Essays in Applied Economics

by

Alan Michael Grant

BS, University of Michigan (2001)

Submitted to the Department of Economics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2007

Author .....
Department of Economics
October, 2006

Certified by .....
Roberto Rigobon
Associate Professor of Economics
Thesis Supervisor

Certified by .....
Robert Gibbons
Sloan Distinguished Professor of Organizational Economics and Strategy
Thesis Supervisor

Accepted by .....
Peter Temin
Elisha Gray II Professor of Economics
Chairman, Departmental Committee on Graduate Studies

Essays in Applied Economics

by

Alan Michael Grant

Submitted to the Department of Economics
in October, 2006, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Economics

Abstract

This dissertation is composed of three chapters, the first demonstrates that natural gas violates many of the simplifying assumptions frequently used in modeling its behavior. Careful analysis of futures contracts written on gas suggests that gas prices are seasonal while returns are non-Gaussian and evidence stochastic volatility. In addition, examination of options prices indicates the intermittent presence of jumps. We find that models which disregard these properties struggle to recover options prices with any precision. Thus, we propose an alternative nonparametric approach to gas options pricing that captures these salient features while also shedding light on the nature of risk aversion embedded in gas markets.

The second chapter presents new estimates and approaches to estimating the home bias puzzle. It uses micro-level data to calculate households' foreign equity exposure as a function of wealth. We find simple estimates have significant errors-in-variables problems and we construct an estimator using grouping to account for this issue. Our estimates still imply low aggregate investment in foreign equity. Finally, we disaggregate the investment decision by incorporating two step decisions that allow households to forgo participating in the market. As a result of the decoupling, we find foreign equity levels closer to that of standard portfolio theories.

The final chapter considers principal-agent models in which the principal cannot measure the output nor the effort level of agents. To model this situation, we use utility models that include identity, justified partly by empirical results from peer-effects, and apply these extended utility functions. In the single agent case, introducing identity amounts to modifying the utility function and does not lead to dramatic results. In the multiple agent case, we find that the addition of identity can lead to more efficient outcomes than cases where identity is ignored. The addition of identity, however, can also lead to counter-intuitive results due to the interactions among agents and may produce second-best outcomes that are worse than the case without identity. Finally the addition of identity can help explain some empirical results that may be difficult to explain with standard models.

Thesis Supervisor: Roberto Rigobon
Title: Associate Professor of Economics

Thesis Supervisor: Robert Gibbons
Title: Sloan Distinguished Professor of Organizational Economics and Strategy

Acknowledgements

First, I would like to thank my advisors, Roberto Rigobon and Robert Gibbons, for all their advice and help. This dissertation would not be possible without their assistance. Second, the dissertation benefited from comments and suggestions from Victor Chernozhukov, Jerry Hausman, Andy Lo, and Whitney Newey. I appreciate all of their ideas and input. Third, I would like to thank Jon Gruber, Sergei Izmalkov, Bill Wheaton, Mark Begley, and Gary King for making this dissertation possible while teaching. Finally, I am grateful to Kai-Uwe Kühn, Richard Canary, and Stephen Salant for encouraging my matriculation to graduate school.

Completing a dissertation is nearly impossible without helpful colleagues and friends. I am thankful for the comments, conversations, and friendship of David Abrams, Michael Anderson, Josh Fischman, and David Matsa. I am also indebted to Erik Ruben, my coauthor for the first two chapters of this dissertation, for his tireless work and effort. I would also like to thank my family for their support. Lastly, words cannot adequately express my thanks to my wife Kimberly for her endless encouragement and companionship.

Contents

List of Figures

List of Tables

Chapter 1

Capturing the Idiosyncrasies of Natural Gas Markets for Better Derivative Pricing

With Erik C. Ruben

Abstract

In this paper, we demonstrate that natural gas violates many of the simplifying assumptions frequently used in modeling its behavior. Careful analysis of futures contracts written on gas suggests that gas prices are seasonal while returns are non-Gaussian, and evidence stochastic volatility. In addition, examination of options prices indicates the intermittent presence of jumps. We find that models which disregard these properties can accurately predict futures prices, but struggle to recover options prices with any precision. Thus, we propose an alternative nonparametric approach to gas options pricing that captures these salient features while also shedding light on the nature of risk aversion embedded in gas markets important for evaluating and calibrating derivatives models.

1.1 Introduction

Over the last thirty years, natural gas markets have come to play an extremely important role in the global economy. In 2005, the United States alone consumed about $260 billion in natural gas.[1] In that same year, greater than $1 trillion in gas

[1]This assumes $10 average gas and 26 trillion cubic feet of consumption. (cf. Energy Information Administration (2004))

futures traded on the New York Mercantile Exchange (NYMEX).[2] A variety of factors including its relative abundance, low cost of transport, and promise as a clean source of fuel, have helped to make natural gas the world's fastest growing commodity as well a major profit driver for leading investment banks.[3] Natural gas is unlikely to lose any of its momentum as interest in national energy independence, reduced environmental impact, and the low cost associated with gas-fired power generation drive producers and consumers towards its further embrace. In addition to growth in the underlying physical market, gas's high price volatility will increasingly motivate market participants to manage their risk by trading in derivatives.

Despite its substantial economic significance, natural gas has received comparatively little attention from researchers in finance. The vast literature in asset pricing has certainly provided some insights into the value of spot prices and derivative contracts written on gas. However, academic work has mostly focused on equities, fixed income securities, and currencies. Commodities, and natural gas in particular, behave quite differently empirically than other asset classes making it difficult to apply, for example, equity derivative models to price natural gas options. Gas prices are clearly seasonal and evidence volatility and jumps which vary through time in a complicated manner. Price levels seem to be related to convenience yields, storage costs, and the price of alternative energy sources such as oil.[4] As documented in this paper, natural gas derivatives have unique properties too; futures on gas are distinguished by a small degree of backwardation, a feature prominent among agricultural and metal commodities, while options on futures display upwards sloping implied volatility wings. Finally, natural gas markets evidence an unusually low degree of geographic integration. As a result, in the United States, where pipeline networks in eastern and western states do not interconnect to a great degree, there is substantial price segmentation across regions.[5] All of these factors suggest that gas's underlying market microstructure and representative stochastic process differ from those of equities and fixed income securities and other commodities as well. These differences have profound implications not only for the understanding of gas prices in their own right, but for the valuation and hedging of both real assets and financial derivative contracts tied to the underlying price of gas.

A collection of recent work, including Pindyck (2001), Pindyck (2004), Gibson and Schwartz (1989), Schwartz (1997), Miltersen and Schwartz (1998), Schwartz and Smith (2000), Todorova (2004b), and Clewlow and Strickland (2000), attempts to capture a few of the stylized facts regarding commodities futures and incorporate them into parametric partial and general equilibrium models. To the extent that the

[2]The value of over-the-counter transactions in futures and other derivative contracts totaled an even greater amount according to industry sources.

[3]See Geman (2005, p. 227).

[4]Convenience yield refers to the benefit associated with directly holding inventory in an underlying asset rather than a derivative contract written on the product. See Pindyck (2004).

[5]See Cuddington and Wang (April 20, 2005).

authors extend their models directly to natural gas, they limit their empirical analysis to futures prices. This is likely due to the fact that exchange-traded gas derivatives were limited before March of 2004 when NYMEX introduced a European options contract on gas futures.[6] In addition to all the other benefits they offer with respect to risk management and market completion, options markets provide rich information about the underlying security's market structure.[7] A spate of recent papers including Jackwerth (2000), Aït-Sahalia and Lo (2000), Carr and Wu (2003b), Carr and Wu (2003a), and Carr and Wu (2004) have exploited the theoretical links between option prices and those of the underlying to deduce important characteristics about investor preferences and admissible stochastic processes.

To date, researchers have made numerous simplifying assumptions in their attempts to model commodities markets. We show that natural gas meaningfully violates many of the idealized conditions commonly imposed. Moreover, these departures, which can be grouped into two classes, are of first order importance when pricing derivatives written on gas. First, there are features such as non-normality, seasonality, and stochastic volatility in returns which are evident from direct analysis of gas spot and futures prices. The second class of deviations are those features, such as low risk aversion and the intermittent presence of jumps, which are only observable via indirect examination using options data. This paper sets out to accomplish the following: (1) leverage specialized empirical tools from the equity options literature along with more standard econometric techniques to document the important features of gas markets often neglected in extant models, (2) demonstrate how the failure to incorporate these features can lead to substantially magnified pricing errors in options markets relative to futures markets, and (3) propose an alternative nonparametric approach to derivative pricing that avoids these pitfalls.

The paper proceeds as follows: In Section 1.2, we furnish an abbreviated survey of the literature to give context to our findings. Section 1.3 is devoted to explaining our data set. Section 1.4 documents important features of natural gas's stochastic process. First, the section offers evidence from spot and futures prices of the idiosyncratic nature of gas's process. Next, it exploits estimation techniques first employed by Carr and Wu to identify the periodic presence of jumps in addition to diffusive behavior. Section 1.5 makes use of an observation by Breeden and Litzenberger (1978) regarding the link between European options prices and state price densities as well as an estimation technique employed by Aït-Sahalia and Lo (2000) and Jackwerth (2000) to provide evidence of another important feature of gas markets: investor risk aversion that is low and relatively constant in wealth. Section 1.6 introduces several approaches to pricing natural gas options representative of the those in the existing literature. It establishes the failure of these candidate models to accurately recover the market prices of options as the result of the misspecifications in the stochastic process and the nature of risk aversion highlighted in the prior two sections of the paper.

[6] NYMEX introduced American options on natural gas futures in 1992.

[7] See Ross (1976).

Section 1.7 offers an alternative nonparametric means of pricing derivatives using a kernel estimator which disregards the common restrictions we have determined to be inaccurate. Consequently, the kernel estimator is significantly more successful in pricing options than the other methods considered. Section 1.8 concludes.

1.2 Background and Literature Review

Some of the earliest research into commodity pricing dates back to work by Kaldor, Working, and Telser who studied the interplay of storage costs and convenience yields and their impact on the relationship between spot and futures prices.[8] Recent efforts have focused on establishing either richer microfoundations or better empirical properties at the cost of reduced form modeling. Among the structural approaches, some of the more notable papers include Sundaresan (1984), Chambers and Bailey (1996), and Routledge, Seppi, and Spatt (2000). Sundaresan (1984) develops an equilibrium model for spot and futures prices in a nonrenewable commodity market characterized by uncertain exogenous discoveries of the resource. The paper finds that in periods between supply shocks, spot prices generate positive excess return as a function of the price elasticity of demand, the mean arrival rate of discoveries, and the degree of enlargement to existing reserves. In times of repeated discoveries, the model predicts discontinuous price declines. Using equilibrium arguments, Sundaresan derives the price of a futures contract as a function of the price elasticity of demand, the spot price, the volatility in reserve levels, and the contract's time to maturity.

Chambers and Bailey (1996) focuses on the determination of spot prices by examining equilibria under various assumptions about the nature of supply shocks. The paper proves the existence of a unique stationary rational expectations equilibrium under three types of disturbances: independent and identically distributed, time dependent, and periodic. It develops testable implications for each model type and conducts an empirical exercise with a variety of agricultural commodities; the paper finds weak support for a model with periodic supply shocks.

Routledge, Seppi, and Spatt (2000) builds on work by Wright and Williams (1989), Chambers and Bailey (1996), and Deaton and Laroque (1996) to develop a competitive rational expectations model of storage. The paper solves for the equilibrium level of inventory in a setting with competitive risk-neutral agents in which "immediate use" consumption value is determined by a mean-reverting Markov process. The inventory rule and shock process together determine the spot and forward price processes. In empirically testing their model with NYMEX crude oil futures, Routledge et al. find that the one-factor version fails to produce the correct conditional and unconditional moments of the data while the two-factor extension has somewhat greater success.

Amongst those papers that start from a set of reduced form assumptions and make use of no-arbitrage arguments, Black (1976) is perhaps the best known. It forgoes

[8] See Kaldor (1939), Working (1948), Working (1949), and Telser (1958).

analysis of spot prices and focuses instead on deriving a closed form expression for the value of commodity options written on futures prices. The method amounts to first, valuing a futures contract as the expected value under the risk-neutral measure of the spot at the time of expiration, and second, replacing the value of the spot in the original Black-Scholes-Merton formula with the discounted value of the futures price. The paper makes the simplifying assumption that futures have a lognormal distribution.

In a series of primarily co-written papers, Eduardo Schwartz has developed several approaches to modeling spot commodity prices as well as futures and options on futures. The first paper in the sequence, Gibson and Schwartz (1989), adapts the two factor partial equilibrium bond pricing model of Brennan and Schwartz (1979) to commodity markets. In this context, the two factors are the spot price of the commodity and the instantaneous convenience yield, which are assumed to evolve according to a geometric Brownian motion and mean reverting diffusion process respectively. As an empirical exercise, the paper looks at weekly oil futures and finds that the model prices short-term futures with reasonable success. Schwartz (1997) extends Gibson and Schwartz (1989) by adding an instantaneous interest rate that also follows a mean-reverting process. More importantly, this iteration in the series shows how to take advantage of the model's inherent Markovness by rewriting it in state space form and estimating the unobserved state variables via the Kalman filter and maximum likelihood technique.

Miltersen and Schwartz (1998) further builds on Schwartz (1997) by deriving an analytical expression for valuing European options on commodity futures in the presence of stochastic interest rates and stochastic convenience yields. In the same *Journal of Finance* issue, Hilliard and Reis (1998) incorporates jumps in the spot process into the framework of Schwartz (1997). The paper manages to endogenize the market price of risk stemming from interest rates but leaves the risk associated with the convenience yield as an exogenous parameter set in equilibrium. Schwartz and Smith (2000) breaks from the tradition of modeling spot prices as draws from a lognormal distribution so that it can capture both the effect of price's long-term impact on supply as well as more immediate deviations from the equilibrium level. To accomplish this, the paper formalizes a two factor model. The first factor follows an Ornstein-Uhlenbeck process which reverts back to zero and is designed to soak up short-term shocks like supply interruptions and demand variation stemming from weather. The second factor, which evolves according to a geometric Brownian motion with drift, incorporates long-term changes to the equilibrium price level resulting from political and regulatory effects, technological improvements related to the discovery and production of the commodity, and expectations regarding exhaustion of the existing supply. As with Schwartz (1997), the two-factor formulation admits an easy state space representation so that Kalman filtering and maximum likelihood can be used for its estimation. Though formally equivalent to Schwartz (1997), Schwartz and Smith (2000) has greater econometric and intuitive appeal. Specifically, the

short-term/long-term model is more "orthogonal" in its dynamics than the approach based on spot prices and convenience yields. In the Gibson and Schwartz (1989) framework, the convenience yield plays a role in the stochastic process for the spot price whereas in the Schwartz and Smith (2000) set-up, the only interaction between factors arises via the correlation of their stochastic increments. Schwartz and Smith argue that this orthogonality is not only cleaner (i.e. the volatility for the price of a futures contracts is equal to the volatility of the sum of short- and long-term factors) but it may make it possible to safely disregard the short-term factor when valuing long-term assets. Such a simplification facilitates extensions to the model like that of a stochastic equilibrium growth rate. Finally, the authors estimate the parameters of the model using prices from oil futures contracts

Todorova (2004b), whose primary aim is achieving a closer fit to natural gas futures data, incorporates explicit seasonal price fluctuations into the framework of Schwartz and Smith (2000). To this end, the paper considers a third "seasonal" stochastic factor as well as various other means of deseasonalizing the data. Clewlow and Strickland (2000) take a nonparametric approach based on principal component analysis to modeling the futures curve; the strategy is flexible enough to take seasonality into account. Examining the case of gas futures, Todorova compares the results generated by her models with those implied by Schwartz and Smith (2000) and the volatility functions model of Clewlow and Strickland (2000). She finds that the three factor model with the stochastic seasonal component produces the highest likelihood amongst all the models considered but that Clewlow and Strickland's methodology has superior out-of-sample prediction performance.

More recently, Doran (2005), building on earlier work in Doran and Ronn (2006), attempts to model natural gas options under various stochastic volatility regimes. Although he does not use actual European option prices, he makes use of a technique pioneered in Barone-Adesi and Whaley (1987) to approximate European prices from traded American options. He finds best out of sample performance in a least absolute deviations sense using a variant of the Bates (1996) stochastic volatility and jumps model where he additionally allows for jumps in the volatility itself.

1.3 Data

In this section, we describe the data used in this paper and highlight some of its salient features. The two types of gas derivatives examined, natural gas futures and European options written on futures, are exchange traded on NYMEX. Yield curves are constructed from data on government securities made available by the US Treasury Department.

1.3.1 Futures Data

Our data on natural gas futures (symbol: NG) consist of daily settlement prices for the natural gas futures contract traded on NYMEX from April 1990 to December 2005.[9] [10] Contracts are priced in dollars per million British thermal units (mmBtu) and obligate the seller to deliver gas to the Henry Hub in Louisiana. The trading unit for the market is 10,000 mmBtu. The data consists of 3,942 days of prices for contracts of 12 different maturities—a one month maturity, two month maturity, and so on up until and including a twelve month maturity contract.[11]

Estimation of parametric pricing models often necessitates the use of synthetic fixed maturity data in order to reduce the dimensionality of the problem; otherwise, contracts would not be comparable through time and hence require the estimation of parameters which change, for example, daily. Thus, we construct a complete futures curve for each trading day using actual prices and interpolate via a cubic spline approximation procedure a constant maturity price series. Throughout the paper, we make special use of our synthetic contracts with maturities that are multiples of a month.[12]

Liquidity in natural gas, as with other commodities, is concentrated in futures rather than spot markets. While the dynamics of futures markets, per se, are not of principal interest in this paper, NYMEX options which play a central role in our investigation, are written on futures and thus our interaction with them is unavoidable. Consequently, we provide some analysis of their characteristics as well. Table 1.1 highlights some basic sample statistics of the shortest maturity contract which can be interpreted as a proxy for the spot price. One observation immediately evident from this table is the increasing average price, a point to which we return in Section 1.6.1 when we detrend the data. In particular, prices rise substantially after the year 2000. In figure 1.1, we illustrate the dynamics of natural gas futures by showing the settlement prices for all of the contracts from March 2004 to December 2005 (i.e. the period for which we have options data) for maturities ranging from one

[9] See `http://www.nymex.com/`.

[10] Settled prices are volume-weighted averages of transactions which occur in the final two minutes of the trading session.

[11] As is standard with futures, the one month contract is not the same instrument across days because the time to expiration changes daily. For example, on January 1st, the contract expiring in February of the same year matures in 30 days while the one month contract on January 2nd matures in 29 days. This feature of futures contracts necessitates some care in modeling since the one month contract is not the same asset through time.

[12] There are, of course, alternatives to constructing a constant maturity price series. One procedure is simply to define the one month contract as the contract expiring the following month. Another is to define the one month contract based on a window wherein it equals the contract which expires the following month if the time to expiry is greater than, say 2 weeks, and equal to the contract expiring in 2 months otherwise. This method easily extends to other contracts. We found that neither of these procedures works well in spot price models and generally produces a poorer fit than that obtained via a spline.

Table 1.1: Natural Gas Futures Sample Statistics
The prices in the table are those of the one month contract and are denominated in dollars.

	Time Period				
Statistic	**Entire Sample**	**90–95**	**95–00**	**00–05**	**3/04–12/05**
Min	1.046	1.046	1.323	1.830	4.570
1st Quartile	1.895	1.510	1.946	3.491	6.146
Median	2.358	1.720	2.287	5.149	6.819
Mean	3.315	1.794	2.579	5.412	7.819
3rd Quartile	4.294	2.085	2.748	6.348	8.117
Max	15.380	3.448	9.980	15.380	15.380
Std. Dev.	2.275	0.376	1.132	2.473	2.582

month to twelve months. The plot demonstrates that prices rise over the time period. In addition, the two prominent diagonal "humps" in the plot clearly testify to strong seasonality as the protrusions in price in the Date-Price plane always correspond to December and January contracts. These humps are also evidence of consistent backwardation in gas markets.

There are several more important features of the futures curve and its evolution through time that deserve mention. First, and not surprisingly, there exists significant correlation between the prices of different contracts on a given day; correlations often exceed 0.9. Second, we examine the realized distributions of returns between pairs of contracts using Kolmogorov-Smirnov tests and find that we cannot statistically reject the null hypothesis that the returns of contracts with different maturities are realizations from the same distribution. Finally, we look at the correlation between continuously compounded daily returns of different contracts and find that returns at time t and time $t+\tau$ have correlations near 0 for values of τ ranging from 30 days to 360 days.

1.3.2 Options Data

We also make use of data on European-style options on natural gas futures (symbol: LN) traded on NYMEX.[13] The options essentially expire at the same time as the underlying futures contract and, as is the case with futures, prices are quoted per mmBTU. One option constitutes the right to buy or sell one futures contract on 10,000 mmBTU of gas. The data consists of daily settlement prices in dollars for call and put options traded on NYMEX from March 2004 (when the options began trading) to December 2005. This corresponds to 455 trading days and 47,408 contracts with a positive trading volume. To allow for simpler pricing models and reduce

[13]We make use of daily "settled" prices as determined by NYMEX's Options Settlement Committee at the end of trading.

Figure 1.1: Natural Gas Futures Prices

The figure plots the daily futures curve from March 1, 2004 until December 31, 2005 using contracts with maturities from 1 to 12 months. Prices are in dollars.

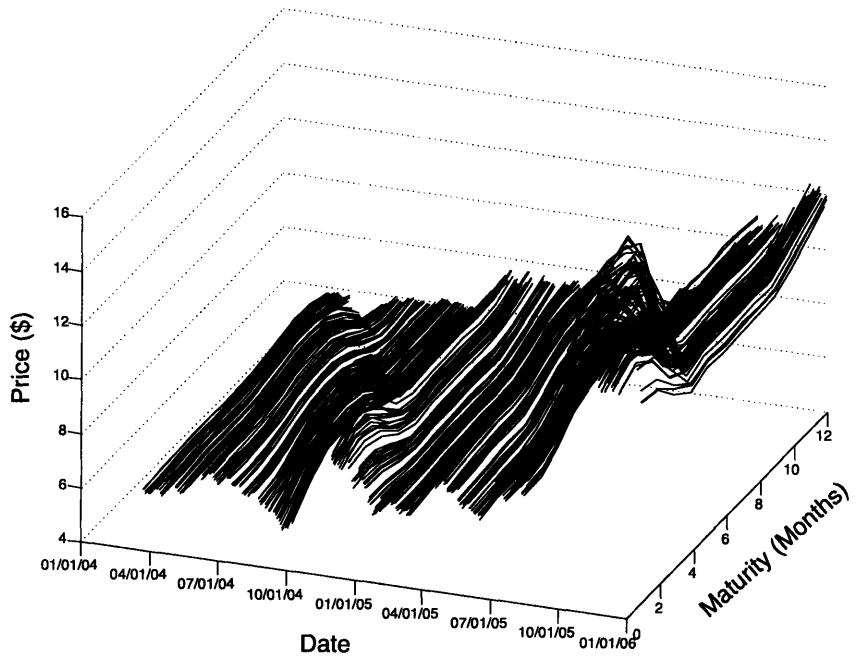


Table 1.2: Option Summary Statistics

Call price is the price of the call in dollars. Also, we have used put-call parity to translate the put prices into calls and the table reports those prices. Implied volatility represents the implied volatility as calculated using the pricing formula from Black (1976), τ represents the time to maturity in years, X represents the strike price in dollars, and r represents the risk-free interest rate used for that option calculated as we have described in Section 1.3.3.

Variable	Mean	Std. Dev.	Min	Median	Max
Call price ($)	1.162	1.010	0.000	0.915	10.710
Implied σ (%)	39.890	12.243	0.049	38.280	159.400
τ (years)	0.753	0.707	0.011	0.564	5.536
X ($)	8.454	3.111	1.000	7.750	99.000
r (%)	3.103	0.836	0.776	3.195	4.557

dimensionality, we use put-call parity to convert the put option prices into call prices. When both puts and calls are traded with the same maturity and strike and both have positive trade volume, we follow a simple decision rule of using the price implied by the derivative with the higher trade volume. Thus we use the actual call price if its volume exceeds that of the put and otherwise use the implied call price of the put. We only include option prices where there is positive trading volume to ensure a higher confidence in the reported price quotes.

For several pricing models, we also need pricing information on the underlying futures contract. Thus, we merge the futures data, interest rate data, and option data to produce a combined data file. This constructed dataset includes 427 trading days from March 2004 to December 2005 and has 38,885 unique option prices. We report various summary statistics for the option data in table 1.2 and table 1.3.

For the pricing models that we introduce later in this paper, we describe the options in terms of the Black (1976) implied volatilities rather than the prices themselves. Since Black (1976) provides a unique one-to-one mapping between prices and implied volatilities, this presents no loss of information. Further, it complies with both industry and academic convention. We are particularly interested in how variables such as moneyness, which is defined as an option's strike price divided by the price of the underlying futures contract, and time to expiry, influence implied volatility and thus prices.

To produce a smooth visualization of the surface implied by the options, we estimate the relationship between time to maturity, moneyness, and implied volatility using a nonparametric series regression.[14] Several representative plots are provided in figure 1.2. As is evident from the plot, the relationship between implied volatility, time to expiry and moneyness is not entirely stable through time yet the overall shape

[14]For these plots, a third order Taylor approximation is used as the approximation function. The results are robust to changing this approximation function.

Table 1.3: Natural Gas Futures and Options Median Volumes by Contract
The Maturity column represents the contract's time to maturity in months. The NG volume column represents the median number of contracts traded over the entire data set from 1990 to 2005 and the LN volume represents the median number of options traded using the entire sample.

Maturity	NG Volume	LN Volume
1	20,700	200
2	9,160	150
3	3,639	150
4	2,033	100
5	1,348	100
6	953	100
7	676	100
8	503	100
9	393	100
10	311	100
11	241	100
12	208	100

is quite persistent.

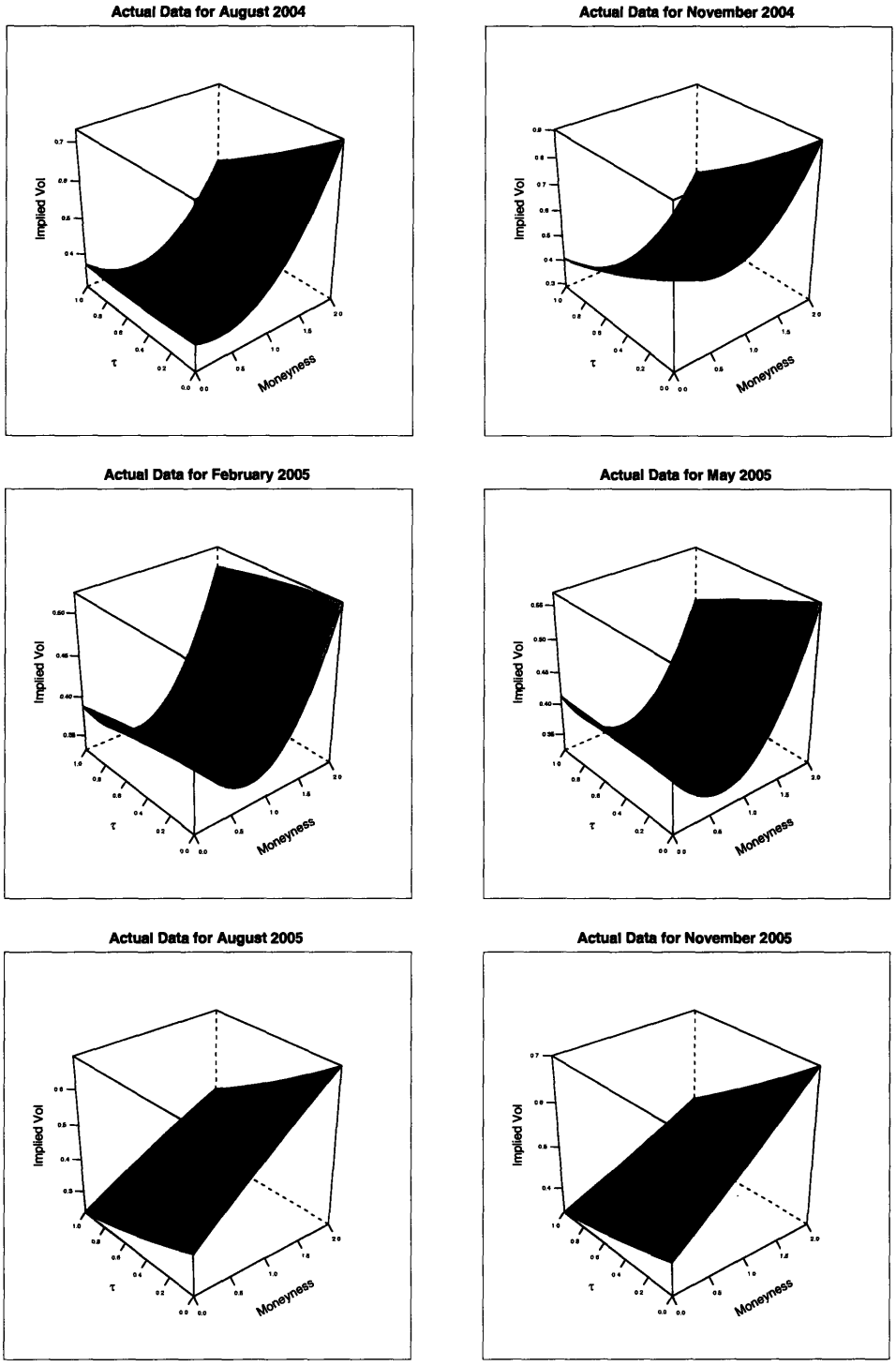
First, the implied volatility is not constant and each month produces a volatility surface entirely different than the plane which arises out of the traditional Black-Scholes-Merton assumptions. Second, over almost any time period, there is a pronounced positive relationship between moneyness and implied volatility. This is the opposite of the relationship we observe in equity index options, and different than the "smiles" that characterize equity options. Also, there is a negative relationship between the time to maturity and implied volatility. Again, this contrasts with what we often observe in equity option markets. These relationships evidenced in the actual data are quite strong and explored in later sections to make inferences about the underlying gas market and evaluate the accuracy with which option models recover market prices.

1.3.3 Interest Rate Data

Our interest rate data consists of daily rate quotes for fixed maturity securities with expirations in 1 month, 3 months, 6 months, and several longer term maturities.[15] The US Treasury Department has made this data available since 1990 thus more

[15]The Treasury Department uses cubic spline interpolation to actually derive these rates. See `http://www.treasury.gov/offices/domestic-finance/debt-management/interest-rate/yieldmethod.html` for more information.

Figure 1.2: Implied Volatility Surface of Actual Prices



than matching the life of our combined futures and options data sets.[16] In estimating various models throughout the paper, we utilize interpolations from cubic splines of the complete fixed-length Treasury rate curve to estimate the risk-free rate associated with a given maturity and trade date. Overall, we find that our models are not especially sensitive to interest rates.

1.4 Analysis Under the Objective Measure: Process Specification

In this section, we use both the futures and the options data to illuminate the dynamics of natural gas. These properties are critical stylized facts that most existing papers, such as those cited in Section 1.2, largely ignore because their introduction makes modeling and estimation a more intractable task, because they are poorly documented, or because they are specific to gas while the author's focus is elsewhere. We view as a central contribution of this paper the notion that correctly capturing these features is particularly significant when one is trying to value derivatives written on gas. We first highlight three features of the gas process that are obtainable from direct analysis of the futures prices and then proceed to analyze those exposed via an examination of options written on those futures.

1.4.1 Evidence from Futures

We begin by utilizing futures data to document the empirical properties which characterize the dynamics driving gas markets. One crucial aspect in which gas markets behave in a similar manner to what we observe in, for example, equity markets involves the pattern in which the distribution of returns evolves over time. Specifically, we test the hypothesis that simple and continuously compounded returns are normally distributed. First, we construct daily, weekly, and monthly returns and generate Q-Q plots against a normal distribution to check for normality. The Q-Q plot shows points that represent the realized quantiles of the actual futures returns and a line that represents the quantiles of a normal distribution with the same mean and standard deviation. If the futures returns were normal, we would expect the points to lie on the line. As with equities, it seems clear from figure 1.3 that short duration returns are non-Gaussian while longer duration returns have a distribution that is closer to normal. We formally test this hypothesis with Shapiro-Wilk tests. We report the results from the Shapiro-Wilk test in table 1.4 and note that the tests reject the hypothesis that returns are normally distributed for short duration returns, but

[16]The data is available online from Treasury's website: `http://www.treasury.gov/offices/domestic-finance/debt-management/interest-rate/yield.shtml`.

Table 1.4: Representative Test of Normality

Maturities are in months and W represents the Shapiro-Wilk test statistic. *** represents significance at the 1 percent level and hence we can reject normality at the 99 percent confidence level.

	Daily	Weekly	Monthly
Maturity	W	W	W
1	0.882***	0.983***	0.991
2	0.888***	0.984***	0.992
3	0.893***	0.985***	0.995
4	0.896***	0.985***	0.996
5	0.896***	0.987***	0.995
6	0.899***	0.988***	0.994
7	0.907***	0.989***	0.993
8	0.919***	0.990***	0.993
9	0.930***	0.991***	0.994
10	0.938***	0.991***	0.996
11	0.941***	0.991***	0.996
12	0.943***	0.992***	0.996

cannot reject normality in monthly returns.[17] These results indicate that the common assumption of a lognormal price process can prove problematic.

A second important property of gas made evident from analysis of futures prices is its randomly time-varying volatility. The accurate pricing of options, an important component of this paper, is closely linked to nature of volatility. More precisely, the Black-Scholes-Merton framework rests heavily on the assumption that an asset's quadratic variation over any finite time interval is deterministic. In the standard case that the underlying asset follows a diffusion process with nonstochastic coefficients, realized variance is deterministic and equal to the integral over time of the squared value of the diffusion coefficients.[18] The presence of stochastic volatility, then, represents a substantial departure from the Black-Scholes-Merton world. We find that natural gas exhibits stochastic volatility and illustrate this fact by calculating rolling 30-day standard deviations. As is evident in figure 1.4, there are significant changes in the estimated volatility through time as well as differences between estimates constructed from contracts of different maturities. Consequently, models which fail to capture gas's time-varying volatility will almost certainly produce incorrect estimates of option prices.

A third and highly distinctive yet poorly modeled feature of natural gas is the

[17]The Kolmogorov-Smirnov test does not work well here because it requires the sample to have no ties in order to generate an exact distribution. We do not meet this requirement and hence must rely on a potentially very inaccurate approximation. Hence we do not report that test statistic.

[18]See Shreve (2004, p. 107) and Rebonato (2004, pp. 97–98) for a more detailed discussion.

Figure 1.3: Representative Q-Q Plots

The first column includes Q-Q plots comparing the sample quantiles of daily, weekly, and monthly returns of the 1-month futures contract against the quantiles of the normal distribution whose mean and variance match the sample mean and variance of the 1 month contracts' returns. The second column offers the same analysis for the 12 month contract.

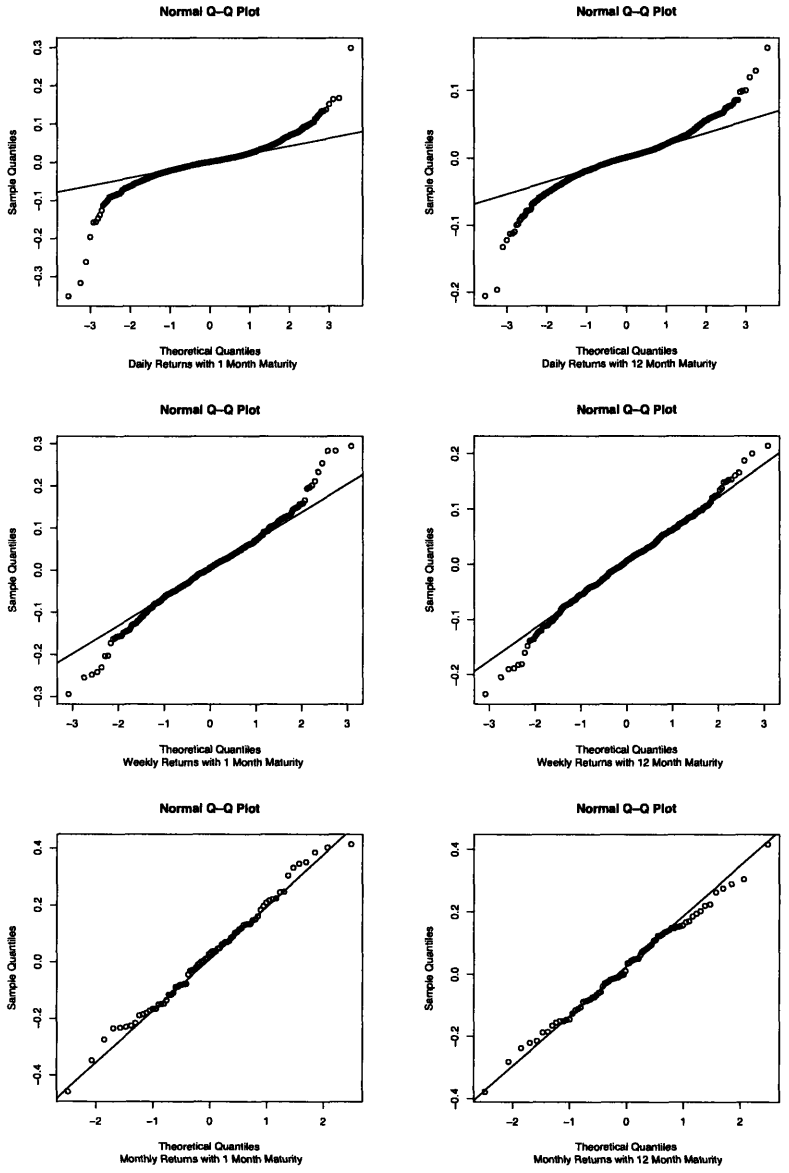


Figure 1.4: Natural Gas Price Volatility

The top panel shows a rolling estimate of the sample standard deviation of the daily price of a 1 month futures contract. The bottom panel shows the analogue for a 12 month futures contract.

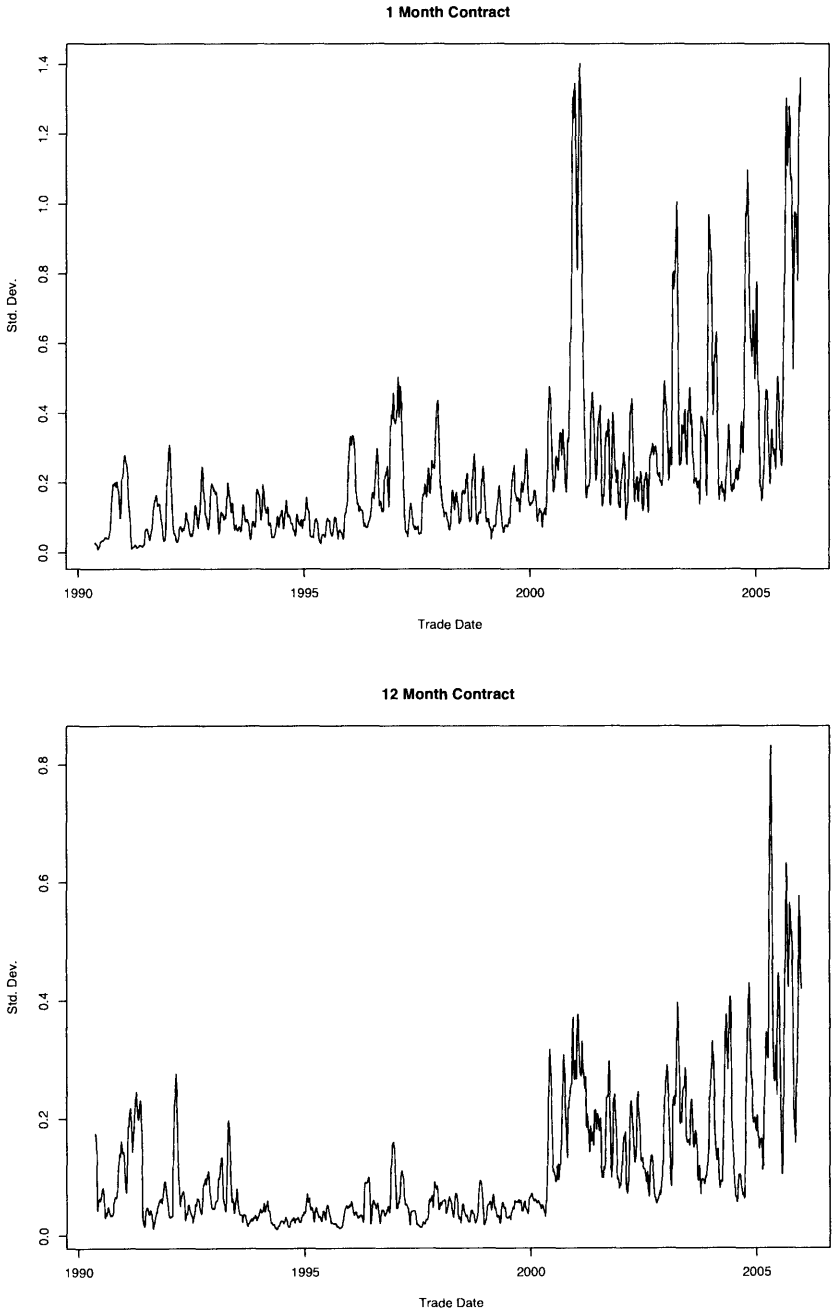


Table 1.5: Average Monthly Price Deviations

The month column represents the period in which a given contract matures. The deviation represents the average deviation over the entire sample where individual deviation is defined as the difference between the price of a contract expiring in a given month on a particular day and the average price for all contracts on that day as a percent of that average price. We aggregate over days to produce the average deviation. These deviations are denominated in dollars.

Month	Deviation	Month	Deviation
January	0.136	July	−0.055
February	0.074	August	−0.050
March	0.006	September	−0.049
April	−0.056	October	−0.029
May	−0.063	November	0.037
June	−0.060	December	0.108

seasonal fluctuation in prices. Figure 1.1 provides visual evidence for the presence of seasonality. Here, we provide further proof by examining the extent to which contracts' prices exceed the average price of all the contracts traded on a given day. More formally, we compute daily price averages and determine the monthly deviation (in percentage terms) from that average. Aggregating over the entire data set and controlling for daily price fluctuations, we identify the seasonal component in natural gas pricing by observing which months, on average, have the highest prices. These average monthly deviations are reported in table 1.5. We find that contracts which expire in December and January are the most costly indicating a strong empirical regularity in the data that must be modeled. We make heavy use of this fact in later sections of the paper when we suggest an alternative procedure for pricing natural gas options.

1.4.2 Evidence from Options

While we have shown that much about gas's representative stochastic process can be learned from direct study of spot and futures' prices, an examination of options can shed further light on the nature of the underlying process. Specifically, it is difficult to identify the presence of jumps if one only observes discretely sampled paths of the underlying asset's price. Unless the sampling frequency is extremely high, wherein market microstructure effects would almost certainly have the unintended consequence of obscuring the result, jump and continuous processes are essentially indistinguishable. Carr and Wu (2003b) addressed a similar problem in equity markets and developed a technique for differentiating between a purely continuous process (PC), pure jump process (PJ), and a combination of the two, or continuous jump process (CJ). Since the presence of jumps in the underlying's process can have substantial

Table 1.6: Asymptotic Behavior of Short-Maturity Options

Process Type	OTM Options	ATM Options
PC	$O(e^{-c/\tau}), c > 0$	$O(\sqrt{\tau})$
PJ	$O(\tau)$	$O(\tau^p), p \in (0,1]$
CJ	$O(\tau)$	$O(\tau^p), p \in (0, \frac{1}{2}]$

impact on the value of derivatives, it is important to identify their existence in order to develop an accurate pricing model. In this section, we employ Carr and Wu's methodology to demonstrate that in fact gas shows evidence of switching between PC and PJ/CJ regimes.

While the interested reader is directed to the original paper for technical details, the basic idea underlying the Carr and Wu's test is simple: short-dated option prices are highly dependent on the presence of jumps. As an example, out-of-the-money (OTM) options with near-term maturities have little chance of recovering any value if the asset on which they are written follows a purely continuous stochastic process. However, if the process admits jumps, the OTM option may retain considerable value depending on the magnitude and frequency of those jumps. Carr and Wu extend this intuition and show via analytical derivation and simulation, the behavior of at-the-money (ATM) and OTM options as the time to maturity approaches zero. There results are summarized in table 1.6 where the $O(\cdot)$ follows the standard Landau notation regarding asymptotic speed.[19] Carr and Wu find that the asymptotic behavior is always exhibited by options maturing within 20 days.

The analysis is nicely captured in term decay graphs which plot the log of the ratio of option prices to maturity, $\frac{C}{\tau}$, against log maturity, $\log\tau$. As the contract approaches expiration, ATM options evidence zero slope in the presence of a finite variation PJ model and a negative slope in the PC or CJ cases where jumps are of infinite variation. In contrast, OTM options are characterized by zero slope in the PJ case and positive slope in the PC case. In order to estimate the slope coefficient, we fit a second-order polynomial

$$\ln\left(\frac{C}{\tau}\right) = a(\ln\tau)^2 + b(\ln\tau) + c$$

to the plots where C is the call price. Consequently, the slope of the graph at a given $\ln\tau$ is given by $2a\ln(\tau) + b$.

Empirical Results

Following the procedure laid out by Carr and Wu (2003b), we estimate the term decay graphs at four log moneyness levels: $k = \ln(K/F) = 0\%, 3\%, 6\%, 9\%$. However, before we produce plots and accompanying polynomial fits, we first filter the data in

[19] $f = O(g)$ should be interpreted as $\limsup_{x \longrightarrow \infty} \frac{f}{g} < \infty$.

Table 1.7: Daily Breakdown of Process Types

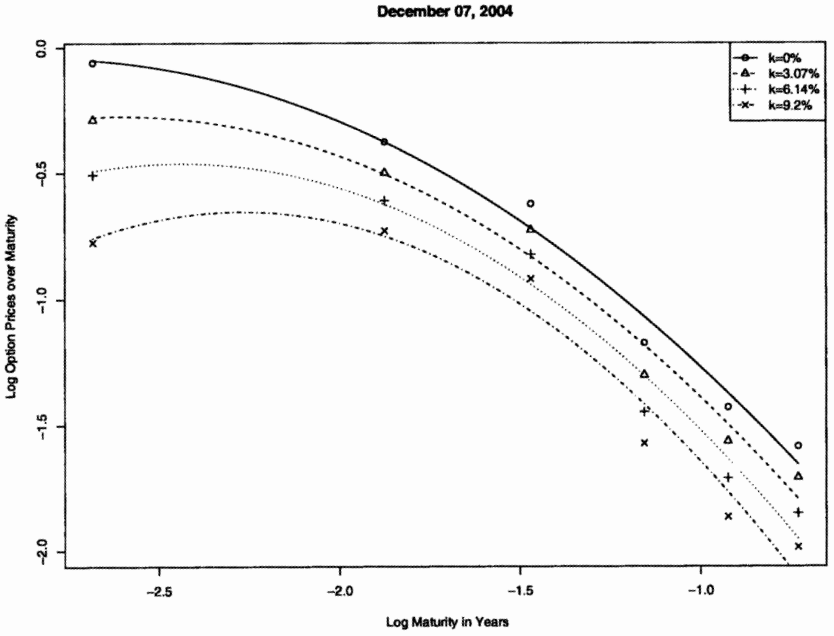
PC, PJ, and CJ refer to Purely Continuous, Pure Jump, and Continuous Jump processes respectively. Figure entries reflect the percentage of days within the associated time period that evidence a given process.

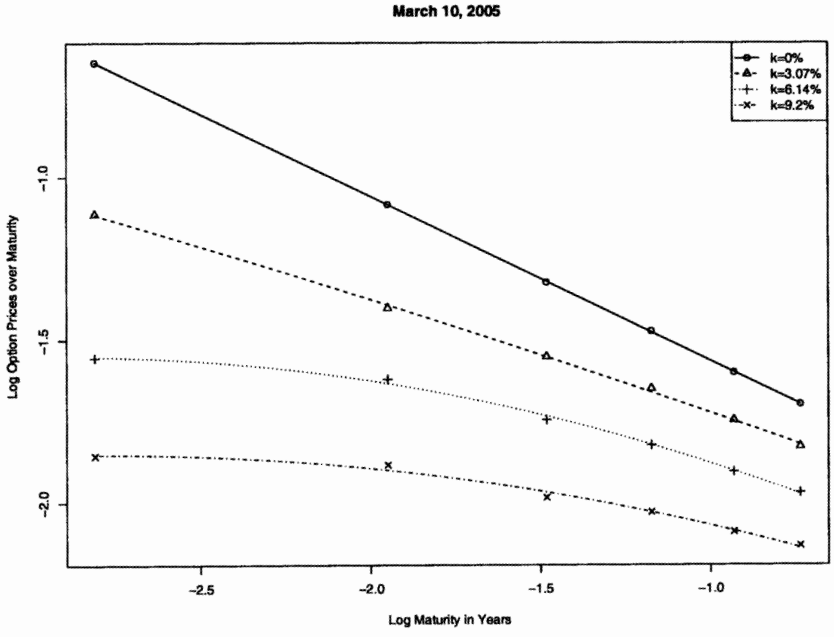
Process	**October-December**	**January-September**
PC	89%	0%
PJ	0%	47%
PJ/CJ	0%	41%
Indeterminate	11%	12%
Total	100%	100%

several ways. In addition to requiring that options contracts have sufficient volume for inclusion, we ensure that there are enough contracts with different strikes for each maturity that interpolation of call prices is possible. Finally, we guarantee that we can fit meaningful regressions to this interpolated options data set by dropping all days in which there are an inadequate number of contracts at the same moneyness level. We display some sample plots in figure 1.5. The top panel suggests that options on December 7, 2004 followed a PC process given the downward sloping OTM plots and flat ATM plot as contracts approached maturity. In contrast, the bottom panel provides evidence that options on March 10, 2005 have a jump component. Given that the plots of the OTM options are not downward sloping, we can largely rule out the possibility of the presence of a PC process on that day.

Table 1.7 summarizes the key result of this section: the stochastic process for natural gas shows evidence of the intermittent presence of jumps. Specifically, the table shows that 89% of the days between October 1 and December 31 in our scrubbed data set follow a PC process while the remaining 11% of the days provide indeterminate evidence for the process type. Conversely, between January 1 and September 30, 88% of the days show evidence of jumps while the remaining 12% of days offer no information. On about half of the days on which we can make inferences, there is evidence of a diffusion component as well. Perhaps the most striking result is that in the October-December period there is not a single day in which the potential for jumps are present while during the January-September period there are zero days on which options prices suggest gas follows a purely continuous process. Finally, one can observe that the periodicity in the presence of jumps overlaps with the seasonality of prices discussed in Section 1.4.1. There we found that prices seem to follow a seasonal pattern wherein spot prices rise in the months of December and January. Carr and Wu concluded from their study of S&P index options that equities too appear to fluctuate between regimes with different combinations of jump and continuous components.

Figure 1.5: Sample Term Decay Plots





1.5 Analysis Under the Risk-Neutral Measure: Risk Aversion

The nature of investors' attitudes toward risk plays a subtle yet important role in the derivation of option pricing models. In the standard Black-Scholes-Merton framework, for example, we derive the price of an option using Girsanov's Theorem. The change in measure, of course, only affects the drift term in the stochastic differential equation governing the evolution of the underlying asset; the diffusion component is unaffected and equal to its analogue under the objective measure. However, as one relaxes the Black-Scholes-Merton assumptions and allows for jumps or stochastic volatility, the measure transformation becomes more complicated and intrusive. If, for example, volatility is thought to be stochastic and follow a mean-reverting process, the measure transformation will affect not only the asset's drift, but the mean-reversion speed and level of the volatility term as well. Likewise, the risk-neutral description of an asset which includes a jump component differs from its real-world counterpart in its jump frequency and jump amplitude. In fact, Rebonato (2004) shows how even in the absence of jumps and stochastic volatility, the introduction of realistic conditions governing supply and demand imbalance can sever the equality between the deterministic volatility term under the risk-neutral and objective measures. To the extent one believes these departures from the Black-Scholes-Merton world are significant enough to impact market prices, one cannot afford to ignore investor preferences when constructing models for option prices.

The precise link between investors' tolerance for uncertainty and the risk-neutral parameters in option models is complex and model-dependent. Nonetheless, estimating risk aversion can play a very important role in helping to evaluate the validity of a derivative pricing model. For example, in the case of a jump-diffusion model, Lewis (2002) derives a closed-form expression linking the risk aversion of a power-utility investor with the real-world and risk-neutral jump frequencies and amplitudes. The paper shows that risk-averse investors perceive negative jumps with greater frequency and amplitude under the risk-neutral measure than under the objective measure. Intuitively, this makes sense as one would expect risk-averse investors to be compensated for bearing greater risk. Rebonato (2004) recommends that practitioners make use of this result in several ways. Starting with some prior on risk aversion, a derivatives trader who finds risk-adjusted jump frequencies of five times per month with downward jumps of 80% should question the soundness of his model unless he imagines the representative investor to be extremely risk-averse. In the event the trader cannot estimate risk aversion with great precision, he can still make use of Lewis's observation by estimating the real world values of jump frequency and amplitude and then choosing bounds on the value of the risk aversion coefficient. Next, he can simply evaluate the corresponding risk-neutral parameters implied by these previous calculations and compare them to those implied by his model. To the extent that these two sets of estimated parameters differ, he might again call into

question his model specification or his parameter estimation procedure.

A careful consideration of risk aversion can also be of help in evaluating stochastic volatility models. However, in contrast with the jump-diffusion case, the implications are more model-specific. Lewis (2000) shows that in models with square-root or GARCH volatility combined with power utility, options prices vary with levels of risk aversion depending on the sign of the correlation between the asset price and volatility. With zero correlation, risk aversion varies inversely with option value but with positive correlation, increases in risk aversion correspond to higher price levels.

The important point is that complex models with many parameters can be difficult to estimate. Thus parameter restrictions informed by an understanding of risk aversion can greatly improve overall model calibration and hopefully the model's ability to recover out-of-sample prices. Given the importance of understanding investor attitudes towards uncertainty in pricing derivatives, we modify approaches taken in Jackwerth (2000) and Aït-Sahalia and Lo (2000) to estimate risk aversion in gas markets.

1.5.1 A Simple Model

Following Constantinides (1982) and Merton (1992), we consider a complete market economy with heterogeneous agents and note that the competitive equilibrium is equal to that arising from a representative investor with utility function $U(\cdot)$. The agent is endowed with one unit of wealth at time t and faces a fixed time horizon T. In equilibrium, the agent holds all of his wealth in gas. His problem, as posed in Jackwerth (2000), is:

$$\max_{W_T} \int U(W_T)P(W_T)dW_T - \lambda\left(\frac{1}{r^{(T-t)}}\int W_T Q(W_T)dW_T - 1\right)$$

where W_T is wealth at time T, λ is the shadow price of the budget constraint, r is the gross interest rate, $Q(\cdot)$ is the risk-neutral probability distribution, $P(\cdot)$ is the objective probability distribution, and S_T is the spot price of gas at time T.

The well-known equilibrium result arising from this simplified version of Merton's optimization problem is

$$U'(S_T) = \frac{\lambda Q(S_T)}{r^{(T-t)}P(S_T)} \tag{1.1}$$

where the time index has been dropped for notational convenience. If we then differentiate equation 1.1 a second time and solve for the coefficient of relative risk aversion, ρ, we find

$$\rho = -\frac{S_T U''(S_T)}{U'(S_T)} = -\frac{\frac{S_T\lambda}{r^{(T-t)}}\left(\frac{(Q'(S_T)P(S_T)-Q(S_T)P'(S_T)}{P^2(S_T)}\right)}{\frac{\lambda Q(S_T)}{r^{(T-t)}P(S_T)}} = \frac{S_T P'(S_T)}{P(S_T)} - \frac{S_T Q'(S_T)}{Q(S_T)}$$

Estimating Risk

We estimate ρ in a two step process. First, we find $Q(\cdot)$ by making use of the observation in Breeden and Litzenberger (1978) that the risk-neutral distribution is equal to the discounted value of the second derivative of a European call option taken with respect to the strike and evaluated at the spot price of the underlying asset.[20] [21] This somewhat surprising result is quite easily understood.[22] Recall that options can be priced under the equivalent martingale measure, $Q(\cdot)$, as

$$C_{T,K}(t,S) = e^{-r(T-t)}\mathbb{E}^Q\{f_K(S_T)|S_t = S\} = e^{-r(T-t)}\int f_K(s)Q(s)ds$$

where S is the spot price, K is the strike price, f_K is $\max(x-K,0)$, t is the initial time, and T is the time of expiry.

Next, note that

$$\frac{\partial f_K(x)}{\partial K} = \left\{ \begin{array}{c} -1 \text{ if } K < x \\ 0 \text{ if } K > x \end{array} \right\}$$

This in turn implies that

$$\frac{\partial^2 f_K(x)}{\partial K^2} = \delta_x(K)$$

where δ_x denotes the Dirac delta function over x. If we then permit the interchange of derivatives and integration, we get our result:

$$\begin{aligned} \frac{\partial^2 C_{T,K}(t,S)}{\partial K^2} &= e^{-r(T-t)}\frac{\partial^2}{\partial K^2}\int f_K(x)Q(x)dx \\ &= e^{-r(T-t)}\int \frac{\partial^2}{\partial K^2} f_K(x)Q(s)dx \\ &= e^{-r(T-t)}\int \delta_x(K)Q(x)dx \\ &= e^{-r(T-t)}Q(K) \end{aligned}$$

where the last equality follows from the definition of the delta function.

$$Q(S_T) = e^{r(T-t)}\left(\frac{\partial^2 C_{T,K}(t,S)}{\partial K^2}\right)\Bigg|_{K=S_T}$$

[20] Although we are dealing with options written on futures in this context, as NYMEX option contracts and their underlying futures expire at the same time, the result carries through unaffected.

[21] Note that there is some inconsistency in the literature as to what is meant by the term "state price density". While Duffie (2001) and Shreve (2004) equate the SPD to the ratio of the risk-neutral density and the objective density multiplied by a risk-free discount factor, other sources such as Aït-Sahalia and Lo (1998) use SPD to mean the risk-neutral density itself. We adhere to the latter convention in this paper.

[22] This nice derivation follows that of Carmona (2004, p. 221).

Next we estimate $P(\cdot)$, the objective distribution of prices, using a kernel density estimator in the spirit of Aït-Sahalia and Lo (2000).[23] [24]

1.5.2 Results and Implications

The estimated risk aversion functions are shown in figure 1.6 along with the 95-percentile confidence intervals for these estimates using the procedures outlined in Aït-Sahalia and Lo (2000).[25] While the optimal bandwidth procedure plays some role in determining both the shape and levels of the plots, two features are prominent and robust. First, the coefficients are small and second, they are meaningfully different than what studies have generally found to be the case in equities markets. We estimate the average value of the relative risk aversion coefficient to be 0.02. The implication is that investors in gas markets are virtually risk-neutral. In contrast, while Hansen and Singleton (1982) and Hansen and Singleton (1984) find that relative risk aversion in equity market varies between -1 and 1, other more recent papers such as Mehra and Prescott (1985), Ferson and Constantinides (1991), Aït-Sahalia and Lo (2000) find evidence for substantially higher risk aversion levels. Mehra and Prescott (1985), for example, cites work done by Fisher Black indicating that the risk aversion coefficient is around 55. Aït-Sahalia and Lo (2000) finds that relative risk aversion in equities is on average about 13. In addition, the authors' nonparametric procedure suggests that the value of the coefficient varies over wealth; risk aversion appears to be as high as 60 at low wealth levels and close to five at average wealth levels. This too suggests a difference between equities and gas as our analysis indicates that risk aversion is essentially constant over wealth and much closer to zero.

Our results seem to offer further evidence for market segmentation. While basic finance theory suggests that there is but one representative investor with a single risk profile, comparisons between the risk aversion levels embedded in gas and equities indicates otherwise. One possible explanation for the low level of risk aversion in gas relates to the inherent market structure. While equities markets boast substantial

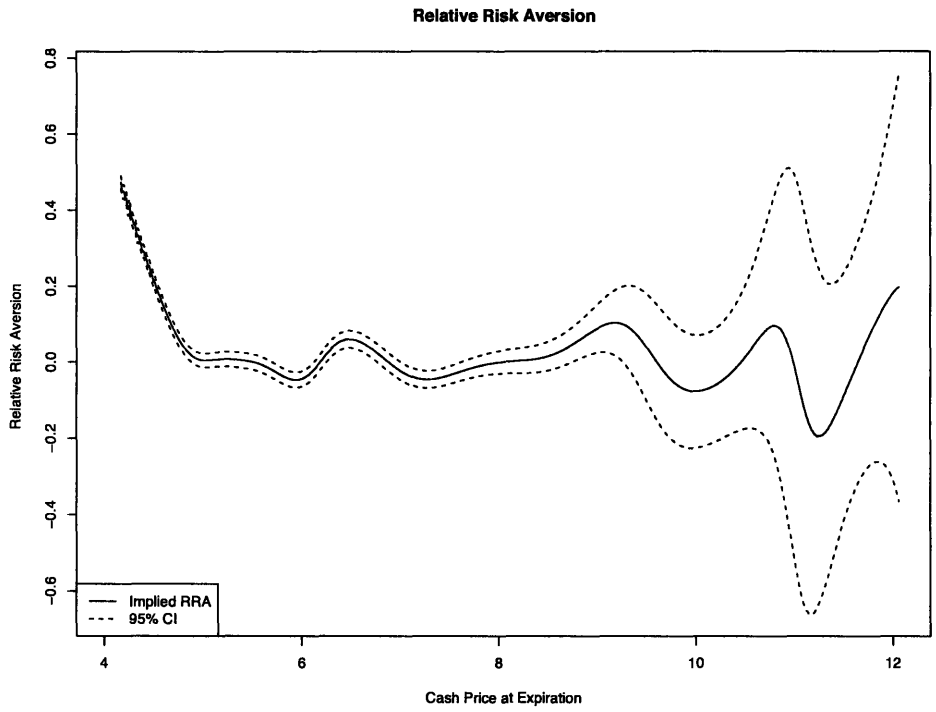
[23] The procedure involves using a kernel density estimator on the time-series of τ-period returns. This density estimator then can easily be transformed into the conditional density of prices. We use this method to estimate the objective probability distribution $P(\cdot)$. For more details, see Section 4 of Aït-Sahalia and Lo (2000).

[24] For the entire nonparametric analysis, we follow Aït-Sahalia and Lo (1998) and choose bandwidths according to the rule they develop which gives the proper rate of convergence of the estimator allowing asymptotic analysis to hold; we choose bandwidth h_x for the estimation of $\pi(x)$ such that $h_x = c_x s(x) n^{-1/(d+2(q+m))}$ where $c_x \equiv \gamma_x/\log(n)$ (γ_x constant), $s(x)$ is the standard deviation of x, n is the number of observations, d is the number of regressors, q is the order of the kernel, and m is the number of derivatives we are estimating.

[25] Using the results of Aït-Sahalia and Lo (2000), $n^{1/2} h_{X/F}^{7/2} h_\tau^{1/2} h_S^{1/2} (\hat{\rho}_t(F_T) - \rho_t(F_T)) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma_\rho^2)$ where $\sigma_\rho^2 \equiv \sigma_{f^{*\prime}}^2 F_T^2/(f^*)^2(F_T)$, $\sigma_{f^{*\prime}}^2 \equiv \left(\frac{\partial C(\hat{\sigma}(\tilde{\mathbf{Y}}),\mathbf{Y})}{\partial \sigma}\right)^2 \sigma_{d^3\sigma^*}^2$, $\sigma_{d^3\sigma^*}^2 \equiv \frac{s^2(\tilde{\mathbf{Y}})(\int_{-\infty}^{\infty}(k_{X/F}^{(3)})^2(\omega)d\omega)(\int_{-\infty}^{\infty}k_\tau^2(\omega)d\omega)(\int_{-\infty}^{\infty}k_S^2(\omega)d\omega)}{\pi(\tilde{\mathbf{Y}})F_t^6}$. Also see Section 1.7 for a more detailed discussion of the kernel estimation approach we employ in order to estimate the underlying densities needed to construct our risk aversion estimates.

Figure 1.6: Plot of Risk Aversion

The figure plots the relative risk aversion estimated via a kernel regression. The solid line represents the estimated coefficient while the dashed line represents the 95 percent confidence interval.



retail investor participation, NYMEX is almost exclusively the domain of institutional investors. A single futures contract often costs tens of thousands of dollars so it comes as no surprise that day-traders and average consumers steer clear of this asset class. One might further conjecture that institutions enjoy greater diversification opportunities, better access to information, and less susceptibility to behavioral biases than that of retail investors. Consequently, they transact in markets with markedly less risk aversion.

1.6 Existing Models

In this section, we examine three popular and representative models from the commodities literature. While the models price futures contracts with reasonable success (see table 1.8), we show that their failure to incorporate the various features highlighted in the previous sections leads to dramatic pricing errors in options markets.[26]

[26]To actually estimate the futures prices in table 1.8, we use the Kalman filtering algorithm for the Schwartz and Smith and Todorova models and the volatility functions for the Clewlow and Strickland

Table 1.8: Errors of Parametric Models in Pricing Futures Contracts
The table reports the root mean squared errors (RMSE) and the RMSE as a percent of the average price (Percent) for the 1 month, 4 month, and 8 month futures contracts using data from 2003–2005 future prices. The RMSEs are denominated in dollars.

Model		1 Month	4 Month	8 Month
Schwartz and Smith	RMSE	3.735	3.144	2.554
	Percent	47.768	38.388	32.900
Todorova	RMSE	3.791	3.169	2.588
	Percent	48.484	38.693	33.338
Clewlow and Strickland	RMSE	0.320	0.250	0.150
	Percent	4.088	3.054	1.936

The first two options models are derived from parametric assumptions of the underlying spot price while the third yields an option price based on modeling the evolution of the forward curve through time.

1.6.1 Modeling the Spot

The first model we consider is developed in Schwartz and Smith (2000) wherein the spot price, S_t, is written as a function of two stochastic factors: an equilibrium price, ξ_t, and χ_t, a short-term deviation around that level. In logarithmic form, the relationship is linear and formulated as $\ln S_t = \chi_t + \xi_t$. Changes in ξ_t, which includes a drift component, reflect long-term shifts in supply and demand as well as the effects of inflation, regulatory developments, and the inevitable improvements in finding, extraction, and distribution technologies. Consequently, the associated SDE of equation 1.2 is that of a standard diffusion process. The short-term deviations around the equilibrium level arise from temporary supply shocks, themselves the result of inclement weather or other hiccups in production or distribution. It is natural, then, that Schwartz and Smith let χ_t follow a mean-reverting Ornstein-Uhlenbeck process as reflected in equation 1.3 where κ is the mean-reversion coefficient. The authors further assume that shocks to ξ_t and χ_t are correlated increments of Brownian Motion which yield equation 1.4 where $\rho_{\chi\xi}$ is the correlation coefficient.

model. The Kalman filter approach involves first estimating the parameters of the process using the entire data set and then computing one day ahead filtered prediction of the futures prices. For the Clewlow and Strickland model, we note that equation 1.13 gives a relationship between today and tomorrow's prices. Using this relationship we can calculate tomorrow's expected price given today's price. For the volatility functions, we again allow mild stationary and reestimate them on a rolling 30 day basis.

$$d\xi_t = \mu_\xi dt + \sigma_\xi dz_\xi \tag{1.2}$$
$$d\chi_t = -\kappa\chi_t dt + \sigma_\chi dz_\chi \tag{1.3}$$
$$d\xi_t d\chi_t = \rho_{\chi\xi} dt \tag{1.4}$$

Schwartz and Smith show that this set-up implies that χ_t and ξ_t have a jointly normal distribution while S_t has a lognormal distribution. In order to derive derivative prices, the authors also evaluate the dynamics of the the two factors under the risk-neutral measure. They reason that the correlation between changes in the state variables and aggregate econòmic wealth is zero and thus that risk adjustment results in simple corrections to the drift terms as reflected in equations 1.5 and 1.6.

$$d\xi_t = (\mu_\xi - \lambda_\xi)dt + \sigma_\xi dz^*_\xi \tag{1.5}$$
$$d\chi_t = (-\kappa\chi_t - \lambda_\chi)dt + \sigma_\chi dz^*_\chi \tag{1.6}$$

Here, the * indicates that the corresponding variable is evaluated under the risk neutral measure while λ_ξ and λ_χ/κ represent the "market price of risk" associated with ξ_t and χ_t respectively. This model together with the assumption that interest rates are independent of spot gas prices allows for the simple calculation of futures prices as the expectation taken with respect to the risk-neutral measure of the future spot price. This leads to equation 1.7,

$$\ln(F_{T,t}) = e^{-\kappa(T-t)}\chi_t + \xi_t + A(T-t) \tag{1.7}$$
$$A(T-t) = (\mu_\xi - \lambda_\xi)(T-t) - (1-e^{-\kappa(T-t)})\frac{\lambda_\chi}{\kappa} + \frac{1}{2}\left((1-e^{-2\kappa(T-t)})\frac{\sigma_\chi^2}{2\kappa} + \sigma_\xi^2(T-t) + 2(1-e^{-\kappa(T-t)})\frac{\rho_{\chi\xi}\sigma_\chi\sigma_\xi}{\kappa}\right)$$

where $F_{T,t}$ is the time t price of a futures contract expiring at time T. Under the risk-neutral model, futures prices are still lognormal with variance σ_ϕ but mean μ_ϕ where

$$\mu_\phi(t,T) \equiv \mathbb{E}^*[\ln(F_{T,t})] = e^{-\kappa(T-t)}\chi_t + \xi_t + (\mu_\xi - \lambda_\xi)(T-t) + (1-e^{-\kappa(T-t)})$$
$$\sigma_\phi(t,T) \equiv \mathrm{Var}^*[\ln(F_{T,t})] = (1-e^{-2\kappa(T-t)})\frac{\sigma_\chi^2}{2\kappa} + \sigma_\xi^2(T-t) + 2(1-e^{-\kappa(T-t)})\frac{\rho_{\chi\xi}\sigma_\chi\sigma_\xi}{\kappa} \tag{1.8}$$

and χ_0 and ξ_0 are initial value of χ_t and ξ_t respectively.

Since the state variables, χ_t and ξ_t, are unobservable, Schwartz and Smith estimate the parameters of the model using MLE where the likelihood function is computed via a Kalman filter. The details of this procedure, which involves first

recasting the model in a discrete state space framework, are outlined in Appendix 1.A.

Finally, the authors apply basic Black-Scholes-Merton methodology to derive a closed form expression for the value of a call option on a futures. Explicitly, the price of a call option with strike K expiring at time t on a futures contract expiring at time T is

$$e^{-r(T-t)}(F_{T,t}\mathcal{N}(d) - K\mathcal{N}(d - \sigma_\phi(t,T))) \tag{1.9}$$

where $d = \frac{\log(F_{T,t}/K)}{\sigma_\phi(t,T)} + \frac{1}{2}\sigma_\phi(t,T)$ and $\mathcal{N}(\cdot)$ is the standard normal cumulative distribution function. We use this pricing formula in conjunction with our MLE estimates of κ, σ_χ, σ_ξ, and $\rho_{\chi\xi}$ to estimate the prices of all of the traded options in our data set. We report our findings in Section 1.6.3.

It is important to note the ways in which this model departs from the realities of the actual natural gas markets given what we documented in Sections 1.4 and 1.5. First, the model asserts that the evolution of both spot and futures prices are lognormal at all horizons. This clearly is at odds with our observation that returns are only Gaussian when calculated on a monthly basis but evidence non-normality over shorter intervals. Second, Schwartz and Smith do not allow for seasonality of returns which is another important feature of the data. Third, their paper imposes constant volatility despite substantial evidence to the contrary, and fourth, makes no provision for the possibility of jumps in the price level. Finally, as the result of these overly, albeit simplified, assumptions, Schwartz and Smith are able to show that the model's formulation under the risk-neutral measure amounts to trivial adjustments to the drift terms of the two state variables. In light of the discussions in Sections 1.4.1, 1.4.2, and 1.5, we can conclude that the true risk-neutral dynamics are likely to be more complex.

Arguably the easily rectified potential shortcoming of Schwartz and Smith's paper is its failure to incorporate the presence of seasonality in the price of gas. To determine whether or not the explicit consideration of seasonality improves the model's ability to recover market prices, we rely on a model introduced in Todorova (2004b). Todorova essentially picks up where Schwartz and Smith leave off and treats seasonality in several different ways. One method she implements with mixed success involves adding a third factor, seasonality, to the Schwartz and Smith framework. While this leads to a nice closed form option pricing formula, the model struggles to recover futures prices due to the substantial number of additional parameters that require estimation.[27] A second approach is to estimate the Schwartz and Smith (2000) model having first deseasonalized the data. To preprocess the data in this fashion, Todorova proposes a procedure outlined by Kendall and Ord (1990) which involves detrending the data and looking at price deviations in order to identify seasonal components. The

[27] While Todorova did not explicitly do so, we derived an expression for the value of a European option based on her three factor model. The expression is straightforward but the addition of a third factor introduces significantly more parameters to estimate. Consequently, the estimation procedure yields extremely unstable results with our data set. Therefore we do not report any results using this method.

process then reintroduces any time trends resulting in a data set with the seasonality removed. To avoid the curse of dimensionality inherent in the first approach, we utilize the deseasonalization strategy in this paper.

To implement Todorova's model, we begin by deseasonalizing the data as previously described. This produces a futures price series that has the seasonal component removed. We then proceed to estimate the model in the framework of Schwartz and Smith (2000) since we assume the deseasonalized data follows the assumptions regarding the dynamics of the process made in that paper. This estimation procedure produces estimates of the parameters of the stochastic process that underlie the deseasonalized data rather than the true stochastic process. However our method of deseasonalization only modifies the underlying process by changing the first moment via adding a nonstochastic constant, therefore neither the distribution nor the variance of the process changes and the option pricing formula derived in equation 1.9 remains valid with the parameters estimated from the deseasonalized price series.

Though Todorova's approach certainly attempts to address the issue of seasonality, it is open to all but one of the same critiques as Schwartz and Smith (2000). Namely, that it fails to incorporate non-normality and stochastic volatility in returns as well as the presence of discontinuities in prices. As a consequence, its formulation for the dynamics under the risk-neutral measure are almost certainly overly simplified.

1.6.2 Forward curve model

In contrast to Schwartz and Smith (2000) and Todorova (2004b) which rely on structural specifications of the underlying spot process, Clewlow and Strickland (2000) derives option prices by modeling the entire forward curve. First, the authors posit a model of the forward curve wherein each contract is a linear combination of n independent sources of uncertainty. Formally,

$$\frac{dF(t,T)}{F(t,T)} = \sum_{t=1}^{n} \sigma_i(t,T) dz_i(t) \tag{1.10}$$

where $F(t,T)$ represents the time t price of a futures contract maturing at time T while $\sigma_i(t,T)$ and $dz_i(t)$ equal that contract's ith volatility function and ith source of risk respectively. Next, the authors extend the model to markets with substantial seasonality in prices by modifying equation 1.10 to incorporate a time dependent spot volatility function. Now,

$$\frac{dF(t,T)}{F(t,T)} = \sigma_S(t) \sum_{t=1}^{n} \sigma_i(T-t) dz_i(t) \tag{1.11}$$

where $\sigma_S(t)$ captures seasonality.[28]

[28]Notice in moving from equation 1.10 to equation 1.11 we have replaced $\sigma_i(t,T)$ with $\sigma_i(T-t)$ which is essentially an assumption on the stationarity of the process; we are assuming that volatility only

We follow the procedure outlined in Appendix 1.C to estimate both the time dependent and individual factor volatility functions. In short, we use the rolling 30 day sample standard deviation to find $\sigma_S(t)$. Estimating the individual volatility functions is more complex and relies on converting the stochastic process in equation 1.11 to a logarithmic form and then discretizing it. This allows us to utilize Principal Component Analysis by constructing a covariance matrix of forward returns. Next, we compute an eigenvector decomposition of this matrix scaled by our estimated spot volatility such that we can recover independent factors that drive the forward curve. A simple transformation of these factors and their associated eigenvalues gives us the discretized volatility functions.[29] One can also use the eigenvalues of the decomposition to choose the number of volatility functions necessary to model the forward curve with desired accuracy. In this paper we use five volatility functions which seem to capture most of the variation in the covariance matrix of returns.

Once we have estimated the volatility functions, it is straightforward to price options assuming interest rates are nonstochastic. The authors derive a closed-form formula for the price of a European call option at time t,

$$c(t,F(t,T),K,T) = e^{-r(T-t)}\left(F(t,T)\mathcal{N}(h) - K\mathcal{N}(h-\sqrt{w})\right), \tag{1.12}$$

where K is the strike price, and both the option and futures mature at time T. Further,

$$h = \frac{\log(F(t,T)/K) + \frac{1}{2}w}{\sqrt{w}}, \quad w = \sum_{i=1}^{n}\left(\int_t^T \sigma_i(u,T)^2 du\right).$$

We calculate one day ahead option prices by first estimating the volatility functions on a 30 day rolling basis to allow for mild nonstationarity.[30] This procedure produces volatility functions for each trading day which in conjunction with equation 1.12, enables us to price any options that trade on that day.

On one level, Clewlow and Strickland (2000) seems less intellectually appealing than the fully parametric approaches taken in the first two papers considered. It is not clear, for example, how to interpret the volatility functions except to understand them as weights on opaque "sources of risk." However, the trade-off is in the model's comparative flexibility as it imposes none of the rigid structure on futures volatility

depends on the length of time until expiration rather that the specific values of t and T. This is an important assumption without which we could not use historical data to estimate the volatility functions. In our actual estimation procedure we do allow for mild non-stationarity by estimating the volatility functions using rolling data.

[29]Since this technique relies on a stochastic process without jumps, we also follow Clewlow and Strickland (2000) and apply what their paper terms a "recursive filter" to remove data points that appear to be generated by a jump. The actual implementation relies on repeated calculations of the sample standard deviation and filtering out observations that exceed an arbitrary threshold scaling of that sample standard deviation. We use 3 standard deviations, but find the results are not extremely sensitive to the choice.

[30]We apply the recursive filter on each of these returns to filter out possible jumps.

included, for example, in equation 1.8. In addition, as with Todorova's model, Clewlow and Strickland's approach explicitly incorporates seasonality, one of gas's important characteristic features. It too, however, fails to permit stochastic volatility and because changes in futures prices are modeled as linear combinations of Brownian increments, returns will necessarily be Gaussian over all horizons. Both of these features contradict the empirical evidence. Finally, by construction, the forward-curve model fails to admit the possibility of jumps even though we have seen that prices behave as if they follow a jump process during certain months of the year.

1.6.3 Empirical Results

We evaluate the models' success in reproducing actual options prices in three ways. First, we compare the models' predictions with the actual prices in terms of root mean-squared error (RMSE). Next, we examine performance by measuring slippage with respect to a delta-hedged portfolio. Finally, we offer a visual exposition of the degree of mispricing by comparing the volatility surfaces implied by the candidate models with the actual implied volatility surfaces calculated in Section 1.3.2.

Table 1.9 reports the RMSE of the different estimators. As one would expect, the models tend to price ITM options better than ATM and OTM contracts. Nonetheless, it is readily apparent that overall the three models price the options quite poorly with the average mispricing often exceeding 100% of the average option price. Although it is not directly observable from the RMSEs, some of the models exhibit quite pronounced tendencies in mispricing. For example, the Clewlow and Strickland (2000) model generally predicts option prices which are too low, while Schwartz and Smith (2000) typically over-prices the options. It is also interesting to note that the models have biases with respect to time to expiration. Table 1.10 shows that the Schwartz and Smith and Todorova models tend to price more distant options accurately while the Clewlow and Strickland approach better recovers the prices of short-term options.

One can also assess a pricing model by evaluating the relative magnitude of the tracking errors associated with a delta-hedged portfolio. Let Δ represent the derivative of an option-pricing formula with respect to the underlying security. For the models in this section, we can calculate this derivative analytically.[31] Given Δ, we can construct a portfolio of the underlying futures and a riskless bond that exactly replicates the option's payoff assuming continuous time. More formally, to construct this portfolio for each option, let $t = 0$ be the time at which the option was first traded

[31] For the nonparametric model introduced in the next section, we must resort to calculating this derivative numerically.

and let

$$V_S(0) = F(0)\Delta(0)$$
$$V_C(0) = -C(0)$$
$$V_B(0) = -(V_S(0) + V_C(0))$$

where $F(t)$ is the futures price, $C(t)$ is the call price on that future, $V_S(t)$ is the values of the futures in the portfolio, $V_B(t)$ is the value of the bonds, and $V_C(t)$ is the values of the call options.[32] By construction, at $t = 0$,

$$V(0) = V_S(0) + V_B(0) + V_C(0) = 0$$

and then we calculate $V(t)$ with

$$V_S(t) = F(t)\Delta(t)$$

and

$$V_B(t) = e^{r_d} V_B(t-1) - F(t)(\Delta(t) - \Delta(t-1)).$$

The tracking error is then defined to be $V(T)$ where T is the date of expiry of the option. Finally, we define a performance measure $\xi = e^{-rT}|V(T)|$ which is the present-value of the tracking error. In table 1.9, we average these tracking errors over different types of options to illustrate how well the various option pricing formulas perform. We find that by this measure as opposed to RMSE, the model in Clewlow and Strickland (2000) enjoys a substantially smaller performance advantage vis-à-vis the models of Schwartz and Smith (2000) and Todorova (2004b); the mean absolute tracking error of Clewlow and Strickand's approach is \$0.23 compared with that of \$0.29 for both Todorova and Clewlow and Strickland.

A final instructive approach to measuring the efficacy of the pricing models is to compare their implied volatility surfaces. As in Section 1.3.2, we use the result from Black (1976) to estimate the implied volatility surfaces.[33] We reproduce representative samples of these surfaces in figures 1.7 and 1.8. These figures clearly show that the parametric models do not seem to yield the regularities we see in the actual data. In particular, the models fail to capture the relationship between moneyness and implied volatility.

[32] We follow the notation of Hutchinson, Lo, and Poggio (1994) for this section.

[33] Just as before, we use a series regression in order to approximate the function for plotting purposes.

Table 1.9: Errors of Models

RMSE represents the root mean squared error of the estimated option price compared to the actual option price. TE represents the mean absolute tracking error of a delta-hedged portfolio. Total represents the error over all options, while ITM, ATM, and OTM represent the errors of in-the-money, at-the-money, and out-of-the-money options respectively. Both the RMSEs and the TEs are denominated in dollars.

Model		RMSE	TE
Schwartz and Smith (2000)			
	Total	2.044	0.292
	ITM	1.922	0.211
	ATM	2.209	0.203
	OTM	2.143	0.365
Todorova (2004b)			
	Total	2.090	0.293
	ITM	1.965	0.211
	ATM	2.257	0.203
	OTM	2.194	0.367
Clewlow and Strickland (2000)			
	Total	0.425	0.233
	ITM	0.367	0.174
	ATM	0.537	0.200
	OTM	0.459	0.291

Table 1.10: Errors of Models by Maturity

Table entries represent the root mean squared errors of the estimated options prices compared to the actual options prices. Maturities, denoted by τ, are in months. SS denotes Schwartz and Smith (2000), T denotes Todorova (2004b), and CS denotes Clewlow and Strickland (2000). The entries in the table are denominated in dollars.

Maturity	SS	T	CS
$\tau < 1$ Month	2.746	2.882	0.159
$1 \leq \tau < 2$	2.475	2.585	0.314
$2 \leq \tau < 3$	2.302	2.387	0.456
$3 \leq \tau < 4$	2.206	2.270	0.499
$4 \leq \tau < 5$	2.116	2.163	0.524
$5 \leq \tau < 6$	2.041	2.075	0.503
$6 \leq \tau < 7$	1.998	2.022	0.401
$7 \leq \tau < 8$	1.940	1.957	0.386
$8 \leq \tau < 9$	1.896	1.908	0.409
$9 \leq \tau < 10$	1.874	1.882	0.447

Figure 1.7: Model Implied Volatility Surfaces for May 2005

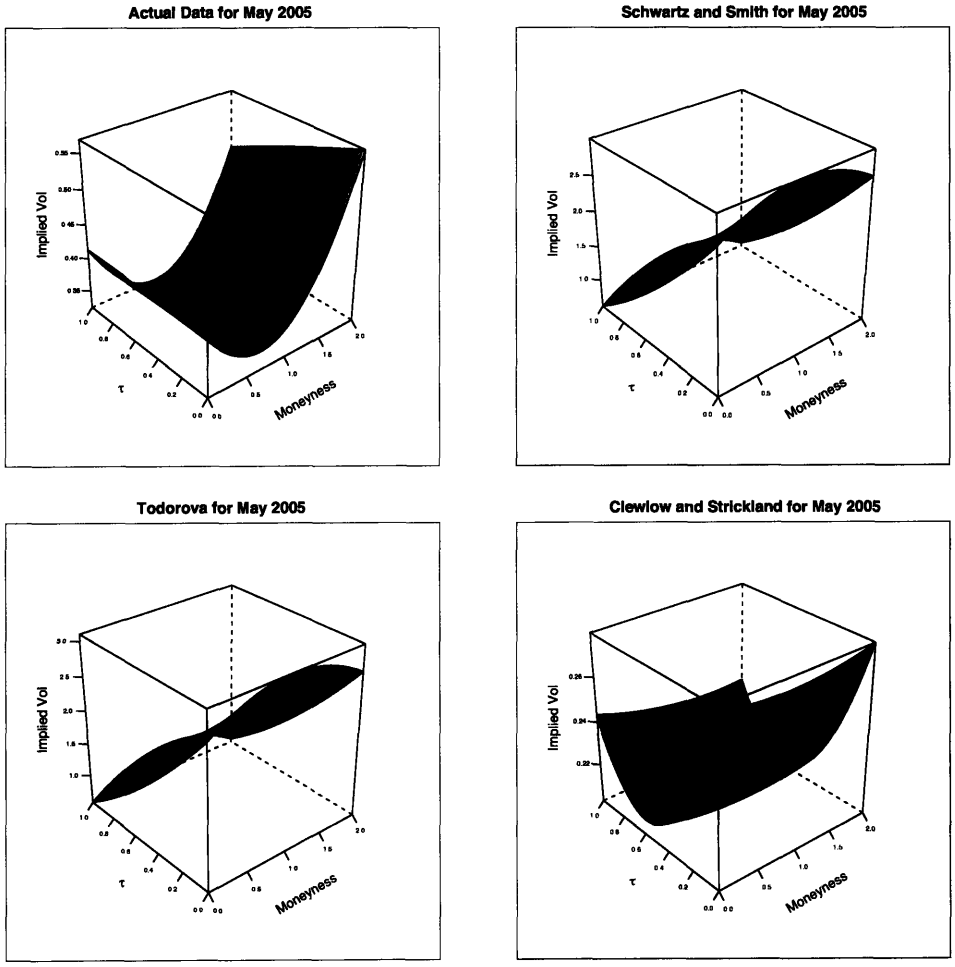
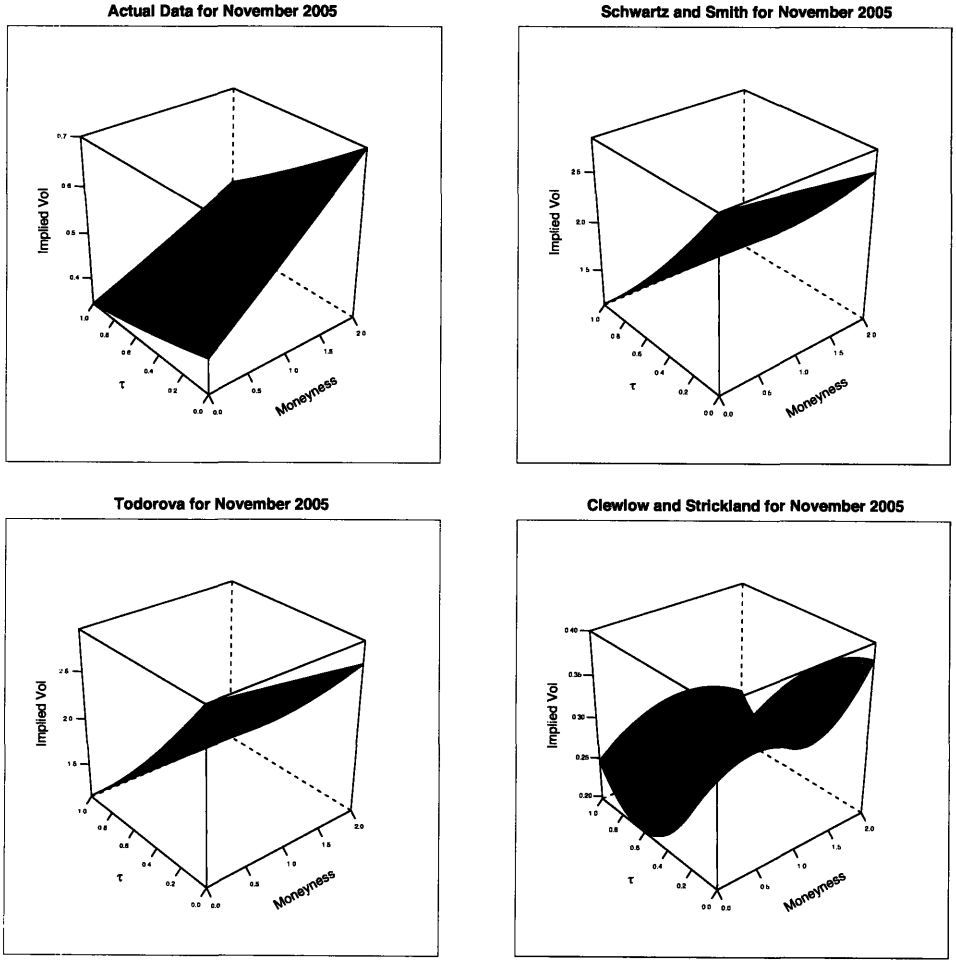


Figure 1.8: Model Implied Volatility Surfaces for November 2005



1.7 Nonparametric Approach

1.7.1 Set-Up

As demonstrated in the previous subsection, the parametric pricing models we have considered neither price options well nor capture the relationships between moneyness, time to maturity, and implied volatility exhibited by the actual data. Thus, we suggest an alternative nonparametric approach to pricing these options inspired by Aït-Sahalia and Lo (1998). The advantage of using a kernel-based regression is that we avoid structural restrictions. As a result, we can theoretically capture all of the important features of gas including the seasonality, non-normality, and time-varying volatility of returns as well as the presence of jumps.

The biggest challenge facing the application of kernel techniques is the curse of dimensionality. With just under 40,000 observations, we reduce the complexity of the problem by estimating σ, the implied volatility of Black (1976), via a kernel regression and then calculate option prices by simply inverting Black's formula. Further, we select as our state variables, the futures price, the strike price, time to maturity, and seasonality.[34] This fourth regressor is a simple measure of seasonality: the month the option expires.[35] In addition, we standardize all of the state variable by de-meaning them and dividing each by their respective standard deviations; this allows us to use a common bandwidth, b, for each element in the kernel's tensor product. As the choice of the kernel function generally does little to affect the outcome of the estimation procedure, we employ a multivariate Gaussian kernel. Formally, we use these assumptions along with the Nadaraya-Watson kernel estimator which yields:

$$\hat{\sigma}(F_{t,\tau},K_t,\tau,S_t) = \frac{\sum_{i=1}^{n} k(\frac{F_{t,\tau}-F_{t_i,\tau}}{b})k(\frac{K_t-K_i}{b})k(\frac{\tau-\tau_i}{b})k(\frac{S_t-S_i}{b})\sigma_i}{\sum_{i=1}^{n} k(\frac{F_{t,\tau}-F_{t_i,\tau}}{b})k(\frac{K_t-K_i}{b})k(\frac{\tau-\tau_i}{b})k(\frac{S_t-S_i}{b})}$$

where $F_{t,\tau}$ is the futures price as of time t of a futures contract expiring in τ periods, K_i represents the strike price, τ_i represents the time to the option's maturity, S_t represents the month that the option expires, and $k(\cdot)$ is the univariate Gaussian kernel. We can then calculate the price of a call option, $\hat{C}$, as

$$\hat{C}(F_{t,\tau},K_t,\tau,r_{t,\tau}) = C_{Black76}(F_{t,\tau},K_t,\tau,r_{t,\tau};\hat{\sigma}(F_{t,\tau},K_t,\tau,S_t))$$

[34] In contrast to traditional parametric econometric approaches, there is no omitted variables bias with kernel techniques. See Aït-Sahalia and Lo (1998, p. 507).

[35] We also tried several other combinations of state variables. For example, we combined the futures price and strike price into a single variable, moneyness, in order to reduce dimensionality. Also, rather than allowing seasonality to enter more generally as the month of option expiry, we tried incorporating seasonality as a dummy variable taking on the value of 1 in the case of December/January expiry and zero otherwise. This seasonality regime was determined based on the findings in Section 1.4. Combinations of these approaches all performed more poorly than the seasonality factor/four state variable set-up we use in the main body of the text.

where $C_{Black76}$ is the price of a call given by the model in Black (1976). As with our implementation of Clewlow and Strickland (2000), we allow for some degree of nonstationarity by recalculating $\hat{\sigma}$ every day on a 30 day trailing basis.

1.7.2 Results

As reflected in figure 1.11, this model leads to lower RMSE and tracking errors than any of the parametric approaches considered in the previous sections.[36] In fact, in terms of RMSE on OTM options, the kernel method improves upon the Clewlow and Strickland (2000) approach by 83% and outperforms the Todorova (2004b) and Schwartz and Smith (2000) models by a 96% margin. In addition, this nonparametric approach suffers no bias with respect to term-structure; the RMSEs are around 0.05 over all maturities. The improvements, as measured by slippage with respect to a delta-hedged portfolio, are also quite substantial. The kernel estimator yields slippages that are 72% better than that of Clewlow and Strickland and 124% better than that of either Schwartz and Smith or Todorova.

In addition, the nonparametric model produces implied volatility surfaces which more closely approximate actual surfaces. We reproduce a representative sample of these implied volatility surfaces in figure 1.9. Comparing these plots with those in figures 1.7 and 1.8, it is immediately clear that the nonparametric approach captures the key features of implied volatility in the dimensions of moneyness and time to maturity that other models do not.

Finally, figure 1.10 and table 1.12 offer an important insight into why the kernel method is more capable of pricing options correctly than the parametric approaches considered. Figure 1.10 plots the SPDs in return space rather than price space. In valuing options using martingale pricing techniques, it is these densities rather than the real-world densities that are relevant. Table 1.12 displays the moments of the risk-neutral distribution of the futures returns at expiry. The Black (1976), Schwartz and Smith (2000), Todorova (2004b), and Clewlow and Strickland (2000) approaches all yield Gaussian distributions while the unrestricted kernel approach allows for negative skewness and kurtosis. Since option prices are far more sensitive to tail distributions than, for example, futures prices, it reasonable to conclude that much of the advantage of using the kernel regression method stems from its inherent ability to capture higher moments of a distribution in a way that parametric models derived from Brownian motions cannot. Further, it is easy to understand why the parametric models are more capable of recovering future price than option prices as noted in Section 1.6.3. With nonstochastic interest rates, futures prices equal the expectation

[36] As is standard with kernel regressions, the results are heavily dependent on the choice of bandwidth b. While there is a literature on "optimal" bandwidth selection, it is not clear that these techniques are more sound than simply applying rules of thumb in small samples. The RMSE of our model varies depending on the choice of bandwidth, but is lower, and in most cases substantially lower, than those of the other models for any reasonable value of b. The results in the table use the bandwidth of $b = 0.25$.

Table 1.11: Errors of Nonparametric Model

RMSE represents the root mean squared error of the estimated option price compared to the actual option price. TE represents the mean absolute tracking error of a delta-hedged portfolio. Total represents the error over all options, while ITM, ATM, and OTM represent the errors of in-the-money, at-the-money, and out-of-the-money options respectively. Both the RMSEs and the TEs are denominated in dollars.

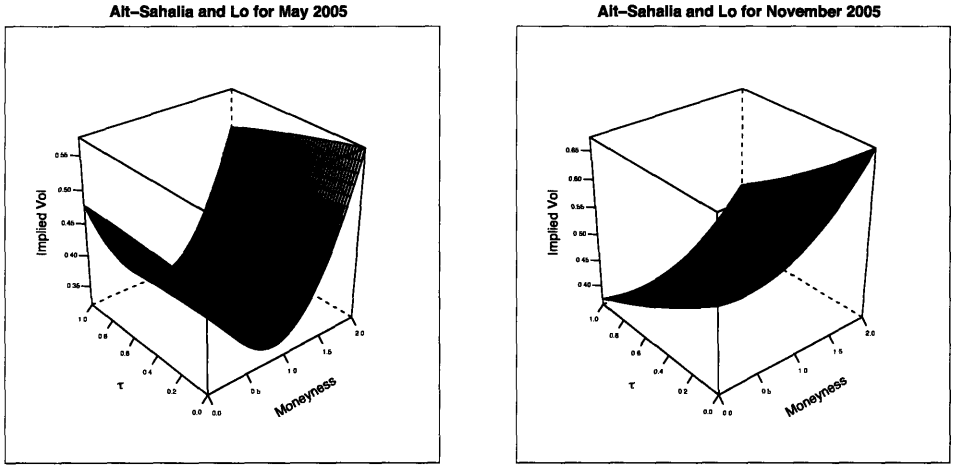
Panel A: RMSE and Tracking Error by Option Type

Model	**RMSE**	**TE**
Aït-Sahalia and Lo (1998)		
Total	0.067	0.184
ITM	0.055	0.140
ATM	0.074	0.212
OTM	0.077	0.211

Panel B: RMSE by Time to Maturity

Maturity	**RMSE**
$\tau < 1$ Month	0.055
$1 \leq \tau < 2$	0.063
$2 \leq \tau < 3$	0.078
$3 \leq \tau < 4$	0.071
$4 \leq \tau < 5$	0.063
$5 \leq \tau < 6$	0.047
$6 \leq \tau < 7$	0.053
$7 \leq \tau < 8$	0.057
$8 \leq \tau < 9$	0.051
$9 \leq \tau < 10$	0.060

Figure 1.9: Nonparametric Estimated Implied Volatility Surfaces for May 2005 and November 2005



under this same risk-neutral density of the spot at the contract's expiry. However, the expectation is applied directly to the spot price rather than the maximum of zero and the difference of the spot price and the strike as is the case with options. Put another way, the futures price is less sensitive to skewness and kurtosis. In sum, the simplifying assumptions used in parametric models examined in this paper, while arguably justifiable when pricing futures, are a substantial liability when trying to price options.

1.8 Conclusion

Though it has received comparatively little attention from economists, natural gas is a remarkably important commodity that quite literally powers our lives. Its significance will surely grow with time as will the associated financial derivatives markets. In recognizing that gas prices behave differently than those of other commodities let alone those of equities, bonds, or currencies, we have attempted in this paper to document gas's unique properties with the aim of building better derivatives models. A direct examination of futures prices revealed evidence of strong seasonality in prices as well as stochastic volatility and the absence of strict normality in returns. Next, we used options prices and a technique borrowed from the recent equity derivatives literature to show that the underlying commodity exhibits regime-switching behavior in its stochastic process; gas prices evolve according to a purely continuous process during the months of October, November, and December, while they evince a combination of pure jumps and jump-diffusions during the remainder of the year. In addition, option prices embed a great deal of information about investors' attitudes towards risk. Using a simple model of investment, we find that investors in gas markets are virtually

Figure 1.10: Estimated SPD-Generated Densities for Continuously Compounded Returns

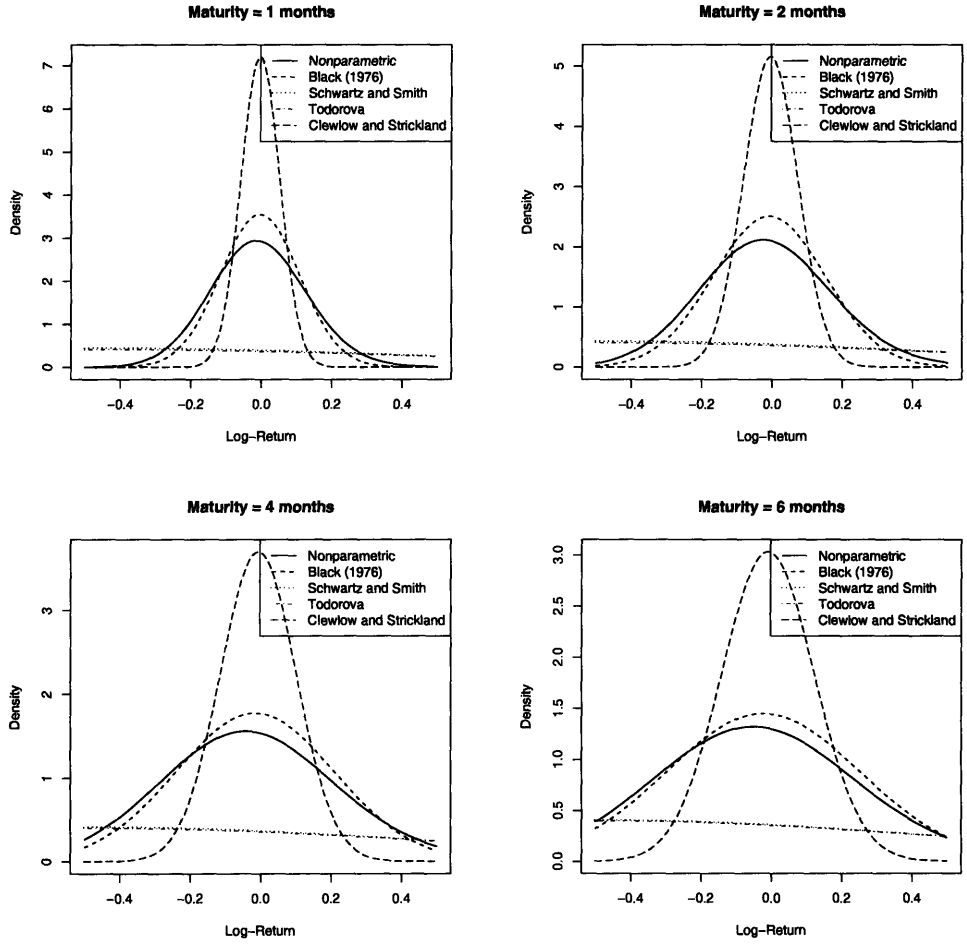


Table 1.12: Moments of SPD-Generated Densities for Continuously Compounded Returns

The table gives the mean, standard deviation, skewness, kurtosis of various models. The nonparametric estimate refers to that derived from kernel density estimation, BS to that from the model of Black (1976) calibrated to to the realized means and variance, SS to that from the model of Schwartz and Smith (2000), T to that from the model of Todorova (2004b), and CS to that from the model of Clewlow and Strickland (2000). The entries in the maturity column are measured in months.

Model	Maturity	Mean	Std. Dev.	Skewness	Kurtosis
Nonparametric	1	−0.006	0.138	−0.190	−0.798
BS	1	−0.004	0.113	0.000	0.000
SS	1	−0.446	0.892	0.000	0.000
T	1	−0.483	0.966	0.000	0.000
CS	1	0.000	0.014	0.000	0.000
Nonparametric	2	−0.015	0.161	−0.141	−1.406
BS	2	−0.008	0.159	0.000	0.000
SS	2	−0.461	0.921	0.000	0.000
T	2	−0.490	0.980	0.000	0.000
CS	2	0.000	0.018	0.000	0.000
Nonparametric	4	−0.026	0.180	−0.101	−1.699
BS	4	−0.016	0.225	0.000	0.000
SS	4	−0.480	0.959	0.000	0.000
T	4	−0.497	0.995	0.000	0.000
CS	4	0.000	0.021	0.000	0.000
Nonparametric	6	−0.028	0.184	−0.117	−1.724
BS	6	−0.023	0.276	0.000	0.000
SS	6	−0.490	0.981	0.000	0.000
T	6	−0.500	1.001	0.000	0.000
CS	6	0.000	0.021	0.000	0.000

risk-neutral over all levels of wealth. This differs from what numerous other papers have observed in equities leading one to conclude that markets are likely segmented. Further, an understanding of risk can prove helpful in calibrating derivatives models as well as evaluating their validity.

We also consider three representative models from the literature and find that while they are effective in forecasting futures, they are unable to accurately recover option prices. We attribute this to their failure to fully incorporate all of the stylized facts highlighted in the first part of the paper. Finally, we introduce an alternative method for pricing gas options based on a kernel regression. We show that this approach, while nonparametric and thus unable to provide the sort of economic intuition of the other models considered, is far more effective in accurately pricing options out-of-sample.

1.A Appendix on Estimating Schwartz and Smith and Todorova Models Using a Kalman Filter

In this approach, we discuss the model of Schwartz and Smith (2000) for concreteness since the extension to that of Todorova (2004b) is quite simple. Calculating options prices using Schwartz and Smith's approach requires that we first estimate the parameters in equations 1.5 and 1.6. This can be accomplished by discretizing each equation, rewriting the dynamic system in state space form, and finally, since the state variables χ and ξ are unobservable, making use of Kalman filtering techniques to construct a likelihood function which we can maximize over our parameters.[37] Comprehensive textbook treatments of the Kalman filter can be found in Hamilton (1994), Zivot and Wang (2003), and Brockwell and Davis (1990). This appendix is meant to serve as a brief overview as it applies to our problem.

We begin with our transition equation which describes the evolution of the state variables

$$\mathbf{x}_t = \mathbf{c} + \mathbf{G}\mathbf{x}_{t-1} + \omega_t \qquad \text{for} \quad t = 1, \ldots, n_T$$

where,

$\mathbf{x}_t \equiv [\chi_t, \xi_t]'$, a 2×1 matrix of unobserved state variables,
$\mathbf{c}_t \equiv [0_t, \mu_t \Delta t]'$, a 2×1 vector,
$\mathbf{G} \equiv \begin{bmatrix} e^{-\kappa \Delta t} & 0 \\ 0 & 1 \end{bmatrix}$, a 2×2 matrix,
$n_T \equiv$ number of time steps,
$\Delta t \equiv$ length of a time step, and

ω_t a 2×1 vector of mean zero, serially uncorrelated, and normally distributed distur-

[37] See Bingham and Kiesel (2004) for more on discretizing continuous functions for maximum likelihood estimation.

bances with $\text{Var}[\omega_t] = \mathbf{W} \equiv \text{Cov}[(\chi_{\Delta t}, \xi_{\Delta t})] = \begin{bmatrix} (1-e^{-2\kappa\Delta t})\frac{\sigma_\chi^2}{2\kappa} & (1-e^{-2\kappa\Delta t})\frac{\rho_{\chi\xi}\sigma_\chi\sigma_\xi}{\kappa} \\ (1-e^{-2\kappa\Delta t})\frac{\rho_{\chi\xi}\sigma_\chi\sigma_\xi}{\kappa} & \sigma_\xi^2\Delta t \end{bmatrix}$.

The measurement equation links the state variables with the observed futures prices. Here we have,

$$\mathbf{y}_t = \mathbf{d}_t + \mathbf{F}_t'\mathbf{x}_t + \mathbf{v}_t, \qquad \text{for} \quad t = 1, ..., n_T$$

where,

$\mathbf{y}_t \equiv [\ln F_{T_1}, ..., \ln F_{T_n}]'$, an $n \times 1$ vector of (log) futures prices with maturities T_i,
$\mathbf{d}_t \equiv [A(T_1), ..., A(T_n)]'$, an $n \times 1$ vector,
$\mathbf{F}_t \equiv \begin{bmatrix} e^{\kappa T_1} & 1 \\ \vdots & \vdots \\ e^{\kappa T_1} & 1 \end{bmatrix}$ an $n \times 2$ matrix, and

$\mathbf{v}_t$, an $n \times 1$ vector of mean zero, serially uncorrelated, normally distributed innovations with $\text{Cov}[v_t] = V$.

Next, the Kalman filtering algorithm is used to compute forecasts $\widehat{\mathbf{x}}_{t|t-1}$ and $\widehat{\mathbf{y}}_{t|t-1}$ where the hat and subscript notation denotes linear projections of $\mathbf{x}$ and $\mathbf{y}$ on their respective vectors of lagged values. If we assume that the initial state, $\mathbf{x}_1$, and the disturbances $\{\omega_t, \mathbf{v}_t\}_{t=1}^T$ are Gaussian, then one can show that $\widehat{\mathbf{x}}_{t|t-1}$ and $\widehat{\mathbf{y}}_{t|t-1}$ are optimal forecasts among any (not just linear) functions of $\mathbf{y}_{t-1}$. Further, the distribution of $\mathbf{y}_t$ is conditionally normal,

$$\mathbf{y}_t|\mathbf{y}_{t-1} \sim \mathcal{N}((\mathbf{F}'\widehat{\mathbf{x}}_{t|t-1}), (\mathbf{F}'\mathbf{P}_{t|t-1}\mathbf{F} + \mathbf{V}))$$

where,

$\mathbf{P}_{t|t-1} \equiv \mathbb{E}[(\mathbf{x}_{t+1} - \widehat{\mathbf{x}}_{t|t-1})(\mathbf{x}_{t+1} - \widehat{\mathbf{x}}_{t|t-1})']$
$\mathbf{P}_{1|0} = \mathbb{E}[(\mathbf{x}_1 - \widehat{\mathbf{x}}_1)(\mathbf{x}_{t+1} - \widehat{\mathbf{x}}_{t|t-1})']$

This in turn allows us to construct the sample log likelihood function

$$\sum_{t=1}^{T} \log f_{\mathbf{y}_t|\mathbf{y}_{t-1}}(\mathbf{y}_t|\mathbf{y}_{t-1})$$

which we can maximize in order to find $\mathbf{c}$, $\mathbf{G}$, $\mathbf{d}$, and $\mathbf{F}$ and in turn κ, λ_x, σ_x, μ_ξ, λ_ξ, σ_ξ, and $\rho_{\chi\xi}$.

1.B Appendix on Todorova and Kendall Deseasonalization

As discussed in Section 1.6.1, Todorova (2004b) is interested in incorporating seasonality into the models of several commodity price processes. To that end, Todorova

modifies the Schwartz and Smith (2000) framework by first fitting the model to a deseasonalized price series and then adding back the seasonal component to the forecasted prices.

We follow the procedure outlined by Todorova (2004a, Appendix E) which in turn is derived from Kendall and Ord (1990, Chapter 4) in order to deseasonalize the futures data. Let $F_{t,\tau}$ denote the mid-month time series of futures that expire in τ months. Since the futures prices seem to exhibit a time trend, that trend must be removed before seasonality can be estimated.[38] Let

$$\hat{m}_t = \frac{1}{12}\left(\frac{1}{2}F_{t-6,\tau} + F_{t-5,\tau} + F_{t-4,\tau} + \cdots + F_{t+4,\tau} + F_{t+5,\tau} + \frac{1}{2}F_{t+6,\tau}\right)$$

be the moving average used to estimate the time trend component $\hat{m}_t$. Given the time trend, we can compute the detrended time series $x_{t,\tau}$ as

$$x_{t,\tau} = F_{t,\tau} - \hat{m}_t.$$

Given the the detrended data it is simple to estimate a monthly seasonal component s_j defined as

$$s_j = \bar{x}_{j,\tau} - \bar{x}$$

where $\bar{x}_{j,\tau}$ is the average detrended price for calendar month j across the years of the sample and $\bar{x}$ is the average detrended price ($\bar{x} = \frac{1}{12}\sum_j \bar{x}_{j,\tau}$). Finally we compute the deseasonalized price series $\tilde{F}_{t,\tau}$ as $F_{t,\tau} - s_j$. We use this deseasonalized, but not detrended, process for estimation using the Kalman filter.

1.C Appendix on Clewlow and Strickland and PCA Analysis

In this appendix, we provide additional details regarding the procedure we used for estimating the Clewlow and Strickland (2000) model. The authors' approach is to model the entire forward curve through time by specifying that the process is composed of n independent volatility functions, each parameterized by the current date and the maturity date, with uncertainty introduced by n number of random shocks which are assumed to be Brownian increments. These assumptions are represented in the stochastic process given by equation 1.10. It is worth noting that this parameterization is fairly general and does not assume any functional forms for the volatility functions save for modelling of shocks as Brownian increments.

To actually estimate these volatility functions, additional assumptions must be made. First, as we have noted, gas exhibits seasonality which we can incorporate

[38] For the mid-month prices, we use the realized price closest to the 15th of each month, so t = April 15, May 15, June 15, ...

into the model quite easily by introducing a time-dependent seasonality adjustment factor which can be proxied by the spot volatility. Second, while we need not introduce functional form restrictions on the volatility functions, we must reduce some of the time dependence of their parameterization in order to estimate them with historical data. To do so, instead of allowing the functions to be dependent on both t and T, the current date and the maturity date respectively, we only allow them to depend on $T-t=\tau$, the time to maturity. The estimation procedure itself determines the number of volatility functions. These two restrictions yield equation 1.11.

At this stage, we have a model for the futures curve evolution with two components to estimate. First, we must estimate the spot volatility and second we must estimate the volatility functions. For the first task we proceed as suggested by the authors and construct rolling estimates of the spot standard deviation by calculating the sample standard deviation of the shortest maturity contract's daily returns on a 30 day rolling basis.

To estimate the volatility functions, we follow the authors who apply Ito's lemma in logarithmic form after the futures prices have been normalized with the rolling volatility as we have previously described. The authors discretize the resulting equation giving us

$$\Delta \log F(t,t+\tau_j) = -\frac{1}{2}\sum_{i=1}^{n}\sigma_i(t,t+\tau_j)^2\Delta t + \sum_{i=1}^{n}\sigma_i(t,t+\tau_j)\Delta z_i \tag{1.13}$$

where Δt is one day, $\log F(t,t+\tau_j)$ is the price of the futures with the jth maturity at time t (τ_j would, for example, be one month or two months), and n represents the number of volatility functions. It is worth noting here that the assumptions we have made so far imply that the left-hand side of equation 1.13, $\Delta \log F(t,t+\tau_j)$, is jointly normally distributed for all of the different maturities. Also, it is clear that the left-hand side is simply the daily continuously compounded returns of the futures contract. Given we are continuing to assume that the stochastic process does not have jumps, before continuing, we take a detour and review a filtering procedure that must be undertaken before proceeding with estimation.

Since it is assumed in the formulation of this model that the underlying stochastic process is one without jumps, returns which appear to violate this assumption can adversely affect estimation. To mitigate this potential problem, we apply a "recursive filter" to the data to remove suspected jumps before estimating the parameters of the stochastic process as suggested by the Clewlow and Strickland. First, we calculate a time series of daily returns for contracts of each maturity as well as the associated sample standard deviations. Next we assert (somewhat arbitrarily) that returns beyond a threshold of three standard deviations constitute a jump and are accordingly eliminated from the data set. Then, we recalculate the sample standard deviations and again eliminate observations associated with potential jumps as previously defined. We repeat this procedure for 10 iterations. As 12 different futures contracts trade

each and every day, we must apply this procedure separately for each contract.

As $\Delta \log F(t, t+\tau_j)$ is simply the daily continuously compounded return of a futures contract, to estimate the volatility functions we begin by constructing the covariance matrix of these returns which we denote Σ_t. To allow some time dependence we estimate these covariance matrices on a rolling 30-day basis. These covariance matrices are then decomposed using an eigenvector decomposition into two pieces: a matrix of eigenvalues and a matrix of eigenvectors. Thus we decompose Σ_t as

$$\Sigma_t = \Gamma_t \Lambda_t \Gamma_t'$$

with

$$\Gamma_t = \begin{pmatrix} v_{11} & \cdots & v_{1n} \\ & \ddots & \\ v_{n1} & \cdots & v_{nn} \end{pmatrix} \text{ and } \Lambda_t = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \lambda_n \end{pmatrix}$$

where Γ_t is a the matrix of eigenvectors and Λ_t is the diagonal matrix of associated eigenvalues. The relative size of the eigenvalues determines the extent to which that eigenvalue and its associated eigenvector (the factor) explains the variance of the sample returns and thus provides a method for choosing an appropriate number of volatility functions; in practice we choose 5 volatility functions for most estimations. The actual volatility functions are recovered by simple algebra from the decomposition as

$$\sigma_i(t, t+\tau_j) = v_{ji}\sqrt{\lambda_i}.$$

Armed with the volatility functions, we can completely characterize the evolution of the forward curve through time. Furthermore due to the assumptions regarding the forward curve given in equation 1.11, we know that $\log F(t, t+\tau_j$ is normally distributed. Since the log-transformed futures prices are normally distributed, we can apply the standard Black-Scholes-Merton approach and derive a closed-form option price which the authors undertake thereby deriving equation 1.12. Since we have estimated the volatility functions at each time t as outlined in this appendix and estimated appropriate interest rates as detailed in Section 1.3.3, it is then straightforward to price the options in the data set using this formula.

Bibliography

AÏT-SAHALIA, Y., AND A. W. LO (1998): "Nonparametric Estimation of State-Price Densities Implicit in Financial Asset Prices," *Journal of Finance*, 53(2), 499–547.

——— (2000): "Nonparametric Risk Management and Implied Risk Aversion," *Journal of Econometrics*, 94(1-2), 9–51.

BARONE-ADESI, G., AND R. E. WHALEY (1987): "Efficient Analytic Approximation of American Option Values," *Journal of Finance*, 42(2), 301–320.

BATES, D. S. (1996): "Jumps and Stochastic Volatility: Exchange Rate Processes Implicit in Deutsche Mark Options," *Review of Financial Studies*, 9(1), 69–107.

BINGHAM, N., AND R. KIESEL (2004): *Risk-Neutral Valuation: Pricing and Hedging of Financial Derivatives*. Springer-Verlag, London.

BLACK, F. (1976): "The Pricing of Commodity Contracts," *Journal of Financial Economics*, 3, 167–179.

BREEDEN, D. T., AND R. H. LITZENBERGER (1978): "Prices of State-Contingent Claims Implicit in Option Prices," *Journal of Business*, 51(4), 621–51.

BRENNAN, M. J., AND E. S. SCHWARTZ (1979): "A Continuous Time Approach to the Pricing of Bonds," *Journal of Banking and Finance*, 3(2), 133–155.

BROCKWELL, P. J., AND R. A. DAVIS (1990): *Time Series: Theory and Methods*. Springer-Verlag, New York, New York, second edn.

CARMONA, R. A. (2004): *Statistical Analysis of Financial Data in S-Plus*. Springer-Verlag, New York.

CARR, P., AND L. WU (2003a): "The Finite Moment Log Stable Process and Option Pricing," *Journal of Finance*, 58(2), 753–77.

——— (2003b): "What Type of Process Underlies Options? A Simple Robust Test," *Journal of Finance*, 58(6), 2581–2610.

——— (2004): "Time-changed Lévy Processes and Option Pricing," *Journal of Financial Economics*, 71(1), 113–41.

CHAMBERS, M. J., AND R. E. BAILEY (1996): "A Theory of Commodity Price Fluctuations," *Journal of Political Economy*, 104(5), 924–957.

CLEWLOW, L., AND C. STRICKLAND (2000): *Energy Derivatives: Pricing and Risk Management*. Lacima Publications, London, England.

CONSTANTINIDES, G. M. (1982): "Intertemporal Asset Pricing with Heterogeneous Consumers and Without Demand Aggregation," *Journal of Business*, 55(2), 253–267.

CUDDINGTON, J. T., AND Z. WANG (April 20, 2005): "Assessing the Degree of Spot Market Integration For U.S. Natural Gas: Evidence from Daily Price Data," Working paper, Georgetown University and Monash University, Forthcoming in *Journal of Regulatory Economics*.

DEATON, A., AND G. LAROQUE (1996): "Competitive Storage and Commodity Price Dynamics," *The Journal of Political Economy*, 104(5), 896–923.

DORAN, J. S. (2005): "Estimation of Risk Premiums in Energy Markets," Working paper, Department of Finance, Florida State University.

DORAN, J. S., AND E. I. . RONN (2006): "Computing the Market Price of Volatility Risk in the Energy Commodity Markets," Working paper, Department of Finance, Florida State University.

DUFFIE, D. (2001): *Dynamic Asset Pricing Theory*. Princeton University Press, Princeton, New Jersey, third edn.

ENERGY INFORMATION ADMINISTRATION (2004): *Annual Energy Review 2004*. Department of Energy, Washington, DC, Report No. DOE/EIA-0384(2004).

FERSON, W. E., AND G. M. CONSTANTINIDES (1991): "Habit Persistence and Durability in Aggregate Consumption," *Journal of Financial Economics*, 29(2), 191–240.

GEMAN, H. (2005): *Commoditites and Commodity Derivatives: Modeling and Pricing for Agriculturals, Metals, and Energy*. John Wiley & Sons, West Sussex, England.

GIBSON, R., AND E. S. SCHWARTZ (1989): "Stochastic Convenience Yield and the Pricing of Oil Contingent Claims," *Journal of Finance Papers and Proceedings*, 45(3), 959–76.

HAMILTON, J. D. (1994): *Time Series Analysis*. Princeton University Press, Princeton, New Jersey.

HANSEN, L. P., AND K. J. SINGLETON (1982): "Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models," *Econometrica*, 50(5), 1269–1286.

——— (1984): "Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models," *Econometrica*, 52(1), 267–268.

HILLIARD, J. E., AND J. REIS (1998): "Valuation of Commodity Futures and Options under Stochastic Convenience Yields, Interest Rates, and Jump Diffusions in the Spot," *Journal of Financial and Quantitative Analysis*, 33(1), 61–86.

HUTCHINSON, J. M., A. W. LO, AND T. POGGIO (1994): "A Nonparametric Approach to Pricing and Hedging Derivative Securities Via Learning Networks," *Journal of Finance Papers and Proceedings*, 49(3), 851–889.

JACKWERTH, J. C. (2000): "Recovering Risk Aversion from Option Prices and Realized Returns," *Review of Financial Studies*, 13(2), 433–51.

KALDOR, N. (1939): "Speculation and Economic Stability," *Review of Economic Studies*, 7(1), 1–27.

KENDALL, M., AND J. K. ORD (1990): *Time Series*. Edward Arnold, London, England, third edn.

LEWIS, A. L. (2000): *Option Valuation under Stochastic Volatility*. Finance Press, Newport Beach, California, USA.

——— (2002): "Fear of Jumps," *Wilmott*, pp. 60–67.

MEHRA, R., AND E. C. PRESCOTT (1985): "The Equity Premium: A Puzzle," *Journal of Monetary Economics*, 15(2), 145–161.

MERTON, R. C. (1992): *Continuous-Time Finance*. Blackwell Publishing, Malden, MA.

MILTERSEN, K. R., AND E. S. SCHWARTZ (1998): "Pricing of Options on Commodity Futures with Stochastic Term Structures of Convenience Yields and Interest Rates," *Journal of Financial and Quantitative Analysis*, 33(1), 33–59.

PINDYCK, R. S. (2001): "The Dynamics of Commodity Spot and Futures Markets: A Primer," *The Energy Journal*, 22(3), 1–29.

——— (2004): "Volatility in Natural Gas and Oil Markets," *The Journal of Energy and Development*, 30(1), 1–19.

REBONATO, R. (2004): *Volatility and Correlation: The Perfect Hedger and the Fox*. John Wiley and Sons Ltd., West Sussex, England, second edn.

ROSS, S. A. (1976): "Options and Efficiency," *Quarterly Journal of Economics*, 90(1), 75–89.

ROUTLEDGE, B. R., D. J. SEPPI, AND C. S. SPATT (2000): "Equilibrium Forward Curves for Commodities," *Journal of Finance*, 55(3), 1297–1338.

SCHWARTZ, E., AND J. E. SMITH (2000): "Short-Term Variations and Long-Term Dynamics in Commodity Prices," *Manangement Science*, 46(7), 893–911.

SCHWARTZ, E. S. (1997): "The Stochastic Behavior of Commodity Prices: Implications for Valuation and Hedging," *Journal of Finance Papers and Proceedings*, 52(3), 923–73.

SHREVE, S. E. (2004): *Stochastic Calculus for Finance II*. Springer-Verlag, New York.

SUNDARESAN, S. M. (1984): "Equilibrium Valuation of Natural Resources," *Journal of Business*, 57(4), 493–518.

TELSER, L. G. (1958): "Futures Trading and the Storage of Cotton and Wheat," *Journal of Political Economy*, 66(3), 233–255.

TODOROVA, M. (2004a): "Essays on Modeling Derivative Claims," Ph.D. thesis, Columbia University.

TODOROVA, M. I. (2004b): "Modeling Energy Commodity Futures: Is Seasonality Part of It?," *Journal of Alternative Investments*, 7(2), 10–31.

WORKING, H. (1948): "Theory of the Inverse Carrying Charge in Futures Markets," *Journal of Farm Economics*, 30(1), 1–28.

——— (1949): "The Theory of the Price of Storage," *American Economic Review*, 39(6), 1254–1262.

WRIGHT, B. D., AND J. C. WILLIAMS (1989): "A Theory of Negative Prices for Storage," *Journal of Futures Markets*, 9(1), 1–13.

ZIVOT, E., AND J. WANG (2003): *Modeling Financial Time Series with S-Plus*. Springer-Verlag, New York, New York.

Chapter 2

Rethinking the Home Bias Puzzle: A Two-Step Approach

With Erik C. Ruben

Abstract

This paper presents new estimates and approaches to estimating the home bias puzzle. We use micro-level data to calculate households' foreign equity exposure as a function of wealth. We find simple estimates have significant errors-in-variables problems and we construct an estimator using grouping to account for this issue. Our estimates still imply low aggregate investment in foreign equity. Finally, we disaggregate the investment decision by incorporating two step decisions that allow households to forgo participating in the market. As a result of the decoupling, we find foreign equity levels closer to that of standard portfolio theories.

2.1 Introduction

In their 1991 paper "Investor Diversification and International Equity Markets," French and Poterba (1991) quantified the strength of one of the most curious and enduring empirical irregularities in open economy macroeconomics: the Home Bias Puzzle (HBP). Their paper presented strong evidence that aggregate equity portfolios in industrialized countries were heavy biased towards domestic stock ownership relative to the predictions of standard portfolio optimization models. While subsequent papers including Bohn and Tesar (1996), have found that the HBP has diminished somewhat since French and Poterba published their results, the magnitudes of the bias in most countries are still far too great to be accounted for using standard explanations.

Most investigations into the HBP have utilized aggregate data. This paper takes a different approach in that we use individual investor level data to examine asset allocation decisions. Moreover, we add to the existing debate over the source of the HBP by disaggregating portfolio selection into two components: first, the binary decision to participate in international markets and second, the conditional foreign asset allocation choices. Whereas previous investigations into the HBP have tried to explain why investors *on average* hold seemingly low shares of foreign assets, we show that *conditional on their choosing to participate in foreign markets*, investors construct portfolios with much higher levels of foreign holdings. Thus, the interesting question regarding foreign asset ownership shifts from one of portfolio share to one of participation in international markets. In addition, we find that while the conditional portfolio allocations to foreign assets are somewhat independent of wealth levels, the participation decision is closely tied to investor affluence.

Using detailed household-level data provided by the Survey of Consumer Finances (SCF), we structure our investigation by first studying the interplay between individuals' portfolio decisions and their levels of net financial wealth. Simple econometric analysis suggests little relationship between wealth and unconditional foreign equity ownership. However, we show that this is the result of an errors-in-variables problem and we discuss the implications and caveats associated with this conclusion. Finally, we decouple the participation and conditional portfolio decisions and provide some evidence that the real empirical "puzzle" is the binary decision of whether or not to invest.

We will proceed as follows. In the next section, we highlight some of the important insights and approaches in the extensive literature and show how they relate to this paper. In Section 2.3, we discuss our data and their relevant statistical properties. Section 2.4 lays out our estimation procedure for the unconditional investment decision and the associated problems in measuring this choice. Section 2.5 estimates the participation decision and conditional portfolio choices. Section 2.6 concludes.

2.2 Literature Review

Even prior to French and Poterba (1991), financial economists[1] asserted that investors hold insufficient foreign assets relative to that suggested by traditional portfolio theory models such as the international version of the capital asset pricing model (CAPM).[2] Most often these older studies used macro level data to estimate the home bias. Often, as is the case with the original French and Poterba (1991) article, the home bias was estimated using accumulated capital flows and valuation adjustments. However there is evidence that these flows are not well suited for estimating the home bias.[3]

[1] See, for example, Levy and Sarnat (1970).

[2] See Sharpe (1964) and Lintner (1965) for the original development of these models.

[3] See Warnock and Cleaver (2003).

In March of 1994 and again in December of 1997, the U.S. government conducted comprehensive studies of its residents' foreign security holdings by surveying major custodians and other large investors.[4] Using this data, Ahearne, Griever, and Warnock (2004) estimates that foreign equities comprise 12 percent of US investors' equity portfolios (using 1997 data). For comparison, French and Poterba (1991) estimate that share to be about 6.2 percent at the end of 1989 and Bohn and Tesar (1996) estimate the share at nearly 8 percent in 1994. Thus there is some evidence that while home bias in the US has lessened substantially over the last twenty years, US foreign equity holdings are still much lower than levels predicted by standard portfolio theory.

Numerous attempts have been made to explain the HBP by loosening typical assumptions like the existence of a representative consumer, riskless borrowing, and complete markets or by more explicitly modeling the gains from diversification.[5] One simple approach is to consider the possibility that domestic equities provide a better hedge for home country specific risks. However models in this spirit generally predict lower levels of home bias than observed. Another form of country specific risk could be related to non-tradable assets such as human capital. For example, Stockman and Dellas (1989), finds an equilibrium in which a country's residents derive little benefit from diversification via claims indexed to nontradable endowments because investors capture all the available gains from diversification by investing in claims linked to tradable goods. However, others have found that this approach is unlikely to provide a sufficient explanation for the HBP. Another version of the nontradables idea was offered by Bottazzi, Pesenti, and van Wincoop (1996) which argued that given the negative correlation between labor income and returns on capital, domestic stocks may provide a better hedge against consumption volatility than international equities. This explanation seems unlikely given the findings of Mankiw and Zeldes (1991) which previously showed that wealthy investors, who constitute a large portion of equity holders, care little about hedging labor income. Lucas (1987) tried a different approach, claiming that observed aggregate consumption volatility is insufficient to justify much international investing in the face of even minor transactions costs. Lucas's paper, however, relied heavily on the assumption of a trend stationary consumption process and at best only explains the HBP for the United States which experiences unusual consumption smoothness. Further, because it only considers aggregate decision-making, Lucas's approach ignores individual heterogeneity and the fact that individuals cannot as easily hedge away idiosyncratic risk and thus may have more motivation to hedge than a representative consumer. There is also a substantial literature that tries to explain the HBP in the context of diversification costs, but many authors have found large gains from international diversification even when taking these costs into account. Another possible explanation relates to simple mismeasurement. Gains from international diversification are derived from historical means and variances, but limited data leads to significant uncertainty regarding

[4]See Griever, Lee, and Warnock (2001) for a discussion.

[5]For a more comprehensive look at the literature, see Lewis (1999) for a recent survey.

these measures and hence the existence of a home bias. Lastly, Ahearne, Griever, and Warnock (2004) believe the underweighting of foreign equity in US portfolios is primarily due to significant information costs.

The present paper does not try to explain the home bias in the traditional sense, but rather redefine the puzzle. To do so, we investigate individual portfolio choices as opposed to the aggregate measures the other studies have used. This allows us to examine the interplay between investors' decision-making process and their individual characteristics including wealth levels.

2.3 Data

The U.S. household level financial data used in this paper are taken from the Federal Reserve's Survey of Consumer Finances. The most recent publicly available version of the SCF, which is conducted every three years, was compiled in 2001. While the Fed has been conducting the survey for decades, it only started collecting data on our primary variable of interest, U.S. household ownership of foreign assets,[6] in 1995. Thus, this paper restricts itself to data from the 1995, 1998, and 2001 surveys. The SCF, which utilizes a dual-frame sample design, offers the most complete obtainable description of U.S. family finances. The survey designers selected about 4,000[7] household participants using a dual-frame methodology. The first group of families were selected from a standard multi-stage area-probability design devised to ensure proper representation of broad characteristics like home ownership. The second group of families were chosen based on Internal Revenue Service data to get disproportionate representation of relatively wealthy Americans. Given the limited survey size, the inclusion of this set of subjects is critical to ensuring proper representation of a concentrated yet significant agglomeration of national wealth and in particular foreign asset ownership.

The SCF has two important features which deserve mention as they have very significant ramifications for performing any sort of econometric analysis on the data. First, the observations do not have equal associated probability and therefore must be weighted before trying to interpret the survey data. The SCF designers constructed the weights using original selection probabilities and frame information as well as information available in the Current Population Survey; the weights sum to the number of households[8] in the sample universe.[9] In addition, the study suffers from missing data. To remedy this, the surveyors opted to impute missing values by

[6]The Fed collects both foreign stock and foreign bond ownership; it does not collect information regarding other forms of foreign financial asset ownership (*e.g.* derivatives, mutual funds, etc.) nor does it provide more detailed information regarding portfolio choices.

[7]The number varies slightly by year.

[8]In 1998, the number of households, and hence sum of weights, was 102.5 million

[9]This is from the survey documentation; see `http://www.federalreserve.gov/pubs/oss/oss2/98/scf98home.html` (Kennickell, 2000)

drawing repeatedly from an estimate of the conditional distribution of the data and then storing these imputations as five successive replicates, or "implicates," of each data record. As a result, the full dataset has five times the actual number of respondents: for example, in 1998 there are 21,545 observations versus 4,305 households. The survey designers argue that multiple imputation promises more efficient estimation than singly-imputed data because it generates multiple outcomes from a stochastic process. In addition, with multiple imputation, users can estimate the level of uncertainty associated with the missing information. Estimation, of course, requires some care since each data record contains five implicates which are not independent observations.

We provide some sample statistics in table 2.1 for the SCF data sets from 1995, 1998, and 2001 that we use in this paper. The table illustrates that weighting is crucial to properly interpreting the data set and any statistics derived from it; the difference between weighted averages and simple averages is substantial. Also, of interest is the relatively small number of observations that actually report owning foreign assets. Although members of these households constitute a significant portion of the total population when weighting is taken into account, the fact that there is only a small number of these observations makes estimation problematic especially if there are errors present.

2.4 Estimation of Unconditional Investment Decision

2.4.1 Simple Models and IV

The first step in offering an explanation for the HBP is to examine the relationship between wealth and foreign stock ownership at the individual household level. Defining $\boldsymbol{y}$ as foreign equity holdings, we are interested in the following regression function:

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i \tag{2.1}$$

where $\boldsymbol{x}_i'$ is a k-dimensional vector specifying the ith observation's characteristics. These characteristics can include various factors (home ownership, trust ownership, sex, etc.), but most importantly include financial wealth. First, let y_i be dollar-denominated foreign stock holdings and x_i net financial wealth and other demographic variables of the household including education, age, and whether the household received professional financial advice. The results are tabulated in table 2.2. These results are striking in that only *FIN* is significant. This is a rather surprising result given that one would think that better informed households, either through education or professional advice, would diversify their portfolio holdings and realize the potential

Table 2.1: Sample statistics of SCF datasets

Statistic	Variable	Implicate	Year 1995	1998	2001
Number of observations			4299	4305	4442
Sum of weights		1	99,010,458	102,548,840	106,495,827
		2	99,010,458	102,548,841	106,495,822
		3	99,010,458	102,548,842	106,495,762
		4	99,010,458	102,548,843	106,495,808
		5	99,010,458	102,548,847	106,495,827
		Average	99,010,458	102,548,842	106,495,809
Observations $FA > 0$		1	236	264	288
		2	235	272	287
		3	234	263	290
		4	238	265	290
		5	239	265	289
		Average	236.4	265.8	288.8
Weighted Mean	Foreign Assets	1	327.5265	1,393.268	1,544.332
		2	323.1702	1,167.291	1,481.277
		3	293.2713	1,104.691	1,516.466
		4	333.4674	1,302.647	1,575.255
		5	399.1033	1,213.458	1,469.048
		Average	335.3077	1,236.271	1,517.276
Weighted Mean	Financial Wealth	1	92,806.6	133,547.2	191,869.5
		2	91,652.16	137,320.6	189,474.2
		3	91,964.38	132,478.4	192,905.4
		4	89,009.1	139,878.0	186,775.0
		5	93,069.14	131,053.4	192,230.0
		Average	91,700.28	134,855.5	190,650.8
Unweighted Mean	Foreign Assets	1	18,094.44	47,711.53	94,184.64
		2	15,588.47	45,271.71	89,610.38
		3	18,712.08	44,969.76	88,969.61
		4	19,853.45	42,882.48	94,104.70
		5	14,833.23	44,804.12	100,138.44
		Average	17,416.33	45,127.92	93,401.55
Unweighted Mean	Financial Wealth	1	1,681,018	2,165,496	2,920,921
		2	1,675,034	2,184,691	2,830,215
		3	1,699,617	2,175,257	2,841,817
		4	1,744,324	2,143,680	2,946,908
		5	1,718,128	2,216,238	2,870,094
		Average	1,703,624.03	2,177,072.29	2,881,990.87

"Observations $FA > 0$" refers to observations who have positive foreign equity holdings.

Table 2.2: OLS regression results: multiple independent variables

	Year		
	1995	1998	2001
Intercept	-336.0259	1973.86	21304.198**
	(1425.6189)	(4462.2167)	(10672.252)
FIN	0.0037452***	0.0150006***	0.0340929***
	(0.0006839)	(0.0020877)	(0.0032153)
Professional Advice	84.449086	-148.343	-3104.597
	(438.67514)	(1580.2427)	(3809.7003)
Education	34.246357	-138.492	-1199.743*
	(80.208039)	(272.35948)	(654.64935)
Age	-3.512104	-17.72655	-176.1148
	(-3.512104)	(44.495354)	(107.34792)

Table 2.3: Results from simple regression.

	Year		
	1995	1998	2001
Intercept	-8.7353	-776.3679	-4,836.696***
	(225.4055)	(785.9162)	(1,856.3869)
FIN	0.00375***	0.0149***	0.0333***
	(0.0007)	(0.0021)	(0.0032)

*** significant at 1% level, ** significant at 5% level, * significant at 10% level. Numbers in parentheses are standard errors. This regression takes into account all five implicates and adjusts the coefficient estimates and standard errors appropriately.[11]

gains that other authors have found with international diversification.[10] Hence, we will proceed by letting our independent variable be only financial wealth (*FIN*). The results are tabulated in table 2.3.

The table clearly shows, not surprisingly, that wealth is strongly related to foreign

[10]We have also looked at numerous other explanatory variables and combinations thereof including use of a professional financial adviser, ownership of a trust, sex, home ownership, and equity in one's company, among others. The results for these regressors are consistent with the results we report in table 2.2: regressions involving wealth and other potential explanatory almost always produce statistically insignificant coefficients estimates for those other factors. This itself is a very interesting result and probably deserves a fuller discussion, but is somewhat outside the scope of this current paper.

[11]See Montalto and Sung (1996) for a discussion of the theory of properly adjusting for multiple implicates in the regression context.

investment level. For example, in 1995, households invested in foreign assets about 0.4 cents of each incremental dollar of wealth. Three years later, investors were allocating about 1.5 cents on the the dollar to international holdings. By 2001, the marginal rate of foreign investment had risen to about 3.3 cents.[12] In sum, the coefficient estimates of marginal foreign investment level, though significant, are quite small and imply foreign equities as a share of total financial assets at a lower level than the original work that established the home bias and thus implying an even larger home bias. There is strong reason to believe, however, that these small coefficient estimates may be partially attributable to measurement error as mismeasurement in explanatory variables yields downwardly biased coefficient estimates.[13] While uncertainty over the existence of a subset of good instruments means we cannot use a standard test to check for the failure or orthogonality, we have two reasons to believe that measurement error presents a particularly acute problem in this dataset. First, anecdotal analysis suggests inconsistencies in the data. For example, there are households in the dataset whose foreign equity holdings exceed their *gross* financial wealth holding. This is clearly a nonsensical result. Second, when we employ an instrumental variables approach to correct for potential measurement error of the covariates using labor income and housing wealth as our instruments, the coefficients increase in magnitude. The results are reported in table 2.4.

Using both labor income and housing as instruments, we find that the coefficient on foreign investment increases approximately fourfold in 1995, twofold in 1998, and 14% in 2001 from the simple regression reported in table 2.3. Thus this standard approach for correcting errors-in-variables does increase the magnitude of the point estimates. There is, however, significant concern with using the instrumental variables approach for if financial wealth is measured with error, it is likely that instruments are as well. This suggests that the instruments are correlated with financial wealth and thus are invalid instruments. Since it seems likely that we have invalid instruments and thus inconsistent point estimates, we will move forward and attempt to estimate the relationship between foreign equities and financial wealth using a different approach. In the spirit of Wald (1940), the basic idea is to average across observations within a wealth range. Assuming no correlation among observations[14] and given our relatively large sample, this strategy would be expected to "average away" much of the measurement error. Essentially averaging across observations reduces random disturbance in magnitude and will hopefully eliminate the errors-in-variables problem thereby allowing us to produce consistent coefficient estimates. There is, of course,

[12]It is worth noting this coefficient changed by a large margin from 1995 to 2001, but further exploration of this result is outside the scope of this paper.

[13]Hausman's "Iron Law;" holds when error is uncorrelated with true value. For a fuller discussion of this result see Greene (2002, Section 9.5.2) and Wooldridge (2002, Section 4.4.2).

[14]This is not much of an assumption in a cross-sectional data set if you believe in random sampling and its correct application; the survey design, however, introduces two-stage samples, but we still have no reason to believe we do not have a random sample and hence uncorrelated disturbances across observations.

Table 2.4: IV regression results

Instruments		Year 1995	1998	2001
Labor Income	Intercept	-10370.8***	-29094.9***	13409.14
		(3128.043)	(6228.006)	(14295.06)
	FIN	0.016311***	0.034093***	0.027756***
		(0.000817)	(0.001781)	(0.003473)
Housing	Intercept	-5246.41*	-35898.3***	-18476.5*
		(2989.843)	(5550.908)	(10822.75)
	FIN	0.013303***	0.037218***	0.038820***
		(0.000741)	(0.001118)	(0.001395)
Both	Intercept	-7591.03***	-34446.9***	-16241.9
		(2938.709)	(5452.211)	(10808.89)
	FIN	0.014679***	0.036551***	0.038044***
		(0.000612)	(0.001037)	(0.001376)

The "Instruments" column denotes which instruments were used in the regression.

a tradeoff here: averaging across observations reduces the number of observations we can use in our regression analysis. Averaging across observations can eliminate much of the uncertainly regarding the actual values of the explanatory variables (if measured without error), but averaging too much will significantly reduce the size of the sample on which we can perform regression analysis and hence give results that, although are asymptotically consistent, have not yet converged to the population values.

2.4.2 Grouping

More formally, our approach is to avoid estimating equation 2.1 directly and instead estimate

$$\tilde{y}_i = \tilde{\boldsymbol{x}}_i' \boldsymbol{\beta} + \tilde{\epsilon}_i \tag{2.2}$$

where the variables are defined as follows. Let $\boldsymbol{X}$ be a $n \times k$ matrix, where n is the number of observations and k the number of explanatory variables.[15] $\boldsymbol{x}_i'$ is a column vector containing the observations of the ith individual. It is important to draw the distinction between implicates and observations; where needed, we will use a second subscript to denote implicates hence the jth implicate for observation i is denoted $\boldsymbol{x}_{i_j}'$. Let $\boldsymbol{Y}$ denote the $n \times 1$ column vector of the associated response variables and $\boldsymbol{w}$

[15]Although we develop the more general case of a k explanatory variables, the empirical results only include a single explanatory variable.

represent the $n \times 1$ column vector giving the weights w_i for each observation i; these weights attempt to estimate the relative effect of each observation on the variance.[16] Let $\tilde{\boldsymbol{x}}_i'$ represent the weighted average of m observations. Thus

$$\tilde{\boldsymbol{x}}_i' = \frac{\sum_{l=1}^{m} w_l \boldsymbol{x}_l}{\sum_{l=1}^{m} w_l}.$$

These averages should be computed separately for each implicate giving rise to a vector (indexed by j) of $\tilde{\boldsymbol{x}}_i'$. Similarly, define

$$\tilde{y}_i = \frac{\sum_{l=1}^{m} w_l y_l}{\sum_{l=1}^{m} w_l}$$

which is simply the weighted average of m dependent variables. Continuing this averaging process we can obtain a matrix $\tilde{\boldsymbol{X}}$ and a column vector $\tilde{\boldsymbol{Y}}$.

We will assume the equation 2.1 satisfies the standard large sample consistency assumptions (*i.e.* the linear model is correct, $\boldsymbol{X}'\boldsymbol{X}$ is non-singular, and $\operatorname{plim}\left(\frac{\boldsymbol{X}'\boldsymbol{\epsilon}}{n}\right) = 0$.) However, we will allow that our sample have a non-spherical covariance matrix and is characterized by heteroscedasticity across observations i[17]. In particular, we suspect the heteroscedasticity will be a positive function of the financial wealth. Thus $\operatorname{var}(\boldsymbol{\epsilon}|\boldsymbol{X}) \neq \sigma^2 \boldsymbol{I}$ and instead

$$\operatorname{var}(\boldsymbol{\epsilon}|\boldsymbol{X}) = \begin{pmatrix} \sigma_1^2 & 0 & 0 \\ & \vdots & \\ 0 & 0 & \sigma_n^2 \end{pmatrix} \tag{2.3}$$

where σ_i^2 is an increasing function of financial wealth. Note that we impose zero covariance between observations since we assume, given the SCF survey methodology, that the data set is a random sample.

If the underlying model expressed in equation 2.1 satisfies the large-sample consistency assumptions, then the the transformed regression model expressed in equation 2.2 also satisfies those assumptions and therefore least squares will provide a consistent estimate of $\boldsymbol{\beta}$. The presence of heteroscedasticity, however, means that OLS applied to the original regression model as well as the transformed model will not be asymptotically efficient even in the class of linear estimators. The motivating point, however, is if the model expressed in equation 2.1 does not satisfy large-sample consistency assumptions which is likely given measurement error and the resulted correlation between the dependent variable and the disturbance. In this case, though, OLS on the transformed model may produce consistent estimates while OLS on the

[16] Recall the our data set is an unequal probability design hence we must take into account the weighting to get accurate results.

[17] Notice there is necessarily heteroscedasticity across observations in $\tilde{\boldsymbol{X}}$

original may not due to errors-in-variables. By averaging away errors in variables, we can construct a transformed model that we can estimate with OLS and produce consistent point estimates.

In this paper we use two alternative approach to grouping observations prior to averaging. The first approach assigns the same number of observations in each grouping. We order the observations by net financial wealth and then, in that ordering assign, the first m observations into the first group, the next m observations into the second group, and continue in this manner for all of the observations. This will produce a data matrix of size $\frac{n}{m} \times k$ and a response vector of size $\frac{n}{m} \times 1$. [18] In is worth noting that the consistency of the averaging approach does not depend on the grouping method. The grouping method will, however, affect whether averaging can help alleviate the errors in variables problem. Indeed by first ordering the data set by wealth and then grouping equal numbers of observations we can appeal to the similarly of observations with similar wealth levels to net out errors in variables in the data. One possible problem with the first approach to grouping is the weighting associated with our dataset. As seen above in the construction of elements of $\tilde{\boldsymbol{X}}$, each group has associated with it a weighted size which is significantly different than the number of observations in each group. In particular, consider a group on the low mean financial wealth and another with a high mean financial wealth, given the financial wealth distribution implied by the data the former group will represent significantly more households than the latter. This fact motivates the second form of grouping we use in the paper. Rather than create groups than have the same number of observations, we can construct groups that would have approximately the same implied size in population terms. In concrete terms, we can construct groups where the sum of the weights in each group is approximately equal.[19] We will use both grouping methods in this paper and the choice of grouping methods does not matter for much of the paper, although we will draw distinction when we use a particular method or if there are significant differences between the methods.

Results

In previous sections we have attempted to estimate the incremental changes in foreign equity holdings for increases in net financial wealth. Thus we have attempted to estimate the following equation:

$$FA_i = \beta_0 + \beta_1(FIN_i) + \epsilon_i \tag{2.4}$$

[18] Integer constraints will cause some difficulty here. Generally the last group will have fewer than m observations in it and the associated average will have to be adjusted accordingly.

[19] Since every observation has an associated population weight different than one, it is not possible to create groups with equal implied population weight. Instead we attempt to construct groups where the associated population weights are as near to each other as possible.

where FA_i is investor i's level of foreign equity holdings (Foreign Assets) and FIN_i is an individual's net financial wealth. We have attempted to estimate this equation using standard OLS and IV approaches with less than satisfactory results which we believe is due to a significant errors in variables problem; when *FIN* is measured with error OLS does not produce consistent coefficient estimates and the increase in coefficient estimates we saw by using IV provides some evidence of this. As we have noted, IV is not without problems and, in particular, may itself be inconsistent, so we have introduced an estimator using averaging in section 2.4.2. We will use this approach to again estimate equation 2.4. This method should reduce errors in variables and hence can produce consistent, but not asymptotically efficient, results. We construct estimates using various group sizes and utilizing the first grouping method, using equal number of observations per group, discussed in section 2.4.2. The results from this approach are summarized in table 2.6.[20]

Table 2.6 shows that the averaging approach doubled the point estimate of 2001 over the simple regression reported in table 2.3 and nearly doubled the IV regression coefficient in table 2.4.[21] These regression results, as well as the alternate quadratic formulation

$$FS_i = \beta_0 + \beta_1(FIN_i) + \beta_2(FIN_i)^2 + \epsilon_i, \tag{2.5}$$

reported in table 2.5 can be visualized in the figures on the following pages. For example, figure 2.1 shows the graph of financial wealth versus the level of foreign equity for both the models in equation 2.4 and equation 2.5 as well as the fitted regression line with the 95 percent confidence interval shaded. The figures illustrate the relatively strong explanatory power of financial wealth which can be seen visually as well as evidenced through the high R^2.[22]

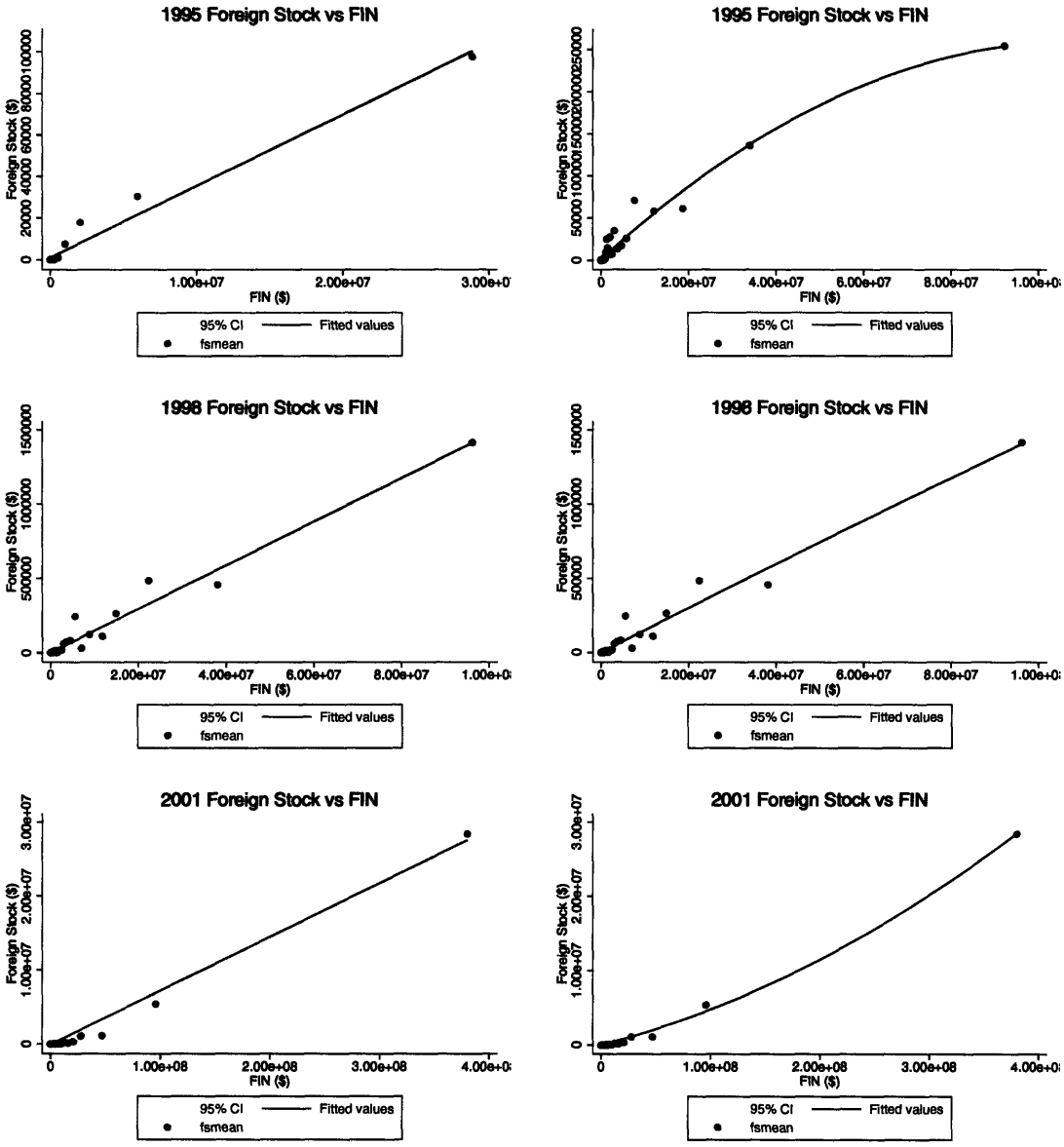
Moreover these results, in particular the rather dramatic changes in coefficient estimates provide strong evidence that our original OLS and IV estimates are not very accurate. OLS, IV, and averaging followed by OLS are all consistent under the standard assumptions and given our relatively large data set, it is unlikely that the estimates should change so dramatically. Given the rather substantial difference in coefficient estimates, we have strong evidence that the consistency assumptions of OLS and IV are not satisfied. This result in particular furthers our believe that this data set and the *FIN* variable exhibit significant errors in variables and thus necessitates the need for more complicated estimation procedures such as we have developed in this section.

[20] We report robust standard errors in the table.

[21] It is worth noting that the other coefficients do not seem to respond in the same way. Still, in the presence of measurement error on explanatory variables, the original regression point estimates reported in table 2.3 are not even consistent so this may be of little practical concern. We do not currently explore the change in the regression coefficients over time.

[22] This result is also robust to logging the data which which results in a more uniform distribution of financial wealth (the original data is nearly exponential in distribution).

Figure 2.1: Level graphs.



The first column is linear models. The second column is quadratic models.

Table 2.5: Averaging OLS regression results

Model		1995	1998	2001
Quadratic	Intercept	391.1065	-614.9783	-31729.07**
		(535.4887)	(2725.747)	(14171.27)
	FIN	0.0047656***	0.0151463***	0.0382529***
		(0.0001819)	(0.0007892)	(0.0016108)
	FIN^2	-2.19e-11***	-5.62e-12	9.63e-11***
		(2.12e-12)	(8.99e-12)	(4.38e-12)

Table 2.6: Averaging OLS regression results by group size

		Year		
Bin Size		1995	1998	2001
5	Intercept	-1142.6175	7911.0542	-77125.7695**
		(3677.528)	(9154.903)	(37633.001)
	FIN	0.006454614***	0.013665547***	0.065552486***
		(0.000358704)	(0.000713654)	(0.002129313)
10	Intercept	-514.903	6903.9853	-95686.9831**
		(1458.548)	(10979.162)	(40786.623)
	FIN	0.005007614***	0.012663343***	0.075210828***
		(0.000138726)	(0.000730178)	(0.001906739)
20	Intercept	1669.2806	12581.1962	-99113.6497
		(1258.341)	(6893.211)	(38171.906)
	FIN	0.003066427***	0.006250884***	0.074315601***
		(0.000165774)	(0.000369908)	(0.001395326)
50	Intercept	1426.1937	14809.7096*	-46235.9569**
		(1099.281)	(8489.548)	(21123.041)
	FIN	0.003161702***	0.005101461***	0.048890705***
		(0.000171369)	(0.000305383)	(0.001784175)
100	Intercept	929.0174	15550.4591	-51119.5422**
		(1014.659)	(11137.517)	(24500.412)
	FIN	0.003656479***	0.004729674***	0.054054914***
		(0.000214935)	(0.000291107)	(0.001599092)

The "Bin Size" refers to the number of observations in each group that are averaged together. This corresponds to m in the notation of section 2.4.2. Each implicate is treated separately; the results are then averaged to produce the results in this table.

Discussion of Results

The averaging approach outlined in the previous section is not without problems. There is a fundamental tradeoff between increased averaging to reduce errors-in-variables and the resulting loss of variability in the transformed observations. In fact, there are at least three effects that we must consider. First, increasing the size of the group can eliminate errors-in-variables due to the law of large numbers since we more observations allow a better estimate of group means. Second, increased averaging reduces the overall variability of the transformed (grouped) data. This problem is particularly acute in our data since most of the variability is in observations with high wealth levels and, since these represent fewer observations, there is a resulting loss of variability in the transformed data. Third, increased grouping reduces the number of data points in the transformed data and hence leads to larger standard errors following estimation. It is clear that groups must be of sufficient size to alleviate errors-in-variables and hence produce consistent point estimates with OLS, but the other two effects both imply too much grouping will not produce good results. Overall, it is unclear the net direction of these three effects. We can, however, illustrate the tradeoffs on our point estimates by looking at the estimated coefficients in equation 2.2 while varying the group size as we have done in figure 2.2. The first panel using the same number of observations per group, while the second uses equal implied population weight per group.

Also, estimating the relationship between wealth and foreign equity investment is somewhat difficult when using levels. It would be beneficial to transform the variables using logarithms and avoid the exponential like distribution of the variables in levels. This transformation is not possible in our data set since, as we saw in section 2.3, almost all of the observations in the data set have foreign equity of zero.

The main result in this section, however, is that we can utilize averaging to produce consistent point estimates of the fraction of the marginal dollar invested in foreign equities. It is worth noting that even in the best case the estimate of this quantity is less than 8 cents on the dollar which, given the linear specification, corresponds to a very small fraction of wealth invested in foreign equities. This result is consistent with the previous literature showing that too little is invested in foreign equities, but given our micro-level data we believe there is more that we can say. In particular, the data illustrates that most households choose foreign equity holdings of zero and, given this, perhaps we can disaggregate our total sample into two parts and look at each part's decision separately as we do in the next section.

2.5 Estimation of Conditional Investment Decision

As we mentioned earlier in section 2.3, only a small number of observations actually have positive foreign asset holdings. We were interested to determine whether we could find characteristics of households that imply households would have positive

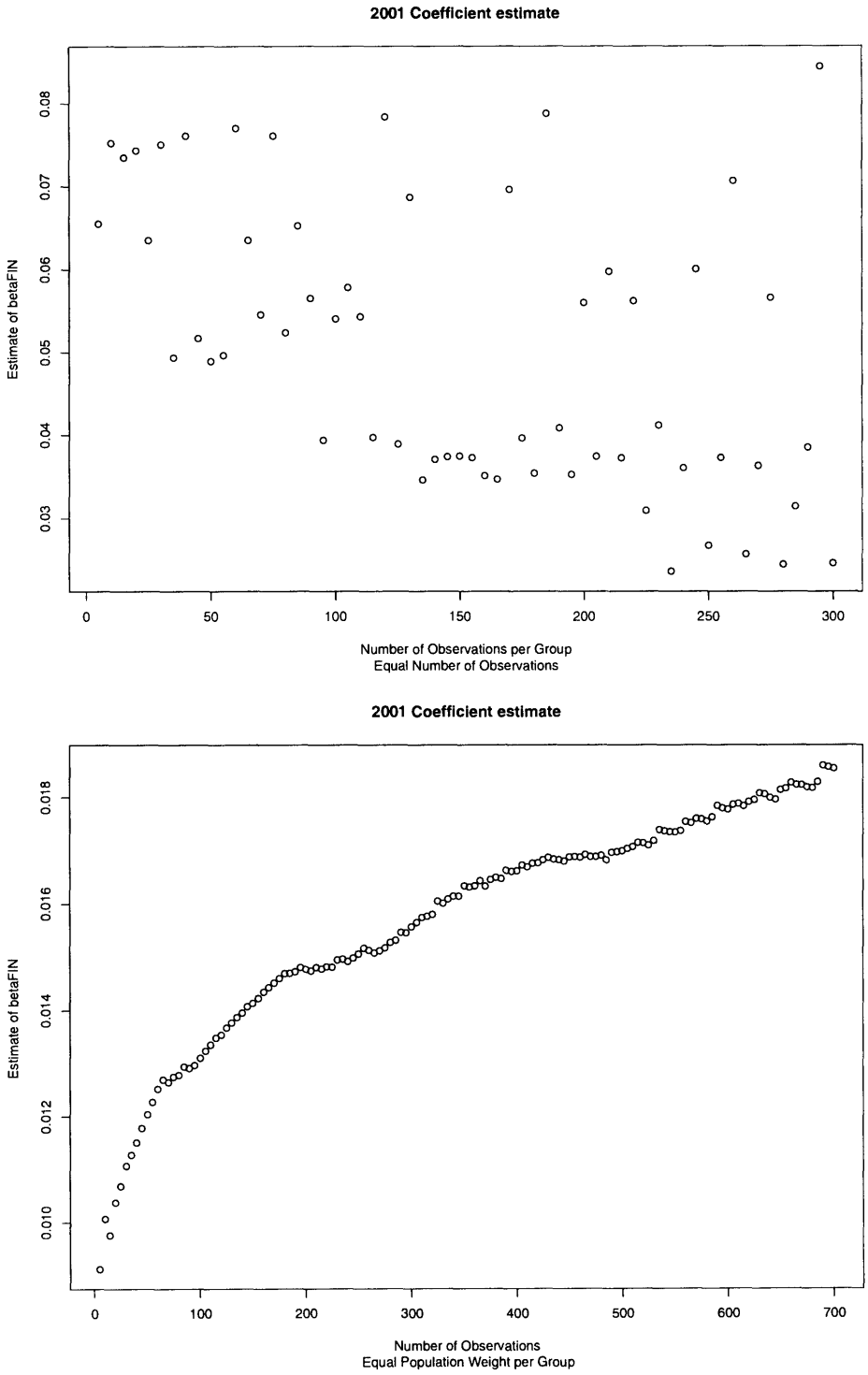


Figure 2.2: Tradeoff between group size and coefficient estimates

Table 2.7: Participation regression results

Model		1995	1998	2001
Linear	Intercept	4.522679***	4.99551***	5.62112***
		(.7132722)	(.7984792)	(9.8008032)
	FIN	6.23e-07***	5.88e-07***	2.81e-07***
		(7.29e-08)	(7.63e-08)	(2.13e-08)
Quadratic	Intercept	3.432893***	3.638101***	4.569739***
		(.57304)	(.6230971)	(.6915298)
	FIN	2.16e-06***	2.12e-06***	8.01e-07***
		(1.95e-07)	(1.80e-07)	(7.86e-08)
	FIN^2	-1.88e-14***	-1.84e-14***	-1.46e-15***
		(2.27e-15)	(2.05e-15)	(2.14e-16)

levels of foreign equity. To that end, we have found that the decision to invest a positive dollar amount in foreign equities is strongly related to financial wealth. Consider the following regression

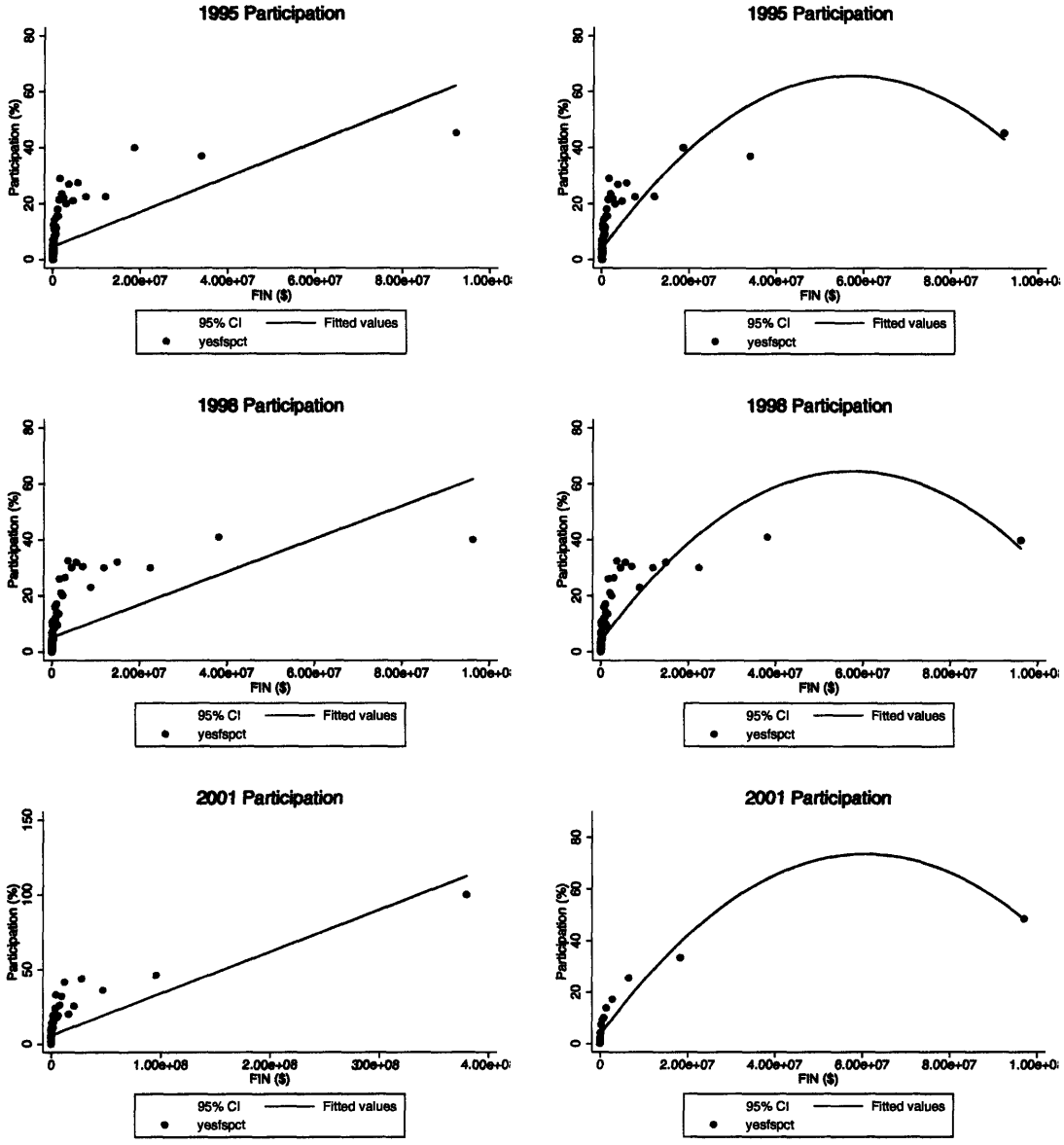
$$participation_i = \beta_0 + \beta_1(FIN_i) + \eta_i \tag{2.6}$$

where *participation* is the fraction (in percent) of a group (in the notation of section 2.4.2) that owns foreign equities and *FIN* is that group's (weighted) mean financial wealth. We report the results in table 2.7 and figure 2.3.[23] The regression results and figures show that participation is greatly influenced by wealth and only high wealth levels induce significant participation. Indeed only in the higher quantiles of the wealth distribution, do over half the quantile's members participate in foreign equity markets. This observation leads to a theory that low aggregate levels of foreign equity holdings are not just a matter of each investor choosing a suboptimal portfolio, but rather the result of many investors failing to participate in the market at all. Hence, perhaps there are significant information costs that investors must incur to participate in foreign equity markets and these fixed costs lead to many investors choosing not to participate at all. This hypothesis leads to the last component of this paper.

As we have established, most households do not participate in the foreign equity market, so the final stage in shedding a new light on the HBP is to disaggregate the international market participation and portfolio allocation decisions. We decouple these choices by only including in the analysis those portfolios which include nonzero levels of foreign equity holdings. We estimate equation 2.2 using this subset of observations and the same averaging approach employed in the previous section in

[23] We also report an alternative quadratic formulation: $participation_i = \beta_0 + \beta_1(FIN)_i + \beta_2(FIN)_i + \eta_i$.

Figure 2.3: Participation Regression Graphs.



The first column is linear models. The second column is quadratic models.

order to properly account for the errors-in-variables problems. As detailed in Table 2.8, we find that conditional investment in foreign assets is substantially higher than the unconditional levels measured in the previous sections. Specifically, in 2001 the conditional investment level is about $.18 or about 2.5 times greater than the unconditional level measured under the averaging approach, and 4.5 times greater than the level measured under the IV approach.[24]

The results in this section, particularly those from 2001, provide evidence that the question of home bias is not so much an issue of levels, but rather one of participation. In 2001, households which actually participated in foreign equity markets (*i.e.* those which made portfolio choices that included positive amount of foreign equity) did so at a rate of 18 cents per dollar. At this rate, those households' foreign equity holdings are more consistent with predictions of standard portfolio choice models. With this result, the home bias question becomes one of participation rather than (conditional) allocation; this, we believe, is our central result.

2.6 Conclusion

This paper is primarily concerned with with foreign asset diversification. Classic results detailing a lack of international diversification relied primarily on macro data; this paper instead relies on micro level data that allows us to look more closely at individual investment decisions. More importantly we consider a richer foreign asset decision rule. This decision rule incorporates the notion that agents first decide whether to invest in foreign assets thereby incurring any fixed costs associated with that action and then contingent on that first decision agents can decide on the level of foreign investment. Using this two-step decision rule and our individual micro data we are able to show that the share of foreign assets is much closer to levels that traditional portfolio theory would predict.

These results however are not consistently strong and moreover the result is only dramatic in 2001. We interpret this as weak evidence that the home bias may be diminishing and more importantly that it may be important to think about foreign asset ownership decisions using more complicated decision rules rather just assuming that all agents are equally able to participate in the foreign equity market.

Furthermore our analysis has omitted some major sources of potential international diversification by not including bonds and mutual funds. The former, although

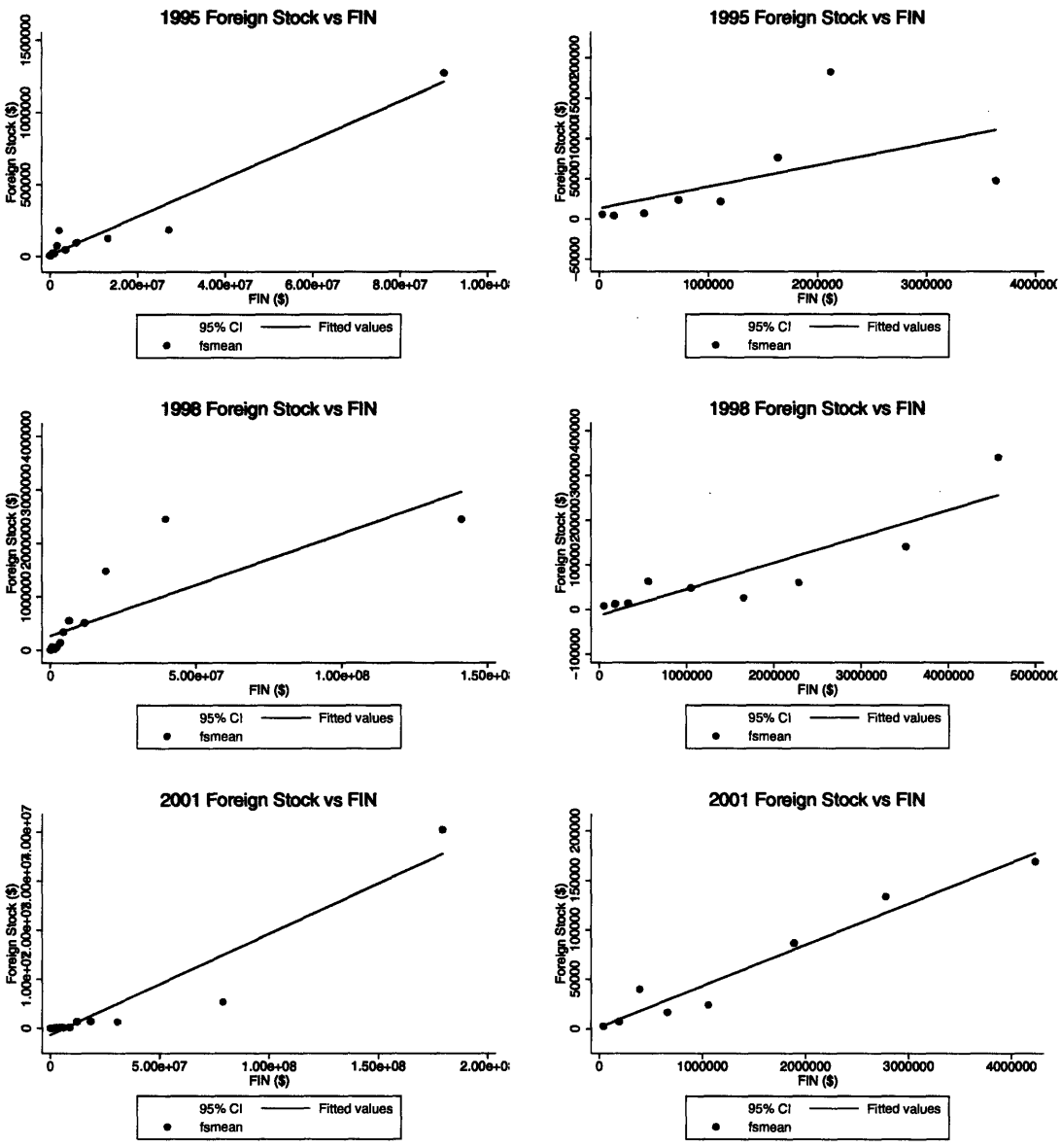
[24]This estimation technique as described in the text has selection issues. Accordingly, one can construct a method to estimate these coefficients taking these selection issues into account. One possible approach is to utilize of a standard Heckman two-stage estimator. First estimate a probit model. Then use the predicted values from this first-stage and average across observations (to correct the errors in variables problem we described earlier in this paper) to estimate the second stage thus producing estimates that take in account the selection issue. Note this approach requires the first stage probit regression to have additional explanatory variables besides financial wealth to have identification in the second stage regression.

Table 2.8: Conditional Foreign Asset Investment in Levels

Implicate		1995	1998	2001
1	Intercept	46131.66**	474336.4*	-1365957
		(17642.96)	(248379)	(946914)
	FIN	0.0073974***	0.007799*	0.2055959***
		(0.0009655)	(0.0043092)	(0.018327)
2	Intercept	47729.69***	-45177.56	-1174927
		(13888.05)	(36474.47)	(905314)
	FIN	0.0075897***	0.0688061***	0.1945395***
		(0.000584)	(0.0011388)	(0.0219754)
3	Intercept	7547.298	397246.2	-941636.5
		(26093.96)	(252567.3)	(759634.2)
	FIN	0.0133184***	0.0106562**	0.1776236***
		(0.0009465)	(0.0042075)	(0.0192358)
4	Intercept	52408.13	266290	-766932.9*
		(30819.98)	(159599.6)	(395787.9)
	FIN	0.0076272***	0.0190496***	0.1572903***
		(0.0018227)	(0.0040166)	(0.0090266)
5	Intercept	75953.85**	-374409.9	-789254.1
		(30627.21)	(313647.5)	(618151.2)
	FIN	0.003985**	0.1034566***	0.1638611***
		(0.0018937)	(0.0072245)	(0.0149347)
Sample Average	Intercept	45954.13	143657	-1007741
		(23814.43)	(202133.6)	(725160)
	FIN	0.00798354	0.0419535	0.17978208
		(0.00124348)	(0.00417932)	0.0166999

Results are levels regression by first eliminating all observations with no foreign assets and then averaging across observations using fixed size groups of 20 observations. This result is from the first set of implicates.

Figure 2.4: Conditional Foreign Asset Ownership in Levels.



The first column is unrestricted linear models. The second column is restricted linear models; we run the regression only on observations with networth below $5 million. The graphs were chosen based on the implicate with the highest coefficient on financial wealth.

in the data included, does little to change our results. The larger omission is that of mutual funds. There is no doubt that this investment vehicle provides a substantial degree of diversification to large numbers of investors, however, we could not include it in our analysis due to data availability. Adequately including mutual funds would require obtaining each household's fund holdings. We do believe though that including mutual funds would only strengthen our results since a larger proportion of individuals who own foreign equities also own mutual funds.

This paper attempts to estimate the propensity of households to invest their financial wealth in foreign equities. First, we note standard OLS estimates are problematic due to significant errors-in-variables problems in the data and the standard approach of using IV is not practical due to invalid instruments. We then construct an estimator by averaging across observations to avoid errors-in-variables and, although these estimates are much larger than the non-averaging approaches, our estimate foreign equity investment levels is quite low. However since we have micro-level data, it is clear than many households are not even participating in the foreign equity market and this fact motivates a two-step investment decsion. We estimate that once households decide to invest in foreign equities, their choices are not nearly as low as implied by aggregate data. The breaking up of the foreign equity investment decision is the central result of this paper.

Bibliography

AHEARNE, A. G., W. L. GRIEVER, AND F. E. WARNOCK (2004): "Information Costs and Home Bias: An Analysis of US Holdings of Foreign Equities," *Journal of International Economics*, 62(2), 313–336.

BOHN, H., AND L. L. TESAR (1996): "U.S. Equity Investment in Foreign Markets: Portfolio Rebalancing or Return Chasing?," *American Economic Review*, 86(2), 77–81.

BOTTAZZI, L., P. PESENTI, AND E. VAN WINCOOP (1996): "Wages, Profits, and the International Portfolio Puzzle," *European Economic Review*, 40(2), 219–254.

FRENCH, K. R., AND J. M. POTERBA (1991): "Investor Diversification and International Equity Markets," *American Economic Review*, 81(2), 222–226.

GREENE, W. H. (2002): *Econometric Analysis*. Prentice Hall, Upper Saddle River, New Jersey, fourth edn.

GRIEVER, W. L., G. A. LEE, AND F. E. WARNOCK (2001): "The US System for Measuring Cross-Border Investment in Securities: A Primer with a Discussion of Recent Developments," *Federal Reserve Bulletin*, 87(10), 633–650.

KENNICKELL, A. (2000): *Codebook for 1998 Survey of Consumer Finances*Board of Governors of the Federal Reserve, Mail Stop 153, Washington, DC 20551.

LEVY, H., AND M. SARNAT (1970): "International Diversification of Investment Portfolios," *American Economic Review*, 60(4), 668–675.

LEWIS, K. K. (1999): "Trying to Explain Home Bias in Equities and Consumption," *Journal of Economic Literature*, 37(2), 571–608.

LINTNER, J. (1965): "The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets," *Review of Economics and Statistics*, 47(1), 13–37.

LUCAS, JR., R. E. (1987): *Models of Business Cycles*. Basil Blackwell, Oxford, UK.

MANKIW, N. G., AND S. P. ZELDES (1991): "The Consumption of Stockholders and Nonstockholders," *Journal of Financial Economics*, 29(1), 97–112.

MONTALTO, C. P., AND J. SUNG (1996): "Multiple Imputation in the 1992 Survey of Consumer Finances," *Financial Counseling and Planning*, 7, 133–146.

SHARPE, W. F. (1964): "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk," *Journal of Finance*, 19(3), 425–442.

STOCKMAN, A. C., AND H. DELLAS (1989): "International Portfolio Nondiversification and Exchange Rate Variability," *Journal of International Economics*, 26(3/4), 271–289.

WALD, A. (1940): "The Fitting of Straight Lines if Both Variables are Subject to Error," *The Annals of Mathematical Statistics*, 11(3), 284–300.

WARNOCK, F. E., AND C. CLEAVER (2003): "Financial Centres and the Geography of Capital Flows," *International Finance*, 6(1), 27–59.

WOOLDRIDGE, J. M. (2002): *Econometric Analysis of Cross Section and Panel Data*. The MIT Press, Cambridge, Massachusetts.

Chapter 3

Using Identity in Principal-Agent Models without Performance Measures

Abstract

In this paper, we consider principal-agent models of organizations in which the principal has no way to measure the output nor the effort level of agents. To model this situation, we appeal to utility models that include identity, justified in part by empirical results from peer-effects, and apply these extended utility functions to the organizational forms we have assumed. We find that in the single agent case the introduction of identity amounts to modifying the utility function and does not lead to dramatic results. In the multiple agent case, we find that the addition of identity can lead to more efficient outcomes than cases where identity is ignored. The addition of identity, however, can also lead to counter-intuitive results due to the interactions among agents and can also produce second-best outcomes that are worse than the case without identity. Finally the addition of identity can help explain some empirical results that may be difficult to explain with standard models.

3.1 Introduction

In the standard literature on organizations and incentives, principals and agents are generally assumed to possess different information: we generally assume the agent is better informed than the principal. Yet often times, even if the principal cannot verify the information or the actions of the agent, he can still implicitly contract on those actions since he can write contracts contingent on realized verifiable outcomes as in the standard moral hazard model where effort is unobservable, but output is verifiable. This paper differs in a substantial way from the standard literature by

instead assuming the principal cannot observe anything regarding the agents actions or the output associated with those actions and thus implies the standard solutions for moral hazard models will no longer be implementable.

The obvious question, then, is if this assumption on observability is sensical. Thus we put forward as the motivating example for the present work the case where managers have little information or expertise in the area in which they are tasked to manage. There are many examples of organizational structures in this vein and often services within an organization are representative examples. Consider most of the service operations of an enterprise including electrical, plumbing, and information technology. Managers may know little to nothing about the relevant area yet are still tasked with motivating and controlling these entities. Since they have little, if any, knowledge of the actions of their subordinates, most types of performance evaluation are essentially useless. Furthermore, since these groups are often service-orientated their effect on observable outputs like production and profitability are difficult to measure. There may, of course, be very coarse performance measures such as whether the lights are on or not, but, as we will see, this can be used to provide essentially no incentives.

If standard incentive contracts cannot be used in these situations, we must turn to other alternatives in order to achieve better second-best outcomes. In fact, by allowing a richer model of utility drawing on both evidence from peer effects and the theoretical modeling of identity, we can construct examples where there are substantive results even without performance measures and incentive contracts. In particular, we will see how adding additional workers or by allowing the principal different actions can affect the incentives of the workers without measuring their performance and still yield, in some cases, efficiency-enhancing results. Of course this result is not universal and the addition of these peer-effects could also lead to worse outcomes, a possibility that is not ruled out in this framework.

The paper proceeds as follows. First, in the next section, we discuss the empirical and theoretical underpinnings of this paper, namely the literature on peer effects and identity. By reviewing the relevant literature, we hope to connect the peer-effects and identity literature together. The section begins by providing empirical evidence for the existence of peer effects amongst agents and then by introducing identity, we will provide a mechanism for which peer effects can be introduced into organizations. Section 3.3 develops the basic model we use throughout the paper. The first part of the section discusses a basic two person, employer-employee relationship looking at how identity can effect the relationship and how each player can affect that relationship. The section then proceeds by adding additional agents and looking at the effect of the identity and effort choices of the players as well as how the additional players will affect the choices the principal must make. The section formalizes and solves various formulations of the model as well as providing discussion of the results. Lastly, section 3.4 concludes.

3.2 Background and Related Literature

Before proceeding to develop a model and discuss its implications, we must first develop and establish the existence of organizations that can be ascribed with the assumptions we are making. Furthermore, we must provide some empirical background to justify the choice in modeling as well as the existence of the effects we rely on as a mechanism that can induce higher effort. For the first task we will rely on mostly anecdotal evidence and posit organizational structures that meet our needs. For the second task we will rely on the economic literature.

3.2.1 Motivation

To begin with, by operating under the assumption that performance measures do not exist, the relevant organizations must be fairly large and therefore the actual group that we will model is a subset of the entire organization. A small organization simply cannot have a lack of performance measures since the firm's profit provides a very clear indication of performance and thus an easy way for owners to provide incentives for all workers.[1] Thus we will generally assume that the organization of interest is a small group in a larger enterprise. This small group cannot produce anything tangible or easily measurable though, since if, for example, the firm could pay workers in the group a piece-rate, we are in the case of an easily measured performance and thus the possibility of standard incentive contracts. Also, the supervisor or manager of the group must have significant authority over the group, but have little ability or knowledge to effectively monitor the group's effort or output. The manager, who we will refer to as the principal, should have preferences that are aligned with the firm and have the power to set wages and make hiring decisions for his group. Lastly the group should have a positive effect on the firm's total output, but that effect should be extremely difficult to independently measure. Again, if output is easily measurable, the assumption regarding the lack of performance measures would not be believable.

Given the criteria we have laid out, do such organizations actually exist? We believe that although this case is not the prototypical organization, there are many examples of groups that meet most of these criteria. First, consider typical service-type groups within larger organizations. These would include divisions that perform maintenance of equipment or facilities, food service, as well as information technology groups. These organizations are certainly necessary to the function of the overall firm, yet in many cases the exact manner in which they affect firm profitability is

[1] What is not clear, however, is what the appropriate threshold is for a small enterprise. For example, a firm with 10 employees is certainly small, yet it may be possible for that organization to exhibit similar organizational problems that we discuss in the text. On the other hand, a firm with only 3 employees would prove very difficult to meet the criteria of the model. We generally will assume firms are larger organizations employing more than 10 employees, yet nothing should be construed to rule out applications to smaller firms if the other criteria are met.

extremely difficult to measure.[2] Furthermore, the supervisor of these groups is often ill-equipped to understand exactly the nature of his subordinates' tasks or is too time constrained to adequately monitor their actions; supervisors can often tell that the workers are doing something, but cannot differentiate productive work from simply acting as though one is working.

Moreover other examples where performance measures are nonexistent or extremely coarse can arise outside of traditional firms and can be evident in other contracting relationships. Consider the case of a homeowner hiring a plumber, a carpenter, an electrician, or other trade, the homeowner can often tell that the job is complete, yet ascertaining the quality of the work or the effort the tradesman expended on the work can be quite difficult. While these examples do not provide an exhaustive list nor establish with proper empirical rigor the existence of organizations we have described, we have shown that groups exist that at least exhibit many of the aspects we require in the model and, having done so, we must proceed to provide some underpinning for the choice of modeling.

3.2.2 Related literature

The models in this paper rely in large measure on the ability of agents to influence other agents' behavior through non-contractual means and thus we require both theory and evidence to support the existence of these means to influence agent behavior. We will rely on two strands of literature to support these models. First, we will look at the peer effects literature and second, we will look at the recent literature on identity. In both cases, we will look at theoretical as well as empirical motivations that can provide basic support for the formulation we will use in the present paper.

The peer effects and social interactions literature is mostly an empirical strand of literature that tries to measure what effects peers have on specific outcomes.[3] Although it is often difficult to measure peer effects due to self-selection and other issues as discussed in Manski (2000), there are many papers that often focus on education, housing, and crime. For example, Sacerdote (2001) uses random roommate assignment at Dartmouth to identify peer effects and he finds that there are some peer effects in GPA as well as in fraternity membership. Katz, Kling, and Liebman (2001) finds that families who received randomized housing vouchers and moved out of high-poverty public housing experienced improvements in multiple measures of

[2]For the IT example, certainly not all IT services would be applicable (like uptime of crucial services), but consider, for one example, desktop support functions and efficiency in making purchasing decisions which are both difficult to measure, yet will affect firm profitability.

[3]For an example of some recent theoretical work on peer-effects, see Battaglini, Bénabou, and Tirole (2005). There is also some older theoretical literature in this vein. For example, Kandel and Lazear (1992) looks at peer pressure in the context of partnerships. At first glance, partnerships would seem to lead to free-rider problems, but the authors introduce peer pressure, shame, guilt, norms, and monitoring to combat this effect. They find peer pressure can have an effect. Their paper differs from ours by not including the psychological concept of identity and allowing that choice to be endogenous.

well-being. Glaeser, Sacerdote, and Scheinkman (1996) uses a model that includes social interactions to help explain the large variance in crime rates across cities which does not seem to be well-explained by other socioeconomic factors.

The existence of social interactions and peer effects is, of course, not restricted to those types of group interactions. For example, Hong, Kubik, and Stein (2004) finds that "social" investors are more likely to invest in the stock market when their peers also invest. There is also some experimental evidence of peer effects. Falk and Ichino (2006) constructs an envelope filling experiment and identifies peer effects since the authors find that when workers are "paired," by working independently in the same room, their output is more similar than when they work in separate rooms. Last, and most important for this paper, there are examples of peer effects in organizations.[4] Jones (1990) looks at the Hawthorne Studies and concludes that workers' productivities were indeed interdependent. In other words, he finds evidence of peer effects. For example, the author finds positive interdependencies between certain pairs of workers. Another compelling example of peer-effects in organizations comes from Ichino and Maggi (2000) which analyzes shirking behaviors in a large Italian bank. Empirically there is a large difference in behavior between employees in the branches in the south versus branches in the north; there is significantly more shirking in southern branches. The paper proceeds and suggests several contributory factors including the workers' backgrounds, sorting, and, most relevant to this paper, group interactions to help explain this differential.[5]

While the peer effects literature we have discussed does provide empirical evidence as to the existence of social interactions or peer effects, it often does not propose mechanisms for the substantial empirical effects they measure. For example, Katz, Kling, and Liebman (2001, page 612) suggests "positive social interactions ... can arise from learning from peers, pure preference externalities ..., stigma effects ..., and physical externalities ...", which is clearly a wide variety of mechanisms. To tractably model the problem we have discussed in Section 3.2.1, we must settle on a mechanism and a more formal model. One approach we could use to introduce social interactions involves the altruism model of Rotemberg (1994). While that model allows social interactions through signals and can produce outcomes approaching the efficient outcome, its approach involves signals and, moreover, cannot cause individuals to choose higher effort unless there is some interdependence in production. While perhaps an extended version of Rotemberg (1994) may provide a starting point, we will proceed down a different avenue since social interactions can also arise in the recent literature on identities. This recent work does, in fact, describe a

[4]There are also several reviews and surveys of social interactions and peer effects, see Glaeser and Scheinkman (1999), Brock and Durlauf (2000), Moffitt (2001), and the previously mentioned Manski (2000).

[5]Mas and Moretti (2006) also finds the existence of peer effects for cashiers in a national supermarket chain. The authors find that introducing high productivity workers causes other workers to increase their own efforts.

mechanism for altering preferences that includes group interactions. This literature uses psychological insights to provide better utility formulations that allows for better models of human behavior; these extended utility formulations can include "identity." By providing a concept of identity, we have a method that allows an individual's utility function to be affected by others in a tractable manner and thus provides much of the theoretical groundwork for this paper.

The economic concept of identity is first developed in Akerlof and Kranton (2000). The authors recognize that traditional utility based models miss important aspects of human behavior and, with the help of psychological theory, seek to develop a more complete model when human interactions are needed. In particular, the authors develop the notion of "identity" and "category." The former concept refers to how an individual looks at himself; it is his self-image. The latter term refers to how we can represent an individual's identity by allowing individuals memberships in social categories; these are the groups that an individual can identify with as well as the groups that he can view others as members of. In the simplest case, the authors propose supplementing utility functions to include identity, that is an individual's utility is a functions of both actions and the individual's self-image. By including identity, the utility function can capture the individual's view of himself as well as how he perceives his actions will affect others' view of him. This last subject captures the notion that categories have associated with them certain prescriptions or actions and thus allows the utility function to capture the norms associated with various social categories. In sum, the addition of identity allows richer interactions between individuals as well as encompassing the individual's own self-image. The addition of identity allows a more complete utility function for individuals that incorporates extensive work in psychology.

While adding identity to utility functions in the manner of Akerlof and Kranton (2000) could be done in standard organizational contexts, the same authors in Akerlof and Kranton (2005) extend their model of identity and specialize it by applying it to organizations. The authors argue that standard monetary incentive schemes often have problems in real organizations and therefore well-functioning organizations should not rely solely on them. The authors argue that workers should be assigned to a job to which they can identify and the addition of these identities is crucial. The authors go on to develop a simple formulation of a utility function that includes identity by allowing workers to be one of two categories: an insider that identifies with the firm and exerts high effort or an outsider that does not identify and consequently exerts low effort. They introduce a simple utility function to characterize this situation

$$U(y,e;c) = \log y - e + I_c - t_c|e^*(c) - e| \tag{3.1}$$

where y is the individual's income, e is his actual effort, c is his social category, I_c is the identity utility associated with that category, $e^*(c)$ is the ideal or norm for a member of category c, and thus the final term in the utility function penalizes a

worker if he deviates from the social norm. By using this utility formulation which captures the notion of identity, the authors look at the effects on standard monetary incentive schemes and find that identity can be a complement or a substitute to monetary incentives depending on how identity interacts with the agent's cost of effort. They also mention that firms may be willing to invest to change a worker's identity as well observing that identity becomes more important if effort is hard to observe. Both of these observations we will address to some degree later in this paper. In the next section, however, we will take the basic model of identity proposed in equation (3.1) and begin to develop a tractable model of the interactions that exist under the circumstances outlined in Section 3.2.1.

3.3 Basic Model

In this section, we will first discuss a basic principal agent type model that introduces the concept of identity which allows us to develop some of the basic results as well as contrasting our approach to the results of classic models. We first address the one agent case and then proceed to multiple agents. In each case, we address the resulting equilibriums of these models as well as addressing some possible extensions incorporating slightly different assumptions.

3.3.1 One-agent case

We will start with the simplest model, that involving a single principal and a single agent. As we have mentioned in both Section 3.1 and Section 3.2.1, the principal is unable to measure the performance of the agent objectively nor subjectively and therefore has little ability to use incentives to solve informational problems caused by the non-aligned objectives of the principal and the agent.[6] Although he has little ability to provide incentives to the agent, the principal and agent are, as usual, assumed to be in at least partial conflict. It is in the principal's interest for the worker to exert a higher effort level than the worker would do so under fixed price contracts since the agent cannot realize all the marginal gains of increased effort and hence works at a socially inefficient level. To add concreteness to this setup consider the following formulation.

Let the principal have a utility function $U_P(\cdot)$ that is increasing in the agent's effort level and decreasing in the agent's wage, thus

$$U_P(e,w,\zeta) = f(e,\zeta) - \kappa(w) \tag{3.2}$$

[6]Although we will carry this assumption throughout the paper, this could be weakened to allow the principal some ability to verify the outcomes resulting from the agent's actions. In this case, there would be interactions between monetary incentives and identity. Without specifying exactly how identity affects the agent's effort choices, we cannot compute comparative statics for the identity model versus a model without identity. We can say, however, the identity will matter.

where we assume e is the agent's effort choice and w is the wage paid to the agent.[7] The function $f(\cdot)$ is increasing and concave in e and $\kappa(\cdot)$ is increasing and convex in w. We will continue with the assumption we developed earlier where our organization is a subunit of a larger organization and the output, $f(\cdot)$, for which the principal derives utility includes the effort choice of the agent as well as a potentially random and uncorrelated vector ζ of other determinants of the firm's output. The agent is assumed to be the standard risk-adverse type with utility increasing in w and decreasing in e with utility function $U_A(\cdot)$ given by

$$U_A(w,e) = u(w) - c(e)$$

where e and w are defined as before and $u(\cdot)$ is increasing and concave while $c(\cdot)$ is increasing and convex. These assumptions, along with those restrictions on the principal's utility function, assure there exists a socially optimal effort level e^*. Notice with a fixed wage contract and $e \in [\underline{e}, \overline{e}]$ the agent would always choose $e = \underline{e}$.[8]

As is evident in this simple framework, the inability of the principal to engage in incentive contracts precludes the agent from ever exerting high effort—as long as the principal has no signal regarding the agent's choice he cannot provide monetary incentives to influence the agent's behavior and hence the agent will exert the low effort level. To resolve this low-effort outcome, consider the case when we have heterogeneous agents each having an identity as we discussed in the previous section. In that case, the agent's utility function can be represented differently and in particular we modify the agent's utility function by including several other terms. To capture the notion of identity we add an identity function which captures the added utility the agent receives from choosing a particular category and a penalty function that reduces the agent's utility if his effort choice—his behavior—differs from the effort choice associated with his chosen category. Specifically, when agent i is of category c_i and the other individuals he interacts with have identity c_{-i}, let $I_{c_i}(c_{-i})$ be the identity function, or the added utility the agent receives when his identity is similar to or different from other individuals. Second, let $\psi(e^*(c_i), e_i)$ be the penalty function where $e^*(c_i)$ is the effort associated with i's identity and e_i is agent i's actual effort choice. The penalty function should be increasing in the distance between the proper action and the actual action and hence penalizes the agent when he performs an action that is not associated with his identity. Thus agent i's utility function is

$$U_i(w, e_i, c_i, c_{-i}) = u(w) - c(e_i) + I_{c_i}(c_{-i}) - \psi(e^*(c_i), e_i). \tag{3.3}$$

When the agent has an identity in this manner, his effort choice can be higher than the choice made by an agent with no identity. If an agent is of a category whose

[7] For now we shall assume the wage is constant, we will come back to this point later.

[8] Assuming that the effort choice is restricted to some range captures the assumption that principal would notice if the agent is doing nothing; $\underline{e}$ is the lowest effort choice that the principal will not realize the agent is shirking.

behavioral norm $e^*(c_i)$ exceeds $\underline{e}$, then that agent's effort choice will generally exceed $\underline{e}$ given, of course, the penalty is severe enough to overcome the convexity of the effort cost. The resulting effort choice then gives a strict improvement to the principal's welfare as the wage is fixed. Thus, in this case, the addition of identity benefits the principal.

Unfortunately, one drawback to the method we have just described arrives when the agent has an identity that is detrimental to the principal. Suppose the agent's identity has associated with it an optimal effort choice of $\underline{e}$, that is the agent is penalized if he works harder than $\underline{e}$, in other words, his category is that of a bad worker.[9] Clearly this is not beneficial to the principal and, in fact, leads to the same outcome that would occur if the agent had no identity at all. Furthermore, and even worse for the principal, identities of this sort would lead to less optimal outcomes even if signals existed as to the agent's effort choice as per standard moral hazard models and would therefore require stronger monetary incentives from the principal to induce higher effort choices.

The preceding result shows that identity in and of itself is not strictly utility enhancing for the principal and indeed it could leave him strictly worse off in the case of observable outputs. Thus in order to be useful to the principal, he must screen potential agents to ascertain their identity before entering into a contract. If the principal can successfully identify those agent's with a "good" identity, he can increase his utility from screening workers based on their identity and only hire good workers. This, of course, assumes that identity is something intrinsic to the agent and not part of the agent's choice set, in fact, we will assume that identity is in the choice set in section 3.3.2. Alternatively, if we assume that identity is not a choice variable, but instead if there existed ways in which to modify a worker's identity and the principal could utilize those methods, the principal would be willing to pay for this enhancement as it can be utility increasing for the principal. Thus in this very simple framework, the addition of agents who have an identity component to their utility functions can lead to better second-best outcomes.[10] On the other hand, adding identity with a single agent is outcome equivalent to modifying the agent's utility function and hence provides little theoretical interest. Moreover, if output is neither measurable nor observable, this leads to a question of the functioning of screening potential workers based on their identities. If the principal cannot monitor the agent, how could he ascertain the agent's identity? The answer, of course, stems from the definition of identity. Identity is not simply the agent's type in the sense that it measures the agent's efficiency, cost of effort, or ability, as typical in hidden information models, rather it is something less tangible, more like the agent's enthusiasm for the job, demeanor, or passion. These traits are probably difficult to measure, but may not be

[9]Notice this case could also arise if the norm exceeds $\underline{e}$ since insufficient penalty functions could still lead to utility maximization at $\underline{e}$.

[10]Of course if one really believes that identity is a key component of utility functions, then one should not compare outcomes with identity to those without as the latter are therefore incorrectly modeled.

unobservable as in the case of output in the framework of this paper.[11]

Although the addition of identity in principal-agent type models leads to some results in the single-agent case, it becomes substantially more interesting when we add additional agents. The addition of additional agents allows interactions among agents and hence allows agents to influence other agents' behaviors. Although with one agent, that agent's identity can influence behavior through his utility function, from an organizational standpoint identity is just another component of the utility function. With multiple agents, however, identity can play a larger role since agents' choices can lead to externalities among the other agents. These peer effects can cause different choices than if we ignore agents' identities.

3.3.2 Multiple agents

To proceed formally, we shall again assume that the principal has the utility function from equation (3.2) and still cannot measure the agent's effort nor has any indication of the agent's output. Thus we are still in the situation where conventional incentive contracts based on performance measures fail. Now suppose that agents have utility functions that incorporate identity; that is, all agents will have higher utility if they perform the "proper" action associated with their respective identities as in equation (3.3). First, consider if agents' identities are fixed throughout the game, then there is little change from the one agent model—principals for whom higher output is of marginally higher value will spend more on screening to hire agents that have identities associated with higher effort if screening is possible.[12] If, on the other hand, agents' identities can change when they interact with their coworkers we can have much more interesting dynamics since interactions between agents' identity choices can lead to different effort choices.[13]

To continue, we shall assume that agents identities are not fixed and indeed for now will assume that an agent chooses his identity as part of his individual optimization program. Some words need to be said regarding this choice of modeling. Allowing agents a choice of identity presents a certain view of identity that differentiates it from the standard notion of "types" in hidden information models, in this view identity is not an unchangeable endowed characteristic of an agent, rather it is a choice that

[11]This naturally leads to the possibility of implementing subjective performance evaluations using identity. In this paper, we hold to the assumption that performance measures are not possible, but allowing identity screening may allow for implementing incentives using subjective evaluations on identity.

[12]Notice that our utility formulation will not lead to the agent choosing a different effort when there are multiple agents compared to when there is only one agent unless the penalty function is changed. If the penalty function is fixed (the identity function is fixed since each agent's category is fixed) and the wage is constant, additional agents will not affect existing agents' effort choices.

[13]Looking at Nash equilibria in choosing social categories and thus categories being chosen endogenously is developed in Shayo (2005). For another example with an application in voting and redistribution, see Lindqvist and Östling (2006).

can be optimized. It is not clear this is always the right choice. For example, if we consider identity to be something like race or gender, then there is clearly not a choice involved. For this model we do not view identity in this manner and instead consider it groups to which agents may decide to identify with.[14] For example, whether a worker identifies with the goals of the company or if an employee is a team player and identifies with his coworkers. These formulations of identity are changeable and thus can conceivably be part of an agent's choice set.[15] [16]

Now assuming that agents can choose both their identity and their effort choice simultaneously, we can analyze the outcomes of the resulting game. To simplify, we will begin with two agents which we shall identify as agent A and agent B. Both will have the same utility function as in equation (3.3), but allowing the functional forms $u(\cdot)$ and $c(\cdot)$ to be agent-dependent. Therefore each agent solves

$$\max_{e_i, c_i} \quad U_i(w, e_i, c_i, c_{-i})$$

for his optimal effort choice $\hat{e}_i$ and category $\hat{c}_i$ holding the other agent's choices fixed. Clearly with the assumptions we have made, the joint solution of these optimization problems is the Nash equilibrium of the game and gives the resulting identity and effort choices of the agents.

The important aspect of this formulation and the resulting equilibrium is the fact the adding or subtracting agents can affect a given agent's effort choice even if his wage or utility function remains unchanged; interactions between agents are important. This has another effect, of course, as the principal must consider that the hiring or firing of agents will affect the effort choices of existing agents and thus hiring decisions are not the simple optimization problem of marginal revenue and marginal cost that is normally considered.

To illustrate this we shall consider a slightly simplified game where agents can choose from a limited number of categories. In particular, we will assume that agents can be one of two categories: G or B; for good workers and bad workers respectively. These identities have associated with them actions and we will assume the action associated with the category G is e^*, the efficient choice of effort, and that $\underline{e}$ is associated with category B thus justifying the notion of a good or bad worker. We will further assume that $e^* > \underline{e}$ so the efficient effort level exceeds the minimum effort. At this stage, we must make additional functional form assumptions on the identity function and the penalty function. We will assume the utility agents receive from

[14] In Akerlof and Kranton (2000) both fixed and changeable categories are considered.

[15] Beyond making identity changeable, we are making another assumption that the agent actually makes the choice rather then identity being imposed. For example, a worker could be indoctrinated into believing in a certain mantra rather than optimally choosing to conform to that identity.

[16] Beyond considering the changeable nature of identity, we can extend the model of identity choice and allow a richer heterogeneity amongst agents by introducing a type-dependent cost to choosing a given identity. For example, a naturally lazy worker would have a higher cost of choosing an identity that requires high effort.

choosing a particular category is determined by the number of other individuals who are also in that category and thus

$$I_{c_i}(c_{-i}) = g(\#\{\text{other individuals of category } c_i\})$$

where $g(\cdot)$ is a convex function of c_i, the ith agent's choice of identity. This identity function formalizes the fact that agents derive utility when they work with like-minded individuals. For the penalty function we shall assume a simple quadratic loss function, so

$$\psi(e_i, e^*_{c_i}, t_{c_i}) = t_{c_i}(e^*_{c_i} - e_i)^2$$

where e_i is the ith individual's effort choice, $e^*_{c_i}$ is the ideal effort for an agent of category c_i (this may not be the efficient choice of effort e^*), and t_{c_i} is a category dependent parameter determining the level of disutility stemming from not acting as per the norm for that category. As usual, the penalty function gives disutility to the agent proportional to how far he deviates from the norm of his social category. Given our assumptions regarding the norms of the two identities, we can already say something about the parameters of the penalty function and the effort choices of the agents. Clearly if e^* is the norm for good workers, then it must be the case that the action e_G chosen by workers of category G satisfies $u(w) - c(e_G) - \psi_G \geq u(w) - c(e) - t_G(e^* - e)^2$ where $\psi_G \equiv t_G(e^* - e_G)^2$ for all $e \neq e_G$ otherwise agents in category G would choose a different effort. We will assume that t_G is large enough to imply $e_G > \underline{e}$, but since the cost of effort is increasing and convex, $e_G < e^*$.[17] A similar argument places no restriction on the parameter t_B since $\underline{e}$ minimizes $c(e)$ over the agent's choice set.

We can illustrate the agents' choices just described graphically. Figure 3.1 on page 99 represents the normal form of the game in which the agents participate. By carefully choosing the size of the penalty parameter we have effectively turned both agents decision into a strictly one-dimensional choice of identity. There are two possible Nash equilibria: that of both agents choosing category B, the bad worker identity, or of both choosing the good worker identity G, however the good worker equilibrium is only possible if the benefit from working with like-minded workers is large enough; specifically if $I(1) \geq c(e_G) - c(\underline{e}) + \psi_G$, that is the benefit from working with similar coworkers exceeds the marginal cost of exerting more effort as well as the penalty associated with the good worker deviating from the efficient output choice.[18] There are also several observations worth noting in this example before moving on to the principal's problem. First, this game, as constructed, amounts to little more than a coordination game since the agents derive additional utility from matching their identity choices. Second, if the agents could coordinate on an outcome, they would

[17] Of course by making t_G arbitrarily large we can make e_G as close to e^* as desired. The key point, of course, is only that e_G is preferred by the principal since $e_G > \underline{e}$.

[18] As we stated earlier, this penalty can be made arbitrarily small by value of the penalty parameter t_G.

		B: G	B: B
A: G		$-c(e_G)+I(1)-\psi_G, -c(e_G)+I(1)-\psi_G$	$-c(e_G)-\psi_G, -c(\underline{e})$
A: B		$-c(\underline{e}), -c(e_G)-\psi_G$	$-c(\underline{e})+I(1), -c(\underline{e})+I(1)$

Figure 3.1: Normal form of identity and effort choice game for two agents.
Agent A's choice of identity is on the rows while agent B's are on the columns. Each cell represents A's payoff followed by B's. Let $\psi_G \equiv t_G(e^* - e_G)^2$ be the penalty associated with the good worker deviating from the efficient choice of effort. Also, the payoffs in the table are excluding the utility from the wage, that is agents actually receive the payoffs in the table plus $u(w)$.

choose the bad worker equilibrium as it is surplus maximizing.

Now that we have characterized the agents' choices and developed some results in a simple two agent case, we will return to the principal and consider his choices. Insofar as providing incentives to the agents, the principal is helpless; holding to the assumptions we have made throughout the paper, the principal cannot measure output nor effort and therefore offers fixed wage contracts. The principal, however, has already made several decisions that affect the form of the organization namely the choice of which agents to hire and how many agents to hire.[19] With respect to which agents to hire, we have already discussed the role of pre-hiring screening of employees, but now that task becomes more difficult since agents can choose their identities endogenously and simultaneously making employee interactions extremely important. Thus we will not consider the principal's choice as to which agents to hire, but now, instead, we will consider the other choice the principal makes, that is, the number of workers to hire.

To make a simplifying assumption, we will consider a very simple case for the principal: whether to hire an additional worker or not. Theoretically the question is quite simple for the principal as he will hire additional workers as long as the increased profit from the worker exceeds the cost of hiring that worker or, in our case, if the value of additional output exceeds the wage rate paid to the worker. There are, however, significant differences between that simple view and the assumptions we have made thus far. First, it is unclear before the worker is hired and, under the assumptions we have made regarding observability of information, even after the hiring how the principal can judge the effectiveness of the hire. More importantly, there is a discontinuity with respect to the hiring of an additional worker since the identity choice of the additional worker will affect both his own effort choice as well as any previously hired workers' identity choices.

To formalize this situation and analyze its outcome, consider the following model. We shall assume the principal is restricted to paying fixed wage contracts and, given this restriction, the cost of hiring n workers is nw where w is the wage paid to each

[19] For now we are assuming a static game which will disallow the principal from trying to modify a workers identity during the game.

worker. Furthermore we shall assume the principal's utility is

$$U_P(\mathbf{e},\zeta,n,w) = f(\mathbf{e},\zeta) - nw$$

where $\mathbf{e}$ is the vector of effort choices made by the agents the principal employs. There are clearly several cases that can confront the principal depending on the identity choice of existing workers and the number of workers. For the simplest case, we will assume the principal is currently employing one worker and is considering hiring an additional worker. We will assume that it is actually economically beneficial to hire an additional worker so $f'(\mathbf{e},\zeta) \geq w$ at $e = e_G$, but it is inefficient at $\underline{e}$. Even with this assumption, though, the actual value of the worker depends not only on his effort, but also his identity choice, as well as the identity choices of the existing workers. Notice with the assumptions we have made regarding the worker's utility function and choice set, the existing worker is necessarily choosing the bad worker identity since $c(\underline{e}) < c(e_G)$ and thus it is actually economically inefficient to be employing that single worker.[20] Now if the principal hires an additional worker and we assume $I(1)$ is large enough, the agents may reach the good worker Nash equilibrium implying it is economically efficient to employ both workers even if it was not efficient to employ the first one. This rather strange result stems from the discontinuity created by the externalities between the agents which can lead to rather surprising results such as this.

Although the results we have just seen are somewhat counterintuitive using standard theory, the addition of a richer utility framework allows these more interesting results to arise. In particular, allowing interactions among agents gives rise to equilibrium outcomes that significantly complicate the principal's task. As we saw in the very simple model we have developed, adding additional agents can cause changes in the effort choices of all the existing agents. In particular, the addition of one agent can lead to many agents changing their effort choice. This result can even in some cases justify hiring more agents than seemingly necessary since the additional agents may be influencing the effort choices of each other. Given the central role that identities can play in interactions among agents and the resulting effect on firm profitability, the question naturally arises how the principal can take advantage of these identities to increase his payoff.

As we have mentioned previously, one method the principal can use to increase his payoff is to screen employees. We mentioned previously that screening employees based on identity may not be impossible since some characteristics associated with identity may be observable. Using the simplifying assumptions we have made so far in our examples, this screening would, of course, be impossible since agents optimally choose their identity and are homogeneous. Richer models, however, could introduce heterogeneity among agents through the identity function, the penalty function, or by

[20] Of course the principal does not realize that employing the worker is profit decreasing since he cannot measure the output nor the effort of the agent.

allowing agents to have an existing identity and by allowing identity changes only at some cost. If an agent must bear a cost to switch identities, then there is some room for screening agents as agents with better initial identities will be more valuable to the principal since they would increase the likelihood that the agents would attain the good worker equilibrium.

Since we allow identity to be a choice, another possible method to modify identity comes from the principal expending effort in order to influence the agents' choices of identity. These methods could lead to changes in both the identity function $I(\cdot)$ and the penalty function $\psi(\cdot)$ which, in many ways, amounts to rewarding good behavior and punishing bad behavior. Just as in the case of screening it is not exactly clear the correct method the principal should use to influence these functions. Identity arises because the individual identifies with something so one method of changing the benefit from belonging to an identity would come from creating an environment conducive to that identity. Perhaps expending effort to cause agents to identify with the objectives of the firm or ways in which good identity employees can better associate together. In the simple model we developed earlier, there would be no place for these methods, however, by generalizing $I(\cdot)$ to be conditional on c_i, allows the principal to potentially influence the identity choices of the agents and can make the good worker equilibrium more likely and possibly surplus-maximizing.[21] Nonetheless a richer formulation of these identity rewards can allow the principal more influence over the game's outcome, but even though the principal need not influence the game using the standard monetary incentive contracts, yet his actions still may lead to more efficient outcomes.

There are several comments that can be made about the equilibrium outcomes predicted by the model we have just developed. Importantly, in the two-player simultaneous move game, we have no method for determining which equilibrium outcome would occur. In fact, if the players could collude, they would choose the bad worker equilibrium. We believe this is partially an artificial result of the simplistic model of identity: introducing heterogeneous identity functions and penalty functions as well as allowing the principal to modify those functions can lead to different equilibriums becoming more or less likely. Another problem can arise from multiplicity of equilibria in general, multiplayer settings with more complex identity functions and identity switching costs. Although, as we prove in the appendix, the simple one-shot game we have developed in this section will have all players choosing the same identity, this is not always the case with more general functions and equilibria could exist in which some workers choose a good worker equilibrium and others choose the bad worker equilibrium. This result is not problematic since it is conceivable that larger organizations would have a variety of different groups each having their own social norm and identity. As we have introduced identity, it is not a method for determining optimal compensation or organization, but rather another factor that must be included to

[21] Clearly in the game we have represented, the good worker equilibrium can never be a dominant strategy equilibrium unless $I_B(c_{-i}) < 0$.

more fully understand employee relationships and their effect on the firm. Further, by adding some realistic extensions to the model, we can often provide some explanation for some empirical results.

With respect to extending the model, we can also introduce another, possibly richer and more realistic, addition to this framework that can be used to provide some explanation for the Ichino and Maggi (2000) result. We shall assume that the identity function is convex in the number of individuals choosing that identity and there is some cost to changing identity that is also increasing in the number of individuals choosing the identity. These assumptions amount to assuming that individuals receive progressively more utility if their social category has more members and the cost from switching away from that category is increasing if the category has many members. These are not unrealistic assumptions. Certainly there may be positive scale in a social category somewhat analogous to network effects. Furthermore, it would not be surprising if trying to switch away from this category would be increasingly costly when there are many members; trying to disassociate from a dominant group could lead to the worker being completely ostracized. Under these assumptions, consider the case of a northern worker moving into a southern branch. The existing equilibrium in the south has workers choosing the low-effort category and hence a higher incidence of shirking. Under the assumptions we have made in this paragraph, the northern worker would have tremendous incentive to also choose the low-effort identity and join the culture of the southern branch. In other words, the worker would have strong incentives to blend in. In a more general setting, these assumptions will prevent small numbers of individuals from changing the culture (the realized equilibrium) of a firm, but also suggest one method that could change the culture: simultaneously import large numbers of individuals that have a different identity.

We can illustrate this situation by slightly modifying the framework we have introduced earlier. First, we will introduce another penalty for the agents. Let γ be the cost borne by an agent when he switches identities. As we described in the last paragraph, this parameter is generally a function that is increasing in the number of agents choosing a different identity from the agent; for simplicity, we will assume it is a constant. The penalty is meant to capture the fact that it is costly for an agent to switch identities. To switch identities he must first begin to view himself differently, as a member of a different category, and, second, others must perceive him as a member of a new category. This addition leads to the slightly modified utility function

$$U_i(w, e_i, c_i, c_{-i}) = u(w) - c(e_i) + I_{c_i}(c_{-i}) - \psi(e^*(c_i), e_i) - \gamma$$

when the agent chooses a category different from what he previously chose. Given the cost to switch identities, it is natural to modify the timing of the game we developed earlier. This modification also fits in well with the empirical example of introducing additional workers into an organization with a well-defined equilibrium. We will consider that the organization is already in an equilibrium and the firm is bringing in a new worker. The new worker already has an identity and once he is hired, he

must decide whether to retain his existing identity or choose to switch identities. The existing workers do not have a choice of identity at this stage; they have already made their choice and we will assume the culture of the firm is fixed temporarily at least.[22] Now consider the case of a southern worker moving north. In that case, the realized equilibrium of the northern branch is the n workers choosing the good identity.[23] The new worker, however, is from the south and choosing the bad worker identity. Now the new worker has a choice between the two: he can continue choosing the bad worker category and receive utility $u(w) - c(\underline{e})$ or switch to the good worker category and receive $u(w) - c(e_G) + I(n) - \psi_G - \gamma$. Again, the agent will find switching identities is beneficial only if the added utility from the identity function is large enough. In other words, the agent will choose to identify with the other employees if the utility he gains from that identification is large enough; if the benefits of camaraderie, friendship, appreciation, self-image, and other non-monetary factors exceed the higher effort cost and cost of choosing a new category, he will change categories.

The preceding formalization also illustrates that different results could be obtained if multiple workers were hired simultaneously. For example, consider hiring m workers from the south and suppose those workers were choosing the bad worker category. Now, it could be the case that the equilibrium for these new workers is to continue choosing the bad worker category and thus their utility function would include $I(m-1)$. This addition makes it less likely the new workers would switch to the good worker identity since in order to do so the identity function must be convex enough to compensate a worker for the switch as well as the utility the worker is already receiving from $I(m-1)$. Moreover, whenever existing workers have an opportunity to choose their identity, they may also be tempted by the new workers' equilibrium and it may be beneficial for the existing worker to switch categories to the bad worker identity. This illustrates one implication of this model that we alluded earlier. A firm can change its culture, the identity choices of its workers, by importing large numbers of workers with a different identity, but this also illustrates the danger and potential fragility of the good worker equilibrium: a firm in that equilibrium could incrementally acclimate new workers, but massive hiring could lead to the bad equilibrium occurring.

Lastly, throughout this subsection, we have assumed identity is a choice variable like effort and thus it is also worth coming back to the case when identity is not a choice variable, that is identity is like individual tastes or ability, an intrinsic element

[22]This assumption is meant to capture the dynamic nature of these identity choices. Since identity is inherently a social concept, the addition of one worker should not have a large effect on the existing social interactions in the firm. This is somewhat at odds with the static outcomes we considered earlier. We do not view these as contradictory, but rather complementary. The static game is more appropriate for small numbers of players newly initiated without a significant history together or a way to illustrate the possible long-term outcomes from a large number of workers. The more dynamic game in this paragraph is the short-run result when there are additions or subtractions to existing groups that have already had a chance to form identities and relationships.

[23]We will assume that all workers in the north are choosing the good worker category.

of an agent that cannot be changed.[24] In this case identity can still play a role, albeit a much smaller role. In this case, as we mentioned earlier, the vector of identities leads to a vector of optimal effort choices of the agents independent of the other agents' choices since the identity benefit function $I(\cdot)$ is fixed with respect to effort. However, since identity is a part of all the agents' utility functions, it will influence a different decision the agents make: namely, whether to work or not. Since the identity benefit directly enters the agents' utility functions, it will change the level of utility and hence whether that level exceeds each agent's reservation utility. In this manner, firing one agent could actually lead to many agents' utilities falling below the reservation threshold even with no changes in the compensation structure. Again this illustrates the potentially large effect that the identity interactions can have among agents and how it further complicates the principal's optimization problems.

3.4 Conclusion

As we have seen in the preceding sections, we can use the foundation laid by peer effects and identity to construct a simple way to model organizations that at first glance seem to have little ability to provide incentives. We saw that some basic intuitions fail in this world, but the model also supports some common sense wisdom. We find that even though monetary incentives are not available, the firm does have other methods to induce effort and interactions among agents play an extremely important role.

To begin with we developed a basic model of a principal and an agent and allowed richer utility formulations that could include identity. In this simplest formulation, we saw the outcomes change little and could be obtained with pre-hiring screening or effective identity-changing behavior by the principal. On the other hand, allowing more agents gives the model richer results and implications. In particular, the addition of multiple agents complicates hiring decisions since it introduces a discontinuity of the marginal revenue of additional workers. This discontinuity can lead to structures that are efficiency enhancing or not and indeed leads to a variety of possible outcomes. We also considered various ways of slightly perturbing the formulation by allowing screening, allowing the principal to affect the agents' identity choices, and having fixed identities and looked at the results of these perturbations on equilibrium outcomes. We also proposed a way to use this formulation to provide a possible explanation for one of the peer effects examples.

Throughout this paper we have tried to use a new utility formulation to capture interactions and outcomes in organizational forms where it is very difficult to provide standard monetary incentives or subjective performance evaluations. Although the more complete utility formulations produce outcomes that vary substantially and are

[24] This would also be similar to introducing a more dynamic game where we disallow identity changes in certain periods, so, in a one-shot game, identity was fixed.

not always efficiency enhancing, there does seem to be evidence that organizational models should include social interactions since there seems to be compelling empirical and theoretical evidence showing its existence. Although we have used identity as a method to provide some structure to organizations with a rather unconventional and admittedly contrived structure, these identity formulations could also be used in more standard organizational forms. Indeed we have illustrated that this small addition to utility functions can give large dividends and lead to outcomes that differ substantially from outcomes given by standard utility formulations.

Appendix

Proposition 1. *When multiple agents play the normal form game illustrated on page 99 and choose between two categories, the resulting equilibrium will have all agents choosing the same identity.*

Proof. Suppose not and n agents choose the good worker identity while m agents choose the bad worker identity. To prevent good workers from deviating, it must be that $I(n-1)-c(e_G)-\psi_G \geq I(m)-c(\underline{e})$ or $I(n-1)-I(m) \geq c(e_G)-c(\underline{e})+\psi_G$. Let $c(e_G)-c(\underline{e})+\psi_G = \eta$. Notice $\eta > 0$ and since $I(\cdot)$ is increasing and convex, this implies $n-1 > m$. Now for one of the m bad workers to not unilaterally deviate requires $I(m-1)-c(\underline{e}) \geq I(n)-c(e_G)-\psi_G$ or $I(m-1)-I(n) \geq -(c(e_G)-c(\underline{e})+\psi_G) = -\eta$. Hence both $I(n-1)-I(m) \geq \eta$ and $I(n)-I(m-1) < \eta$. Since $I(\cdot)$ is increasing, $I(n-1) < I(n)$ so $I(n-1)-I(m-1) < I(n)-I(m-1)$ and thus $I(n-1)-I(m-1) < I(n)-I(m-1) < \eta \leq I(n-1)-I(m)$ implying that $I(n-1)-I(m-1) < I(n-1)-I(m)$ or $I(m-1) \geq I(m)$ and thus violating the assumption that $I(\cdot)$ is increasing. Hence all agents must choose the same identity in equilibrium. □

Bibliography

AKERLOF, G. A., AND R. E. KRANTON (2000): "Economics and Identity," *Quarterly Journal of Economics*, 115(3), 715–53.

——— (2005): "Identity and the Economics of Organizations," *Journal of Economic Perspectives*, 19(1), 9–32.

BATTAGLINI, M., R. BÉNABOU, AND J. TIROLE (2005): "Self-Control in Peer Groups," *Journal of Economic Theory*, 123(2), 105–34.

BROCK, W. A., AND S. N. DURLAUF (2000): "Interactions-Based Models," Technical Working Paper 258, NBER, http://www.nber.org/papers/T0258.

FALK, A., AND A. ICHINO (2006): "Clean Evidence on Peer Effects," *Journal of Labor Economics*, 24(1), 39–57.

GLAESER, E. L., B. SACERDOTE, AND J. A. SCHEINKMAN (1996): "Crime and Social Interactions," *Quarterly Journal of Economics*, 111(2), 507–48.

GLAESER, E. L., AND J. A. SCHEINKMAN (1999): "Measuring Social Interactions," Working paper, Princeton University.

HONG, H., J. D. KUBIK, AND J. C. STEIN (2004): "Social Interaction and Stock Market Participation," *Journal of Finance*, 59(1), 137–63.

ICHINO, A., AND G. MAGGI (2000): "Work Environment and Individual Background: Explaining Regional Shirking Differentials in a Large Italian Firm," *Quarterly Journal of Economics*, 115(3), 1057–90.

JONES, S. R. G. (1990): "Worker Interdependence and Output: The Hawthorne Studies Reevaluated," *American Sociological Review*, 55(2), 176–90.

KANDEL, E., AND E. P. LAZEAR (1992): "Peer Pressure and Partnerships," *Journal of Political Economy*, 100(4), 801–817.

KATZ, L. F., J. R. KLING, AND J. B. LIEBMAN (2001): "Moving to Opportunity in Boston: Early Results of a Randomized Mobility Experiment," *Quarterly Journal of Economics*, 116(2), 607–54.

LINDQVIST, E., AND R. ÖSTLING (2006): "Ethnic Identity and Social Class," Working paper, Stockholm School of Economics.

MANSKI, C. F. (2000): "Economic Analysis of Social Interactions," *Journal of Economic Perspectives*, 14(3), 115–136.

MAS, A., AND E. MORETTI (2006): "Peers at Work," Mimeo, UC Berkeley and NBER.

MOFFITT, R. A. (2001): "Policy Interventions, Low-Level Equilibria, and Social Interactions," in *Social Dynamics*, ed. by S. N. Durlauf, and H. P. Young, chap. 3, pp. 45–82. The MIT Press, Cambridge, Massachusetts.

ROTEMBERG, J. J. (1994): "Human Relations in the Workplace," *Journal of Political Economy*, 102(4), 684–717.

SACERDOTE, B. (2001): "Peer Effects with Random Assignment: Results for Dartmouth Roommates," *Quarterly Journal of Economics*, 116(2), 681–704.

SHAYO, M. (2005): "A Theory of Social Identity with an Application to Redistribution," Ph.D. thesis, Princeton University.