrade market.

2. OPTIMAL ALLOCATION

In this section, we first provide a result of DD characterizing the best allocation with no frictions (i.e., without private information or hidden retrade), which is often referred to as the first-best allocation. This allocation provides the highest social welfare among all allocations that are resource feasible, i.e., it maximizes (5) subject to (4). Next, we present a result of Farhi, Golosov, and Tsyvinski (2009) showing that the first-best allocation remains feasible even with the frictions of private θ and hidden retrade. The first-best allocation thus remains optimal in this environment, even with these two frictions present.

Optimal Allocation with no Frictions

Let us start out by noting that given the infinite impatience of the agents of type $\theta = 0$, it is never efficient in this economy to have the impatient types consume a positive amount at date 2. Likewise, given the complete patience of type $\theta = 1$ and $\hat{R} > 1$, it is never efficient to have the patient types consume a positive amount at date 1.

Lemma 1 *If $c = \{c_1(0), c_2(0), c_1(1), c_2(1)\}$ maximizes (5) subject to (4), then $c_2(0) = c_1(1) = 0$.*

Proof. In the Appendix. ■

Below, we will often write c_1 for $c_1(0)$ and c_2 for $c_2(1)$, silently assuming $c_2(0) = c_1(1) = 0$, and refer to (c_1, c_2) as an allocation. With

these notational shortcuts, the social welfare function (5) can be written simply as

$$\pi u(c_1) + (1 - \pi)u(c_2), \tag{8}$$

the aggregate resource constraint (4) as

$$\pi c_1 + (1 - \pi)\frac{c_2}{\hat{R}} \leq e, \tag{9}$$

and first-best allocation can be defined as a maximizer of (8) subject to just (3), i.e., ignoring the incentive constraint (7). Further, from (1) and (2) we have $c_1 = \frac{s}{\pi}$ and $c_2 = \frac{x}{1-\pi}\hat{R}$. If no initial wealth is to be wasted, we must have $x = e - s$. We can thus express any resource-feasible allocation (c_1, c_2) as a function of the initial liquid investment s alone:

$$(c_1, c_2) = \left(\frac{s}{\pi}, \frac{e - s}{1 - \pi}\hat{R} \right)$$

with $s \in [0, e]$. The social welfare function (8) can thus be written as

$$\pi u\left(\frac{s}{\pi}\right) + (1 - \pi)u\left(\frac{e - s}{1 - \pi}\hat{R}\right). \tag{10}$$

Denote this function by $W(s)$. The first-best planning problem is reduced here to finding a level of liquid investment s in $[0, e]$ that maximizes $W(s)$. Denote such a level by s^* . The corresponding level of illiquid investment is $x^* = e - s^*$ and the first-best optimal allocation is $(c_1^*, c_2^*) = \left(\frac{s^*}{\pi}, \frac{e-s^*}{1-\pi}\hat{R}\right)$.

Proposition 1 (*Diamond and Dybvig*) *The social welfare function $W(s)$ has a unique maximizer s^* in $[0, e]$. The maximizer satisfies*

$$\pi e < s^* < \pi \frac{\hat{R}}{\pi \hat{R} + 1 - \pi} e. \tag{11}$$

Proof. In the Appendix. ■

The two inequalities in (11) imply that the first-best consumption allocation (c_1^*, c_2^*) satisfies

$$e < c_1^* < \frac{\hat{R}e}{\pi \hat{R} + 1 - \pi}, \tag{12}$$

$$\hat{R}e > c_2^* > \frac{\hat{R}e}{\pi \hat{R} + 1 - \pi}. \tag{13}$$

The right inequalities above show that the first-best allocation does not provide full insurance, $c_1^* < \frac{\hat{R}e}{\pi \hat{R} + 1 - \pi} < c_2^*$. The reason for this is

that first-period consumption is more expensive to provide than second-period consumption. At the full-insurance allocation

$$c_1 = c_2 = \frac{\hat{R}e}{\pi\hat{R} + 1 - \pi}, \quad (14)$$

marginal utility of consumption is the same at both dates, but by giving up $\varepsilon > 0$ units of consumption at date 1 the planner can deliver $\hat{R}\varepsilon > \varepsilon$ units of consumption at date 2. Such a reallocation would therefore increase overall expected welfare, and so full insurance is not optimal.

The left inequality in (11) implies that the first-best allocation gives a larger present value of consumption to impatient agents than to patient ones. Indeed, discounting consumption at date 1 and 2 at, respectively, the rate of return of the short- and long-term asset, and using the left inequalities in (12) and (13), shows

$$\frac{c_1^*}{1} > e > \frac{c_2^*}{\hat{R}}. \quad (15)$$

The optimality of this unequal allocation of the present value of consumption follows because relative risk aversion of the utility function $u(c)$ larger than 1 means that as consumption c increases, marginal utility of consumption $u'(c)$ drops fast (faster than $1/c$). Liquid investment $s = \pi e$ gives a final consumption allocation $(c_1, c_2) = (e, \hat{R}e)$, where the present value of both types' consumption is the same (and equal to the per capita initial endowment):

$$\frac{c_1}{1} = e = \frac{c_2}{\hat{R}}. \quad (16)$$

At this allocation, however, $c_2 = \hat{R}e > e = c_1$, so the marginal utility of c_2 is low and the marginal utility of c_1 is high. By increasing the liquid investment s at date 0 above $s = \pi e$, say by $\varepsilon > 0$, the planner gives up the return $\hat{R}\varepsilon$ but is able to increase consumption in the high marginal utility state, i.e., at date 1. On balance, this is an improvement because $u'(c_1)$ is sufficiently high relative to $u'(c_2)$ and \hat{R} [that is, $\varepsilon u'(e) > \hat{R}\varepsilon u'(\hat{R}e)$].

Alternatively, we can express this intuition using the elasticity of substitution of the utility function u . With zero elasticity of substitution (Leontief preferences), the full insurance allocation (14) would be optimal. With unit elasticity of substitution (logarithmic preferences), the allocation (16) spending the same amount on each good would be optimal. Under the DD assumption of the elasticity of substitution larger than zero but smaller than one, it is optimal to make c_1 and c_2 closer to one another than under logarithmic preferences, but not go all the way to full insurance.

Optimal Allocation with Private Shocks and Retrade

Having characterized the optimal allocation in the first-best version of the DD environment, we now ask what the optimal allocation is with private information and a hidden retrade market, i.e., with the addition of the IC constraint (7).

With realizations of θ being private information and with agents having access to retrade, Farhi, Golosov, and Tsyvinski (2009) show that the first-best allocation is incentive compatible, i.e., remains feasible and thus optimal. This result is obtained as follows. The retrade interest rate R associated with the optimum (i.e., the shadow interest rate at the first-best), denoted by R^* , is

$$R^* = \frac{c_2^*}{c_1^*}. \quad (17)$$

First, let us check that with “endowments” (c_1^*, c_2^*) , the interest rate $R = R^*$ is an equilibrium interest rate in the hidden market. Note that from $c_2^* > c_1^*$ we get $R^* > 1$ and from $\frac{c_1^*}{1} > e > \frac{c_2^*}{R}$ we get $R^* < \hat{R}$, so $1 < R^* < \hat{R}$. Suppose the impatient types enter the hidden market with an endowment vector $(c_1^*, 0)$ and patient types enter with $(0, c_2^*)$. The impatient agent has no income at $t = 2$, so he cannot borrow in this hidden market (for there is nothing he could pay back with). Also, this agent wants to consume his income c_1^* irrespective of the interest rate. Thus, the impatient type’s utility is maximized with the quantity of zero traded at the interest rate R^* . A patient agent could borrow against his date-2 endowment c_2^* and consume at date 1, but $R^* > 1$ implies he would not want to do it, as his marginal utility of consumption is the same at either date and he can consume only $\frac{c_2^*}{R^*} < c_2^*$ if he decides to use the hidden market and consume at date 1. This confirms that consumption (c_1^*, c_2^*) and interest rate R^* are an equilibrium in the retrade market (with zero quantity traded in equilibrium).

Now consider potential deviations in the revelation of θ combined with retrade. The first-best allocation is immune to these deviations because at the interest rate R^* the present value of each type’s endowment is the same. Indeed, the impatient types could claim endowment $(0, c_2^*)$ and borrow against c_2^* in order to consume at date 1, but doing so would give them $\frac{c_2^*}{R^*} = c_1^*$ units of consumption, so there is no gain for them from doing so. As well, the patient types could claim endowment $(c_1^*, 0)$ and save at the market interest rate R^* . But doing so gives them final consumption $R^*c_1^* = c_2^*$ so, again, no gain. This confirms

that the first-best allocation is incentive compatible in the model with private information and hidden retrade.

Note that although the possibility of hidden retrade does not change the optimal allocation, it does change the IC constraint. With just private information about the liquidity shock θ (without retrade), the IC constraint would be $c_2 \geq c_1$. The first-best allocation satisfies this constraint as a strict inequality simply because $c_2^* > c_1^*$. With the hidden retrade market, however, the IC constraint holds only as an equality because $\frac{c_2^*}{R^*} = c_1^*$.

Next, we move on to discuss market provision of liquidity in this environment.

3. COMPETITIVE EQUILIBRIUM WITH INCOMPLETE MARKETS

The remainder of this article is devoted to studying competitive equilibrium outcomes under three different market arrangements, and comparing these outcomes with the optimal allocation (c_1^*, c_2^*) .

In this section, we discuss a simple incomplete-markets model of trade, in which agents invest directly in the two assets and subsequently trade them (i.e., there are no intermediaries, no state-contingent contracts). This natural model of trade is a point of departure for Diamond and Dybvig (1983). DD start their analysis of market provision of liquidity by considering this incomplete market structure. They conclude that the equilibrium level of liquidity is too low, i.e., there is a market failure. We briefly review this result in this section and move on to showing in the next section that with hidden retrade this conclusion generalizes to any market structure (even when state-contingent contracts and/or intermediaries are taken into consideration).

The simple market structure is as follows. At date 0, each agent invests directly in the two assets subject to $s + x \leq e$. At date 1, after agents find out their type θ , they trade the long-term asset for cash at a market-determined price p . In addition to the market for the long-term asset, agents have access at date 1 to a market for one-period IOUs.¹⁰ A formal statement of the agents' optimization problem and competitive equilibrium in this economy is given in the Appendix. Note that this market structure is incomplete: There are no contracts for provision of consumption conditional on θ .

A simple arbitrage argument shows that in any equilibrium of this trading arrangement the date-1 cash price p of a unit of the long-term

¹⁰ As we will see, however, the (hidden) IOU market will not be active here, nor imposing any binding constraints on the equilibrium allocation.

asset must be 1. This argument is as follows. The fact that a market for the long-term asset exists at date 1 makes the long-term asset de facto liquid and thus a perfect substitute, at date 0, for the short-term asset. The return from holding the long-term asset for one period, therefore, must be the same as the return from investing in the short-term asset. The date-1 price of the long-term asset must therefore be $p = 1$, or else there is an arbitrage.

Indeed, if $p > 1$, all agents want to invest their initial resources in the long-term asset only, as investing a unit of resources in that asset and selling it at date 1 yields p , while investing in the short-term asset yields 1. In this case, however, nobody has cash at date 1 and thus aggregate demand for the long-term asset is zero. This level of demand is inconsistent with the equilibrium price p being positive. Similarly, if $p < 1$, all agents want to invest exclusively in the short-term asset at date 0, as investing a unit of resources in the long-term asset is dominated by investing this unit in the short-term asset and then buying the long-term asset at date 1 at price $p < 1$. This, however, means that supply of the long-term asset at date 1 is zero while demand is positive, as the patient types are willing to buy at $p < 1$. Thus, $p < 1$ cannot be an equilibrium price, either.¹¹

The only price p consistent with equilibrium, therefore, is $p = 1$. At this price, the return from holding the short- and the long-term asset from date 0 to date 1 is the same, so agents are indifferent between investments s and x . At date 1, the impatient agents want to sell their holdings x of the illiquid asset. With $p = 1$, the patient agents want to hold on to their x and spend their cash s to purchase additional units of the long-term asset, as the return on this investment, $\frac{\hat{R}}{p} = \hat{R}$, exceeds their required rate of return, 1. Aggregate supply of the long-term asset to the market at date 1 is therefore πx and the supply of cash is $(1 - \pi)s$. The market-clearing condition, thus, is

$$\pi x p = (1 - \pi)s,$$

where, by the arbitrage argument given above, $p = 1$. The date-0 budget constraint implies

$$x = e - s.$$

Solving the above two conditions, we obtain

$$s = \pi e, \quad x = (1 - \pi)e. \tag{18}$$

¹¹ Strictly speaking, these corner investment strategies are not arbitrages because they are not self-financing. But they could be turned into arbitrages if agents could short the expensive asset at date 0.

This solution is unique, so there exists only one equilibrium. In equilibrium, consumption of the impatient types is $c_1 = s + px = \pi e + 1(1 - \pi)e = e$, while the patient types consume $c_2 = \left(x + \frac{s}{p}\right) \hat{R} = \left((1 - \pi)e + \frac{\pi e}{1}\right) \hat{R} = e\hat{R}$. Let us denote the unique equilibrium consumption bundle by (\hat{c}_1, \hat{c}_2) . We have just shown that

$$(\hat{c}_1, \hat{c}_2) = (e, \hat{R}e). \quad (19)$$

In the hidden retrade market, there is no active trade. The equilibrium retrade interest rate is $R = \hat{R}$. At this rate, agents choose not to alter their consumption allocation (\hat{c}_1, \hat{c}_2) by either borrowing or lending. The hidden retrade market has no impact on the equilibrium outcome here because the (regular, “non-hidden”) date-1 market for the long-term asset already offers a riskless return $\frac{\hat{R}}{p} = \hat{R} = R$. The hidden IOU retrade market is thus redundant.

A key property of the DD environment is that the equilibrium allocation of consumption, (\hat{c}_1, \hat{c}_2) , is inefficient. That is, this allocation yields lower ex ante welfare than the optimal allocation c^* . Clearly, the right inequalities in (12) and (13) tell us that $c_1^* > \hat{c}_1$ and $c_2^* < \hat{c}_2$. Since, by Proposition 1, the optimum (c_1^*, c_2^*) is a unique welfare maximizer, equilibrium allocation (\hat{c}_1, \hat{c}_2) is indeed inefficient.

As we saw in Section 2, optimal allocation calls for a present-value transfer from the patient types to the impatient types. In equilibrium with incomplete markets, however, each agent consumes the worth of his own initial endowment, e , i.e., there are no present value transfers between types, and insurance markets are missing. Moreover, it is easy to see that an intervention by a benevolent planner/government can improve welfare without introducing any new markets. If the planner forces each agent to invest $(s, x) = (s^*, x^*)$ at date 0 and allows free trade at date 1, the market price for the long-term asset will be $p = p^*$, the retrade market rate will be $R = R^*$, and the equilibrium consumption allocation will be (c_1^*, c_2^*) .¹²

In sum, the equilibrium investment in the liquid asset is too low relative to the optimum, $s = \pi e < s^*$, i.e., free trade leads to underprovision of liquidity.

¹² In the language of the incomplete-markets literature, equilibrium (\hat{c}_1, \hat{c}_2) is constrained-inefficient.

4. COMPETITIVE EQUILIBRIUM WITH CONTINGENT CONTRACTS

In this section, we allow for state-contingent contracts. We review the following important result. Jacklin (1987) points out that when retrade is allowed, an arbitrage argument similar to the one used in the previous section implies that markets will underprovide liquidity, even when fully state-contingent contracts are allowed. With retrade, thus, the market failure shown in the previous section for the simple incomplete-markets model continues to hold for all feasible models of trade in the DD environment, including the intermediation economy of Diamond and Dybvig (1983).

Consider the following general model of trade with fully state-contingent contracts, direct investment, and retrade.¹³ In addition to directly investing in the two assets, agents can contract with intermediaries and access the hidden IOU market. Intermediaries, or banks, make available to agents at date 0 a state-contingent contract (ξ_1, ξ_2) . Under this contract, which can be thought of as a deposit contract, the agent can obtain from the intermediary, at the agent's discretion, either ξ_1 at date 1 or ξ_2 at date 2 (but not both). Let us normalize the price of this contract to e , i.e., an agent who accepts a contract deposits his whole initial wealth with a bank. Also, as before, agents can borrow from and lend to each other privately in the hidden retrade market at date 1.

Under this market structure, an agent has the following choices to make. At date 0, he decides whether to deposit his wealth e with a bank or to invest directly in assets s and x . If he deposits, after he learns his type θ , he chooses whether to withdraw at date 1 or 2, and how much, if at all, to borrow or lend in the hidden retrade market at the market rate R . If the agent chooses not to deposit at date 0 but rather to invest directly, he selects a portfolio (s, x) . At date 1, after he learns his type and his cash investment s matures, the agent decides how much to borrow or lend in the retrade market at the market rate R .

Competition among banks (existing or potential entrants) drives banks' profits to zero and forces each active bank to offer the same contract (namely, the contract that maximizes the ex ante expected utility of the representative agent, for otherwise agents would deposit with a different bank). Since intermediation is an activity with constant returns to scale in this model, it is without loss of generality to assume

¹³ For a formal statement a version of this economy see Section 3.1 of Farhi, Golosov, and Tsyvinski (2009) or Allen and Gale (2004).

that a single large bank operates in equilibrium (the market, however, is perfectly contestable).

The bank's contract design problem is similar to the social planning problem in that in both cases the objective is to maximize the agent's expected utility. There is, however, a key difference. The planner can control date-0 investment, which enables her to have an (indirect) impact on the retrade market interest rate R . The bank cannot force the agents to deposit, which means it must act competitively, i.e., take prices as given. In particular, the bank takes as given the retrade market interest rate R .

Given this difference, it is not hard to see that the optimal allocation (c_1^*, c_2^*) cannot be an equilibrium allocation. If (c_1^*, c_2^*) were to be an equilibrium allocation, the interest rate R in the hidden retrade market would have to be equal to the shadow rate R^* given in (17), for otherwise agents would use that market to trade away from this allocation. But R^* cannot be an equilibrium retrade interest rate because the fact that R^* is strictly smaller than \hat{R} creates an arbitrage opportunity. This arbitrage opportunity is similar to the one that in the incomplete-markets model discussed in the previous section pinned down the secondary-market asset price p at 1.

The arbitrage strategy, described in Jacklin (1987), calls for investment $x = e$ at date 0. If the agent executing this arbitrage is patient, i.e., his $\theta = 1$, he consumes nothing at date 1 and $\hat{R}e > c_2^*$ at date 2. If he turns out impatient, i.e., his $\theta = 0$, he can access the retrade market and borrow at rate R^* , which gives him date-1 consumption $\frac{\hat{R}e}{R^*} > c_1^*$. In either case, thus, he consumes more than (c_1^*, c_2^*) , which shows that (c_1^*, c_2^*) , with its shadow interest rate R^* , cannot be an equilibrium allocation of consumption.

What allocation can be a market equilibrium allocation in this model? The Jacklin arbitrage strategy pins down the interest rate in the retrade market at $R = \hat{R}$. With this interest rate, it is easy to check (or consult Allen and Gale [2004] or Farhi, Golosov, and Tsyvinski [2009]) that the equilibrium allocation (19) from the incomplete-markets model discussed in the previous section is a unique equilibrium allocation, also here in the richer model with fully state-contingent contracts.¹⁴

Why is the planner able to do better than the market in this model? The planner makes the Jacklin arbitrage strategy infeasible for the

¹⁴ This conclusion applies to all conceivable market structures in which the Jacklin arbitrage strategy remains feasible. In particular, when the hidden retrade market is included, it applies to the general competitive private information model of Prescott and Townsend (1984) in which agents trade lotteries over allocations subject to incentive compatibility constraints.

agent by controlling initial investment (s, x) . In the planning problem, although the agent has unfettered access to the retrade market, the agent does not have private control over his initial investment. The initial investment choice is publicly observable and therefore can be controlled by the planner/government. The Jacklin arbitrage strategy calls for the all-long investment $(s, x) = (0, e)$ at date 0. By forcing/choosing investment $(s, x) = (s^*, e - s^*)$, the planner eliminates this arbitrage. Moreover, this choice of date-0 investment pins down the amount of resources available at dates 1 and 2 and, thus, also the interest rate in the hidden retrade market, which with liquid investment s^* is $R = R^*$. In a competitive market economy, by contrast, firms have to respect the agents' freedom to not contract with them but instead to invest directly (or set up another firm that will do the investing for them, as in Farhi, Golosov, and Tsyvinski [2009]). Intermediaries thus cannot make the Jacklin arbitrage strategy infeasible for the agents. Having to respect this arbitrage condition, the best allocation they can provide is $(\hat{c}_1, \hat{c}_2) = (e, \hat{R}e)$ with the associated retrade market interest rate $R = \hat{R}$.

To recap, the planner internalizes the fact that her control of the initial investment changes the price in the equilibrium of the retrade market. Firms, in contrast, take all prices as given, including those in the retrade market. The discrepancy constitutes a pecuniary externality in this model and the equilibrium allocation is inefficient.

Efficiency Without Retrade

The Jacklin arbitrage strategy is clearly impossible to execute if arbitrageurs do not have access to the hidden retrade market. Absent retrade, competitive equilibrium with state-contingent contracts would be efficient. Indeed, if the retrade market is shut down, the value function $\tilde{V}(c, R; \theta)$ defined in (6) reduces to $\tilde{V}(c, R; \theta) = \max_{\tilde{\theta}} u(c_1(\tilde{\theta}) + \theta c_2(\tilde{\theta}))$, which no longer depends on R . The incentive constraint (7), therefore, no longer depends on a price.¹⁵ This means that there is no pecuniary externality. The welfare theorems of Prescott and Townsend (1984) apply, and competitive equilibrium is efficient. In particular, it can be implemented as a banking equilibrium of Diamond and Dybvig (1983) with the equilibrium deposit contract $(\xi_1, \xi_2) = (c_1^*, c_2^*)$.

The theoretical results we reviewed in this section suggest that retrade generates a pecuniary externality and leads to equilibrium

¹⁵ In particular, given Lemma 1, the impatient types will never misrepresent their type and the patient types' incentive constraint reduces to $c_2 \geq c_1$.

underprovision of liquidity. In practice, banks and other financial intermediaries have ample access to various retrade markets. Therefore, one might be tempted to take as an implication of this theory the prediction that markets will fail to provide sufficient liquidity. In the next section, we present a simple version of the analysis of Kilenthong and Townsend (2011) showing that this conclusion would be premature: If harnessed inside appropriate venues, retrade can be consistent with efficient functioning of markets in the provision of liquidity.

5. COMPETITIVE EQUILIBRIUM WITH SEGREGATED EXCHANGES

In this section, we consider the model of Kilenthong and Townsend (2011), in which a market-maker eliminates the Jacklin arbitrage by segmenting the retrade market and pricing entry into market segments as a function of the investment portfolio held by agents entering a given segment. With the Jacklin arbitrage eliminated, the pecuniary externality causing market failure is eliminated as well. The resulting equilibrium is efficient. We supplement the analysis of Kilenthong and Townsend (2011) by characterizing explicitly how the equilibrium exchange entry fees depend on the fundamentals of the exchange and on the portfolio of the agent (equation [27] and Figure 2). We conclude with a discussion of an important difference between the environment with pecuniary externality studied in the previous sections and the environment without it that we study here. The segregated-exchanges equilibrium is efficient, but, effectively, it requires that agents commit *ex ante* to not using the hidden retrade market *ex post*. Whether or not retrade leads to a pecuniary externality and inefficiency of market outcomes, therefore, depends on the practical feasibility of such a commitment.

Trade Inside Segregated Exchanges at Date 1

Before we define the general equilibrium concept with segregated exchanges proposed by KT, we describe in this subsection segregated exchanges, their fundamentals, and internal prices.

A segregated exchange is a competitive market for the long-term asset that opens at date 1 after types θ are realized. A defining characteristic of such an exchange is a set of fundamentals determining the market price p at which the long-term assets will be traded. The fundamentals and the price must be consistent: Given the fundamentals in an exchange, the price p must indeed be a competitive equilibrium price in that exchange. In the DD economy at hand, the level of the

cash asset investment s held by each member of an exchange is a sufficient description of the fundamentals in the exchange. Thus, we will index exchanges by $S \in [0, e]$, where S represents the level of liquid investment held by each agent entering the exchange. Note that this definition assumes identical asset holdings by all exchange members. We will see later that this assumption is without loss of generality in the present environment.

Equilibrium price in exchange S

Let us derive an equilibrium consistency condition between fundamentals S and price p in the exchange $S \in [0, e]$. It is a simple equilibrium pricing condition in a competitive market with all agents holding the same portfolio of assets $(s, x) = (S, e - S)$ and experiencing shocks θ drawn from the same distribution. We will denote the equilibrium price in exchange S by $p(S)$.

The equilibrium condition for consistency between S and p is

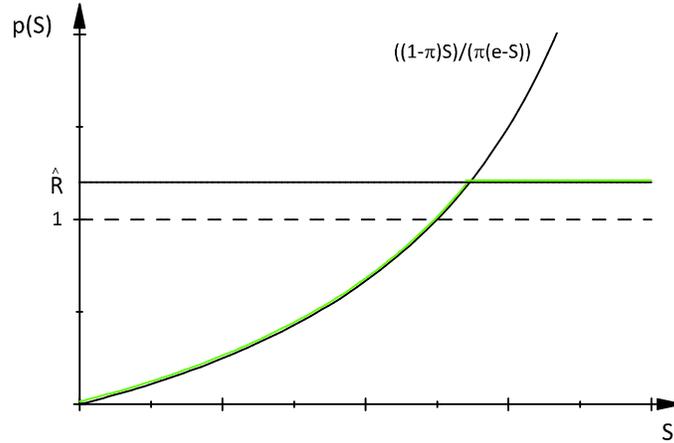
$$p(S) = \min \left\{ \frac{(1 - \pi) S}{\pi (e - S)}, \hat{R} \right\}. \tag{20}$$

This condition is derived as follows. The equilibrium price of the illiquid asset is determined by supply and demand in exchange S in the same way as it was determined in the incomplete-markets model of Section 3. At date 1, the impatient agents want to sell their long-term asset in the market at any price. They supply $\pi(e - S)$ units of the long-term asset to the market. The behavior of the patient agents depends on the price p . If $p > \hat{R}$, a short position in the asset gives them a positive return, so patient agents want to sell their holdings of the asset, just like the impatient ones. This cannot be an equilibrium, as demand for the asset is zero and supply is positive. Thus, in any equilibrium, $p \leq \hat{R}$. With $p \leq \hat{R}$, a long position in the asset gives patient agents a non-negative return (strictly positive if $p < \hat{R}$). With any such price, the patient agents are willing to buy the long-term asset. They demand $(1 - \pi)\frac{S}{p}$ units. Thus, the equilibrium price $p(S)$ solves $\pi(e - S) = (1 - \pi)\frac{S}{p(S)}$, which gives us

$$p(S) = \frac{(1 - \pi) S}{\pi (e - S)}, \tag{21}$$

provided that $p(S) \leq \hat{R}$. Solving $\hat{R} = \frac{(1-\pi)S}{\pi(e-S)}$ for S , we get a threshold

$$\bar{S} = \frac{\pi \hat{R} e}{\pi \hat{R} + 1 - \pi}. \tag{22}$$

Figure 1 Equilibrium Asset Price p in Exchange S 

For all $S \geq \bar{S}$, the equilibrium price is flat at \hat{R} .¹⁶ Combining this restriction with (21) gives us the consistency condition (20).

Figure 1 illustrates the derivation of the consistency condition (20) graphically. When S is small and $e - S$ is large, there is a large quantity of the illiquid asset in the market, supplied by the impatient agents, and very few units of the consumption good (cash), supplied by the patient agents, and so the price of the asset is low.¹⁷ In exchanges with higher S , the proportion of cash to units of the asset in the market is higher, so the price $p(S)$ is higher. This is true up to the threshold \bar{S} . In exchanges with S larger than \bar{S} , the price $p(S)$ remains flat at \hat{R} and the patient types are indifferent between buying and selling the asset. The price of the asset cannot exceed \hat{R} , as at a price higher than \hat{R} the patient agents would switch from buying to selling the asset. As we see, the range of prices that can be consistent with some fundamentals $S \in (0, e]$ is

$$0 < p \leq \hat{R}. \quad (23)$$

¹⁶ Note that \bar{S} is the same threshold that in Proposition 1 results with the full-insurance allocation (an upper bound on s^*).

¹⁷ We will exclude the exchange $S = 0$ from our analysis. In this exchange, the supply of resources at date 1 would be zero and thus welfare of the impatient agents would be extremely low. No agent would want to enter this exchange at date 0.

Markets at Date 0 and Equilibrium Definition

In this subsection, we use the segregated exchanges to define the KT notion of competitive equilibrium with segregated retrade.

At date 0, agents choose their investments s and x and join segregated exchanges. Each agent can physically join one exchange. Exchanges are defined by their fundamental level of liquid investment S . Associated with each exchange is an entry fee pricing any deviations of the investment portfolio of an agent wishing to join a given exchange from that exchange's fundamentals. If an agent joins an exchange S with liquid investment s , the amount of shortage of his liquid asset relative to the exchange fundamentals is $S - s$. Upon entry, the agent is charged a fee proportional to the amount of shortage of liquid investment in his portfolio. The price per unit of shortage in exchange S is $\delta(S)$. Thus, an agent entering exchange S with liquid investment s is charged an entry fee of $\delta(S)(S - s)$. This charge is assessed by the exchange as of the time of entry, i.e., at date 0. The unit price $\delta(S)$ can be positive or negative. Note that if $\delta(S) > 0$ and an agent joins exchange S with liquid investment $s > S$, the entry fee is negative, so the exchange makes a payment to the agent.

In sum, at date 0 agents choose investment portfolios (s, x) and exchange membership S subject to the budget constraint

$$s + x + \delta(S)(S - s) \leq e. \quad (24)$$

If, for example, an agent decides to join exchange S and go all-long, i.e., invest $s = 0$ and $x = e$, then the price for this shortage would be $\delta(S)S$. Clearly, public observability of the agent's portfolio is important for the assessment of fees. In particular, agents cannot avoid fees by "window dressing" or changing the composition of their portfolio after the fees are assessed but before the shock θ is realized and exchanges open for business.

What if an agent chooses not to join an exchange? The decision not to join is equivalent to joining an exchange in which the price of any "deviation" or "shortage" relative to the "fundamentals" is zero. Thus, not joining a segregated exchange is equivalent to maintaining access to the free exchange in which $\delta = 0$. As we will see shortly, the exchange $S = \pi e$ will have $\delta = 0$. This exchange corresponds to the incomplete-markets model of Section 3, where, as we saw earlier, all agents choose investment $s = \pi e$ at date 0. It is natural to default all agents who do not join a different exchange into this one. The model with segregated exchanges, therefore, nests the simple incomplete-markets model as a special case in which there is only one secondary market for the long-term asset, and access to this market is free.

Let us now discuss the agents' objective function as of date 0. Agents maximize

$$E[V_1(s, x, S; \theta)], \quad (25)$$

where $V_1(s, x, S; \theta)$ is the indirect utility function as of date 1, i.e., the value the agent can get in exchange S with an asset portfolio (s, x) and a liquidity shock realization θ . The indirect utility function

$$\begin{aligned} V(s, x, S; \theta) &= \max u(c_1 + \theta c_2), \\ & \text{s.t.} \\ & c_1 + p(S)n \leq s, \\ & n \geq -x, \\ & c_2 \leq (x + n)\hat{R}, \end{aligned} \quad (26)$$

where n is the agent's net demand at date 1 in the market for the illiquid asset inside exchange S .

Next, we define competitive equilibrium with segregated exchanges.

Definition 1 (*Kilenthong and Townsend*) *A price system $(p(\cdot), \delta(\cdot))$, ex ante investment and exchange membership choices s, x, S , value functions $V_1(\cdot, \cdot, \cdot; \theta)$ for $\theta \in \{0, 1\}$, and a consumption allocation (c_1, c_2) are an equilibrium with segregated exchanges if*

1. *expectations are correct: For each S , price $p(S)$ satisfies the consistency condition (20) and value functions $V_1(\cdot, \cdot, \cdot; \theta)$ solve (26);*
2. *agents optimize ex ante: Taking prices $(\delta(\cdot), p(\cdot))$ and value functions $V_1(\cdot, \cdot, \cdot; \theta)$ as given, agents' choices s, x, S maximize their ex ante utility (25) subject to the budget constraint (24);*
3. *market clearing: Consumption allocation (c_1, c_2) is an equilibrium allocation of consumption in the exchange S .*

Note that this definition does not allow for mixed strategies. In general, mixed strategies may be useful, as agents face a discrete choice of exchange membership. As the theorem presented next makes clear, in the environment at hand it is without loss of generality to restrict attention to equilibria in pure strategies, where all agents, being ex ante identical, join the same exchange.¹⁸

¹⁸ In excluding random exchange assignments, this definition follows Definition 4 in Kilenthong and Townsend (2014a).

Efficient Equilibrium with Segregated Exchanges

Theorem 1 *Prices $p(S)$ as in (20) and*

$$\delta(S) = \min \left\{ 1 - \frac{\pi}{1 - \pi} \left(\frac{e}{S} - 1 \right), 1 - \hat{R}^{-1} \right\}, \tag{27}$$

ex ante investment and membership choices $s = s^$, $x = e - s^*$, $S = s^*$, and consumption allocation $(c_1, c_2) = (c_1^*, c_2^*)$ are a competitive equilibrium with segregated exchanges.*

The rest of this subsection is devoted to proving this theorem. We need to check the three equilibrium conditions in Definition 1.

We start by characterizing value functions (26). For $\theta = 0$, the optimized value of (26) is

$$V_1(s, x, S; 0) = u(s + p(S)x). \tag{28}$$

Clearly, the impatient agents want to sell their holdings x of the long-term asset at any price $p(S)$ and consume all their wealth at date 1, as they have no use for consumption at date 2. At price $p(S)$, an impatient agent can afford consumption $c_1 = s + p(S)x$, which gives us (28).

The patient type’s value as of date 1 is

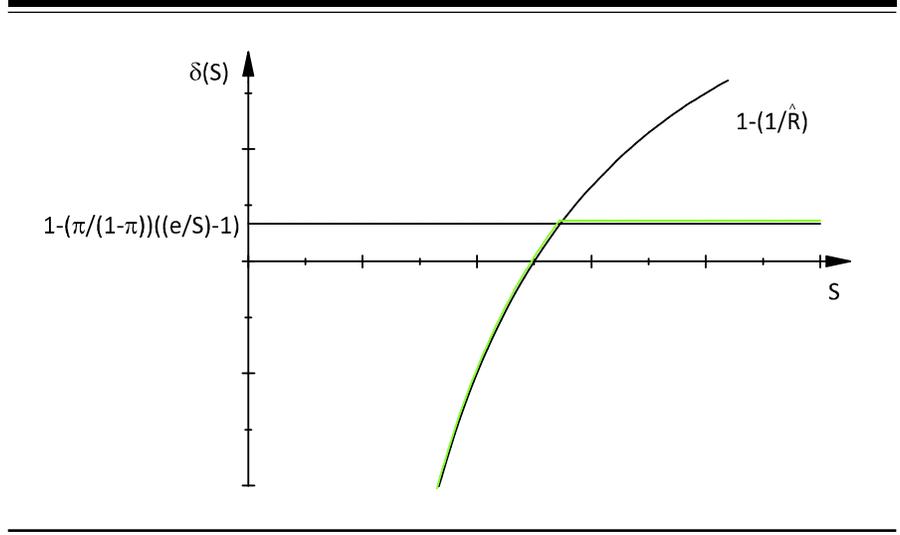
$$V_1(s, x, S; 1) = u \left(\left(x + \frac{s}{p(S)} \right) \hat{R} \right). \tag{29}$$

To see that this is the case, note that in each exchange S patient agents are happy to buy the long-term asset at date 1 because, by (23), $p(S) \leq \hat{R}$ in all exchanges S . This means that the rate of return on this investment, $\frac{\hat{R}}{p(S)}$, exceeds the patient type’s rate of time preference, which is 1. A patient agent’s demand for the long-term asset is $n = \frac{s}{p(S)}$, his consumption at date 1 is $c_1 = 0$, and consumption at date 2 is $c_2 = \left(x + \frac{s}{p(S)} \right) \hat{R}$. These quantities substituted to (26) with $\theta = 1$ give us (29).

We can now confirm that with value functions (28) and (29) the first equilibrium condition (correct expectations) is satisfied, as these value functions and prices $p(S)$ defined in (20) are consistent with agents’ optimization at date 1. Note that the general pattern of behavior at date 1 is the same in all exchanges. The impatient types sell and the patient types buy the long-term asset. The exchanges are different only in the composition of demand and supply, which gives rise to different equilibrium prices at which the asset is traded in each exchange.

In order to check the second equilibrium condition (agents’ optimization ex ante), we now study the agents’ behavior at date 0. Substituting the indirect utility functions (28) and (29) into the objective (25), we express the ex ante expected utility function of the

Figure 2 Unit Liquidity Shortage Price δ in Exchange S



representative agent as

$$\pi u(s + p(S)x) + (1 - \pi)u\left(\left(x + \frac{s}{p(S)}\right) \hat{R}\right).$$

This expression gives the agent’s expected value of being in exchange S with assets s and x . The representative agent chooses investment portfolio (s, x) and exchange membership S to maximize this value subject to the date-0 budget constraint (24).

The structure of the portfolio fees $\delta(S)$ charged upon exchange entry is a key part of the budget constraint. Figure 2 graphs against S the unit liquid asset shortage price $\delta(S)$ given in (27). As we argue, these prices support the efficient equilibrium.

It is easy to check directly in (27) that $1 - \frac{\pi}{1-\pi} \left(\frac{e}{S} - 1\right) < 1 - \hat{R}^{-1}$ for all $S < \bar{S}$, where \bar{S} is, as before, given in (22). Thus,

$$\delta(S) = \begin{cases} 1 - \frac{\pi}{1-\pi} \left(\frac{e}{S} - 1\right) & \text{for } S \leq \bar{S}, \\ 1 - \frac{1}{\hat{R}} & \text{for } S \geq \bar{S}. \end{cases}$$

Note that $\delta(S)$ is increasing. This means that the portfolio charge per unit of liquidity shortage is higher in exchanges with higher fundamental liquidity S . Substituting in (27) $S = \pi e < \bar{S}$, we check that $\delta(\pi e) = 0$. Thus, the exchange with $S = \pi e$ is a (unique) free-entry exchange, where portfolio charges are zero for all portfolios (s, x) . In exchanges with $S > \pi e$, $\delta(S) > 0$, i.e., agents are subject to a positive charge for shortage of liquidity in their portfolio. For all $S < \pi e$,

$\delta(S) < 0$, i.e., portfolio charges are positive if the long-term investment x is less than $e - S$.

We now can study the agents' date-0 problem of choice of investment (s, x) and exchange membership S . For each exchange S , we need to determine the investment portfolio (s, x) the agent will choose conditional on joining S and the consumption pair (c_1, c_2) he will be able to afford inside S . This will give us the ex ante expected value of joining S , which we will then use to determine the agent's most preferred exchange membership decision and thus the solution to his utility maximization problem.

We start by examining the exchanges with $S \geq \bar{S}$. What ex ante value can the representative agent obtain if he plans on joining one of these exchanges? All exchanges $S \geq \bar{S}$ have the same entry fees and long-term asset prices:

$$\begin{aligned} \delta(S) &= 1 - \hat{R}^{-1}, \\ p(S) &= \hat{R}. \end{aligned} \tag{30}$$

Given a portfolio (s, x) , an agent in exchange $S \geq \bar{S}$ can afford consumption

$$c_1 = s + \hat{R}x$$

if impatient, or

$$c_2 = \left(x + \frac{s}{\hat{R}}\right) \hat{R} = s + \hat{R}x$$

if patient. As we see, the agent is fully insured against the liquidity shock θ in any exchange $S \geq \bar{S}$, as his optimal consumption in any such exchange is independent of the realization of the liquidity shock θ . His ex ante expected utility is therefore simply

$$u(s + \hat{R}x). \tag{31}$$

With the entry fee of $\delta(S) = 1 - \hat{R}^{-1}$ per unit of liquidity shortage, the agent's ex ante budget constraint (24) can be written as

$$s + \hat{R}x \leq \hat{R}e - (\hat{R} - 1)S. \tag{32}$$

Comparing the agent's objective (31) and his budget constraint (32), we see that the agent is indifferent between all portfolios (s, x) on the budget line $s + \hat{R}x = \hat{R}e - (\hat{R} - 1)S$. This is because any such portfolio gives the agent the same ex ante utility of $u(\hat{R}e - (\hat{R} - 1)S)$. Since $\hat{R} > 1$, this value is decreasing in S . Thus, among all exchanges $S \geq \bar{S}$, exchange \bar{S} is the best one for the agent.

Next, let us consider the choices of an agent who plans on joining one of the exchanges with $S \leq \bar{S}$. The prices this agent faces are

$$\begin{aligned}\delta(S) &= 1 - \frac{\pi}{1-\pi} \left(\frac{e}{S} - 1 \right), \\ p(S) &= \frac{(1-\pi)S}{\pi(e-S)}.\end{aligned}\tag{33}$$

Thus, given a portfolio (s, x) , in exchange S the agent can afford consumption

$$c_1 = s + \frac{(1-\pi)S}{\pi(e-S)}x$$

if impatient, or

$$c_2 = \left(x + \frac{s}{\frac{(1-\pi)S}{\pi(e-S)}} \right) \hat{R} = \left(s + \frac{(1-\pi)S}{\pi(e-S)}x \right) \frac{\pi(e-S)}{(1-\pi)S} \hat{R}$$

if patient. Unlike in the previous case, these consumptions are not identical. They are, however, directly proportional to $s + \frac{(1-\pi)S}{\pi(e-S)}x$. Substituting these consumption values into the ex ante expected utility function, we have

$$\pi u \left(s + \frac{(1-\pi)S}{\pi(e-S)}x \right) + (1-\pi)u \left(\left(s + \frac{(1-\pi)S}{\pi(e-S)}x \right) \frac{\pi(e-S)}{(1-\pi)S} \hat{R} \right).\tag{34}$$

With the entry fee $\delta(S)$ given in (33), the agent's ex ante budget constraint (24) can be rewritten, after some algebra, as

$$s + \frac{(1-\pi)S}{\pi(e-S)}x \leq \frac{S}{\pi}.$$

Comparing this budget constraint and the agent's objective (34) we see that here, as in the previous case, the agent is indifferent between all portfolios (s, x) on the budget line $s + \frac{(1-\pi)S}{\pi(e-S)}x = \frac{S}{\pi}$ as any such portfolio gives him the same expected utility value of

$$\pi u \left(\frac{S}{\pi} \right) + (1-\pi)u \left(\frac{e-S}{1-\pi} \hat{R} \right).$$

Finally, we observe that this objective function, representing the agent's utility from joining exchange S , is mathematically the same as the objective function (10) in the social welfare maximization problem studied in Proposition 1. As we saw there, this objective is maximized by a unique $s^* < \bar{S}$. Thus, exchange $S = s^*$ is a unique maximizer in the

agent’s utility maximization problem we study here.¹⁹ To simplify the notation, we will use S^* to denote the exchange $S = s^*$.

The last equilibrium condition that we need to check is to confirm that (c_1^*, c_2^*) is an equilibrium allocation of consumption in exchange S^* with the asset price $p(S^*)$. For the pair (c_1^*, c_2^*) to be resource-feasible in exchange S^* , agents must enter this exchange carrying the investment portfolio $(s^*, e - s^*)$. Portfolio $(s^*, e - s^*)$ is (weakly) optimal for an agent joining exchange S^* because, as we saw earlier, conditional on joining an exchange, agents are indifferent among all portfolios (s, x) on the budget line. Finally, since the asset price $p(S^*)$ satisfies the consistency condition (20), the market for the long-term asset inside the exchange S^* does clear.

We conclude that the prices and quantities specified in Theorem 1 are indeed a competitive equilibrium with segregated exchanges. This equilibrium is efficient, as the equilibrium consumption bundle is exactly the optimal consumption bundle (c_1^*, c_2^*) .

Discussion

In the two equilibrium concepts without segregated exchanges that we discussed in Sections 3 and 4, arbitrage pinned at $p = 1$ the equilibrium price in the secondary market for the long-term asset or, equivalently, the retrade market interest rate at $R = \hat{R}$. In the model with segregated exchanges, agents trade the long-term asset in the secondary market inside the exchange S^* at the equilibrium price

$$p(S^*) = \frac{(1 - \pi) s^*}{\pi (e - s^*)} = \frac{(1 - \pi) \pi c_1^*}{\pi (1 - \pi) \frac{c_2^*}{\hat{R}}} = \hat{R} \frac{c_1^*}{c_2^*} = \frac{\hat{R}}{R^*} > 1.$$

Why does arbitrage not force $p(S^*)$ down to 1 in the segregated exchanges model?

The Jacklin arbitrage strategy is infeasible in the segregated exchange model because of the entry fees ex ante and the separation of agents in different exchanges ex post. The Jacklin arbitrage strategy calls for the all-long initial investment $(s, x) = (0, e)$ and a subsequent sale of the long-term asset, or borrowing against it, in case the agent attempting arbitrage turns out needing funds at date 1. But which exchange should the arbitrageur join at date 0? If he defaults to the entry-fee-free market $S = \pi e$, he does not receive the favorable asset price $p(S^*) > 1$ but only the arbitrage-free price $p(\pi e) = 1$, so no arbitrage profit can be made in this exchange. If the arbitrageur joins

¹⁹ Note in particular that the right inequality in (11) implies that exchange $S = s^*$ dominates the exchange $S = \bar{S}$ and thus also all exchanges $S \geq \bar{S}$.

exchange S^* , he must pay the entry fee of $\delta(S^*)S^*$. This fee offsets exactly the profit he makes selling the long-term asset at the high price $p(S^*)$, thus eliminating the overall profitability of this attempt at arbitrage. The entry fee offsets exactly the asset sale profit because, conditional on joining an exchange, agents are indifferent between all feasible portfolio choices. In particular, the arbitrageur joining exchange S^* with the all-long portfolio $(0, e)$ does no better than an agent entering this exchange with the equilibrium portfolio $(s^*, e - s^*)$. Similarly, if the arbitrageur with portfolio $(0, e)$ joins any other exchange S , he is exactly as well off as an agent joining S with the fundamentals-consistent portfolio $(S, e - S)$. Thus, the arbitrageur joining S obtains the ex ante expected utility value of $W(S)$. As we saw in Proposition 1, this value is maximized at $S = S^*$. No arbitrage attempt therefore can be successful.

The agents' ability to commit to not trading across exchanges ex post is key in eliminating the Jacklin arbitrage. The segregated exchanges mechanism lets each agent join only one exchange. In addition, it requires that agents sign off their right to trade freely with the counterparty of their choice. Instead, it requires that agents commit to trading only with other members of the exchange they belong to. If agents do not have the ability to contractually give away their freedom to trade without counterparty restrictions, an impatient arbitrageur residing in the entry-fee-free exchange $S = \pi e$ can easily convince a patient agent in exchange S^* to buy the long-term asset from him rather than in exchange S^* because he can sell for less than $p(S^*)$ and still make a profit. As agents anticipate this at date 0, price expectations embedded in $p(S)$ are not credible and the equilibrium breaks down. Thus, the restriction of participation to one exchange only and the assumption of the agents' ability to commit to not step out of their exchanges ex post are crucial.

In the KT equilibrium, segregated exchanges can therefore be thought of as a commitment device allowing the agents to promise credibly to not access the hidden IOU market. Clearly, if in the KT model agents could access the hidden IOU retrade market *after* they trade in segregated exchanges, the equilibrium with segregated exchanges supporting the optimal asset price $p(S^*)$ would collapse. The argument for it is the same as in Section 4. The optimal allocation (c_1^*, c_2^*) is consistent with free access to the retrade market only if the interest rate in this market equals $R^* = c_2^*/c_1^*$. But with this interest rate, the Jacklin arbitrage can again be executed by investing all long, joining the entry-fee-free exchange $S = \pi e$, and not trading in this exchange

but rather borrowing in the IOU market if liquidity is needed at date 1.²⁰

In the banking model discussed in Section 4, the intermediary designing the state-contingent deposit contract cannot put any restrictions on retrade between depositors and non-depositors. The market-making firm in the segregated exchanges model, in contrast, can. In particular, an agent who did not join exchange S and subject his portfolio to the entry fee $\delta(S)$ cannot retrade with agents who did join exchange S . This additional power given to the market-maker in the segregated exchanges model makes her equally as effective as the social planner in Section 4 in controlling agents' investment at date 0. Unlike the planner, the market-maker does not control this investment directly but rather sets up prices (i.e., exchange entry fees) to induce efficient investment.

As we see, the model with segregated exchanges, where retrade does not lead to a pecuniary externality, requires a different economic environment than the models in Sections 3 and 4, where access to hidden retrade causes an externality. The segregated exchanges model requires that agents have the ability to commit themselves to refrain from trading in the *hidden* retrade market, which effectively makes this model equivalent to the model with observable trades that we discussed in Section 4.²¹ If such commitment can be made credible, e.g., by physically separating agents ex post, then all agents would choose to extend it ex ante. If, however, it is a feature of the environment that such a commitment cannot be made credible, as in Farhi, Golosov, and Tsyvinski (2009), access to hidden retrade makes the Jacklin arbitrage strategy feasible, the pecuniary externality exists, and markets fail to provide sufficient liquidity in equilibrium.

Clearly, the cap-and-trade mechanism will not be successful at limiting greenhouse gas emissions if firms can emit completely privately/anonymously, i.e., without anyone observing it. If they can, the price of the right to emit one tonne of CO₂ will be zero. In the KT model, retrade is analogous to observable emissions that can be priced. In the pecuniary externality model, hidden retrade is analogous to anonymous emissions that cannot be priced or internalized with a cap-and-trade scheme.

Are then segregated exchanges a solution to the pecuniary externality problem caused by retrade? Segregated exchanges do not solve

²⁰ Better yet, the arbitrageur could join one of the exchanges $S < \pi e$, where $\delta(S) < 0$, which means with $s = 0$ he would get a payment from the exchange upon entry.

²¹ That the segregated exchanges model requires a different environment than the unfettered hidden retrade model is clear from Table 1 on page 1,046 in Kilenthong and Townsend (2011).

the pecuniary externality problem, but they show that retrade does not have to lead to one. The literature on pecuniary externalities with complete markets and retrade assumes that agents have unfettered access to an anonymous, hidden retrade market and cannot do anything to make credible an *ex ante* promise to refrain from accessing this market *ex post*. The segregated exchanges model assumes that such a commitment is possible. The segregated exchanges model, therefore, does not solve the pecuniary externality problem associated with anonymous, hidden retrade. Instead, it points out that retrade by itself does not imply the existence of a pecuniary externality. The model shows that retrade can be accounted for within the competitive market framework without violating efficiency, provided that a sufficiently rich market structure, including markets for exchange membership, is allowed for.

In addition, the KT model shows that exclusivity and *ex post* trade restrictions can be socially valuable. Their role can be to serve as a commitment device that agents may be able to use to help them refrain from the “harmful,” hidden retrade activity and still be able to engage in efficient, priced retrade.

6. CONCLUSION

The literature we review makes it clear that in the Diamond-Dybvig economy, the agents’ access to retrade is key in understanding whether markets are efficient or require government intervention. The theory makes a distinction between two kinds of retrade: the “priced” kind and the “hidden” kind. Hidden, anonymous retrade leads to a pecuniary externality and market failure. Priced retrade, harnessed into access-controlled segregated exchanges with exchange- and portfolio-dependent entry fees does not cause market failure.

The observation of retrade itself in present-day financial markets does not therefore imply that markets are inefficient or efficient in providing liquidity. To answer the question of efficiency, one must assess which of the two kinds of retrade discussed in the model is a better reflection of reality. Kilenthong and Townsend (2014a) suggest that the assumption of restricted retrade is a good one in financial markets. In other applications, for example in the problem studied in Kehoe and Levine (1993) where pecuniary externalities result from workers’ unrestricted access to spot labor markets, this assumption may be more problematic, as firms may lack the commitment to deny employment to workers who have defaulted on some financial obligations in the past. Given these theoretical predictions and their implications for the efficacy of government intervention, empirical research identifying

the nature of retrade and the existence or nonexistence of pecuniary externalities is needed.

APPENDIX

Proof of Lemma 1

Suppose allocation c is optimal with $c_2(0) > 0$ and define an allocation $\hat{c} = \{\hat{c}_1(0), \hat{c}_2(0), \hat{c}_1(1), \hat{c}_2(1)\}$ as follows:

$$\begin{aligned} \hat{c}_1(0) &= c_1(0), & \hat{c}_1(1) &= c_1(1), \\ \hat{c}_2(0) &= 0, & \hat{c}_2(1) &= c_2(1) + \frac{\pi}{1-\pi}c_2(0). \end{aligned}$$

Allocation \hat{c} is feasible because at each date $t = 1, 2$ it uses the same amount of resources as allocation c . Indeed:

$$\begin{aligned} & \pi \left(\hat{c}_1(0) + \frac{\hat{c}_2(0)}{\hat{R}} \right) + (1 - \pi) \left(\hat{c}_1(1) + \frac{\hat{c}_2(1)}{\hat{R}} \right) \\ &= \pi c_1(0) + (1 - \pi) \frac{c_2(1) + \frac{\pi}{1-\pi}c_2(0)}{\hat{R}} \\ &= \pi \left(c_1(0) + \frac{c_2(0)}{\hat{R}} \right) + (1 - \pi) \left(c_1(1) + \frac{c_2(1)}{\hat{R}} \right) \\ &\leq e. \end{aligned}$$

Allocation \hat{c} , however, attains a higher value of the objective (5) because it provides the same utility $u(c_1(0))$ to the impatient type and a higher utility $u(c_1(1) + c_2(1) + \frac{\pi}{1-\pi}c_2(0)) > u(c_1(1) + c_2(1))$ to the patient type. This contradicts the supposed optimality of c .

To prove that $c_1(1) = 0$, suppose that c is optimal with $c_1(1) > 0$ and define an allocation $\hat{c} = \{\hat{c}_1(0), \hat{c}_2(0), \hat{c}_1(1), \hat{c}_2(1)\}$ as follows:

$$\begin{aligned} \hat{c}_1(0) &= c_1(0), & \hat{c}_1(1) &= 0, \\ \hat{c}_2(0) &= c_2(0), & \hat{c}_2(1) &= c_2(1) + \hat{R}c_1(1). \end{aligned}$$

Allocation \hat{c} is feasible because it costs the same in present value terms as the feasible allocation c . Indeed:

$$\begin{aligned} & \pi \left(\hat{c}_1(0) + \frac{\hat{c}_2(0)}{\hat{R}} \right) + (1 - \pi) \left(\hat{c}_1(1) + \frac{\hat{c}_2(1)}{\hat{R}} \right) \\ &= \pi \left(c_1(0) + \frac{c_2(0)}{\hat{R}} \right) + (1 - \pi) \frac{c_2(1) + \hat{R}c_1(1)}{\hat{R}} \\ &= \pi \left(c_1(0) + \frac{c_2(0)}{\hat{R}} \right) + (1 - \pi) \left(c_1(1) + \frac{c_2(1)}{\hat{R}} \right) \\ &\leq e. \end{aligned}$$

Table 1 Allocation Mechanisms, Frictions, and Outcomes

Allocation Mechanism	Private Shocks θ	Hidden IOU Retrade Market	Market Structure	Allocation	Page
First-best planning problem	Absent	Absent	No markets—planner chooses allocation	(c_1^*, c_2^*)	313
Planning problem with private θ	Present	Absent	No markets—planner chooses allocation	(c_1^*, c_2^*)	321
Planning problem with private θ and hidden retrade	Present	Present	No markets—planner chooses allocation	(c_1^*, c_2^*)	315
Incomplete markets model	Present	Present	Ex ante: direct investment (s, x) Ex post: market for the long-term asset and the hidden IOU market	(\hat{c}_1, \hat{c}_2)	316
Market model with state-contingent contracts, no hidden IOU retrade markets	Present	Absent	Ex ante: deposit contract (ξ_1, ξ_2) with price e Ex post: no markets	(c_1^*, c_2^*)	321
Market model with state-contingent contracts and hidden IOU retrade markets	Present	Present	Ex ante: deposit contract (ξ_1, ξ_2) with price e Ex post: hidden IOU retrade market	(\hat{c}_1, \hat{c}_2)	319
Market model with segregated exchanges	Present	Absent	Ex ante: continuum of segregated exchanges Ex post: retrade inside exchanges, but not across	(c_1^*, c_2^*)	322

Allocation \hat{c} , however, attains a higher value of the objective (5) than c , because it provides the same utility $u(c_1(0))$ to the impatient type and a higher utility $u(c_2(1) + \hat{R}c_1(1)) > u(c_1(1) + c_2(1))$ to the patient type. This contradicts the supposed optimality of c . QED

Proof of Proposition 1

Since $W''(s) = \frac{1}{\pi}u''\left(\frac{s}{\pi}\right) + \frac{1}{1-\pi}\hat{R}^2u''\left(\frac{e-s}{1-\pi}\hat{R}\right) < 0$, we have that $W'(s) = u'\left(\frac{s}{\pi}\right) - \hat{R}u'\left(\frac{e-s}{1-\pi}\hat{R}\right)$ is continuous and strictly decreasing. The existence of a unique solution to $W'(s) = 0$ in $(0, e)$ thus follows from the fact that $\lim_{s \rightarrow 0}W'(s) \rightarrow \infty$ and $\lim_{s \rightarrow e}W'(s) \rightarrow \infty$.

For the two bounds on s^* , it is sufficient to show that $W'(\pi e) > 0$ and $W'\left(\pi e \frac{\hat{R}}{\pi\hat{R}+1-\pi}\right) < 0$. We first note that relative risk aversion everywhere strictly greater than one implies that the function $f(\alpha) = \alpha u'(\alpha e)$ is strictly decreasing. Indeed, with $f'(\alpha) = u'(\alpha e) + \alpha e u''(\alpha e)$ we have that $f'(\alpha) < 0$ follows from $1 < \frac{-\alpha e u''(\alpha e)}{u'(\alpha e)}$. Now, evaluating W' at $s = \pi e$, we have $W'(\pi e) = u'(e) - \hat{R}u'(e\hat{R}) = f(1) - f(\hat{R}) > 0$, where the strict inequality follows from f strictly decreasing and $\hat{R} > 1$. To show $W'\left(\pi e \frac{\hat{R}}{\pi\hat{R}+1-\pi}\right) < 0$, note that with $s = \pi \frac{\hat{R}e}{\pi\hat{R}+1-\pi}$ we have $\frac{s}{\pi} = \frac{e-s}{1-\pi}\hat{R} = \frac{\hat{R}e}{\pi\hat{R}+1-\pi}$. Therefore,

$$\begin{aligned} W'\left(\pi \frac{\hat{R}e}{\pi\hat{R}+1-\pi}\right) &= u'\left(\frac{\hat{R}e}{\pi\hat{R}+1-\pi}\right) - \hat{R}u'\left(\frac{\hat{R}e}{\pi\hat{R}+1-\pi}\right) \\ &= (1 - \hat{R})u'\left(\frac{\hat{R}e}{\pi\hat{R}+1-\pi}\right) \\ &< 0, \end{aligned}$$

where the inequality follows from $u' > 0$ and $\hat{R} > 1$. QED

Formal Definition of Incomplete-Markets Equilibrium in Section 3

At date 0, each agent chooses an investment portfolio (s, x) . Agents solve

$$\begin{aligned} &\max_{(s,x) \geq (0,0)} \mathbb{E}[V_1(s, x; \theta)] \\ &s.t. \\ &s + x \leq e, \end{aligned} \tag{35}$$

where $V_1(s, x; \theta)$ is the indirect utility function representing the value the agent can obtain at date 1 if he holds investments (s, x) and receives realization θ of the liquidity shock. This indirect utility function is defined as follows:

$$\begin{aligned} V_1(s, x; \theta) = & \max_{(c_1, c_2) \geq (0, 0), n, b} u(c_1 + \theta c_2), \\ & s.t. \\ & c_1 + pn + b \leq s, \\ & n \geq -x, \\ & c_2 \leq (x + n)\hat{R} + bR, \end{aligned} \quad (36)$$

where n represents net purchases of the long-term asset in the asset market at date 1 and b represents expenditures on the IOUs in the hidden retrade market. Let $n(\theta; s, x, p)$ denote net demand for the long-term asset of a type- θ agent.

Competitive equilibrium consists of initial investments s and x , value functions $V(s, x; \theta)$, a date-1 price p for the long-term asset, and a gross interest rate R in the hidden retrade market such that (i) given p and R , value functions solve (36); (ii) given V , investment choices s and x solve (35); and (iii) the date-1 market for the long-term asset clears, $\mathbb{E}[n(\theta; s, x, p)] = 0$, and R is an equilibrium interest rate on the hidden retrade market.

REFERENCES

- Allen, Franklin, and Douglas Gale. 2004. "Financial Intermediaries and Markets." *Econometrica* 72 (July): 1,023–61.
- Bianchi, Javier. 2011. "Overborrowing and Systemic Externalities in the Business Cycle." *American Economic Review* 101 (December): 3,400–26.
- Bisin, Alberto, and Piero Gottardi. 2006. "Efficient Competitive Equilibria with Adverse Selection." *Journal of Political Economy* 91 (June): 485–516.
- Dávila, Julio, Jay H. Hong, Per Krusell, and José-Víctor Ríos-Rull. 2012. "Constrained Efficiency in the Neoclassical Growth Model with Uninsurable Idiosyncratic Shocks." *Econometrica* 80 (November): 2,431–67.

- Diamond, Douglas W., and Philip H. Dybvig. 1983. "Bank Runs, Deposit Insurance, and Liquidity." *Journal of Political Economy* 114 (June): 401–19.
- Di Tella, Sebastian. 2014. "Optimal Financial Regulation and the Concentration of Aggregate Risk." Stanford Graduate School of Business Working Paper.
- Ellerman, A. Denny, and Barbara K. Buchner. 2007. "The European Union Emissions Trading Scheme: Origins, Allocation, and Early Results." *Review of Environmental Economics and Policy* 1 (1): 66–87.
- Ennis, Huberto M., and Todd Keister. 2010. "On the Fundamental Reasons for Bank Fragility." Federal Reserve Bank of Richmond *Economic Quarterly* 96 (1): 33–58.
- Farhi, Emmanuel, Mikhail Golosov, and Aleh Tsyvinski. 2009. "A Theory of Liquidity and Regulation of Financial Intermediation." *The Review of Economic Studies* 76 (3): 973–92.
- Geanakoplos, John D., and Heraklis M. Polemarchakis. 1986. "Existence, Regularity, and Constrained Suboptimality of Competitive Allocations when the Asset Market is Incomplete." In *Essays in Honor of Kenneth Arrow*, Volume 3, edited by Walter P. Heller, Ross M. Starr, and David A. Starrett. New York: Cambridge University Press, 65–95.
- Golosov, Mikhail, and Aleh Tsyvinski. 2007. "Optimal Taxation with Endogenous Insurance Markets." *The Quarterly Journal of Economics* 122 (2): 487–534.
- Greenwald, Bruce C., and Joseph E. Stiglitz. 1986. "Externalities in Economies with Imperfect Information and Incomplete Markets." *The Quarterly Journal of Economics* 101 (May): 229–64.
- Jacklin, Charles J. 1987. "Demand Deposits, Trading Restrictions, and Risk Sharing." In *Contractual Arrangements for Intertemporal Trade*, edited by Edward C. Prescott and Neil Wallace. Minneapolis: University of Minnesota Press, 26–47.
- Kehoe, Timothy J., and David K. Levine. 1993. "Debt-Constrained Asset Markets." *The Review of Economic Studies* 60 (October): 865–88.
- Kilenthong, Weerachart T., and Robert M. Townsend. 2011. "Information-Constrained Optima with Retrading: An Externality and Its Market-Based Solution." *Journal of Economic Theory* 146 (May): 1,042–77.

- Kilenthong, Weerachart T., and Robert M. Townsend. 2014a. "Segregated Security Exchanges with Ex Ante Rights to Trade: A Market-Based Solution to Collateral-Constrained Externalities." Working Paper 20086. Cambridge, Mass.: National Bureau of Economic Research (May).
- Kilenthong, Weerachart T., and Robert M. Townsend. 2014b. "A Market Based Solution to Price Externalities: A Generalized Framework." Working Paper 20275. Cambridge, Mass.: National Bureau of Economic Research (July).
- Lorenzoni, Guido. 2008. "Inefficient Credit Booms." *The Review of Economic Studies* 75 (3): 809–33.
- Mas-Colell, Andreu, Michael D. Whinston, and Jerry R. Green. 1995. *Microeconomic Theory*. New York: Oxford University Press.
- Prescott, Edward C., and Robert M. Townsend. 1984. "Pareto Optima and Competitive Equilibria with Adverse Selection and Moral Hazard." *Econometrica* 52 (January): 21–45.
- Stiglitz, Joseph E. 1982. "The Inefficiency of the Stock Market Equilibrium." *The Review of Economic Studies* 49 (April): 241–61.
- Viner, Jacob. 1932. "Cost Curves and Supply Curves." *Zeitschrift für Nationalökonomie* 3: 23–46.