The Quality of Commodity Markets

Von der Wirtschaftswissenschaftlichen Fakultät der
Gottfried Wilhelm Leibniz Universität Hannover
zur Erlangung des akademischen Grades

Doktor der Wirtschaftswissenschaften
– Doctor rerum politicarum –

genehmigte Dissertation

von

M.Sc. Tobias Lauter
geboren am 27.07.1990 in Friedrichshafen

2023

Referent: Prof. Dr. Marcel Prokopczuk
Koreferentin: Prof. Dr. Judith Christiane Schneider
Tag der Promotion: 16.07.2023

Eigenständigkeitserklärung - Declaration of Original Ownership

Hiermit bestätige ich, dass ich die vorliegende Arbeit selbstständig ohne Hilfe Dritter verfasst und keine anderen als die angebenen Hilfsmittel benutzt habe. Die Stellen der Arbeit, die dem Worlaut oder dem Sinn nach anderen Werken entnommen sind, wurden unter Angaben der Quelle kenntlich gemacht.

I hereby confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Datum	(Tobias Lauter)

Abstract

This thesis studies the quality, that is, both the liquidity and price efficiency, of commodity futures and options markets. Chapter 1 introduces the subject and provides an overview. Chapter 2 identifies the most suitable low-frequency proxies for analyzing commodity market quality. We use an 11-year sample of millisecond time-stamped order book data and examine the correlation of high-frequency liquidity and price efficiency measures with their low-frequency proxies measured with daily or 5-minute Time-and-Sales (TAS) data. We find that for liquidity, the volatility-over-volume measures are the best proxies for bid–ask spread and price impact. The correlation of price efficiency measures with their daily-frequency counterparts is low. Moderately correlated proxies can be achieved by using 5-minute data. Chapter 3 studies commodity futures markets quality using the previously identified best proxies. We investigate the impact of two major changes: (1) The influx of index investors after 2004 (financialization) and (2) the introduction of side-by-side trading of open-outcry and electronic limit order books around mid-2006 (electronification). Our sample consists of daily measures of liquidity and intraday informational efficiency spanning the years 1996 to 2018. We find that market quality has improved over the sample period, including and especially during the years of financialization and electronification. These improvements appear to be more pronounced in commodities that are part of a major index. We further employ different data sets of aggregate trader positions data curated by the Commodity Futures Trading Commission (CFTC) but find no evidence of a harmful effect of index trading activity on commodity market quality. Despite a sharp increase in long open interest of commodity index traders (CITs) in soybean meal in January 2013 when it was added to the Bloomberg Commodity Index (BCOM), the quality of the soybean meal futures market did not worsen. Finally, our comprehen-

sive data set enables us to compare market quality around index roll days across the pre- versus post-financialization regime. Consistent with our previous findings, market quality during index roll days did not worsen, but appears to have improved slightly. Overall, the results show that commodity financialization was not harmful to market quality, but rather coincided with improvements. The switch to electronic limit order books had positive effects. Motivated by the evidence from Chapter 2, that market efficiency is noisy, and lacking evidence of harmful index trading using weekly aggregate position data in Chapter 3, we combine 5-minute WTI ETF, options, and futures data in Chapter 4 in order to be able to detect very short-lived inefficiencies in a less noisy almost model-free way. This allows us to study the role of ETF-related trading in New York Mercantile Exchange (NYMEX) West Texas Intermediate (WTI) crude oil futures and options markets. We detect and model put–call–parity deviations in short-term at-the-money (ATM) and their underlying futures at the 5-minute frequency between January 2010 and October 2021. Then, we relate those to ETF-related futures trading. Our findings suggest that those trades are likely informed, but are not timed to exploit arbitrage opportunities. This implies that average financial investors have temporary price impact but more likely due to adverse selection risk of market makers than due to inventory risk induced by large directional trades. Our results highlight the use of ETFs as an alternative for informed trading even in highly liquid markets. Chapter 5 concludes and lays out open questions and possible paths for future research.

Keywords: Commodity Markets, Market Quality, Futures, Options, Liquidity, Market Efficiency

Zusamenfassung

Diese Dissertation untersucht die Qualität, d.h. sowohl die Liquidität als auch die Preiseffizienz, von Warentermin- und Optionsmärkten. Kapitel 1 führt in das Thema ein und gibt einen Überblick. In Kapitel 2 werden die am besten geeigneten Niederfrequenz-Proxies für die Analyse der Qualität von Rohstoffmärkten identifiziert. Wir verwenden eine 11-jährige Stichprobe von Orderbuchdaten im Millisekundenbereich und untersuchen die Korrelation von hochfrequenten Liquiditäts- und Preiseffizienzmaßen mit niederfrequenten Schätzwerten, die mit täglichen oder 5-minütigen Time-and-Sales-Daten (TAS) gemessen werden. Wir stellen fest, dass für die Liquidität Volatilität-über-Volumen-Maße die besten Näherungswerte für die Geld-Brief-Spanne und den Preiseinfluss sind. Die Korrelation der Preiseffizienzmaße mit ihren Pendants in der Tagesfrequenz ist gering. Mäßig korrelierte Näherungswerte können durch die Verwendung von 5-Minuten-Daten erreicht werden. Kapitel 3 untersucht die Qualität der Warenterminmärkte unter Verwendung der zuvor ermittelten besten Schätzern. Wir untersuchen die Auswirkungen von zwei wichtigen Veränderungen: (1) Der Zustrom von Indexinvestoren nach 2004 (Finanzialisierung) und (2) die Einführung des Parallelhandels von Parkett- und elektronischen Limit-Orderbüchern um Mitte 2006 (Elektronifizierung). Unsere Stichprobe besteht aus täglichen Messungen der Liquidität und der Intraday-Informationseffizienz, die die Jahre 1996 bis 2018 umfassen. Wir stellen fest, dass sich die Marktqualität im Untersuchungszeitraum verbessert hat, auch und gerade in den Jahren der Finanzialisierung und Elektronifizierung. Diese Verbesserungen scheinen bei Rohstoffen, die Teil eines großen Index sind, stärker ausgeprägt zu sein. Darüber hinaus verwenden wir verschiedene Datensätze mit aggregierten Positionen, die von der Commodity Futures Trading Commission (CFTC) kuratiert werden, finden aber keine Hinweise auf eine schädliche Auswirkung von

Indexhandelsaktivitäten auf die Qualität der Rohstoffmärkte. Trotz eines starken Anstiegs des offenen Interesses von Rohstoffindexhändlern (CITs) an Sojaschrot im Januar 2013, als dieses in den Bloomberg Commodity Index (BCOM) aufgenommen wurde, hat sich die Qualität des Sojaschrot-Futures-Marktes nicht verschlechtert. Schließlich ermöglicht es uns unser umfassender Datensatz, die Marktqualität an den Indexrolltagen vor und nach der Finanzialisierung zu vergleichen. In Übereinstimmung mit unseren früheren Ergebnissen hat sich die Marktqualität an den Indexrolltagen nicht verschlechtert, sondern leicht verbessert. Insgesamt zeigen die Ergebnisse, dass die Finanzialisierung der Rohstoffmärkte der Marktqualität nicht geschadet hat, sondern vielmehr mit Verbesserungen einherging. Die Umstellung auf elektronische Limit-Orderbücher hatte positive Auswirkungen. Motiviert durch den Nachweis aus Kapitel 2, dass die Markteffizienz verrauscht ist, und den ausbeleibender Hinweise auf schädlichen Indexhandel unter Verwendung wöchentlicher aggregierter Positionsdaten in Kapitel 3, kombinieren wir in Kapitel 4 5-Minuten-WTI-ETF-, Options- und Futures-Daten, um in der Lage zu sein, sehr kurzlebige Ineffizienzen auf eine weniger verrauschte, fast modellfreie Weise zu messen. Auf diese Weise können wir die Rolle des ETF-Handels auf den Rohöl-Futures- und Optionsmärkten der New York Mercantile Exchange (NYMEX) für West Texas Intermediate (WTI) untersuchen. Wir ermitteln und modellieren Put-Call-Paritätsabweichungen bei kurzfristigen at-the-money (ATM) und ihren zugrunde liegenden Futures im 5-Minuten-Takt zwischen Januar 2010 und Oktober 2021. Anschließend setzen wir diese mit dem ETF- und Futures-Handel in Beziehung. Unsere Ergebnisse deuten darauf hin, dass diese Trades wahrscheinlich informiert sind, aber nicht zur Ausnutzung von Arbitragemöglichkeiten getätigt werden. Dies bedeutet, dass durchschnittliche Finanzinvestoren einen vorübergehenden Einfluss auf den Preis haben, der jedoch eher auf das Risiko der negativen Auswahl der Market Maker zurückzuführen ist als auf das Bestandsrisiko,

das durch große direktionale Trades entsteht. Unsere Ergebnisse unterstreichen die Verwendung von ETFs als Alternative für den informierten Handel selbst auf hochliquiden Märkten. Kapitel 5 schließt mit einem Ausblick auf offene Fragen und mögliche Wege für zukünftige Forschung.

Schlagwörter: Rohstoffmärkte, Marktqualität, Terminmärkte, Optionen, Liquidität, Markteffizienz

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Commodity derivatives markets are of central interest to producers, consumers, and regulators. They facilitate the hedging of commodity-related risks, while also providing a basis for business decisions (Black, 1976). As classical consumption assets they affect the cost of living for all humans taking part in a globalized economy.[1] Depending on the degree of development of a nation, commodity-dependent costs like food, transportation, electricity or heating can make up a substantial part of peoples' consumption basket. Price spikes can give rise to social unrest and disrupt societies, as it was the case in the Arab Spring. Thus, regulators are particularly interested in the functioning of commodity markets and in the US, for example, its market oversight is separate from that of securities markets.

The nature as a consumption good has been questioned after commodity markets underwent substantial changes. More or less freely organized commodity spot and derivatives markets have been existing since centuries.[2] Especially the latter ones were mostly populated by specialists rather than the common consumer. This changed during the start of the 21st century

[1] Native tribes in the rain forests are probably the only exemptions.

[2] For example, the Babylonian Code Hammurabi dates back to 1750 BC and regulates forward trading in ancient Mesopotamia.

when regulations were relaxed, interest in commodity markets as a means of diversification of investment portfolios arose and market access became easier as commodity (index) ETFs became popular and exchanges introduced electronic limit order books. This was coined the financialization of commodity markets (e.g., Tang and Xiong, 2012).

The role of these new financial traders–especially passive long-only commodity index traders (CITs)–and their influence on the quality of commodity derivatives markets became a controversial topic in both the public media and academic literature. They were accused of being responsible for excess volatility, increased co-movement (Tang and Xiong, 2012), and disruptions of the price formation process leading to inefficient pricing. This thesis focuses on the latter. CITs are commonly regarded as uninformed investors that invest for reasons unrelated to the fundamental value of the commodities. Theoretical predictions of an increase of uninformed trading can be derived from classical mictrostructure theory. In the Glosten and Milgrom (1985) model, for example, an increase in the share of uninformed traders results in lower bid-ask spreads set by the market maker, but the convergence of the trade price to the fundamental value is slowed down. That means, liquidity improves but prices are less efficient. If the market maker is not risk-neutral, however, inventory effects of uninformed traders that unanimously buy or sell could lead to a wider bid–ask spread. Goldstein and Yang (2022) develop a model of commodity financialization that predicts positive relationship between financial trading and liquidity but a hump-shaped (inverse-U) one with price efficiency. Opposite theoretical predictions for predictable uninformed trading exist as well: The Sunshine-Trading Effect of Admati and Pfleiderer (1991) suggests that credibly uninformed liquidity demand is met with additional liquidity while Brunnermeier and Pedersen (2005) point out that such trades could be profited of in a predatory fashion. Theoretical channels of how financial traders might impact commodity market quality therefore ex-

ist, but their predictions are not clear-cut, which is why empirical analysis is required and performed in the chapters of this thesis.

We use tools that have been developed in the context of microstructure theory of (equity) markets and employ it on commodity derivatives data. Since the number of developed measures is large, Chapter 2 serves as a pre-study that identifies the most appropriate measures of commodity market quality. Since bid-ask spreads and intraday volumes had not been recorded in open-outcry trading which was the dominant market structure until about 2008, measures are required that are able to capture bid-ask spreads, price impact, and price efficiency from trade prices alone. To do so, we conduct a horse race of proxies with low data requirements and compare them using correlations with measures computed from data that includes bid-ask prices and volume. The approach follows Goyenko et al. (2009) who study liquidity measurement for stock markets. Marshall et al. (2012) conduct a similar study for commodity futures, but the majority of their sample period spans periods when most trading was conducted in the pits. Our sample consists of 11 years of quote and volume data and we also study price efficiency measures and include newly developed measures. One class of these more recently derived measures, volatility-over-volume (VoV) ratios (Kyle and Obizhaeva, 2016; Fong et al., 2018), turns out to be superior to the widely-used Amihud (2002) measure in capturing liquidity. VoV–measures exhibit the highest correlations for single commodities and aggregate commodity market quality in the time-series dimension. Mincer and Zarnowitz (1969) regressions confirm these results but highlight the importance of computing the measures for each commodity individually and scaling them in order to make their level interpretable. Our evidence on market efficiency suggests that these measures are noisy and that a valid measurement requires an intraday sampling frequency.

Chapter 3 builds on the previous insights and uses the best measures to construct a panel of commodity market quality that exhibits differences in

the degree of financialization. Our study is more comprehensive than previous evidence, because it unites the following properties: (1) In the time-series dimension, the panel includes the time before the start of the financialization. (2) Some market quality measures, especially the market efficiency measures, are estimated from intraday trade data which are different from using settlement prices (which are volume-weighted average prices). (3) The staggered introduction of side-by-side electronic and pit trading allows us to measure the impact of this change introduced by most commodity changes starting around mid-2006. Except for Raman et al. (2020), electronification effects have not been studies in conjunction with those of commodity financialization.

Our results hint at an improvement of commodity market quality after 2004–a common breaking point for the start of the financialization. Commodity market quality exhibits a significant positive shift after 2004 and especially after traders migrated to electronic limit order markets. This shift is more pronounced among commodities that have been part of a major broad commodity index throughout the entire sample period. Instead of a sudden shift in the level of commodity market quality, we also estimate a regime-conditional linear trend model that confirms the improvement during both the financialization and electronification period.

Predictable and unpredictable trades of CITs might have different impacts on commodity market quality which is why the last part of Chapter 3 studies if they have an effect. First, we decompose weekly aggregate open interest data using ARIMA models with seasonal components and estimate panel regressions. We also use a predictable change to CIT activity when soybean meal was added to a major commodity index. Finally, we study predictable roll trades and compare patterns in market quality before and after 2004. In line with the results of Bessembinder et al. (2016), we find evidence for a Sunshine Trading Effect. All results point towards no or a slight positive

effect of CIT trading on commodity market quality.

Previous results have shown that CITs are not harmful to commodity market quality. In Chapter 4, we thus study the role of ETF-related trading in West Texas Intermediate (WTI) sweet crude oil markets at the 5-minute frequency. Eglite et al. (2023) show that ETFs are used to conceal insider trading while Israeli et al. (2017) show that market quality of stocks declines with increased ETF ownership. In order to measure market quality with less noise, fewer assumptions about the data generating process, and at a higher frequency, we rely on options data. This approach has not been used in the literature on commodity financialization. We employ measures derived from put–call–parity (PCP) that are almost model-free: absolute differences in implied volatilities, the occurrence of arbitrage opportunities, and available arbitrage profits. We use the Lee and Ready (1991) algorithm to sign every single trade in short-term WTI futures and two related ETFs. Almost 3% in futures trading is linearly related to ETF trading. Then, we regress differences in implied volatility on absolute (non-) ETF-related order imbalances (OIB) and explore why they are increased when ETF-related absolute OIB are high. Positive subsequent returns following large absolute ETF-related OIB hint at information-based trading while higher price impact due to larger order sizes is unlikely. Our results highlight that (index) ETF trading cannot simply be classified as uninformed noise. ETFs might be a trading vehicle that could be preferred by informed traders over futures as they are more granular and easier to handle.

Overall, the results of this thesis show that the presence of financial traders and low barriers of entry is beneficial to the quality of commodity markets. While some newly arrived investors surely are uninformed, they appear to be accompanied by well-informed agents that provide liquidity and make informed investment decisions.

This thesis proceeds as follows. Chapter 2 identifies the most appropriate

commodity market quality measures that are then used in Chapter 3 to infer the impact of index investors. Chapter 4 use commodity options and futures to study the impact of ETF-related order imbalances at a high frequency. Finally, Chapter 5 summarizes the main findings and outlines potential paths for future research.

Chapter 2

Measuring Commodity Market Quality*

2.1 Introduction

O'Hara and Ye (2011) define market quality as '*a market's ability to meet its dual goals of liquidity and price discovery. In general, markets with lower transaction costs are viewed as higher quality, as are markets in which prices exhibit greater efficiency*' (p. 463). The computation of spreads, price impact or intraday autocorrelations requires high-frequency data that is both computationally and literally expensive. Additionally, for some markets, quote data is simply not available. For commodity futures, quotes were not recorded in open-outcry but only in more recent electronic limit order markets. Thus, if one wishes to study long-term commodity market quality, one has to rely on low-frequency proxies to measure both aspects.

In this paper, we aim to identify the best low-frequency proxies for measuring market quality. This is important, because in many situations high-quality intraday data is not available, whereas low-frequency, i.e., daily data

*This chapter is based on the article "Measuring Commodity Market Quality" authored by Tobias Lauter and Marcel Prokopczuk, Journal of Banking & Finance 145, 2022, 106658.

is. We estimate low-frequency proxies and conduct a horse-race-type study in the style of Goyenko et al. (2009) and Fong et al. (2017) for equity markets or Schestag et al. (2016) for bond markets. Marshall et al. (2012) conduct such a study for commodity market liquidity that is most similar to ours. However, their sample period ranges from 1996 to August 2008 and thus covers a period when electronic trading was mostly limited to overnight hours. Most volume, however, was still in the pits until around 2008. Thus, quote data was only available for overnight-trading hours. The reliability of their results is therefore contingent on overnight liquidity in electronic limit order books being highly correlated with the liquidity in the pits. Today, both day and night trading is mainly conducted electronically and market makers use algorithms that supply liquidity almost around the clock. After 2008, overnight and day market quality are likely to be closely related. Before, however, it is unlikely that market makers in the pits were the same as those submitting orders in the electronic overnight market. Given this substantial change of the markets' structure, our study is thus more than a simple up-date of the sample period. Moreover, Marshall et al. (2012) focus on liquidity only and do not consider informational efficiency at all.

Thus, we make the following contributions to the literature: First, we employ a sample that entails 11 years of reliable quote data generated during the main trading hours. The length of the sample allows a valid measurement of time-series variation in benchmarks and proxies as opposed to cross-sectional variation alone. Second, when measuring liquidity, we also include new measures of liquidity developed over recent years. Third, we do not only focus on liquidity but also study market efficiency to provide a complete picture of market quality as defined by O'Hara and Ye (2011). Fourth, we also provide guidance on obtaining effective spread proxies for the most traded US-commodities to be used in analyses in which not only correlations but also levels are relevant. Finally, we apply the same low-frequency proxies on

5-minute Time-and-Sales (TAS) data. This sampling frequency represents a middle ground between tick data and daily data. As it turns out, this is necessary to reach acceptable correlations of proxies and benchmarks of informational efficiency.

Our research aims at enabling investigations involving either or both the liquidity and price efficiency of commodity markets. Such studies might investigate the impact of financialization (e.g., Tang and Xiong, 2012), changes in market structure, like the switch from open-outcry to electronic limit order books (e.g., Shah and Brorsen, 2011; Raman et al., 2020), the influence of algorithmic trading on commodity market quality (Hendershott et al., 2011, for example, study the influence of algorithmic trading on the market quality of equity markets), or the impact of margin requirements on liquidity (Daskalaki and Skiadopoulos, 2016). Studies of systematic price efficiency (Rösch et al., 2017) that include or focus on commodity markets require valid proxies. In the realm of asset pricing, the search for liquidity premia is a common objective which requires suitable liquidity proxies.[1] The impact of (short-term) mispricings on the real economy (e.g., Brogaard et al., 2019) is another interesting research avenue that requires valid price efficiency measures.

In order to identify the most suitable proxies, we collect and pre-process terabytes of high frequency tick-by-tick data of major commodity futures markets since 2008. Then, we calculate high frequency measures of liquidity and price efficiency for every trading day and every commodity in the sample using millisecond time-stamped trade and quote data. We average these measures in each commodity-month to obtain a panel of monthly high-frequency measures as a benchmark. Then, we calculate monthly proxies

[1]For example, Szymanowska et al. (2014), Daskalaki et al. (2014), and Fernandez-Perez et al. (2019) use the Amivest or Amihud ratio measure to perform commodity sorts on liquidity. Liquidity estimates are also necessary to compute returns net of transaction costs, e.g., in the context of portfolio construction (see, e.g., Daskalaki and Skiadopoulos, 2011; Daskalaki et al., 2017).

using daily data. The most appropriate selection criterion depends on the objective at hand. For asset pricing, cross-sectional rank-correlation across low- and high-frequency measures are most relevant. For studies of aggregate market quality, time-series correlation of a measure averaged across commodities is most relevant. Finally, for studies of individual commodity market quality, individual time-series correlations are more important. Thus, we elect the best method in each category, to accommodate the requirements of various research designs.

Since bid–ask spread estimates are important to compute after-cost returns, we asses the ability of the proxies to capture both the level and variation in bid–ask spreads by estimating benchmark–proxy regressions in the spirit of Mincer and Zarnowitz (1969).

Lastly, we compare the performance of several proxies calculated using 5-minute TAS data. This is sensible because of three reasons. First, some research designs, e.g. in the context of financialization, those that include weekly aggregate positions data curated by the Commodity Futures Trading Commission (CFTC) or those that study index roll days, require sub-monthly frequencies in market quality estimates while some proxies are impossible to estimate or are too noisy to be estimated from a single data point. Second, pricing errors might be short-lived and thus not be measurable with daily settlement prices. Third, using prices aggregated to a 5-minute frequency drastically reduces the computational requirements compared to tick-by-tick data. Depending on the proxy, measures can be computed within minutes. This approach represents a middle ground between terabytes of individual trades, that requires days or weeks to process, and daily data, that are analyzed within seconds. Most proxy measures we employ have been developed for an arbitrary sampling frequency so we compare an almost identical selection of measures.

Based on our results, we can give the following recommendations to

researchers who wish to study the quality of commodity futures markets using low-frequency data:

(1) Spreads are best captured by the VoV(Spread) measure of Kyle and Obizhaeva (2016) and Fong et al. (2018).

(2) Price impact is best captured by VoV(λ).

(3) In contrast to the Amihud (2002) measure, daily VoV–measures are able to capture daily variation in liquidity

(4) Our proxies for informational efficiency are not correlated to their benchmarks. To approximate monthly price efficiency from non-overlapping data, we recommend using 5-minute TAS data instead of daily data. Both the variance ratio of Smith (1994) and the pricing error volatility (MA(1)) of Hasbrouck (1993) exhibit moderate correlations with their respective benchmark-variants (see Table 2.9). If intraday data are not available, spread or impact proxies are valid alternatives to capture Hasbrouck's pricing error volatility.

(5) In order to obtain unbiased effective spreads estimates, we recommend using the VoV(Spread) measure and then mapping it to the correct level and variance using the parameters provided in Table 2.8.

2.2 Measuring Market Quality

In this section, we provide an overview of the liquidity and efficiency measures employed.

As high-frequency liquidity benchmarks, for spreads, we use time-weighted relative quoted spreads (twRQS) and volume-weighted relative effective spreads (vsRES), and for price impact, we use volume-weighed 5-minute relative price

impact (vwRPI), as well as the slopes (λ) of 5-minute mid-quote returns regressed on signed dollar–volume (Kyle, 1985) or on the signed square-root of volume (λ^{root}) following Hasbrouck (2009).

As low-frequency liquidity proxies, we employ the return–covariance-based Roll (1984) measure (Roll) with its variations with absolute covariances (RollAbs) as in Easley et al. (2021) and a variant that is estimated using a Gibbs sampler (RollGibbs; Hasbrouck, 2004), the Effective Tick measure (EffTick; Holden, 2009) which relies on price clustering, the Corwin and Schultz (2012) measure (HighLow) which requires high and low prices, the Abdi and Ranaldo (2017) measure which extends it by incorporating closing prices, two volatility-over-volume measures (VoV(Spread) and VoV(λ)) (Kyle and Obizhaeva, 2016; Fong et al., 2018), the Amihud (2002) measure which is absolute returns over dollar–volume, the slope of returns regressed on lagged volume signed with lagged returns as in Pástor and Stambaugh (2003), several of the aforementioned spread measures divided by volume (Goyenko et al., 2009), and finally the inverse of volume (1oV).

As high-frequency benchmarks for price efficiency, we decompose log-prices into a stationary and a random walk component estimated using a vector-auto-regression ($\sigma_s^{VAR(5)}$) of 1-minute returns and signed dollar–volume with 5 lags or a moving-average process ($\sigma_s^{MA(1)}$) with a single lag (Hasbrouck, 1993), the AMIM measure by Tran and Leirvik (2019), absolute deviations from unity of 1-minute to 30-minute mid-quote return variance ratios (Lo and MacKinlay, 1988), and absolute first-order autocorrelations of 1-minute mid-quote returns.

As low-frequency proxies, we employ the same measures but use the microstructure-adjusted variance ratio by (Smith, 1994) instead of simple variance ratios.

Since a simple average has been shown to be a powerful combination technique in the forecasting literature (Clemen, 1989), we also compute an

average of all spread proxies (AvgSpread), price impact proxies (AvgImpact), and of all price efficiency proxies (AvgEff). To neutralize differences in scale, we standardize each proxy for each commodity before averaging.

Table 2.1 provides a summary of the measures with their data requirements. For details on their computation, see the appendix.

Table 2.1: **An Overview of Benchmark and Proxy Measures with Data Requirements**

This table we provides an overview of the benchmark (left panel) and proxy (right panel) measures we employ. The codes for required data are: S = Settlement Price, V = Volume, O = Opening Price, H = High Price, L = Low Price, B = Bid Price, A = Ask Price, T = Trade Price.

High-frequency Benchmark Measures			Low-frequency Proxy Measures		
Measure	Paper	Required Data	Measure	Paper	Required Data
			Panel A: Spread		
twRQS		B, A	Amihud	Amihud (2002)	S, V
vwRES		B, A, T, V	VoV(Spread)	Kyle and Obizhaeva (2016), Fong et al. (2018)	H, L, V, S
			Roll	Roll (1984)	S
			RollAbs	Easley et al. (2021)	S
			RollGibbs	Hasbrouck (2004)	S
			EffTick	Holden (2009)	S (, H, L, O)
			HighLow	Corwin and Schultz (2012)	H, L
			AbdiRanaldo	Abdi and Ranaldo (2017)	O, H, L, S
			Panel B: Price Impact		
vwRPI		B, A, T, V	Amihud	Amihud (2002)	S, V
λ	Kyle (1985)	B, A, T, V	1oV	Lou and Shu (2017)	V
			$VoV(\lambda)$	Fong et al. (2018)	H, L, V, S
λ^{root}	Hasbrouck (2009)	B, A, T, V	RolloV	Goyenko et al. (2009)	S, V
			RollAbsoV	Goyenko et al. (2009)	S, V
			RollGibbsoV	Goyenko et al. (2009)	S, V
			EffTickoV	Goyenko et al. (2009)	S, V (, H, L, O)
			HighLowoV	Goyenko et al. (2009)	H, L, V
			AbRaoV	Goyenko et al. (2009)	O, H, L, S, V
			PastorStambaugh	Pástor and Stambaugh (2003)	S, V
			Panel C: Price Efficiency		
$\sigma_s^{VAR(5)}$	Hasbrouck (1993)	T, V	$\sigma_s^{MA(1)}$	Hasbrouck (1993)	S
AMIM	Tran and Leirvik (2019)	B, A	AMIM	Tran and Leirvik (2019)	S
VR_{30}	Lo and MacKinlay (1988)	B, A	VR_2	Lo and MacKinlay (1988)	S
$\lvert AR1\rvert$	Chordia et al. (2008)	B, A	VR^{Smith}	Smith (1994)	T
			$\lvert AR1\rvert$	Chordia et al. (2008)	S

2.3 Method of Comparison

We calculate high-frequency measures for each commodity futures contract and every day from January 2008 to December 2018. Some of the intraday measures are calculated using every single quote or trade updates on the tape, others rely on data that is aggregated to a 1-minute frequency. Then, we aggregate each of the measures to a monthly frequency either by taking a simple average or a volume-weighted average for measures whose estimates are volume weighted, e.g., the volume-weighted relative effective spread (vwRES). This provides us with monthly benchmarks for spreads, impact, and efficiency.

In the next step, we estimate low-frequency proxies that rely on daily data for each commodity-month. We then compare these measures using Pearson (Spearman) correlation coefficients of each measure–proxy pair in the time-series (cross-section).

We use relative liquidity measures, like relative effective spreads or relative price impact instead of dollar or tick-multiple measures, because the margin requirements are periodically adjusted by the exchanges and set proportionally to the current futures price, so the cost of a collateralized futures position is also proportional to the current price.

For many decades, commodity futures were traded face to face in open-outcry markets – the pits – where traders would shout and use hand signals to submit offers, bids, and market orders. Bids and offers were executable as long as they were in the 'mouth of the trader' and only very few were recorded. High-frequency recordings of pit activity were limited to time-stamped trade prices, if the price was different from the previous one (so-called TAS data). In 1992, the Chicago Mercantile Exchange (CME) introduced GLOBEX, an electronic limit order trading platform which was used during off-hours, when pits were closed.

Around mid-2006, exchanges like the New York Mercantile Exchange

(NYMEX), the Commodity Exchange (COMEX), and the Chicago Board of Trade (CBOT) (all three now belonging to the CME Group) introduced side-by-side trading during pit trading hours. The Intercontinental Exchange (ICE) followed in 2007. After this change, volume gradually shifted to electronic trading.

By 2008, a major fraction of the volume had migrated from the pits to the electronic limit order platform. Thus, quote data for most commodity futures start in mid-2006 and are reliable from 2008 onward.[2] This is why we choose January 2008 as the starting point of our sample.

A major concern of this study is if correlations of measures estimated using limit order book data are a valid method for choosing the best proxies in open-outcry markets. Both market designs have in common that traders can post bid and ask orders as well as pick them up with market orders. However, in the pits, orders remain only valid for a short period of time – as long as they are in the 'mouth of the trader'. This way, orders cannot walk the book. In the pits, trading is not anonymous but traders can make a name for themselves and often know each other by name. Their roles and employers are also visible to other market participants. On the Chicago Mercantile Exchange (CME) floor for example, traders wear colored jackets and badges. The measures we compare are, however, not by design or assumption only valid in electronic limit order markets. Both VoV(Spread) and VoV(λ) by Fong et al. (2018), for example, are implementations of the Kyle and Obizhaeva (2016) microstructure invariance hypothesis which is a meta-microstructure hypothesis abstract from any market structure. The Roll (1984) model assumes a constant spread with marketable orders hitting standing bids and asks – no

[2] https://www.cmegroup.com/education/files/globex-retrospective-2012-06-12.pdf,
https://www.cmegroup.com/media-room/press-releases/2006/8/02/nymex_to_offer_sidebysidetradingofphysicallydeliveredenergyfutur.html,
https://www.cmegroup.com/media-room/press-releases/2006/8/01/cbot_launches_electronictradingofagfuturesduringdaytimehours.html.

matter if posed orally or electronically. The same applies to the measures by Corwin and Schultz (2012) and Abdi and Ranaldo (2017). The Effective Tick measure similarly relies on price clustering alone without assuming a specific market structure. The Amihud (2002) measure builds on Kyle's lambda – a measure of price impact. In fact, the model of Kyle (1985) does not include a bid–ask spread but price impact can be measured irrespective of the exact trading mechanism. Open-outcry and limit order book markets share some important features. Both are order-driven auction markets. That means, the interpretation of trade prices, which is the basis for the proxies we employ, is similar. They are prices of trades executed by market orders hitting standing bid or ask prices. The proxies we test are robust to sample splits in both the time-series and the cross-sectional dimension (detailed results can be found in the Appendix), i.e., they work in times of slower and more recent faster markets and for commodities ranging from highly liquid energy to more illiquid commodities like oats and rough rice. This makes us confident to believe that the proxies are robust to market design and speed.

2.4 Data

We include the largest US-based commodity futures markets in our analysis: From the energy sector, these are New York Mercantile Exchange (NYMEX) WTI crude oil (CL), heating oil (HO), natural gas (NG), Intercontinental Exchange (ICE) (EU) natural gas (NGLNM), Brent crude oil (LCO), and gas oil (LGO). For grains, we use Chicago Board of Trade (CBOT) soybeans (S), corn (C), wheat (W), Kansas City Board of Trade (KCBT) hard red winter wheat (KW), CBOT soybean meal (SM), soybean oil (BO), rough rice (RR), and oats (O). For metals, we consider Commodity Metals Exchange (COMEX) gold (GC), silver (SI), copper (HG), NYMEX platinum (PL), and palladium (PA). Softs are represented by ICE (US) cotton (CT), sugar No.11

(SB), coffee (KC), cocoa (CC), Chicago Mercantile Exchange (CME) lumber (LB), as well as ICE (US) orange juice. From the livestock sector, we use CME live cattle (LC), lean hogs (LH), and feeder cattle (FC). Table B.1 provides an overview. For each of these commodities, we use the contract with the highest volume, which is often the front-month contract, but can sometimes be the 5th-nearest contract as it is the case for soybean meal or soybean oil. To identify the most liquid contract, we use a 5-day moving average of volume. Overnight periods are excluded from the sample due to low volume. The high-frequency data is sourced from Refinitiv's Datascope Select, formerly Thomson Reuters Tick History (TRTH/SIRCA). Each trade and quote update is labeled with a millisecond time-stamp and represents a new row of the data, which results in billions of rows and terabytes data.

First, we pre-process the data to remove erroneous entries, like one-sided quotes, trades without volume, or negative bid–ask spreads. Some of the measures we employ require us to flag if a trade was initiated by a buy or sell order. We assign trade directions using the Lee and Ready (1991) algorithm as is standard in the literature.

We obtain daily (open-, high-, low-, and settlement-) price and volume data from Thomson Reuters Datastream. Settlement prices are average prices during a certain time-interval of the day. This makes the use of some measures problematic like, e.g., variants of Roll's measure that rely on the last price being a trade executed at the bid or ask.

For each commodity, we treat outliers in daily high-frequency estimates by removing those values whose log absolute distance from the centered 22-day moving median exceeds a value of one. After calculating monthly proxies and aggregating daily high-frequency measures to a monthly frequency, we apply the same procedure with a 12-month moving median. Upon visual inspection, this method appears to be able to effectively identify outliers when the time-series exhibit considerable heteroskedasticity while retaining

true spikes.

Descriptive Statistics Table 2.3 provides averages for all benchmark measures for each commodity in our sample. Considerable heterogeneity in market quality across commodities is evident. For example, most energy and precious metals exhibit low spreads and high levels of price efficiency. For the others, the level of market quality is lower with some niche commodities (e.g., rough rice (RR), oats (O) or lumber (LB)) exhibiting considerable lower levels of market quality.[3]

[3]Sample averages of benchmark measures can be found in the Appendix.

Table 2.2: **Commodity Futures Considered**

This table gives an overview of the commodities we consider in our analysis. NYMEX = New York Mercantile Exchange, ICE = Intercontinental Exchange, CBOT = Chicago Board of Trade, KCBT = Kansas City Board of Trade, COMEX = Commodity Exchange, CME = Chicago Mercantile Exchange. CME, NYMEX, CBOT, KCBT, and COMEX are all part of CME Group.

Sector	Exchange	Commodity	Ticker
Energy			
	NYMEX	WTI Crude Oil	CL
	NYMEX	Heating Oil	HO
	NYMEX	Natural Gas	NG
	ICE (EU)	Natural Gas	NGLNM
	ICE (EU)	Brent Crude Oil	LCO
	ICE (EU)	Gas Oil	LGO
Grains			
	CBOT	Soybeans	S
	CBOT	Corn	C
	CBOT	Wheat	W
	KCBT	Hard Red Winter Wheat	KW
	CBOT	Soybean Meal	SM
	CBOT	Soybean Oil	BO
	CBOT	Rough Rice	RR
	CBOT	Oats	O
Metals			
	COMEX	Gold	GC
	COMEX	Silver	SI
	COMEX	Copper	HG
	NYMEX	Platinum	PL
	NYMEX	Palladium	PA
Softs			
	ICE (US)	Cotton	CT
	ICE (US)	Sugar 11	SB
	ICE (US)	Coffee	KC
	ICE (US)	Cocoa	CC
	CME	Lumber	LB
Livestock			
	CME	Live Cattle	LC
	CME	Lean Hogs	LH
	CME	Feeder Cattle	FC

Table 2.3: **Average Benchmark Measures**

This table provides average benchmark values of all commodities in our sample for the period 2008 to 2018. Displayed units: twRQS, vwRES, and vwRPI in bps; λ in bps per million USD, λ^{root} in bps per root-million USD; $\sigma_s^{VAR(5)}$ in % p.a.. AMIM, VR_{30}, and $|AR_1|$ are unchanged. All measures are inverse measures of market quality, i.e., higher values indicate lower liquidity or price efficiency.

		Spreads		Price Impact			Efficiency			
Category	Ticker	twRQS	vwRES	vwRPI	λ	λ^{root}	$\sigma_s^{VAR(5)}$	AMIM	VR_{30}	$\|AR1\|$
Energy	CL	1.96	2.56	1.90	0.64	2.34	15.82	-0.19	0.31	0.05
	HO	2.77	2.94	1.86	1.05	2.40	14.50	-0.16	0.29	0.05
	NG	4.00	5.05	3.76	2.07	4.72	20.55	-0.12	0.33	0.06
	NGLNM	36.45	22.42	13.74	0.98	3.55	55.18	-0.07	0.58	0.10
	LCO	2.42	2.98	1.44	0.11	0.69	10.59	-0.07	0.24	0.04
	LGO	4.79	4.91	2.13	0.05	0.81	11.50	-0.08	0.29	0.06
Grains	S	6.82	3.14	2.70	0.69	2.23	13.73	-0.08	0.46	0.06
	C	10.50	6.60	5.52	1.01	3.12	18.90	-0.06	0.48	0.08
	W	12.90	5.89	5.22	2.44	5.12	22.21	-0.07	0.46	0.07
	KW	23.05	6.15	5.64	6.16	7.41	21.61	-0.04	0.49	0.07
	SM	12.78	4.34	3.65	2.46	4.38	17.29	-0.05	0.52	0.07
	BO	10.72	3.91	3.33	1.96	3.48	14.12	-0.06	0.55	0.07
	RR	56.40	20.25	16.98	3392.50	210.65	60.93	-0.11	0.64	0.09
	O	71.54	28.10	23.07	120.52	43.80	77.32	-0.11	0.52	0.09
Metals	GC	1.16	1.48	1.20	0.27	1.01	7.00	-0.12	0.36	0.06
	SI	3.80	3.78	3.39	1.99	3.61	14.37	-0.14	0.32	0.06
	HG	3.00	2.75	2.69	1.90	3.11	11.23	-0.21	0.30	0.05
	PL	6.91	5.62	5.46	4.03	4.55	16.11	-0.16	0.36	0.06
	PA	14.27	12.00	10.35	12.67	10.43	32.57	-0.12	0.45	0.08
Softs	CT	5.21	5.32	3.58	2.63	4.70	21.22	-0.07	0.47	0.07
	SB	7.01	8.36	6.07	1.79	4.30	21.87	-0.15	0.34	0.06
	KC	6.30	7.01	3.45	1.53	4.72	21.21	-0.18	0.40	0.06
	CC	6.87	6.85	4.44	3.12	6.00	21.70	-0.15	0.44	0.06
	LB	28.91	26.94	22.58	61.87	27.95	67.24	-0.10	0.54	0.10
Livestock	LC	3.68	3.60	3.56	2.10	3.38	11.79	-0.10	0.35	0.07
	LH	5.98	5.81	5.44	5.68	7.08	19.13	-0.08	0.37	0.07
	FC	8.42	6.85	6.76	6.22	6.19	20.78	-0.09	0.38	0.07

Aggregate Market Quality In order to obtain an impression of time-variation in aggregate market quality (AMQ), we compute two indices which are depicted in Figure 2.1. The index in Panel A is the first principal component of all standardized market quality benchmark measures and all commodities (except RR due to missing data). The index in Panel B is created using repeated averaging. First, we average the standardized benchmark measures across all 28 commodities and measures for the categories liquidity and price efficiency. Then, we take an average of the obtained liquidity and price efficiency indices. Finally, we multiply the resulting index by -1 to obtain a measure of AMQ.

Both indices behave very similar (correlation of 99%) and exhibit a positive trend as well as a sharp negative spike in late 2008. During this time, commodity prices collectively spiked and subsequently declined sharply until 2009. Commodity volatility (measured using an exponentially moving average process of daily squared GSCI returns with a decay parameter of 0.94) also spiked during this time and exhibits a correlation of around -89% with AMQ during our sample period. In line with the findings of Rösch et al. (2017), AMQ is also highly correlated with funding liquidity approximated by the TED spread (correlation of around 79%) and the VIX (correlation of 85%).

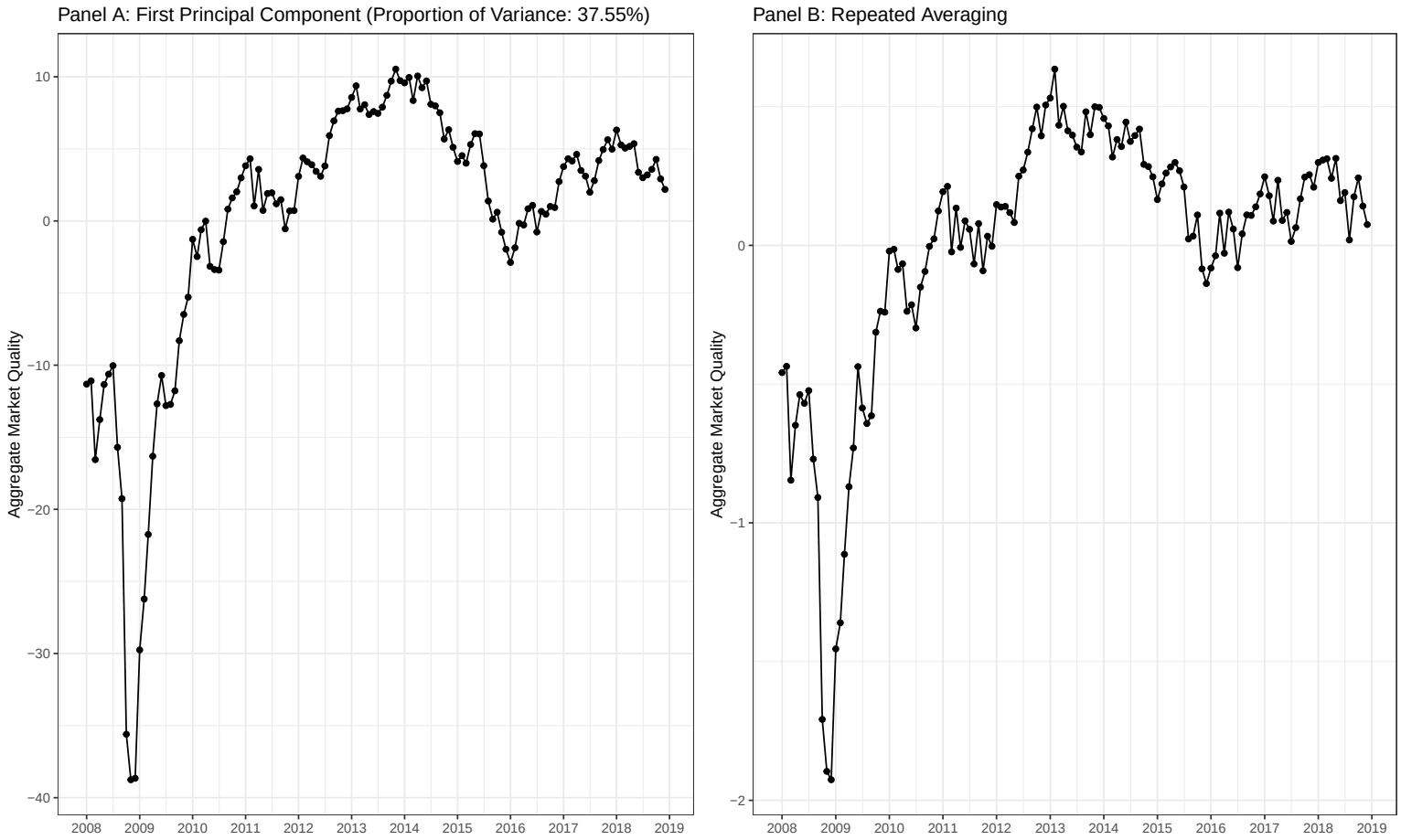

Figure 2.1: **Aggregate Market Quality**

This figure shows monthly estimates of aggregate commodity market quality. Panel A shows the first principal component of all measures and commodities. Panel B shows an index generated using repeated averaging. For each commodity, we first scale each benchmark measure in our sample to zero mean and unit variance. Then, for liquidity and efficiency, we average across commodities and measures. We then take the average of the resulting liquidity and efficiency time-series. Finally, we multiply the resulting index by minus 1 to obtain aggregate market quality (AMQ).

2.5 Main Results

Time-Series Correlations We start by calculating the Pearson time-series moment-correlation of each benchmark–proxy pair and for each commodity. The results are shown in Table 2.4. We average the coefficients across commodities and list them in the columns AvgCor. **Bold** numbers indicate the highest average correlation higher than 0.4 as well as any correlation coefficient that is not different from the highest at a 5% significance level. Following Goyenko et al. (2009), we use a t-test in the spirit of Fama and MacBeth (1973) after transforming the correlation coefficients using the Fisher z-transform $z = \frac{ln(1+\rho)}{1-\rho}$, where ρ is the correlation coefficient so they follow a normal distribution. Additionally, we select a winner among proxies for each of the 28 commodity based on the highest correlation coefficient (that is at least 0.4). A proxy is also considered a winner, if it is not significantly different from the best proxy at the 5% confidence level using the test by Zou (2007). We count the number of wins across commodities report it under $\#wins$.[4] Since correlations of price efficiency proxies are low, we count the number of positive and significant (5% level) coefficients under $\#sig$.

Panel A shows the the results for spread benchmark–proxy pairs. The coefficients imply that VoV(Spread) and the Amihud measure exhibit the highest average correlation with both benchmarks, twRQS and vwRES. VoV(Spread) is the proxy with the highest correlation for 21 and 25 out of 28 commodities. Measures based on the Roll model perform worst. This might be due to the fact that settlement prices are not actual last trade prices executed at the bid or ask but weighted averages during certain time-intervals of the day.

The results for price impact are shown in Panel B. VoV(λ) performs best. Its average correlations are the highest and its correlations are highest

[4]In the Appendix, we provide commodity-by-commodity time-series correlations for vwRES, vwRPI, and $\sigma_s^{VAR(5)}$ with their respective proxies. $\#wins$ corresponds to the number of **bold** font entries by column.

Table 2.4: **Time-Series Correlations of Proxies and Benchmarks**

This table shows results for the time-series correlation of market quality proxies and benchmarks. The row indicates the proxy measure estimated each month from daily data and the column indicates the benchmark estimated every day from intraday data and aggregated to a monthly frequency. We calculate the correlation of a benchmark–proxy pair for each commodity. AvgCor indicates the average correlation coefficient. ***Bold*** *numbers indicate the highest coefficient that is greater than* 0.4*. Those that are not different from it at a* 5% *significance level are also in* ***bold*** *font. We use a t-test of Fisher z-transformed coefficients in the spirit of Fama and MacBeth (1973). #wins indicates the number of commodities for which the proxy exhibits the highest correlation greater than 0.4 or is not significantly different from the highest one using the test by Zou (2007) with a 5% confidence level. #sig is the number of commodities for which the proxy's correlation coefficient is positive and significantly different from zero at the 5% level.*

Panel A: Spread

	twRQS		vwRES	
	AvgCor	#wins	AvgCor	#wins
Amihud	**0.684**	12	**0.685**	13
VoV(Spread)	**0.731**	21	**0.740**	25
Roll	0.301	1	0.307	1
RollAbs	0.334	0	0.349	0
RollGibbs	0.437	0	0.439	0
EffTick	0.509	7	0.519	8
HighLow	0.511	1	0.502	2
AbdiRanaldo	0.520	2	0.511	1
AvgSpread	**0.663**	3	**0.672**	7

Panel B: Price Impact

	vwRPI		λ		λ^{root}	
	AvgCor	#wins	AvgCor	#wins	AvgCor	#wins
Amihud	**0.608**	13	**0.796**	13	0.800	7
$1oV$	0.467	3	0.638	2	0.583	0
$VoV(\lambda)$	**0.650**	22	**0.818**	21	**0.864**	25
RolloV	0.377	0	0.554	0	0.562	2
RollAbsoV	0.490	2	0.650	1	0.645	0
RollGibbsoV	**0.560**	3	**0.759**	6	0.748	0
EffTickoV	0.502	6	0.667	1	0.622	0
HighLowoV	**0.581**	12	**0.790**	10	0.775	5
AbdiRanaldooV	0.555	7	0.745	7	0.740	6
PastorStambaugh	−0.008	0	0.052	0	0.040	0
AvgImpact	**0.617**	15	**0.826**	22	0.814	10

Panel C: Efficiency

	$\sigma_s^{VAR(5)}$		AMIM		VR_{30}		$\lvert AR1\rvert$	
	AvgCor	#sig	AvgCor	#sig	AvgCor	#sig	AvgCor	#sig
$\sigma_s^{MA(1)}$	0.296	20	0.008	2	−0.000	4	0.027	4
AMIM	0.028	1	0.017	0	0.003	1	0.023	2
VR_2	−0.019	3	0.013	0	0.031	2	0.011	0
$\lvert AR1\rvert$	−0.022	1	0.011	1	−0.002	1	−0.016	1
AvgEff	0.098	9	0.019	2	0.010	1	0.013	1
AvgSpread	**0.787**	28	0.023	3	0.012	4	0.056	4
AvgImpact	**0.717**	28	0.032	3	0.062	5	0.101	5

for most commodities. When daily high and low prices are not available, the Amihud measure is the best choice. Combining multiple proxies by averaging does not improve the correlation beyond the best single proxy for any benchmark – neither for spread nor for impact proxies. We also compute proxy–benchmark correlations by year (detailed results are tabulated in the Appendix) that show that VoV–measures perform either at par are superior to the Amihud measure in each of the years in the sample.

Panel C presents results for price efficiency measures. Of the single measures, only the daily $\sigma_s^{MA(1)}$ measure is able to capture some variation in $\sigma_s^{VAR(5)}$. The average efficiency proxy is also virtually uncorrelated with all the benchmarks. However, spread and impact measures are able to capture some of the variation in $\sigma_s^{VAR(5)}$. This might be the case because both liquidity and $\sigma_s^{VAR(5)}$ are correlated with volatility. Another possible explanation is the connection between liquidity and price efficiency through arbitrage activity (Chordia et al., 2008). These results show that it appears to be difficult to approximate time-series variation in commodity price efficiency at lower frequencies with existing approaches.

Time-Series Correlation of Aggregate Benchmarks and Proxies Next, we study the best way to approximate aggregate (systematic) market quality. Chordia et al. (2000) find that there exists commonality in liquidity in the equity market. Some of the asset pricing literature claims that there exists a risk premium for the covariance of asset returns with market-wide liquidity (Amihud, 2002; Pástor and Stambaugh, 2003; Acharya and Pedersen, 2005; Sadka, 2006). Marshall et al. (2013) find temporal systematic liquidity in commodities that is separate from systematic liquidity in equities. There also appears to exist systematic price efficiency in stocks (Rösch et al., 2017). In each of these studies, researchers face the decision of choosing the most appropriate proxy for building a liquidity or efficiency factor. Thus, we provide

some guidance on picking a proxy for commodities and compute market-wide market quality benchmarks and proxies to compare their correlations.

First, we scale each measure for each commodity to zero mean and unit standard deviation. Then, for each month, we compute cross-sectional averages for each measure. This way, we obtain a single time-series for every measure. Finally, we compute correlation estimates for each benchmark–proxy pair.

We report the estimates in Table 2.5. The correlations of spread measures in Panel A suggest that all proxies are reliable. As for individual commodities, VoV(Spread) is the best spread proxy. For price impact proxies in Panel B, the results are similar. The VoV(λ) measure is the best proxy but almost all proxies can be used to approximate aggregate price impact. The correlations of efficiency measures in Panel C suggest that aggregate variance ratios and the absolute autocorrelation coefficient are mildly correlated with spread and price impact proxies and highly correlated with $\sigma_s^{VAR(5)}$. Among the efficiency measures, only $\sigma_s^{MA(1)}$ is correlated with $\sigma_s^{VAR(5)}$ and to a low degree with $|AR1|$.

Overall, the results suggest that it is easier to approximate aggregate market quality, especially liquidity. Aggregate spread and price impact proxies are almost all highly correlated with the benchmarks. Systematic variance ratios and autocorrelation measures are difficult to capture at a low frequency. These improvements in correlation likely occur, because averaging reduces noise in proxy estimates and benchmarks. Finally, we can see that the winners in the aggregate are the same as for individual commodities.

Cross-Sectional Correlations In asset pricing studies, portfolio sorts are very common and also used while searching for commodity risk factors (see, e.g., Yang, 2013). To simplify the choice of the most appropriate market quality measures to sort commodities, we provide some guidance in this section.

Table 2.5: **Time-Series Correlations of Aggregate Proxies and Benchmarks**

This table shows the Pearson moment-correlations of proxies for aggregate market quality with different benchmarks. The row indicates the proxy measure estimated each month from daily data and the column indicates the benchmark estimated every day from intraday data and aggregated to a monthly frequency. Each month, we average measures and proxies across commodities. Then, we calculate the correlation of each measure–proxy pair and report them in this table. ***Bold*** *numbers indicate the highest correlation coefficient and all those that are not different from the highest at the significance 5% level. We employ the test by Zou (2007) to test for differences in the correlation coefficients.*

Panel A: Spread

	twRQS	vwRES
Amihud	0.928	**0.897**
VoV(Spread)	**0.956**	**0.913**
Roll	0.659	0.658
RollAbs	0.769	0.756
RollGibbs	0.830	0.792
EffTick	0.918	0.883
HighLow	0.890	0.849
AbdiRanaldo	0.908	0.855
AvgSpread	0.939	**0.902**

Panel B: Price Impact

	vwRPI	λ	λ^{root}
Amihud	0.886	**0.949**	0.943
$1oV$	0.760	0.849	0.830
$VoV(\lambda)$	**0.915**	**0.950**	**0.960**
RolloV	0.777	0.827	0.836
RollAbsoV	0.854	0.917	0.915
RollGibbsoV	0.858	0.932	0.924
EffTickoV	0.848	0.921	0.906
HighLowoV	0.873	**0.947**	0.941
AbdiRanaldooV	0.885	**0.939**	0.931
PastorStambaugh	0.241	0.270	0.246
AvgImpact	0.879	**0.946**	0.940

Panel C: Efficiency

	$\sigma_s^{VAR(5)}$	AMIM	VR_{30}	$\lvert AR1\rvert$
$\sigma_s^{MA(1)}$	0.693	0.130	0.109	0.332
AMIM	0.062	−0.122	−0.042	0.075
VR_2	−0.136	−0.075	−0.120	−0.054
$\lvert AR1\rvert$	−0.109	−0.152	−0.073	−0.070
AvgEff	0.221	−0.063	−0.065	0.092
AvgSpread	**0.951**	0.138	0.189	0.360
AvgImpact	0.883	0.098	0.209	0.317

For each benchmark–proxy pair, we calculate the Spearman rank correlation in a given month. We repeat this procedure every month to obtain a time-series of monthly cross-sectional correlations. Then, we average across months and report them under AvgCor. In every of the 132 months, we select the winner as the proxy with the highest correlation coefficient, or the one that is not significantly different from it at the 5% confidence level using a Fisher z-based test described in Sheshkin (2004).

The results are shown in Table 2.6. Among the spread proxies in Panel A, the VoV(Spread), Amihud, and EffTick measures exhibit the highest correlation with the benchmark. They also emerge as the winners in most of the 132 months. All other measures do not seem to capture cross-sectional variation in spreads very well.

Results for price impact in Panel B show that the VoV(λ), EffTickoV, and HighLowoV measures perform best. Except for the PastorStambaugh measure, the correlations of all measure–proxy pairs are in a similar range. In the cross-section, 1oV also seems to be a valid price impact proxy. So it is likely that the similarly high correlation of all measures except the Pástor and Stambaugh (2003) measure is driven by differences in volume. Thus, volume appears to be a main driver for cross-sectional variation in liquidity.

Price efficiency correlations in Panel C show that liquidity proxies are able to capture $\sigma_s^{VAR(5)}$ reasonably well. Impact proxies also appear to be mildly correlated with other efficiency benchmarks except for AMIM.

In order to assess how sorted portfolio returns are affected by the choice of measure, we construct monthly rebalanced equally-weighted long-short portfolios. For all five liquidity benchmarks and three liquidity proxies (Amihud, VoV(Spread), VoV(λ)), we form illiquid minus liquid (IML) factors from the top and bottom 3 commodities sorted by the respective measure. We find that the IML-factor returns sorted by proxies are highly correlated (> 0.8) among themselves and their correlations with benchmark-sorted factor re-

Table 2.6: **Cross-Sectional Correlations of Proxies and Benchmarks**

This table shows results of cross-sectional rank correlations. The row indicates the proxy measure estimated each month from daily data and the column indicates the benchmark estimated every day from intraday data and aggregated to a monthly frequency. We calculate the correlation of a benchmark–proxy pair in each month. AvgCor indicates the average correlation coefficient. ***Bold*** *numbers indicate the highest coefficient that is greater than* 0.4*. Those that are not different from it at a* 5% *significance level are also in* ***bold*** *font. We use a t-test of Fisher z-transformed coefficients in the spirit of Fama and MacBeth (1973). #wins indicates the number of commodities for which the proxy exhibits the highest correlation greater than 0.4 or is not significantly different from the highest one using the Fisher z-based test described in Sheshkin (2004) with a 5% confidence level. #sig is the number of commodities for which the proxy's correlation coefficient is positive and significantly different from zero at the 5% level.*

Panel A: Spread

	twRQS		vwRES	
	AvgCor	#wins	AvgCor	#wins
Amihud	**0.757**	132	**0.808**	131
VoV(Spread)	**0.758**	132	**0.825**	131
Roll	0.074	14	0.159	17
RollAbs	0.104	5	0.175	7
RollGibbs	0.114	2	0.214	13
EffTick	0.700	107	0.628	81
HighLow	0.065	2	0.134	4
AbdiRanaldo	0.166	10	0.266	24
AvgSpread	0.550	79	0.638	81

Panel B: Price Impact

	vwRPI		λ		λ^{root}	
	AvgCor	#wins	AvgCor	#wins	AvgCor	#wins
Amihud	0.770	132	0.836	130	0.869	127
$1oV$	0.744	131	0.840	130	0.829	117
$VoV(\lambda)$	0.762	132	0.835	130	**0.894**	131
RolloV	0.682	119	0.771	118	0.804	122
RollAbsoV	0.741	132	0.818	128	0.843	124
RollGibbsoV	0.768	132	0.841	130	0.870	128
EffTickoV	**0.803**	132	0.741	113	0.783	102
HighLowoV	0.751	132	**0.855**	128	0.877	130
AbdiRanaldooV	0.753	130	0.802	129	0.861	129
PastorStambaugh	0.023	14	0.020	12	0.021	7
AvgImpact	0.769	132	0.832	129	0.865	128

Panel C: Efficiency

	$\sigma_s^{VAR(5)}$		AMIM		VR_{30}		$\lvert AR1\rvert$	
	AvgCor	#sig	AvgCor	#sig	AvgCor	#sig	AvgCor	#sig
$\sigma_s^{MA(1)}$	0.203	23	0.004	3	−0.019	3	−0.001	3
AMIM	0.019	9	0.045	6	0.015	4	0.024	5
VR_2	0.000	5	0.012	2	0.020	4	0.003	8
$\lvert AR1\rvert$	−0.028	2	0.022	2	−0.020	4	−0.017	5
AvgEff	0.068	9	0.036	2	0.003	4	0.004	4
AvgSpread	**0.727**	130	0.047	8	0.116	20	0.180	25
AvgImpact	**0.753**	132	0.088	3	0.312	51	**0.423**	83

turns are very similar (in the range of 0.63 to 0.8). For cross-sectional asset pricing studies, we would thus expect similar conclusions when commodities are sorted by the Amihud or a VoV–measure.

We also inspect the stability of cross-sectional correlations over time. Detailed results in the form of cross-sectional correlations over time for each proxy can be found in the Appendix. Overall, correlations are stable over time. Together with the temporal stability of time-series correlations (detailed results are also in the Appendix), this can be interpreted as an indication that our identified best proxies are robust to changes in the composition of market participants and market design and therefore also most suitable for approximating liquidity pre-2008.

Summary of Main Results Overall, our results show that VoV measures are the most reliable in capturing individual, aggregate and cross-sectional variation in liquidity. The Amihud measure is in some way also a volatility-over-volume measure with the absolute return in the numerator. However, during a high-volume high-volatility day, a low absolute return can give a false impression of low price impact. Thus, the Amihud measure is more prone to produce extreme values that require outlier-treatment. In our sample, about 4.8% of the monthly Amihud values are identified as outliers and excluded from the analysis – versus 0% for VoV(Spread) and 0.2% for VoV(λ). The volatility measure by Parkinson (1980) over volume as suggested by Fong et al. (2018) appears to capture volume-induced volatility (price impact) better than absolute returns. The use of high and low prices enables the measure to capture intraday volatility more reliably, which seems to give it an edge over the Amihud measure. If high and low prices are not available, one can estimate the volatility from end-of-day prices. We estimate daily volatility using an exponentially moving average model with a decay of 0.94 and find that it performs at par compared to the Amihud measure but pro-

duces fewer extreme values (VoV(Spread) with EWMA–volatility: 0.3% are outliers, VoV(λ) with EWMA–volatility: 3.1% are outliers). Thus, we recommend using VoV measures to approximate liquidity in commodity futures, especially when high and low prices are available.

Our results complement and extend previous horse-races of liquidity proxies (Goyenko et al., 2009; Marshall et al., 2012). They find that the EffTick and Amihud measure are the best liquidity proxies. Our average time-series correlations are slightly higher compared to those in Marshall et al. (2012). Time-series correlations of aggregate liquidity (equally weighted portfolios) in Goyenko et al. (2009) are also around 0.9 for spreads but lower for price impact (around 0.3 to 0.5). Average cross-sectional correlations are also slightly lower for equities compared to commodities. We show that new measures developed by Kyle and Obizhaeva (2016) and Fong et al. (2018) perform better. Our sample spans a time during which financial (and commodity) markets underwent changes in terms of market access, speed, and especially the rise of algorithmic trading. We show that these proxy measures, some of which were developed for slow dealer markets, remain to serve their purpose and are still useful.

We also add to the literature by studying efficiency measures at different frequencies. Our analysis suggests that price efficiency is notoriously hard to capture with low-frequency data. The pricing error volatility (e.g., estimated using a VAR(5) as in $\sigma_s^{VAR(5)}$) appears to be distinct from other measures like variance-ratios. It is correlated with volatility and liquidity measures, which are in turn also correlated with volatility. Other efficiency benchmarks are also correlated with liquidity proxies, but to a lesser degree. We can only recommend using $\sigma_s^{MA(1)}$ as a proxy when estimating systematic commodity price efficiency. For all other research designs, a daily frequency appears to be too low.

2.6 Further Analysis

2.6.1 Levels of Bid–Ask Spreads

In this section, we provide results for researchers wishing to incorporate the actual level of bid–ask spreads into their analysis. We also present estimates in order to allow mapping spread proxies to the level of their benchmarks. For example, a study of after-cost risk premia requires reasonable estimates of ideally both the variation and the level of the bid–ask spread. The recommendations in this section aim at enabling researchers to accurately estimate it without having to acquire and process intraday data and facilitate level-sensitive analyses of samples starting before 2008.

First, we provide time-series averages of inverse market quality measures estimated from quote data for each commodity in Table 2.3. For the estimation of after-cost returns, vwRES averages can be be employed as constants for the time period (2008–2018) they are measured in.

In order to enable analyses that require both the level and variation of the bid–ask spread, we quantify their linear relationships with our spread proxies. We estimate a linear regression with the quote-based benchmark as the dependent and the daily-data-based proxy as the independent variable (including a constant) in the spirit of Mincer and Zarnowitz (1969). We estimate this for each commodity and benchmark–proxy pair and report cross-sectional statistics. Table 2.7 shows the average constant, its cross-sectional standard error along a test statistic against the null that the mean is 0. It also includes the average slope with its relative standard error[5], a test statistic against the mean being 1, and the average R^2. An ideal proxy would result in a constant of 0, a slope of 1 and a small error. Of all proxies we consider, only the VoV–measures exhibit constants that are on average close to 0.

[5]We define the relative standard error as the regular standard error divided by the mean. We use it in order to make it more easily comparable across proxies

For quoted spreads, its average slope is also close to 1, along with that of the Effective Tick measure. Since effective spreads are lower than quoted spreads, the slopes are lower. Overall, the results suggest that most proxies are on a different scale than the benchmark measures. That means, they would require scaling to be used in an analysis involving the actual level of the bid–ask spread. Consistent with our correlation results, VoV(Spread) exhibits the smallest errors as indicated by the highest R^2. Overall, VoV–measures appear to be best suited for capturing both the level and the variation in bid–ask spreads of commodity futures as its constant is close to 0, its slope is close to 1 with smaller (relative) errors and higher R^2 compared to other proxies. However, the cross-sectional (relative) errors of the constant and slope are not negligible. Thus, we provide estimates for the constant and slope of vwRES–VoV(Spread) regressions by commodity in Table 2.8. In order to obtain the most accurate effective spread proxy–estimates using daily data, we suggest computing the VoV(Spread) measure and then scaling it for each commodity individually by multiplying it by the slope and adding the constant in bps.

Table 2.7: **Benchmark–Proxy Regressions**

This table shows the result of benchmark–proxy regressions. For each proxy–benchmark pair and commodity, we estimate OLS regressions with the benchmark as the dependent variable. Then, we compute the cross-sectional average constant (in bps), slope and R^2 across all 28 commodities. SE(Const) is the cross-sectional standard error of the average constant in bps and Rel SE(Slope) is the relative cross-sectional standard error of the average slope which is the standard error divided by the average slope.

Benchmark	Proxy	Avg Const (bps)	SE(Const) (bps)	t(Const=0)	Avg Slope	Rel SE(Slope) (%)	t(Slope=1)	Avg R^2 (%)
twRQS	Amihud ($\times 10^8$)	9.216	2.644	3.486	0.304	17.702	-12.955	53.194
	VoV(Spread)	0.251	1.946	0.129	0.961	20.891	-0.196	58.151
	Roll	11.256	2.842	3.960	0.021	36.050	-129.203	12.952
	RollAbs	10.272	2.731	3.761	0.027	29.739	-123.037	15.093
	RollGibbs	8.087	2.322	3.483	0.098	27.502	-33.301	23.259
	EffTick	8.116	2.691	3.016	1.020	19.085	0.101	30.352
	HighLow	6.964	2.144	3.248	0.120	22.872	-31.940	30.048
	AbdiRanaldo	9.828	2.363	4.158	3.467	24.838	2.865	30.698
vwRES	Amihud ($\times 10^8$)	5.325	0.946	5.628	0.203	22.172	-17.753	50.776
	VoV(Spread)	-0.393	1.037	-0.379	0.591	17.641	-3.926	57.856
	Roll	6.705	1.099	6.103	0.014	38.119	-180.912	13.121
	RollAbs	6.023	1.032	5.835	0.018	28.212	-191.723	15.954
	RollGibbs	4.815	0.853	5.642	0.062	25.391	-59.728	23.071
	EffTick	4.655	0.898	5.186	0.713	20.374	-1.971	31.355
	HighLow	4.328	0.877	4.936	0.074	23.377	-53.850	28.904
	AbdiRanaldo	6.009	0.975	6.161	2.091	22.910	2.278	29.283

Table 2.8: **vwRES–VoV(Spread) Regressions by Commodity**

This table shows the constant, slope and R^2 of regressions in which vwRES is the dependent variable and VoV(Spread) is the independent variable.

Category	Ticker	Const (bps)	Slope	R^2 (%)
Energy	CL	0.26	0.58	75.19
	HO	0.03	0.48	64.62
	NG	2.00	0.40	36.18
	NGLNM	-12.10	2.95	76.51
	LCO	0.89	0.49	62.72
	LGO	1.68	0.56	52.58
Grains	S	1.60	0.34	51.66
	C	4.77	0.31	11.88
	W	3.08	0.32	33.17
	KW	2.67	0.27	43.92
	SM	1.82	0.30	57.58
	BO	1.01	0.37	56.89
	RR	-4.47	0.17	80.48
	O	-1.52	0.56	62.45
Metals	GC	0.22	0.47	86.76
	SI	0.99	0.43	76.73
	HG	0.38	0.35	90.02
	PL	-3.47	0.78	82.69
	PA	-2.59	0.73	90.57
Softs	CT	0.74	0.40	67.59
	SB	6.22	0.21	8.83
	KC	3.39	0.34	17.92
	CC	1.81	0.37	78.67
	LB	-18.68	1.26	38.29
Livestock	LC	-0.88	0.72	66.72
	LH	-0.17	0.55	61.84
	FC	-9.57	1.55	60.85

2.6.2 Time-and-Sales Data

Using daily instead of millisecond time-stamped quote data is a step that drastically reduces computation time from weeks to minutes. However, there is a middle ground. Intraday data for commodity futures were recorded when the contracts were traded in the pits – Time-and-Sales (TAS) data. In this

section, we test if it possible to capture market quality using 5-minute TAS data with the same or similar proxy measures.

We transform our tick-by-tick intraday data that include volume and quotes to TAS data by taking 5-minute snapshots of trade prices and count the number of trades in each interval. Volume and and quotes are discarded. Thus, the data are in the format as if they were recorded in the pits which also allows to draw some conclusions about the validity of the proxies across market designs.

With 5-minute data, computation time for most measures is still around several minutes. Only the RollGibbs measure requires several days to estimate. Besides possibly yielding more accurate estimates, this compromise also yields daily estimates that do not suffer from extreme levels of autocorrelation as a rolling-window approach of daily measures.

Another advantage of 5-minute data is that the end-of-period prices are actual trade prices instead of averages like the settlement price. This might lead to an increase in the accuracy of Roll estimators. The fact that we now observe trade prices that hit standing bid or ask limit orders complicates measurements of price efficiency. In contrast to daily settlement prices, intraday trade prices suffer from a bid–ask bounce that induces auto-covariance into the trade price process. The $\sigma_s^{MA(1)}$ requires only trade prices and should therefore be more exact compared to when estimated from daily data. However, we do not consider AMIM, $VR_{2/10/30}$ or $|AR1|$ because of the bid–ask bounce. Instead, we resort to the adjusted variance ratio of Smith (1994).

For estimation, we treat each 5-minute interval like we treated daily data before. We record the last price, the high, and the low price. Since TAS data do not include recordings of volume, we use the number of trades as a proxy. VoV measures can be computed on a daily basis, so we include two versions each: one is computed using 5-minute data with the number of trades; the other is computed from daily high, low, and volume data. For comparison, we

also include estimates of the Amihud (2002) measure computed from daily single data-points. We estimate each proxy for each commodity and each trading day. We also compute VoV–measures that use the same information as the Amihud (2002) measure, i.e., settlement prices and volume. We employ an exponentially weighted moving average (EWMA) process with a decay factor of 0.94 to obtain daily volatility estimates and compute VoV(Spread, Daily, EWMA) as well as VoV(λ, Daily, EWMA).

As for monthly data, we compute pairwise time-series Pearson moment-correlations and report cross-sectional averages that we test against the best of each benchmark. Table 2.9 reports the results.

Average correlations of spread measures in Panel A show similar results to lower frequency data. VoV(Spread) measured using daily data emerges as being sufficient to capture daily variations in spreads. That means, instead of using 5-minute TAS data, researchers can rely on the daily VoV(Spread) to obtain a daily spread proxy.

Panel B shows average correlations for price impact-measure pairs. The results suggest that the daily VoV(λ) measure is the best daily price impact proxy. Therefore, as for spreads, we recommend researchers to approximate price impact on a daily frequency using the daily version of VoV(λ).

Taken together, the results for spreads and price impact imply that VoV–measures estimated from daily data are able to capture variation in liquidity at a daily frequency significantly better than the daily Amihud (2002) measure. This is a simple result of the fact that some days might exhibit high volatility but near-zero returns, which leads to an underestimation of liquidity when a single data point is used to compute the Amihud (2002) measure. VoV–measures, even if they are computed from daily returns and volume alone (VoV(Daily, EWMA)), are able outperform the daily Amihud (2002) measure in terms of correlation with the benchmarks.

Results for efficiency proxies in Panel C suggest that $\sigma_s^{MA(1)}$ exhibits

Table 2.9: **Time-and-Sales Proxies**

This table shows cross-sectional averages of time-series correlation coefficients of daily (monthly) benchmark measures estimated from intraday data including quotes (columns) and daily (monthly) proxies estimated from 5-minute TAS data (rows). We add versions of some proxy measures which are calculated from daily data. Average correlations of monthly averages are reported in the columns 'Monthly'. ***Bold*** *numbers indicate the highest coefficient that is greater than* 0.4*. Those that are not different from it at a* 5% *significance level are also in* ***bold*** *font. We use a t-test of Fisher z-transformed coefficients in the spirit of Fama and MacBeth (1973).*

Panel A: Spread

	Daily		Monthly	
	twRQS	vwRES	twRQS	vwRES
Amihud	**0.580**	**0.458**	**0.652**	**0.591**
Amihud(Daily)	0.321	0.308	0.577	0.556
VoV(Spread)	**0.585**	**0.602**	**0.699**	**0.727**
VoV(Spread, Daily)	**0.550**	**0.577**	**0.728**	**0.735**
VoV(Spread, Daily, EWMA)	**0.562**	0.498	**0.648**	0.646
Roll	0.290	0.367	0.536	0.586
RollAbs	0.355	0.388	**0.637**	0.653
RollGibbs	0.474	0.508	**0.653**	**0.684**
EffTick	0.453	**0.502**	**0.663**	**0.783**
HighLow	0.312	0.412	0.482	0.543
AbdiRanaldo	0.439	0.514	**0.605**	0.679
AvgSpread	**0.561**	**0.612**	**0.711**	**0.754**

Panel B: Price Impact

	Daily			Monthly		
	vwRPI	λ	λ^{root}	vwRPI	λ	λ^{root}
Amihud	0.256	0.524	0.514	**0.548**	0.676	0.654
Amihud(Daily)	0.229	0.335	0.370	0.535	0.685	0.690
1oV	0.089	0.275	0.217	0.310	0.417	0.359
VoV(λ)	0.345	**0.641**	**0.723**	**0.660**	**0.771**	0.819
VoV(λ, Daily)	**0.404**	**0.620**	**0.708**	**0.650**	**0.844**	**0.882**
VoV(λ, Daily, EWMA)	0.257	**0.639**	**0.661**	**0.570**	**0.766**	0.804
RolloV	0.127	0.360	0.333	0.514	0.607	0.576
RollAbsoV	0.232	0.540	0.528	**0.642**	0.775	0.743
RollGibbsoV	0.266	**0.658**	0.627	**0.678**	**0.823**	0.788
EffTickoV	0.232	0.491	0.423	**0.602**	0.660	0.594
HighLowoV	0.251	**0.583**	0.570	**0.647**	0.744	0.730
AbdiRanaldooV	0.313	**0.644**	**0.666**	**0.674**	**0.803**	0.806
PastorStambaugh	0.039	0.000	−0.012	0.000	−0.038	−0.047
AvgImpact	0.316	**0.669**	**0.693**	**0.644**	**0.812**	**0.827**

Panel C: Efficiency

	Daily				Monthly			
	$\sigma_s^{VAR(5)}$	AMIM	VR_{30}	$\lvert AR1\rvert$	$\sigma_s^{VAR(5)}$	AMIM	VR_{30}	$\lvert AR1\rvert$
$\sigma_s^{MA(1)}$	0.326	0.077	0.149	0.026	0.733	0.069	0.145	0.074
VR^{Smith}	0.021	0.053	0.219	0.017	0.145	0.115	0.286	0.110
AvgEff	0.205	0.081	0.232	0.027	0.610	0.104	0.251	0.103
AvgSpread	**0.595**	0.034	0.007	0.017	**0.890**	0.023	0.023	0.040
AvgImpact	0.508	0.027	0.029	0.047	0.827	0.050	0.044	0.095

a medium correlation with $\sigma_s^{VAR(5)}$. Also, VR^{Smith} as well as the average efficiency proxy are able to capture VR_{30} to some degree.

Lastly, to enable a 'fair' comparison of monthly proxies estimated from daily data with daily proxies estimated from 5-minute TAS data, we aggregate the latter to a monthly frequency by taking simple averages. We report the results in along the daily results in Table 2.9. The results for spreads in Panel A suggest that all spread proxies work well but not significantly better than VoV(Spread) estimated from daily data. The same conclusion can be drawn from results for price impact in Panel B. Average correlations of spread and price impact measures are similar in size compared to those in Table 2.4 estimated from daily data. High correlations of TAS-based liquidity proxies and their benchmarks imply that the measures are expected to be valid when estimated from data generated in the pits.

Correlations of price efficiency measures in Panel C show that it is possible to approximate price efficiency using TAS data at a medium frequency. The average correlation of $\sigma_s^{MA(1)}$ with $\sigma_s^{VAR(5)}$ is high and the correlation of VR^{Smith} with VR_{30} is in a medium range. Thus, we suggest researchers who wish to approximate commodity price efficiency to use intraday (e.g., 5-minute) TAS data instead of daily data. Temporal aggregation can help in reducing noise and increasing the benchmark–proxy correlation to medium levels. We also recommend using both $\sigma_s^{MA(1)}$ and VR^{Smith} as proxies since they appear to capture different aspects of the price process.

2.7 Conclusion

In this paper, we identify the best low-frequency proxies for liquidity and informational efficiency of commodity futures markets.

In contrast to previous studies, we have 11 years of high-frequency data available. Our sample includes periods during which algorithmic trading rose

to become a central part of market activity. Our findings suggest that most proxies remain valid in a modern market environment. Some of the measures we employ have not been included in previous studies of commodity liquidity including VoV measures which we find to be the best liquidity proxies. Instead of solely focusing on the liquidity aspect of market quality, we also include price efficiency into our analysis which appears to be notoriously hard to capture using low-frequency data. We also study the usefulness of TAS-based proxies – a data type available since 1996 in Datascope that represents a middle ground in terms of computational cost but enables the estimation of price efficiency proxies that exhibit a medium to high correlation with their benchmarks.

Based on our results we recommend the following:

(1) Spreads are best captured by the VoV(Spread) measure of Kyle and Obizhaeva (2016).

(2) Price impact is best captured by VoV(λ) of Fong et al. (2018).

(3) In contrast to the Amihud (2002) measure, VoV–measures computed from daily data are able to capture daily variation in liquidity.

(4) To approximate price efficiency, we recommend using both the variance ratio of Smith (1994) and the pricing error volatility (MA(1)) of Hasbrouck (1993) estimated using 5-minute TAS data. If intraday data is not available, spread or impact proxies are valid alternatives to capture Hasbrouck's pricing error volatility.

(5) In order to obtain unbiased effective spreads estimates, we recommend using the VoV(Spread) measure and then mapping it to the correct level and variance using the parameters provided in Table 2.8.

Our findings provide guidance for researchers and policy makers who aim to investigate commodity market quality and face the decision of picking

an appropriate measure given their research design: individual commodities, systematic market quality or cross-sectional sorts. Asset pricing studies often involve sorted portfolios or aggregate measures while individual commodity market quality is needed for investigations of changes in market structure like the financialization of commodity markets, the emergence of electronic trading and the subsequent rise of algorithmic trading.

A Appendix

This appendix provides additional details on the computation of the market quality measures we employ and the accuracy of commodity market quality proxies. Sections A.1 and A.2 explain how we estimate liquidity and price efficiency, respectively. Proxy averaging is explained in Section A.3. Sections A.4 and A.5 show that VoV–measures are always at par or superior to the Amihud measure when applying cross-sectional or time-series sample splits. Section A.6 shows that cross-sectional correlations exhibit considerable stability over time. Section A.7 tests more advanced forecast combination techniques to combine proxies none of which is superior to simple averages or the single best proxies. Section A.8 provides insights that indicate that the low correlations of price efficiency proxies estimated from daily data are due to noise. We provide averages of proxy measures for our sample in Table A.1. Finally, we show the time-series evolution and cross-sectional scale of three selected benchmark measures (vwRES, λ^{root}, and $\sigma_s^{VAR(5)}$) in Figure A.1.

Table A.1: **Average Proxy Measures**

This table shows average values of proxy measures for each commodity in the sample 2008–2018.

Ticker	Amihud ($\times 10^8$)	VoV(Spread)	Roll	RollAbs	RollGibbs	EffTick	HighLow	AbdiRanaldo	1oV ($\times 10^6$)	VoV(λ) ($\times 10^3$)	RolloV ($\times 10^8$)
BO	10.13	7.94	106.46	101.30	45.48	4.23	55.16	0.87	9.36	10.36	9.63
C	4.06	5.96	131.21	115.03	56.45	8.47	59.33	1.16	3.30	6.76	4.12
CC	42.53	13.46	143.17	131.80	58.40	6.29	64.21	1.26	34.43	22.84	45.49
CL	0.95	3.94	182.62	162.87	74.59	2.27	82.33	1.89	0.61	3.77	0.85
CT	26.77	11.30	109.82	116.27	53.40	3.33	59.80	1.17	23.34	17.94	16.42
FC	38.07	10.56	65.30	69.21	31.77	2.84	33.49	0.34	55.66	15.78	20.79
GC	0.61	2.70	69.15	73.36	35.09	1.32	40.51	0.46	0.80	2.13	0.25
HG	7.03	6.86	115.60	120.25	55.23	2.44	65.54	1.04	4.57	9.09	2.50
HO	3.88	6.03	150.95	128.32	61.64	1.17	71.58	1.45	2.47	7.09	4.02
KC	17.83	10.59	151.99	140.42	69.34	5.17	71.69	1.46	13.99	15.62	19.77
KW	38.43	13.08	148.19	142.23	59.95	7.67	66.88	1.45	27.90	21.81	44.17
LB	909.57	35.69	126.21	131.64	61.25	8.67	52.88	1.56	711.30	97.88	782.99
LC	7.73	6.18	74.55	70.24	34.06	3.26	33.89	0.38	11.21	7.01	7.69
LCO	1.31	4.28	136.39	147.07	71.21	2.65	74.50	1.71	0.85	4.27	1.03
LGO	3.60	5.79	131.31	122.62	56.21	5.05	67.74	1.78	2.53	6.63	3.35
LH	27.91	10.80	141.36	117.00	64.99	4.99	49.62	1.03	23.11	16.42	30.35
NG	5.24	7.64	227.53	192.88	97.17	4.28	98.36	2.96	2.61	9.69	5.06
NGLNM	43.39	11.67	171.97	142.99	69.63	18.45	43.17	1.37	23.28	19.33	44.41
O	1998.78	53.06	134.53	130.76	60.44	17.83	76.72	1.37	1601.40	177.65	1973.89
OJ	2986.83	42.67	135.57	151.89	60.13	8.85	59.46	1.58	2274.91	139.27	755.46
PA	236.00	20.06	194.61	147.40	61.36	3.28	75.19	1.48	127.52	47.53	333.32
PL	47.43	11.60	137.74	113.51	47.18	1.91	53.31	0.84	32.41	19.45	47.01
RR	68631.58	145.72	98.87	104.36	47.45	9.78	50.79	0.81	69606.87	813.48	55844.21
S	2.02	4.61	111.01	99.44	46.96	19.18	52.13	0.75	1.92	4.51	1.99
SB	15.09	10.19	148.71	138.48	70.13	9.98	75.97	1.86	10.96	14.85	14.79
SI	4.99	6.57	145.08	139.67	64.31	3.11	72.72	1.61	3.52	8.19	2.49
SM	10.47	8.39	131.66	115.99	57.90	5.52	63.56	1.08	8.36	11.33	10.32
W	9.78	8.77	160.47	147.34	65.96	6.67	75.79	1.77	6.82	12.15	11.06

(continued)

Table A.1: **Average Proxy Measures (Continued)**

Ticker	RollAbsoV ($\times 10^8$)	RollGibbsoV ($\times 10^8$)	EffTickoV ($\times 10^{10}$)	HighLowoV ($\times 10^8$)	AbRaoV ($\times 10^{10}$)	PastorStambaugh ($\times 10^{10}$)	$\sigma_s^{MA(1)}$ (% p.a.)	AMIM	VR_2	$\lvert AR1 \rvert$
BO	8.94	4.09	38.00	4.91	8.16	158.59	3.05	-0.25	0.16	0.16
C	3.53	1.76	27.57	1.85	3.65	-4.64	3.80	-0.27	0.14	0.15
CC	40.89	18.60	200.81	20.26	41.56	155.87	4.61	-0.21	0.16	0.16
CL	0.76	0.35	0.93	0.38	0.92	1.12	5.58	-0.27	0.12	0.16
CT	25.15	11.48	64.29	12.84	25.82	-207.76	4.67	-0.24	0.15	0.15
FC	30.65	14.53	172.35	16.55	16.68	-226.15	2.34	-0.28	0.15	0.15
GC	0.43	0.21	0.76	0.25	0.28	-0.66	1.92	-0.34	0.13	0.14
HG	7.29	3.29	12.05	3.82	5.60	-8.83	3.88	-0.32	0.14	0.16
HO	3.14	1.58	2.57	1.88	3.85	-11.99	4.32	-0.28	0.16	0.16
KC	17.13	8.54	67.00	8.84	19.16	85.31	4.67	-0.38	0.15	0.15
KW	38.90	15.90	213.91	18.17	39.18	65.14	4.66	-0.22	0.17	0.19
LB	747.20	349.19	5188.43	304.55	923.92	642.32	5.33	-0.28	0.14	0.17
LC	7.05	3.51	36.22	3.51	4.04	-55.63	2.38	-0.33	0.15	0.16
LCO	1.23	0.60	2.26	0.64	1.34	13.09	4.75	-0.20	0.17	0.18
LGO	2.95	1.43	12.28	1.71	4.96	22.72	4.51	-0.17	0.15	0.17
LH	24.16	13.81	111.44	10.52	21.65	117.04	4.02	-0.25	0.12	0.14
NG	4.38	2.25	9.91	2.34	7.03	-82.93	6.21	-0.20	0.14	0.16
NGLNM	33.63	15.69	563.12	9.38	33.16	140.97	6.22	-0.21	0.14	0.15
O	1566.38	730.62	22516.51	891.52	1733.25	10257.50	4.74	-0.24	0.13	0.14
OJ	882.34	339.45	4669.01	333.98	778.88	2391.00	5.93	-0.29	0.15	0.15
PA	187.96	77.23	639.36	87.17	218.83	-615.19	5.35	-0.31	0.17	0.15
PL	33.55	15.97	69.98	16.53	30.65	-313.34	4.51	-0.18	0.19	0.19
RR	46278.12	24499.31	527018.70	26283.06	39831.19	-311760.59	3.61	-0.20	0.14	0.15
S	1.78	0.86	36.62	0.97	1.46	-8.89	3.12	-0.26	0.13	0.15
SB	13.62	6.84	100.97	7.39	17.69	43.69	4.64	-0.33	0.14	0.15
SI	3.61	1.94	9.95	2.22	4.46	-23.65	5.14	-0.23	0.19	0.17
SM	9.69	4.79	44.01	5.25	9.76	39.23	3.67	-0.23	0.14	0.16
W	9.61	4.18	42.76	4.86	11.20	0.85	4.70	-0.27	0.16	0.18

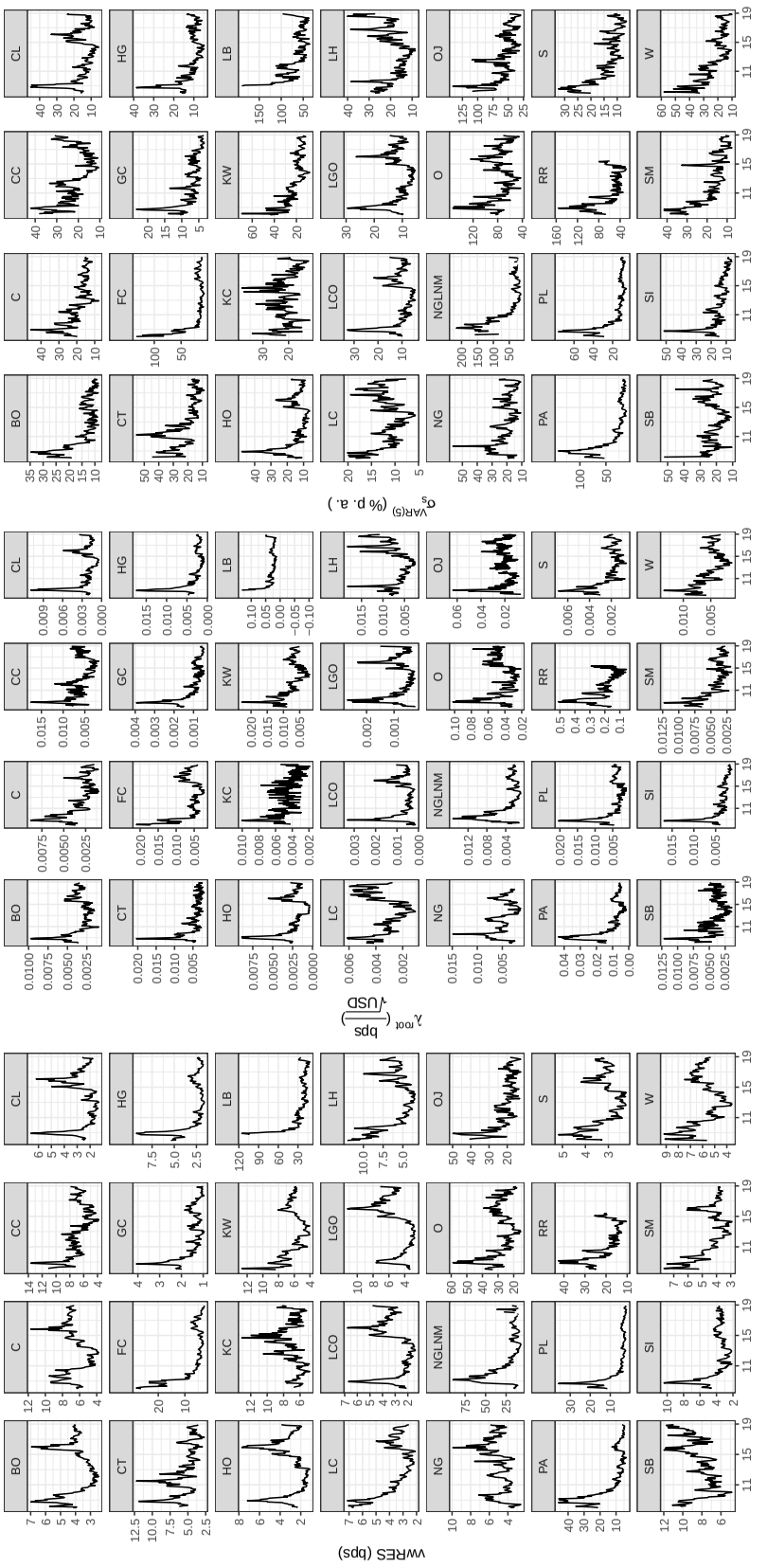

Figure A.1: **vwRES, λ^{root}, and $\sigma_s^{VAR(5)}$**

This table shows monthly estimates of vwRES, λ^{root}, and $\sigma_s^{VAR(5)}$ in bps, bps per root–USD–volume, and % p.a., respectively.

A.1 Liquidity Measures

Liquidity: High-Frequency Benchmark Measures

Time-Weighted Relative Quoted Spread (twRQS) We estimate the relative quoted spread as

$$RQS_k = \frac{A_k - B_k}{M_k} \tag{A.1}$$

where A_k is the ask price of the k^{th} quote, B_k is the bid price, and M_k is the mid-quote price $M_k = (A_k + B_k)/2$. We calculate the quoted spread of each quote recorded and compute a daily estimate by weighting it by the duration it was active, but no longer than 5 minutes. This way, we obtain a daily estimate of time-weighted relative quoted spreads (twRQS).

Quoted spreads, however, can be a misleading metric when depth is low and orders walk the book. Moreover, issues arise when trades are concentrated during times when quoted spreads are low. This is why we also consider the effective spread, which is closer to the true trading costs incurred by an investor trading more than a single contract at a time.

Volume-Weighted Relative Effective Spread (vwRES) We estimate the relative effective spread of the k^{th} trade of a given day as

$$RES_k = 2\, Q_k\, \frac{P_k - M_k}{M_k} \tag{A.2}$$

where Q_k is a binary variable that is $+1$ for a buyer-initiated trade and -1 for a seller-initiated trade flagged using the Lee and Ready (1991) algorithm, P_k is the k^{th} trade price, and M_k is the prevailing mid-quote price. To compute the daily average, we weight by volume to obtain a daily estimate for the volume-weighted relative effective spread (vwRES).

Volume-Weighted Relative Price Impact (vwRPI) Brennan and Subrahmanyam (1996) argue that liquidity should rather be measured by the price impact of order flow instead of spreads. Thus, following Goyenko et al. (2009), we also calculate price impact as another facet of liquidity. For every trade, we calculate the relative price impact as

$$RPI_k = Q_k \frac{M_{k+5min} - M_k}{M_k}, \tag{A.3}$$

where M_{k+5min} is the mid-quote price five minutes after the k^{th} trade. We weight by volume when aggregating to a daily frequency. This measure is intended to capture the permanent price impact from private information being revealed through trading or market makers adjusting quotes to limit inventory risk, as well as transitory effects when prevailing market depth is low or not replenished within 5 minutes (low resiliency). For each day, we form a volume-weighted average.

Kyle's Lambda (λ) We also measure price impact using two parametric approaches. The functional form of the first approach is based on the results in Kyle (1985). We estimate a regression of the form (see, e.g., Chordia et al., 2008)

$$mr_t = \alpha + \lambda\, OIB_t + \epsilon_t, \tag{A.4}$$

where t denotes 1-minute intervals, mr_t is the mid-quote return calculated from the prevailing mid-quote price at the end of the 1-minute interval, and OIB_t (order imbalance) is the buy minus sell volume within the time-interval measured in USD. Volume is the number of contracts traded times the futures price in USD times the size of a contract. We estimate this regression for each day in our sample to obtain a daily estimate of λ as a measure of price impact.

Root Lambda (λ^{root}) The theoretical results of Kyle (1985) imply a linear relationship between order imbalance and returns. In empirical studies, however, a concave functional form with signed root-volume is also often assumed (e.g., Hasbrouck, 2009; Collin-Dufresne and Fos, 2015). We estimate this alternative lambda using a regression of the form

$$mr_t = \alpha + \lambda^{root}\ sign(OIB_t)\ \sqrt{|OIB_t|} + \epsilon_t, \tag{A.5}$$

where $sign(OIB_t)$ takes the value $+1$ if OIB_t is positive and -1 if it is negative.

Liquidity: Low-Frequency Proxy Measures

To approximate liquidity at daily frequencies, we use measures that differ with respect to data requirements. We use three different versions of the Roll (1984) model which can be estimated from trade prices alone. The Effective Tick approach by Holden (2009) can also be estimated from daily trade prices, but we also compute it with open, high, and low prices. The measures by Corwin and Schultz (2012) and Abdi and Ranaldo (2017) require high and low (and close or settlement) prices. Lastly, we estimate proxies for price impact that require volume data. These are the volatility over volume measures of Kyle and Obizhaeva (2016) and Fong et al. (2018), the well known ratio of Amihud (2002), the regression-based measure of Pástor and Stambaugh (2003), as well as spread-over-volume measures of Goyenko et al. (2009). In the following, we provide a short description of each proxy.

Roll Measure (Roll) Roll (1984) proposes a model in which a constant spread s arises from order processing costs alone. s can easily be estimated from trade prices at any sampling frequency. In his model, the true or fair

(log) price m_t follows a random walk

$$m_t = m_{t-1} + u_t, \tag{A.6}$$

where u_t is a zero-mean disturbance term. Observable (log) trade prices follow

$$p_t = m_t + s\, Q_k. \tag{A.7}$$

where Q_k is $+1$ (-1) for buyer-initiated (seller-initiated) trades. He shows that the spread s can be estimated as

$$s = 2\,\sqrt{-Cov[\Delta p_t, \Delta p_{t-1}]}, \tag{A.8}$$

where $\Delta p_t = p_t - p_{t-1}$. Empirically, returns can exhibit positive autocorrelation, which is why we drop estimates when the auto-covariance turns out to be positive.

Roll Measure with Absolute Covariances (RollAbs) We also include a variant of the Roll model following Easley et al. (2021) that handles positive covariances by taking the absolute value as

$$s = 2\,\sqrt{|Cov[\Delta p_t, \Delta p_{t-1}]|}. \tag{A.9}$$

Roll's Measure Estimated with a Gibbs-Sampler (RollGibbs) Hasbrouck (2004) proposes using a Gibbs-sampler to estimate the Roll model (and extensions). The sampler sequentially updates and draws from full conditional distributions including drawing vectors of trade indicators Q_t. Each month and for each commodity we estimate the spread using 1000 sweeps but discard the first 200 as a burn-in.

Effective Tick (EffTick) Holden (2009) proposes a spread estimator that relies on the observation that prices tend to cluster on certain multiples of ticks. He assumes that prices that cluster at, e.g., quarters, cannot be generated by a spread of whole dollars. We assume different clustering for different minimum tick-regimes (decimal, fractional). Since the method relies on price clustering alone, we include all prices that are available: high, low, open, and settlement prices. For a minimum tick-size of one cent, we assume that prices tend to cluster at either \$0.01, \$0.05, \$0.1, \$0.25 or whole dollars. So $S = (0.01, 0.05, 0.1, 0.25, 1)$. Let j denote the j^{th} element in S and J the number of elements in S. For decimal minimum tick sizes of, e.g., \$0.001 or \$0.1, we assume that clustering occurs at the same multiples of the minimum tick. Following the notation of Holden (2009), we define $A = (100, 20, 10, 4, 1)$, $B = (80, 8, 8, 3, 1)$, and $O_{jk} = 0\ \forall j, k$ except for $O_{2,1} = 20$, $O_{3,2} = 10$, $O_{4,2} = O_{4,3} = 2$, and $O_{5,4} = 1$. For a fractional regime with a minimum tick-size of $\$\frac{1}{8}$ we set $S = (\frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1)$. So $A = (8, 4, 2, 1)$, $B = (4, 2, 1, 1)$, and $O_{j,k} = 0$ except for $O_{2,1} = 4$, $O_{3,2} = 2$, and $O_{4,3} = 1$. For a minimum tick-size of $\frac{1}{4}$, we assume $S = (\frac{1}{4}, \frac{1}{2}, 1, 2)$, so A, B and O are identical to the eights scheme. When the minimum tick-size is 0.05, we set $S = (0.05, 0.1, 0.25, 0.5, 1)$, so that $A = (0.05, 0.1, 0.25, 0.5, 1)$, $B = (8, 8, 2, 1, 1)$ and $O_{j,k}$ is zero except for $O_{2,1} = 10$, $O_{3,1} = 2$, $O_{3,2} = 2$, $O_{4,3} = 2$, and $O_{5,4} = 1$. Holden (2009) defines N_j as the number of prices that can be generated by the S_j but not by elements of S that are larger than S_j. The empirical probabilities are then given by

$$F_j = \frac{N_j}{\sum_{j=1}^{J} N_j} \forall j = 1, ..., J. \tag{A.10}$$

The unconstrained probabilities are

$$U_j = \left\{ \begin{array}{ll} \frac{A_1 F_1}{B_1}, & \text{for } j = 1 \\ \frac{A_j F_j}{B_j} - \sum_{k=1}^{j-1} \frac{O_{j,kF_k}}{B_k}, & \text{for } j = 2, ..., J \end{array} \right\} \tag{A.11}$$

and the constrained probabilities are computed iteratively for $j = 1, ..., J$ as

$$\hat{\gamma}_j = \left\{ \begin{array}{ll} min(max(U_j, 0), 1), & \text{for } j = 1 \\ min(max(U_j, 0)1 - \sum_{k=1}^{j-1} \hat{\gamma}_k), & \text{for } j = 2, ..., J \end{array} \right\} \tag{A.12}$$

which are used to compute the spread estimate relative to the price by

$$\text{EffTick} = \frac{\sum_{j=1}^{J} \hat{\gamma}_j S_j}{\bar{P}}. \tag{A.13}$$

High–Low Spread Measure (HighLow) Corwin and Schultz (2012) propose an estimator that relies on the assumption that the high price likely being a trade that hit the ask and the low price likely being a trade that lifted the bid. Their estimator for the spread is

$$s_t = \frac{2(e^{\alpha_t} - 1)}{1 + e^{\alpha_t}} \tag{A.14}$$

with

$$\begin{aligned} \alpha_t = \frac{\sqrt{2\beta_t} - \sqrt{\beta_t}}{3 - 2\sqrt{2}} - \sqrt{\frac{\gamma_t}{3 - 2\sqrt{2}}}, \quad \beta_t = \sum_{j=0}^{1} [log(\frac{H_{t+j}}{L_{t+j}})], \\ \text{and } \gamma_t = log(\frac{max(H_t, H_{t+1})}{min(L_t, L_{t+1})}), \end{aligned} \tag{A.15}$$

where H_t and L_t are high and low prices of time-interval t, respectively. The measure requires only two days (or other intervals) of data to compute an estimate of the spread. We average across all daily estimates in a given month.

High–Low–Close Measure (AbdiRanaldo) Abdi and Ranaldo (2017) build on the idea of high and low prices but include the last price. We use their two-day corrected version because they argue that it exhibits a higher correlation with their high-frequency benchmarks. It is computed as

$$s_{two-day-corrected} = \frac{1}{N}\sum_{t=1}^{N} s_t,\ s_t = \sqrt{max(4(c_t - \nu_t)(c_t - \nu_{t+1}), 0)}, \quad \text{and } \nu_t = \frac{l_t + h_t}{2}, \tag{A.16}$$

where c_t, l_t, and h_t are close, low, and high log-prices, respectively. N is the number of trading days in the month.

Volatility over Volume Spread Measure (VoV(Spread)) Kyle and Obizhaeva (2016) derive volatility over volume as an illiquidity measure of the form

$$VoV(Spread) = (\frac{\sigma^2}{Volume})^{\frac{1}{3}}. \tag{A.17}$$

where $Volume$ is the daily dollar volume of a contract. Daily dollar volume is the number of contracts traded times the futures settlement price times the size of the contract. We follow Fong et al. (2018) and estimate σ using the estimator proposed by Parkinson (1980)

$$\sigma = \sqrt{8/\pi}\ log(High/Low) \tag{A.18}$$

and set the invariant proportionality factor to 8, i.e., we multiply VoV(Spread) by 8.

Volatility over Volume Impact Measure (VoV(λ)) Building on the results by Kyle and Obizhaeva (2016), Fong et al. (2018) propose a proxy for

root-volume impact. The estimator has the form

$$VoV(\lambda) = \frac{\sigma}{Volume^{1/2}}, \tag{A.19}$$

where σ is estimated again using $\sigma = \sqrt{8/\pi}\ log(High/Low)$ following Parkinson (1980). VoV(Spread) and VoV(λ) are highly correlated with an average Pearson time-series correlation of 0.98 and an average cross-sectional Spearman correlation of 0.99 in our sample. Both can therefore be used as spread and price impact proxies.

Amihud Amihud (2002) proposes using a low-frequency version of Kyle's lambda in absolute terms to estimate the price impact of order flow

$$\text{Amihud} = \frac{|r|}{Volume}. \tag{A.20}$$

$|r|$ is the absolute log-return on a day and $Volume$ is the daily dollar Volume. We calculate this fraction for every trading day and take a simple average to obtain a monthly estimate.

Pástor Stambaugh (PastorStambaugh) Pástor and Stambaugh (2003) suggest using a regression of returns on signed volume to capture price impact. Even though they recommend not using it for single assets, we include it in our analysis for completeness. In their model, volume obtains the same sign as the return r_t of that particular day t. We estimate the regression

$$r_t = \theta + \phi\ r_{t-1} + \gamma\ sign(r_{t-1})\ Volume_{t-1} + \epsilon_t, \tag{A.21}$$

where γ is the price impact measure and $Volume_t$ is the volume at day t.

Spread–Impact Measures Following Goyenko et al. (2009), we also include price impact versions of different relative spread measures. They call these 'Extended Amihud Proxies'. These are simply calculated as

$$\text{Extended Amihud Proxy} = \frac{Spread}{Volume}. \tag{A.22}$$

We use the Roll, RollAbs, EffTick, HighLow, AbdiRanaldo, and RollGibbs estimates for the spread to compute such impact proxies which we name, e.g., RolloV for Roll over volume. Lou and Shu (2017) point out that the pricing component of the Amihud ratio is mainly attributable to volume in the denominator. To address concerns that price impact proxies might simply be driven by volume, we also include $1oV = \frac{1}{Volume}$ as a proxy in our analysis.

Proxies We Do Not Consider Several other liquidity proxies have been proposed that build on the number of days without a trade occurring (see, e.g., the FHT measure by Fong et al. (2017), or Zeros by Lesmond et al. (1999)). However, in our application, none of the commodity contracts exhibits a sufficient number of zero-trade days which is why we omit these measures from our study.

A.2 Efficiency

Efficiency: High-Frequency Benchmark Measures

Pricing Error Volatility from a VAR(5) ($\sigma_s^{VAR(5)}$) Hasbrouck (1993) proposes a VAR model to estimate the joint dynamics of price changes and order imbalance. The model decomposes the price process into a random-walk and a transitory component. He assumes that the log-price process is

$$p_t = m_t + s_t, \tag{A.23}$$

where m_t is the efficient price that follows a random walk and s_t is a stationary pricing error that can arise from any source. Hasbrouck (1993) suggests using the pricing error variance σ_s as a proxy for market efficiency.

Each day, we estimate a bi-variate vector auto-regressive model with 5 lags, VAR(5). We employ actual trade returns r_t and order imbalance OIB_t sampled at the 1-minute frequency. The VAR(5) model is

$$r_t = \sum_{i=1}^{5} a_i r_{t-i} + \sum_{i=1}^{5} b_i OIB_{t-i} + \nu_{1,t} \tag{A.24}$$

$$OIB_t = \sum_{i=1}^{5} c_i r_{t-i} + \sum_{i=1}^{5} d_i OIB_{t-i} + \nu_{2,t}. \tag{A.25}$$

We then invert the model into a vector moving average model (VMA) of the form

$$r_t = \sum_{k=0}^{\infty} a_k^* \nu_{1,t-k} + \sum_{k=0}^{\infty} b_k^* \nu_{2,t-k} \tag{A.26}$$

$$OIB_t = \sum_{k=0}^{\infty} c_k^* \nu_{1,t-k} + \sum_{k=0}^{\infty} d_k^* \nu_{2,t-k}. \tag{A.27}$$

Next, we truncate the number of VMA parameters at 11 lags. The return process is then given by

$$r_t = \theta_1 \nu_{1,t} + \theta_2 \nu_{1,t} + \Delta s_t, \tag{A.28}$$

where $\theta_1 = \sum_{i=0}^{11} a_i^*$ and $\theta_2 = \sum_{i=0}^{11} b_i^*$ and the pricing error is

$$s_t = \sum_{j=0}^{11} \alpha_j \nu_{1,t-j} + \sum_{i=0}^{11} \beta_j \nu_{2,t-j}$$

when imposing the Beveridge and Nelson (1981) restriction. Finally, α_j and

β_j are given by

$$\alpha_j = -\sum_{k=j+1}^{11} a_k^*, \quad \beta_j = -\sum_{k=j+1}^{11} b_k^*. \tag{A.29}$$

A lower bound for the pricing error variance is then

$$\sigma_s^2 = \sum_{j=0}^{11} \begin{bmatrix} \alpha_j & \beta_j \end{bmatrix} Cov(\nu) \begin{bmatrix} \alpha_j \\ \beta_j \end{bmatrix}. \tag{A.30}$$

In the literature, σ_s is sometimes used as a measure of price efficiency (see, e.g., Hendershott and Moulton, 2011; Rösch et al., 2017) but also as a measure of price impact (see, e.g., Collin-Dufresne and Fos, 2015). Hasbrouck (1993) refers to it as a measure of market quality.

Pricing Error Volatility from a MA(1) ($\sigma_s^{MA(1)}$) Hasbrouck (1993) also proposes using a moving average process for returns alone to infer the pricing error variance. Thus, we estimate a MA(1) model of the form

$$r_t = \epsilon_t - a\epsilon_{t-1}. \tag{A.31}$$

After imposing the Beveridge and Nelson (1981) restriction, a lower bound for the pricing error volatility can be estimated as

$$\hat{\sigma_s} = \sqrt{a^2\sigma_\epsilon^2}. \tag{A.32}$$

Adjusted Market Inefficiency Magnitude (AMIM) Tran and Leirvik (2019) propose a measure for the level of market inefficiency which they call 'Adjusted Market Inefficiency Magnitude', short AMIM. To implement their measure, we first estimate an AR(10) model for the mid-quote returns mr_t

(differences in log mid-quote prices) of a given day:

$$mr_t = \alpha + \sum_{q=1}^{10} \beta_q mr_{t-q} + \epsilon_t. \tag{A.33}$$

Then, they propose to standardize the vector of coefficients by calculating

$$\hat{\beta}^{standard} = L^{-1}\hat{\beta}, \tag{A.34}$$

where $\hat{\beta}$ is a column vector of the coefficients and L is a lower triangular matrix from the Choleski decomposition of the covariance matrix of $\hat{\beta}$. AMIM is then calculated as

$$AMIM_t = \frac{MIM_t - R_{CI}}{1 - R_{CI}}, \tag{A.35}$$

where

$$MIM_t = \frac{\sum_{j=1}^{q} |\hat{\beta}_{j,t}^{standard}|}{1 + \sum_{j=1}^{q} |\hat{\beta}_{j,t}^{standard}|} \tag{A.36}$$

and $R_{CI} = 0.9184596$ for 10 lags.

Variance Ratio (VR_{30}) Lo and MacKinlay (1988) propose variance ratios to test if a process follows a random walk. We approximate price efficiency as

$$|1 - \frac{\sigma^2(mr_{k,t})}{k\sigma^2(mr_t)}|, \tag{A.37}$$

where $mr_{k,t}$ refers to overlapping mid-quote returns aggregated to a k-minute frequency while mr_t are mid-quote returns at the 1-minute frequency. We estimate variance ratios for $k = 30$. We follow O'Hara and Ye (2011) and subtract the ratio from one and take the absolute value, such that a high

value corresponds to higher degrees of inefficiency, as it is the case for the other measures as well.

Absolute First-Order Autocorrelation ($|AR1|$) A simple way to measure deviations from the random walk is to calculate first-order autocorrelation. This is related to the variance ratio measure as well as the AMIM measure, that incorporates multiple lags. Nevertheless, using AR(1) coefficients is a common proxy for market efficiency (see, e.g., Chordia et al., 2008; Boehmer et al., 2021). Let mr_t denote 1-minute mid-quote returns. We estimate the AR(1) process

$$mr_t = \alpha + \beta mr_{t-1} + \epsilon_t \tag{A.38}$$

and use $|\beta|$ as a proxy for market efficiency.

Efficiency: Low-Frequency Proxy Measures

Since most of the measures can be computed on any sampling frequency, we employ the same methods to estimate low-frequency proxies. Only $\sigma_s^{VAR(5)}$ cannot be estimated without order imbalance. We compute monthly estimates with daily settlement prices. We use σ_s^{MA1}, AMIM, $|AR1|$, and a variance ratio with $k = 2$ (VR_2).

Microstructure-Adjusted Variance Ratio (VR^{Smith}) Smith (1994) develops a class of variance ratio estimators that account for microstructure effects. We use the version proposed in his paper that incorporates the model by Blume and Stambaugh (1983) whose analytic solution is described in the appendix of his paper. The adjusted variance ratio is

$$V_k = \frac{\frac{1}{T}\sum_{t=1}^{T}(ln(P_t) - ln(P_{t-k}) - k\mu)}{km_2} - 1, \tag{A.39}$$

where P_t is a trade price,

$$\mu = \frac{1}{T}\sum_{t=1}^{T} ln(P_t) - ln(P_{t-1}), \text{ and} \tag{A.40}$$

$$m_2 = \frac{1}{T}\sum_{t=1}^{T} (ln(P_t) - ln(P_{t-1}) - \mu)^2 - 2\sigma_\delta^2, \tag{A.41}$$

where in turn

$$\sigma_\delta^2 = \frac{\frac{j}{T}\sum_{t=1}^{T}(ln(P_t) - ln(P_{t-1}) - \mu) - \frac{1}{T}\sum_{t=1}^{T}(ln(P_t) - ln(P_{t-j}) - j\mu)^2}{2j - 2}. \tag{A.42}$$

We set $j = 2$ and $k = 4$. This estimator allows us to use 5-minute trade price data to compute a variance ratio estimate that is not overshadowed by bid–ask bounce induced autocorrelation which could substantially bias standard variance ratios.

A.3 Average of Proxies (Avg)

We also consider proxy combinations which potentially reduce noise of the different approaches. In the forecasting literature, the simple mean of different estimates has been shown to be a useful combination method (see, e.g., Clemen, 1989). This approach does not require estimating weights. Our proxies, however, are on different scales. The Amihud measure, for example, does not have an intuitive unit of measurement. Thus, we first standardize each proxy for each commodity. Then, we compute the simple average across all proxies. We compute an average of all spread proxies (AvgSpread), of all price impact proxies (AvgImpact), and of all (inverse) proxies for market efficiency (AvgEff).

Standardizing in the time-series dimension eradicates all cross-sectional differences between commodities. This is why we take a rank-based approach when studying the performance of aggregate measures cross-sectionally. We

first calculate the rank of each commodity for each measure. For example, the commodity with the lowest spread according to the Amihud measure is assigned rank one. Then, we average ranks across measures in every commodity-month.

A.4 Time-Series Correlations by Commodity

To get an idea of the cross-sectional distribution of the correlations, we provide individual estimates for each commodity in Tables A.4, A.2, and A.3 for the vwRES, vwRPI, and $\sigma_s^{VAR85)}$, respectively. We test each correlation coefficient against the null of zero correlation with the benchmark and indicate significance by stars. For each commodity, we test the correlation against the null that it is equal to the correlation of the best proxy using the test by Zou (2007). If we cannot reject such a test at the 5% level, we print the correlation coefficient in **bold** font.

Table A.2: **Time-Series Correlations of vwRPI with Proxies by Commodity**

*This table shows Pearson correlation coefficients of a single benchmark with different proxes (column) for different commodities (row). We test if each correlation coefficient is different from zero. ***, **, and * refer to significance at the 1, 5, and 10% level, respectively.* ***Bold*** *numbers indicate the highest coefficient exceeding 0.4 and any other in that row that is not different at a 5% significance level using the test by Zou (2007).*

	Amihud	$1oV$	$VoV(\lambda)$	RolloV	RollAbsoV	RollGibbsoV	EffTickoV	HighLowoV	AbdiRanaldooV	PasStam	AvgImpact
BO	**0.535*****	0.458***	**0.573*****	0.544***	0.457***	**0.535*****	0.410***	**0.474*****	**0.542*****	0.086	**0.578*****
C	0.408***	**0.520*****	0.312***	0.342***	0.338***	0.386***	**0.461*****	**0.450*****	**0.407*****	−0.259*	**0.457*****
CC	**0.694*****	0.599***	**0.694*****	0.610***	0.535***	**0.644*****	0.472***	**0.693*****	0.665***	0.042	**0.679*****
CL	0.735***	0.473***	**0.853*****	0.639***	0.735***	0.820***	0.566***	0.777***	0.648***	0.543***	0.813***
CT	**0.758*****	0.455***	**0.743*****	−0.137	0.580***	0.643***	0.314***	0.634***	**0.711*****	−0.019	0.678***
FC	**0.749*****	0.617***	**0.785*****	0.269*	0.466***	0.716***	**0.743*****	**0.764*****	0.666***	−0.226*	**0.770*****
GC	0.751***	0.398***	**0.864*****	0.191	0.704***	0.760***	**0.805*****	**0.847*****	0.750***	−0.247**	**0.846*****
HG	0.900***	0.889***	**0.941*****	0.413***	0.877***	0.891***	0.831***	0.890***	0.765***	−0.528***	0.905***
HO	**0.799*****	0.708***	**0.802*****	0.565***	0.546***	0.746***	0.627***	**0.789*****	0.655***	−0.115	**0.768*****
KC	0.320***	0.306***	0.331***	0.126	0.257***	**0.401*****	0.309***	0.263***	0.228**	0.240*	0.351***
KW	**0.498*****	0.234***	**0.504*****	0.349***	**0.447*****	0.434***	0.387***	**0.452*****	**0.539*****	−0.391***	**0.464*****
LB	0.555***	0.312***	**0.651*****	0.222	0.431***	0.493***	0.396***	0.392***	0.506***	−0.023	0.517***
LC	**0.728*****	0.581***	**0.770*****	0.481***	0.548***	0.692***	0.620***	0.671***	0.705***	0.154	**0.763*****
LCO	**0.723*****	0.542***	**0.751*****	0.368***	0.591***	0.695***	**0.715*****	**0.743*****	**0.677*****	0.603***	**0.737*****
LGO	**0.593*****	**0.570*****	**0.590*****	0.409***	0.375***	0.503***	0.461***	**0.576*****	0.525***	−0.006	**0.563*****
LH	0.647***	0.585***	**0.731*****	0.403***	0.460***	0.525***	0.599***	**0.717*****	0.614***	0.166	**0.700*****
NG	0.615***	0.381***	**0.698*****	0.438***	0.401***	0.552***	0.512***	0.552***	0.547***	−0.371***	**0.643*****
NGLNM	0.654***	0.375***	**0.742*****	0.301**	0.654***	0.678***	0.586***	0.565***	0.658***	−0.080	0.633***
O	0.336***	0.061	**0.584*****	0.402***	0.227**	0.186**	0.276***	0.259***	0.414***	0.144	0.401***
OJ	0.190*	0.225**	**0.422*****	0.193	0.374***	0.250***	0.130	0.280***	0.175*	0.232*	0.372***
PA	0.785***	0.771***	**0.821*****	0.635***	0.703***	0.757***	0.685***	0.740***	0.676***	−0.197*	0.775***
PL	**0.907*****	0.779***	0.853***	0.788***	0.841***	0.741***	0.774***	0.768***	0.793***	−0.343***	0.844***
RR	0.637***	0.147	**0.754*****	0.297**	0.152	0.332***	0.236**	0.471***	0.517***	0.221	0.501***
S	**0.515*****	**0.486*****	**0.500*****	0.269**	0.275***	0.412***	**0.462*****	**0.500*****	0.381***	−0.074	**0.500*****
SB	**0.417*****	0.364***	**0.413*****	0.399***	**0.401*****	0.394***	0.365***	0.388***	0.273***	0.000	**0.439*****
SI	0.845***	0.739***	0.853***	0.380***	0.768***	0.855***	**0.837*****	**0.879*****	**0.788*****	0.338***	**0.887*****
SM	0.279***	0.218**	0.315***	0.335***	0.249***	0.282***	0.170*	0.364***	0.296***	0.047	0.309***
W	**0.458*****	0.276***	0.362***	0.315***	0.324***	0.357***	0.312***	0.372***	**0.421*****	−0.157	0.391***

Table A.3: **Time-Series Correlations of $\sigma_s^{VAR(5)}$ with Proxies by Commodity**

*This table shows Pearson correlation coefficients of a single benchmark with different proxes (column) for different commodities (row). We test if each correlation coefficient is different from zero. ***, **, and * refer to significance at the 1, 5, and 10% level, respectively.* ***Bold*** *numbers indicate the highest coefficient exceeding 0.4 and any other in that row that is not different at a 5% significance level using the test by Zou (2007).*

	$\sigma_s^{MA(1)}$	AMIM	VR_2	$\lvert AR1\rvert$	AvgEff	AvgSpread	AvgImpact
BO	0.361^{***}	-0.096	0.025	-0.109	0.093	$\mathbf{0.822}^{***}$	0.749^{***}
C	0.137	-0.053	-0.336^{***}	-0.176^{*}	-0.186^{**}	$\mathbf{0.782}^{***}$	0.683^{***}
CC	0.241^{***}	-0.109	-0.057	-0.119	-0.010	$\mathbf{0.785}^{***}$	0.725^{***}
CL	0.560^{***}	0.101	-0.114	0.080	0.224^{***}	$\mathbf{0.916}^{***}$	0.884^{***}
CT	0.285^{***}	-0.062	-0.098	-0.069	0.007	$\mathbf{0.733}^{***}$	0.488^{***}
FC	-0.011	0.170^{*}	0.068	0.045	0.090	$\mathbf{0.535}^{***}$	$\mathbf{0.573}^{***}$
GC	0.320^{***}	0.095	0.063	-0.044	0.120	$\mathbf{0.858}^{***}$	$\mathbf{0.836}^{***}$
HG	0.513^{***}	0.153	0.116	0.020	0.237^{***}	$\mathbf{0.894}^{***}$	0.828^{***}
HO	0.176^{*}	0.059	-0.179^{*}	-0.079	0.024	$\mathbf{0.873}^{***}$	$\mathbf{0.863}^{***}$
KC	0.327^{***}	0.047	0.079	0.146	0.214^{**}	$\mathbf{0.581}^{***}$	0.419^{***}
KW	0.335^{***}	0.143	0.196^{**}	0.252^{***}	0.286^{***}	$\mathbf{0.817}^{***}$	0.712^{***}
LB	0.256^{***}	-0.010	-0.046	0.064	0.091	$\mathbf{0.814}^{***}$	$\mathbf{0.805}^{***}$
LC	0.058	-0.108	-0.012	-0.173^{*}	-0.071	$\mathbf{0.733}^{***}$	0.553^{***}
LCO	0.414^{***}	0.104	-0.240^{***}	0.058	0.118	$\mathbf{0.930}^{***}$	$\mathbf{0.910}^{***}$
LGO	0.167^{*}	0.005	-0.094	-0.014	0.033	$\mathbf{0.898}^{***}$	$\mathbf{0.882}^{***}$
LH	0.158^{*}	-0.068	-0.071	0.106	0.032	$\mathbf{0.822}^{***}$	0.745^{***}
NG	0.346^{***}	-0.008	0.055	-0.011	0.172^{**}	$\mathbf{0.807}^{***}$	0.670^{***}
NGLNM	0.636^{***}	0.142	0.279^{***}	0.085	0.401^{***}	0.799^{***}	$\mathbf{0.876}^{***}$
O	0.063	0.058	0.064	-0.056	0.015	$\mathbf{0.664}^{***}$	0.558^{***}
OJ	0.321^{***}	-0.031	0.022	0.082	0.123	$\mathbf{0.612}^{***}$	0.452^{***}
PA	0.457^{***}	-0.200^{**}	-0.004	-0.120	0.056	$\mathbf{0.904}^{***}$	$\mathbf{0.918}^{***}$
PL	0.609^{***}	0.145	-0.075	-0.249^{***}	0.193^{**}	$\mathbf{0.890}^{***}$	$\mathbf{0.899}^{***}$
RR	-0.037	-0.012	-0.179	-0.103	-0.045	$\mathbf{0.687}^{***}$	$\mathbf{0.690}^{***}$
S	0.342^{***}	0.198^{**}	0.086	-0.134	0.171^{**}	$\mathbf{0.751}^{***}$	0.683^{***}
SB	0.197^{**}	0.004	-0.102	-0.038	0.008	$\mathbf{0.610}^{***}$	0.486^{***}
SI	0.228^{**}	-0.137	-0.096	-0.087	-0.034	$\mathbf{0.881}^{***}$	0.800^{***}
SM	0.414^{***}	0.092	-0.119	-0.154^{*}	0.084	$\mathbf{0.840}^{***}$	$\mathbf{0.810}^{***}$
W	0.411^{***}	0.156	0.244^{***}	0.173^{*}	0.308^{***}	$\mathbf{0.788}^{***}$	0.575^{***}

The estimates in Table A.4 show that for most commodities, VoV(Spread) is the best proxy. Nevertheless, there are differences between commodities in how easily spreads can be approximated using low-frequency measures. For example, spreads of NYMEX WTI (CL) futures can be approximated reasonably well by a multitude of proxy measures, while for example variation in the spread of CBOT Corn (C) futures can only be approximated by few measures.

The individual results for relative price impact (vwRPI) in Table A.2 are similar. The VoV(λ) measure emerges as the best proxy for almost all commodities. It is able to approximate price impact for all commodities except for CBOT Corn (C). The Amihud measure is able to produce significantly correlated estimates but is for most commodities inferior in terms of correlation compared to VoV(λ). Based on these findings, we thus recommend using VoV measures to approximate spreads and price impact of individual commodities.

Table A.4: **Time-Series Correlations of vwRES with Proxies by Commodity**

*This table shows Pearson correlation coefficients of a single benchmark with different proxies (column) for different commodities (row). We test if each correlation coefficient is different from zero. ***, **, and * refer to significance at the 1, 5, and 10% level, respectively.* ***Bold*** *numbers indicate the highest coefficient exceeding 0.4 and any other in that row that is not different at a 5% significance level using the test by Zou (2007).*

	Amihud	VoV(Spread)	Roll	RollAbs	RollGibbs	EffTick	HighLow	AbdiRanaldo	AvgSpread
BO	0.684***	**0.754***	0.391***	0.331***	0.448***	0.383***	0.558***	0.531***	0.657***
C	**0.477***	0.345***	−0.135	−0.078	0.027	**0.546***	0.170*	0.213**	0.274***
CC	**0.861***	**0.887***	0.418***	0.426***	0.476***	0.337***	0.554***	0.565***	0.744***
CL	0.775***	**0.867***	0.607***	0.653***	0.751***	0.653***	0.678***	0.651***	0.828***
CT	0.744***	**0.822***	0.204	0.292***	0.542***	0.516***	0.715***	0.750***	**0.779***
FC	**0.761***	**0.780***	−0.025	0.032	0.091	**0.714***	0.342***	0.317***	0.608***
GC	0.826***	**0.931***	0.164	0.489***	0.624***	0.720***	0.792***	0.547***	0.841***
HG	0.921***	**0.949***	0.626***	0.746***	0.789***	0.715***	0.740***	0.827***	0.907***
HO	0.740***	**0.804***	0.520***	0.497***	0.649***	0.213**	0.690***	0.623***	0.708***
KC	0.279***	**0.423***	0.223*	0.298***	0.338***	−0.075	**0.429***	0.374***	**0.431***
KW	**0.645***	**0.663***	0.232**	0.396***	0.316***	0.411***	0.463***	0.523***	0.577***
LB	**0.631***	**0.619***	**0.426***	0.375***	0.493***	**0.524***	0.299***	**0.547***	**0.628***
LC	**0.801***	**0.817***	0.069	0.134	0.328***	0.559***	0.351***	0.408***	0.676***
LCO	0.686***	**0.792***	0.491***	0.586***	0.690***	0.474***	**0.742***	0.696***	**0.780***
LGO	0.665***	**0.725***	0.388***	0.315***	0.482***	**0.658***	0.563***	0.571***	0.682***
LH	0.706***	**0.786***	0.236**	0.296***	0.278***	0.674***	0.480***	0.520***	0.709***
NG	**0.627***	**0.601***	0.331***	0.207**	0.353***	0.355***	0.169*	0.335***	0.512***
NGLNM	0.856***	**0.875***	0.228*	0.593***	0.676***	**0.835***	0.468***	0.536***	0.816***
O	0.501***	**0.790***	0.404***	0.293***	0.518***	0.548***	0.535***	0.538***	**0.783***
OJ	0.047	0.516***	0.496***	0.386***	0.319***	**0.523***	0.466***	0.532***	**0.637***
PA	0.884***	**0.952***	0.446***	0.648***	0.529***	0.775***	0.703***	0.590***	0.918***
PL	**0.934***	**0.909***	0.623***	0.536***	0.624***	0.690***	0.659***	0.721***	0.882***
RR	0.791***	**0.897***	0.288*	0.249**	0.377***	0.171	0.517***	0.502***	0.725***
S	**0.709***	**0.719***	0.173	0.229***	0.432***	0.562***	0.574***	0.555***	0.635***
SB	**0.404***	0.297***	−0.025	0.046	0.021	0.227***	−0.058	−0.119	0.142
SI	**0.884***	**0.876***	0.307***	0.184**	0.401***	**0.841***	0.552***	0.433***	0.737***
SM	**0.751***	**0.759***	0.194*	0.284***	0.397***	0.497***	0.483***	0.596***	0.656***
W	**0.593***	**0.576***	0.296***	0.335***	0.334***	**0.494***	0.422***	0.421***	**0.551***

Table A.3 reports the individual correlations of all price efficiency proxies with $\sigma_s^{VAR(5)}$. $\sigma_s^{MA(1)}$ is the only single proxy that is significantly correlated with the high-frequency benchmark. However, correlations are rather low. Average spread and impact proxies are significantly correlated with the benchmark of almost all commodities.

A.5 Time-Series Correlations by Year

In our main analysis, we have shown that volatility-over-volume measures are superior to the oft-used Amihud (2002) measure in terms of correlation with their benchmarks. As a robustness check, we compute proxy–benchmark time-series correlations for each year in our sample. The average correlation coefficients are reported in Tables A.5 for spreads, in Table A.6 for price impact, and in Table A.7 for price efficiency. In every year between 2008 and 2018, the VoV(Spread) measure is either the single best spread proxy or insignificantly worse than the best. In about half of the years, the Amihud measure exhibits a significantly lower average correlation. The picture is the same for price impact measures. In every year, $VoV(\lambda)$ outperforms or performs at par with the Amihud measure. This confirms that the VoV measures are indeed consistently superior in capturing time-series variation in commodity liquidity when compared to other measures, including the popular Amihud measure. The year-by-year results for price efficiency proxies estimated from daily data also affirm our previous results.

Table A.5: **Time-Series Correlations of Spread Proxies and Benchmarks by Year**

This table shows average Pearson moment-correlation coefficients of benchmark–proxy pairs of bid-ask spreads. We calculate the correlation of a benchmark–proxy pair for each commodity and average it across all commodities. ***Bold*** *numbers indicate the highest coefficient that is greater than* 0.4*. Those that are not different from it at a* 5% *significance level are also in* ***bold*** *font. We use a t-test of Fisher z-transformed coefficients in the spirit of Fama and MacBeth (1973).*

Panel A: twRQS

	Amihud	VoV(Spread)	Roll	RollAbs	RollGibbs	EffTick	HighLow	AbdiRanaldo	AvgSpread
2008	**0.660**	**0.713**	0.278	0.274	0.487	0.428	0.427	0.479	**0.623**
2009	**0.664**	**0.677**	0.377	0.309	0.355	0.297	0.449	0.400	0.599
2010	**0.497**	**0.538**	0.032	0.155	0.209	0.102	0.254	0.261	0.369
2011	0.366	**0.494**	0.186	0.177	0.332	0.137	0.265	0.371	**0.448**
2012	0.333	**0.452**	0.004	0.081	0.216	0.240	0.256	0.394	0.334
2013	0.383	**0.492**	0.169	0.187	0.223	0.134	0.234	0.229	0.356
2014	**0.474**	**0.637**	0.387	0.275	0.375	0.194	0.445	0.336	**0.503**
2015	**0.471**	**0.566**	0.192	0.180	0.238	0.252	0.373	0.296	**0.432**
2016	**0.516**	**0.568**	0.214	0.247	0.316	0.190	0.344	0.327	0.498
2017	**0.407**	**0.592**	0.371	0.282	0.256	0.122	0.331	0.325	0.478
2018	0.545	**0.683**	0.274	0.162	0.320	0.145	0.417	0.337	0.492

Panel B: vwRES

	Amihud	VoV(Spread)	Roll	RollAbs	RollGibbs	EffTick	HighLow	AbdiRanaldo	AvgSpread
2008	**0.663**	**0.758**	0.313	0.379	0.511	0.353	0.442	0.512	0.657
2009	0.745	**0.830**	0.373	0.357	0.406	0.354	0.474	0.458	0.683
2010	**0.562**	**0.610**	0.002	0.172	0.215	0.191	0.266	0.243	0.399
2011	0.421	**0.627**	0.180	0.184	0.323	0.257	0.331	0.364	**0.491**
2012	0.494	**0.630**	0.204	0.228	0.315	0.208	0.371	0.444	**0.509**
2013	0.431	**0.584**	0.347	0.305	0.316	0.070	0.247	0.268	0.445
2014	0.559	**0.718**	0.441	0.341	0.373	0.203	0.448	0.381	0.563
2015	0.395	**0.465**	0.126	0.228	0.209	0.238	0.310	0.298	**0.425**
2016	**0.438**	**0.485**	0.162	0.195	0.300	0.282	0.327	0.239	**0.457**
2017	0.335	**0.429**	0.151	0.085	0.131	0.034	0.173	0.107	0.237
2018	**0.529**	**0.601**	0.314	0.203	0.266	0.155	0.299	0.236	0.405

Table A.6: **Time-Series Correlations of Price Impact Proxies and Benchmarks by Year**

This table shows average Pearson moment-correlation coefficients of benchmark–proxy pairs of price impact. We calculate the correlation of a benchmark–proxy pair for each commodity and average it across all commodities. ***Bold*** *numbers indicate the highest coefficient that is greater than* 0.4*. Those that are not different from it at a* 5% *significance level are also in* ***bold*** *font. We use a t-test of Fisher z-transformed coefficients in the spirit of Fama and MacBeth (1973).*

Panel A: vwRPI

	Amihud	$1oV$	$VoV(\lambda)$	RolloV	RollAbsoV	RollGibbsoV	EffTickoV	HighLowoV	AbdiRanaldooV	PastorStambaugh	AvgImpact
2008	**0.524**	0.323	**0.583**	0.370	**0.455**	**0.481**	0.398	**0.516**	**0.500**	0.146	**0.541**
2009	**0.477**	0.347	**0.559**	0.258	0.365	**0.444**	0.298	**0.421**	**0.426**	−0.009	**0.492**
2010	0.269	0.212	0.323	0.028	0.140	0.271	0.204	0.292	0.315	0.033	0.328
2011	0.217	0.043	0.354	0.183	0.129	0.197	0.119	0.257	0.266	−0.080	0.230
2012	0.250	0.045	0.373	0.053	0.114	0.229	0.127	0.296	0.237	−0.054	0.241
2013	0.129	−0.058	0.259	0.182	0.108	0.103	0.018	0.131	0.180	−0.041	0.161
2014	0.347	0.184	**0.448**	0.254	0.239	0.312	0.234	0.337	0.337	−0.134	0.354
2015	0.389	0.331	0.361	0.181	0.272	0.310	0.269	0.257	0.267	−0.132	0.350
2016	0.387	0.274	0.369	0.192	0.241	0.277	0.264	0.335	0.283	−0.003	0.377
2017	0.277	0.065	0.263	0.311	0.165	0.263	0.036	0.182	0.241	−0.097	0.237
2018	0.350	0.152	**0.404**	0.226	0.234	0.290	0.161	0.284	0.244	0.083	0.327

Panel B: λ

	Amihud	$1oV$	$VoV(\lambda)$	RolloV	RollAbsoV	RollGibbsoV	EffTickoV	HighLowoV	AbdiRanaldooV	PastorStambaugh	AvgImpact
2008	**0.795**	0.682	**0.790**	0.552	0.612	**0.775**	0.652	**0.796**	**0.732**	0.106	**0.837**
2009	**0.708**	0.548	**0.802**	0.539	0.488	0.596	0.487	0.658	0.573	−0.099	**0.728**
2010	**0.508**	**0.505**	**0.517**	0.194	0.302	0.404	**0.421**	**0.447**	0.272	0.036	**0.551**
2011	**0.453**	0.269	**0.514**	0.178	0.251	**0.419**	0.322	**0.470**	**0.405**	−0.046	**0.496**
2012	0.451	0.272	**0.596**	0.281	0.332	0.438	0.233	0.465	0.368	0.103	**0.555**
2013	0.247	0.021	0.389	0.216	0.224	0.218	0.087	0.223	0.228	−0.107	0.270
2014	**0.521**	0.279	**0.633**	0.333	0.359	0.397	0.301	0.506	0.383	−0.161	**0.531**
2015	**0.506**	**0.481**	**0.458**	0.257	0.332	0.349	0.291	**0.439**	0.342	−0.163	**0.529**
2016	**0.544**	0.362	**0.639**	0.385	0.333	0.464	0.407	0.559	0.443	0.107	**0.603**
2017	**0.424**	0.218	**0.542**	0.091	0.162	0.363	0.172	**0.420**	0.362	0.043	0.383
2018	0.516	0.305	**0.720**	0.361	0.315	0.598	0.196	0.611	0.507	0.189	**0.606**

(continued)

Table A.6: **Time-Series Correlations of Price Impact Proxies and Benchmarks by Year (Continued)**

Panel C: λ^{root}

	Amihud	$1oV$	$VoV(\lambda)$	RolloV	RollAbsoV	RollGibbsoV	EffTickoV	HighLowoV	AbdiRanaldooV	PastorStambaugh	AvgImpact
2008	0.748	0.567	**0.818**	0.506	0.595	0.752	0.602	**0.765**	0.708	0.165	**0.804**
2009	0.729	0.506	**0.839**	0.553	0.501	0.612	0.462	0.668	0.612	−0.093	0.742
2010	**0.556**	0.396	**0.638**	0.181	0.325	0.442	0.375	0.464	0.361	0.055	**0.576**
2011	0.489	0.193	**0.650**	0.171	0.264	0.447	0.272	0.497	0.484	−0.050	0.508
2012	0.553	0.204	**0.723**	0.318	0.336	0.523	0.248	0.506	0.474	0.032	0.602
2013	0.363	0.066	**0.558**	0.277	0.303	0.317	0.109	0.322	0.356	−0.159	0.371
2014	0.657	0.314	**0.790**	0.435	0.423	0.466	0.321	0.595	0.496	−0.210	0.638
2015	**0.617**	**0.471**	**0.617**	0.386	0.394	0.420	0.318	**0.528**	0.409	−0.090	**0.630**
2016	0.604	0.346	**0.732**	0.427	0.372	0.471	0.338	0.570	0.471	0.005	0.636
2017	0.522	0.178	**0.705**	0.139	0.236	0.403	0.142	0.445	0.444	0.067	0.460
2018	0.613	0.279	**0.783**	0.433	0.335	0.605	0.235	0.636	0.556	0.173	0.660

A.6 Temporal Stability of Cross-Sectional Correlations

We also inspect the stability of cross-sectional correlations over time. Instead of averaging the correlation coefficients across months, we plot them in Figure A.3 and Figure A.2. To ease the visual inspection, we estimate and plot a LOESS-regression for each benchmark–proxy pair.[6]

[6]LOESS stands for Locally Estimated Scatterplot Smoothing (see, e.g., Cleveland et al., 1993).

Table A.7: **Time-Series Correlations of Price Efficiency Proxies and Benchmarks by Year**

This table shows average Pearson moment-correlation coefficients of benchmark–proxy pairs of market efficiency. We calculate the correlation of a benchmark–proxy pair for each commodity and average it across all commodities. ***Bold*** *numbers indicate the highest coefficient that is greater than* 0.4*. Those that are not different from it at a* 5% *significance level are also in* ***bold*** *font. We use a t-test of Fisher z-transformed coefficients in the spirit of Fama and MacBeth (1973).*

Panel A: $\sigma_s^{VAR(5)}$

	$\sigma_s^{MA(1)}$	AMIM	VR_2	$\lvert AR1\rvert$	AvgEff	AvgSpread	AvgImpact
2008	0.241	0.150	0.021	0.030	0.168	**0.631**	**0.555**
2009	0.196	0.070	−0.004	0.018	0.138	**0.638**	**0.611**
2010	0.164	0.094	0.051	−0.051	0.089	**0.456**	0.389
2011	0.100	0.067	0.039	0.017	0.086	**0.428**	0.287
2012	−0.000	−0.079	−0.068	−0.054	−0.040	**0.500**	**0.419**
2013	0.199	0.070	0.057	0.057	0.126	**0.447**	0.323
2014	0.254	0.000	0.030	0.065	0.113	**0.562**	**0.467**
2015	0.204	0.032	0.089	0.062	0.118	**0.521**	**0.413**
2016	0.147	−0.029	0.049	0.085	0.079	**0.563**	**0.456**
2017	0.182	−0.068	−0.097	−0.048	−0.025	**0.457**	0.286
2018	0.086	−0.057	−0.008	0.068	0.041	**0.484**	**0.419**

Panel B: AMIM

	$\sigma_s^{MA(1)}$	AMIM	VR_2	$\lvert AR1\rvert$	AvgEff	AvgSpread	AvgImpact
2008	0.046	0.082	0.098	0.054	0.117	−0.119	−0.187
2009	−0.104	−0.118	−0.062	−0.087	−0.130	0.004	0.006
2010	−0.121	0.035	−0.223	0.047	−0.116	−0.008	0.076
2011	0.006	0.002	−0.021	−0.059	−0.021	−0.080	−0.020
2012	−0.107	−0.052	−0.006	−0.039	−0.008	−0.071	−0.051
2013	−0.032	0.157	−0.093	−0.015	0.011	0.029	0.035
2014	0.095	0.001	0.115	0.044	0.078	0.009	0.005
2015	0.021	−0.082	0.069	0.049	0.019	0.086	0.065
2016	0.114	0.044	0.067	0.063	0.073	−0.045	−0.078
2017	0.032	0.032	0.029	0.058	0.097	−0.004	0.072
2018	0.002	0.063	−0.002	−0.008	0.015	0.022	0.025

Panel C: VR_{30}

	$\sigma_s^{MA(1)}$	AMIM	VR_2	$\lvert AR1\rvert$	AvgEff	AvgSpread	AvgImpact
2008	0.121	0.066	0.045	0.002	0.085	0.143	0.144
2009	0.007	0.042	0.087	0.049	0.035	−0.016	0.003
2010	−0.159	0.162	−0.027	−0.087	−0.082	−0.149	−0.019
2011	−0.101	0.036	−0.024	0.023	0.000	−0.083	−0.064
2012	−0.070	−0.077	−0.076	−0.129	−0.112	−0.095	−0.031
2013	−0.084	0.065	−0.031	−0.040	−0.034	−0.079	0.069
2014	0.077	0.019	0.109	0.043	0.131	0.032	0.078
2015	0.130	0.111	0.094	0.045	0.119	−0.065	0.010
2016	0.081	0.161	0.118	0.140	0.165	0.018	−0.047
2017	0.035	−0.031	−0.031	0.005	−0.066	−0.027	−0.039
2018	−0.191	−0.127	−0.156	−0.147	−0.207	−0.131	−0.039

Panel D: $\lvert AR1\rvert$

	$\sigma_s^{MA(1)}$	AMIM	VR_2	$\lvert AR1\rvert$	AvgEff	AvgSpread	AvgImpact
2008	0.026	−0.042	0.004	0.016	0.004	−0.030	−0.071
2009	0.001	−0.019	−0.042	−0.017	−0.018	0.016	0.116
2010	−0.022	0.019	−0.078	−0.059	−0.056	−0.089	0.008
2011	0.050	0.067	0.061	0.055	0.098	0.043	0.018
2012	−0.043	0.126	0.007	0.067	0.068	−0.030	0.003
2013	0.031	0.049	0.019	−0.047	0.025	−0.047	0.023
2014	−0.016	−0.096	−0.032	−0.125	−0.086	−0.125	−0.092
2015	0.014	0.028	−0.010	−0.003	0.011	−0.146	−0.118
2016	−0.008	0.199	0.049	0.060	0.057	0.074	0.053
2017	0.018	−0.006	0.004	0.008	−0.043	−0.083	0.088
2018	−0.086	0.022	−0.064	−0.036	−0.058	−0.056	−0.095

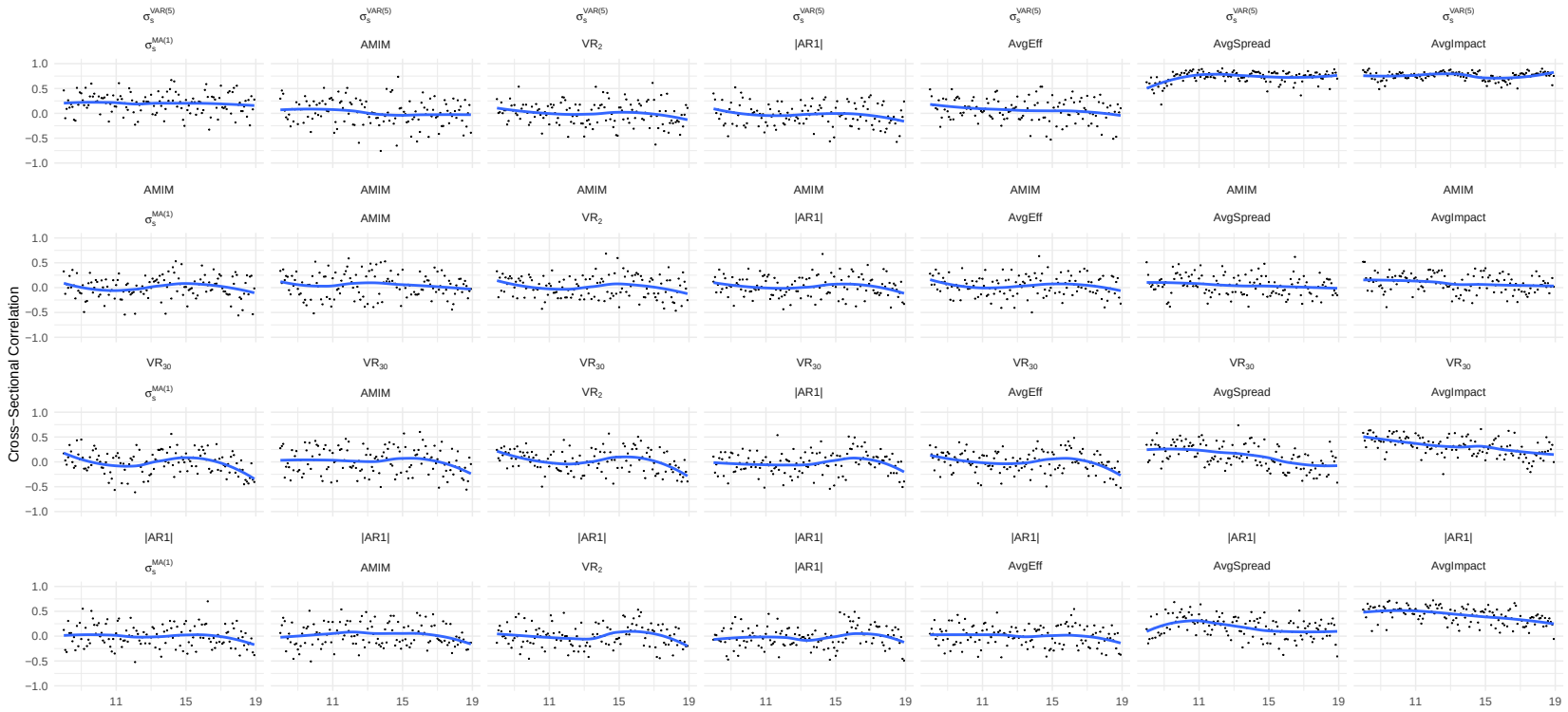

Figure A.2: **Temporal Stability of Cross-Sectional Correlations: Price Efficiency**

This figure shows monthly cross-sectional correlations for each benchmark–proxy pair. To ease the visual inspection of the estimates, we estimate a LOESS-regression for each benchmark–proxy pair.

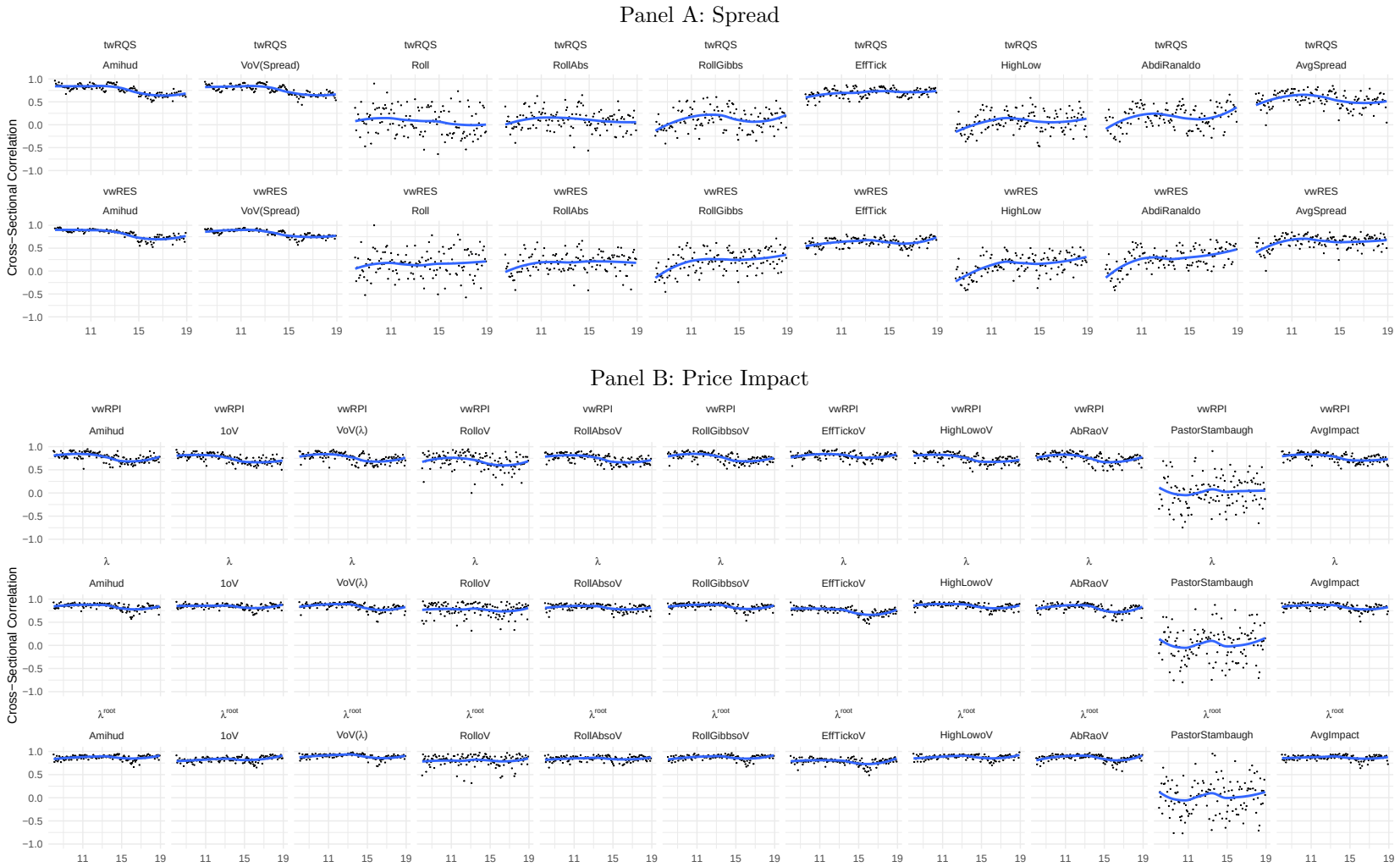

Figure A.3: **Temporal Stability of Cross-Sectional Correlations: Liquidity**

This figure shows monthly cross-sectional correlations for each benchmark–proxy pair. To ease the visual inspection of the estimates, we estimate a LOESS-regression for each benchmark–proxy pair.

The results for spreads in Panel A of Figure A.3 suggest that the cross-sectional correlation of the best proxies, VoV(Spread) and Amihud, is fairly stable over time. Correlations declined only slightly. The EffTick measure started at a lower correlation in 2008. At the end of the sample in 2018, its correlation with both benchmarks is at the same level as the ones with VoV(Spread) and Amihud. The correlations of other measures increased during the second half of the sample.

In Panel B of Figure A.3, we plot the cross-sectional correlations of price impact benchmarks and proxies by month. The correlations of all measures except for the PastorStambaugh measure are at a similar level. Their correlations with vwRPI slightly declined during the sample period. Correlations with λ-benchmarks are more stable and do not exhibit a clear trend.

Finally, we show the monthly cross-sectional correlations for price efficiency benchmark–proxy pairs in Figure A.2. The correlation of $\sigma_s^{MA(1)}$ with $\sigma_s^{VAR(5)}$ is positive and stable but on a low level. Correlations of the spreads and price impact proxies with $\sigma_s^{VAR(5)}$ appears to be stable over time. AMIM is not captured by any proxy. Cross-sectional variation in $VR_{2/10/30}$ and $|AR1|$ can be captured by impact proxies but correlations decline during the sample period.

These results imply that even after increased market speed with the rise of algorithmic trading (AT), traditional proxy measures are able to capture cross-sectional liquidity. This hints at these measures being robust to changes in market design. The question is, whether these proxies also work when commodity futures were mainly traded in pits instead of electronic limit order markets. High cross-sectional correlations of volume-based liquidity proxies during the first half of the sample might be interpreted as an indication that this is the case and that these proxies are also most suitable for approximating liquidity pre-2008.

A.7 Proxy Combinations

Our results suggest that forming simple, equally weighted averages of standardized proxies produces estimates that often perform at par with the best single-proxy measure. The fact that combining several proxies reduces noise is well known in the forecasting literature with simple averages having the advantage of not requiring parameter estimation (Stock and Watson, 2004). We test if it is possible to obtain superior market quality proxies by combining different individual proxies. We also compare these to the already tested simple average (SA). First, we standardize all proxy and benchmark measures across the entire sample. Since most of the methods require data for parameter estimation, we repeatedly split the sample into a training and a testing sample. For all years from 2008 to 2018, we leave out one year, use all the other years to estimate parameters, and then forecast the year we left out (leave-one-year-out cross-validation).

We use the following methods to combine the proxies. (1) The simple average (SA) is an equally weighted average across all standardized proxies. No estimation is required. (2) Bates and Granger (1969) (BG) introduced the idea of forecast combination. Their method calculates the weights based on the diagonal of the mean squared error (MSE) matrix. (3) The approach suggested by Newbold and Granger (1974) (NG) is also based on the same matrix of mean squared errors as (BG) but imposes the condition that the weights must add up to one. The extension by Hsiao and Wan (2014) allows non-diagonal MSE matrices. (4) The trimmed eigenvalue forecast combination (TEFC) of Hsiao and Wan (2014) discards the worst predictors in terms of RMSE and estimates optimal combination weights based on the eigenvectors of the mean squared prediction error matrix. (5) The constrained least squares (CLS) forecast combination suggested by Bates and Granger (1969) determines the weights by performing ordinary least squares without inter-

cept and with the constraint that all weights must add up to one. (6) The least absolute deviation (LAD) method is also called median regression and estimates a linear regression by minimizing the sum of absolute deviations. (7) The random forest (Breiman, 2001) (RF) is an aggregate of regression trees where both the sample for each tree and the proxies at each split are bootstrapped. This way, individual trees are less correlated and the RF is less prone to overfitting. We grow 1,000 regression trees each of which randomly samples one third of the proxies. (8) Boosted stumps (BS) are gradient-boosted single-node regression trees estimated with 100 boosting iterations.

Except for SA, we fit each model using the training sample and use the testing sample to compare realized to predicted values. We pool all commodities into one sample for estimation to ensure that sufficiently many data points are available for parameter estimation. Then, we compute time-series Pearson moment-correlations for each commodity and report the average in Table A.8. Again, we identify the best (combined) proxy and test the average correlation of all other proxies against the one of the winner using a t-test in the spirit of Fama and MacBeth (1973). We highlight the best (combined) proxy and all those whose average correlation is not different from it at the 5% level in bold font.

In Panel A, we report the results for spreads. Most of the proxy combinations exhibit a very similar average correlation with the benchmark compared to the best proxy VoV(Spread). None of the methods that we employ appears to be clearly superior to the other.

In Panel B, we report the correlations for price impact (combined) benchmark–proxy pairs. Again, no combination method is significantly different from the best single proxy VoV(Spread) or other combination methods.

The correlations for price efficiency in Panel C show that no combination method is able to achieve a correlation with $\sigma_s^{VAR(5)}$ that is higher than that with $\sigma_s^{VAR(5)}$.

Table A.8: **Proxy Combinations**

This table shows average Pearson time-series correlation estimates. First, we split the sample into a training sample (2014.01–2018.12) and a testing sample (2008.01–2014.12). We standardize all proxy and benchmark measures in each sub-sample. Then, we estimate different models to combine the proxies using the pooled training sample. Finally, we compute correlations of combined proxies with their benchmarks for each commodity and report averages. ***Bold*** *numbers indicate the highest correlation exceeding 0.4 and those that are not significantly different from the highest at a 5% level using a standard t-test of Fisher z-transformed coefficients. Besides the single measures we use different combination techniques. Their codes are: SA = Simple Average, TEFC = Trimmed Eigenvalue Forecast Combination (Hsiao and Wan, 2014), CLS = Constrained Least Squares, BG = Bates and Granger (1969), LAD = Least Absolute Deviation, NG = Newbold and Granger (1974), RF = Random Forest, BS = Boosted Stumps.*

Panel A: Spreads

	twRQS	vwRES
Amihud	**0.704**	**0.699**
VoV(Spread)	**0.744**	**0.750**
Roll	0.279	0.275
RollAbs	0.336	0.341
RollGibbs	0.435	0.439
EffTick	0.518	0.530
HighLow	0.519	0.511
AbdiRanaldo	0.530	0.515
SA	0.674	0.677
BG	**0.717**	**0.719**
NG	**0.758**	**0.762**
TEFC	**0.728**	**0.738**
CLS	**0.758**	**0.762**
LAD	**0.756**	**0.756**
RF	**0.748**	**0.747**
BS	**0.747**	**0.750**

(continued)

Table A.8: **Proxy Combinations (Continued)**

Panel B: Price Impact

	vwRPI	λ	λ^{root}
Amihud	**0.615**	**0.796**	0.810
1oV	0.480	0.641	0.590
$VoV(\lambda)$	**0.660**	**0.825**	**0.882**
RolloV	0.366	0.492	0.516
RollAbsoV	0.492	0.642	0.646
RollGibbsoV	**0.574**	0.759	0.761
EffTickoV	0.524	0.666	0.640
HighLowoV	**0.598**	0.791	0.791
AbdiRanaldooV	0.567	0.744	0.746
PastorStambaugh	0.004	0.044	0.029
SA	**0.629**	**0.824**	0.826
BG	**0.640**	**0.840**	**0.858**
NG	**0.652**	**0.849**	**0.887**
TEFC	**0.621**	**0.843**	**0.882**
CLS	**0.652**	**0.850**	**0.885**
LAD	**0.654**	**0.845**	**0.884**
RF	**0.625**	**0.836**	**0.878**
BS	**0.628**	**0.835**	**0.871**

Panel C: Efficiency

	$\sigma_s^{VAR(5)}$	AMIM	VR_{30}	$\lvert AR1\rvert$
$\sigma_s^{MA(1)}$	0.269	0.008	0.005	0.038
AMIM	0.007	−0.013	−0.003	0.011
VR_2	−0.032	0.016	0.025	0.009
$\lvert AR1\rvert$	−0.035	0.019	0.000	−0.012
SA	0.076	0.012	0.010	0.015
BG	0.104	0.012	0.010	0.015
NG	0.203	0.006	0.007	0.020
TEFC	0.131	0.013	0.007	0.022
CLS	0.203	0.006	0.007	0.020
LAD	0.209	−0.066	−0.050	−0.079
RF	0.194	0.029	0.027	−0.013
BS	0.227	−0.015	−0.025	−0.062

Overall, proxy combination does not significantly improve the correlation beyond that of the best single proxy. We thus recommend using VoV measures instead. More involved combination techniques do not seem to yield a noteworthy benefit.

A.8 The Noise in Price Efficiency Measures Estimated at Different Frequencies

Our analysis shows that the efficiency estimates of commodity futures prices is sensitive to the sampling frequency. Proxies estimated from daily data are not correlated to their benchmarks computed from intraday data, except for pricing error volatility.

In this section, we explore possible reasons and implications. First, both daily and intraday price efficiency could exhibit a low signal-to-noise ratio (SNR) if commodity futures price efficiency does not vary much over the sample period. Second, either one could exhibit a low SNR if one is able to detect inefficiencies, but the other is not. Third, both could have a high SNR but long- and short-lived price inefficiencies might be driven by different market forces and therefore be largely unrelated. For example, Chordia et al. (2005) show that inefficiencies in equity markets are short-lived with an average time-span of 5 to 60 minutes. Then, measures based on daily data cannot detect these. On the other hand, price inefficiencies that are longer-lived than a whole trading day (e.g. momentum) could be entirely different. Intraday measures of price efficiency might not be able to capture these. For example, evidence from other asset classes suggests that inventory effects last up to two months in equities (Hasbrouck and Sofianos, 1993; Subrahmanyam, 2008) and at least one day in equity options (Muravyev, 2016).

To explore this, we compute SNRs for benchmark, TAS, and daily proxy measures and compare their cross-sectional averages. First, we aggregate

the benchmark and TAS measures to a monthly frequency by taking simple averages. For each measure, we estimate the signal as a local non-linear trend using a LOESS-regression (Cleveland et al., 1993). We use second-degree polynomials and smoothing parameters (α) ranging from 0.1 and 0.9 in steps of 0.1. A high α corresponds to a smoother LOESS-estimate. For each commodity and α, we compute the R^2. Then, we transform them to SNRs using the relation

$$SNR = \frac{R^2}{1 - R^2}. \tag{A.43}$$

We report cross-sectional averages for different values of α in Table A.9. The second column indicates the data from which the measures were computed. Unsurprisingly, the measures computed from quote data (Benchmark) exhibit the highest SNR for all measures and all degrees of smoothing. The SNR of σ_s is the highest among all measures. It is possible that there is less variation in weak-form price efficiency than in its semi-strong variant. This could explain why we see higher SNRs and correlations in pricing error volatilities but not in purely autocorrelation-based measures like variance ratios, AMIM or the absolute first-order autocorrelation. The SNR of the σ_s proxy estimated from TAS data decreases with an increase in α. For all other measures, the SNR of both TAS and daily proxies deteriorates faster with increased smoothing. The SNR of TAS measures is not far from the SNR of its benchmarks, while price efficiency proxies computed from daily data appear to be noisier. Overall, these results suggest that the low correlation of daily price efficiency proxies is likely due to high levels of noise in daily measures, which can be reduced by increasing the sampling frequency and using TAS data. Thus, intraday price efficiency appears not to be pure noise, but to exhibit trends that can be captured if measured at high frequencies. Price efficiency estimated from daily data, however, requires longer time-series than a month to estimate.

The low correlation of daily and intraday price efficiency measures has

implications for regulators as well as investors, who are advised to measure and monitor price efficiency at different frequencies to react accordingly. Several papers measure commodity futures price efficiency using proxies that rely on daily settlement prices (Kim, 2015; Brogaard et al., 2019; Bohl et al., 2021). Their results might therefore be affected by high levels of noise, which might explain why some of their results are different from those using intraday data (Chen and Chang, 2015; Bessembinder et al., 2016; Raman et al., 2020).

Table A.9: **Signal-to-Noise Ratios of Price Efficiency Measures**

This table shows average SNRs of LOESS-regressions with different values for the smoothing parameter α. The second column, Data, indicates the frequency at which the price efficiency measure was computed.

α	Data	σ_s	VR	AMIM	$\lvert AR1 \rvert$
0.100	Benchmark	7.899	1.741	1.346	1.165
	TAS	3.633	1.064	1.116	1.049
	Daily	0.586	0.433	0.443	0.412
0.200	Benchmark	3.691	0.972	0.621	0.615
	TAS	1.978	0.488	0.468	0.462
	Daily	0.265	0.172	0.155	0.166
0.300	Benchmark	2.843	0.744	0.473	0.473
	TAS	1.506	0.349	0.312	0.329
	Daily	0.189	0.113	0.099	0.113
0.400	Benchmark	2.267	0.622	0.382	0.382
	TAS	1.216	0.264	0.240	0.254
	Daily	0.146	0.082	0.069	0.084
0.500	Benchmark	1.903	0.555	0.324	0.326
	TAS	1.019	0.224	0.198	0.208
	Daily	0.120	0.064	0.053	0.067
0.600	Benchmark	1.668	0.506	0.274	0.284
	TAS	0.910	0.201	0.169	0.181
	Daily	0.103	0.053	0.044	0.058
0.700	Benchmark	1.495	0.463	0.237	0.247
	TAS	0.837	0.184	0.151	0.161
	Daily	0.091	0.046	0.038	0.051
0.800	Benchmark	1.381	0.434	0.214	0.221
	TAS	0.777	0.169	0.137	0.146
	Daily	0.082	0.038	0.033	0.043
0.900	Benchmark	1.296	0.406	0.193	0.197
	TAS	0.726	0.155	0.121	0.131
	Daily	0.074	0.032	0.027	0.036

Chapter 3

Financialization, Electronification, and Commodity Market Quality*

3.1 Introduction

Markets are places where supply meets demand and prices are formed. A high-quality market is liquid and forms efficient prices (O'Hara and Ye, 2011). Market quality is time-varying and driven by a multitude of factors that include the market's design and the composition of participants. Uninformed traders play a central role in microstructure theory. In models, they randomly submit buy or sell orders and are required for a market to exist in the first place (Akerlof, 1970; Bagehot, 1971). Their share determines the ease with which informed traders can hide, which affects both the speed of convergence to the fair price and the width of the bid–ask spread (e.g., Glosten and Milgrom, 1985). However, it is possible that uninformed traders unanimously

*This chapter is based on the Working Paper "Financialization, Electronification, and Commodity Market Quality" by Tobias Lauter, Marcel Prokopczuk, and Stefan Trück, 2023.

trade in one direction, which results in order imbalances creating inventory risk that affects prices set by risk-averse liquidity suppliers (e.g., Stoll, 1978). Additionally, unanticipated order imbalances might be mistaken for informed trading. This can happen when the activity of uninformed traders overlaps, e.g., due to correlated fund flows from investors or shared portfolio rebalancing triggers. In this paper, we investigate the influence of both unanticipated and anticipated trading of uninformed market participants on market quality. We do this by studying commodity index traders (CITs) that are accused of disrupting price formation by buying and selling large volumes unanimously.[1] In the 2000s, commodity futures markets saw a substantial change in first trader composition and then market design that provides an opportunity to study their impact on market quality.

First, the composition of market participants changed. After regulatory relaxations following the Commodity Futures Modernization Act (CFMA) in 2000, commodity index ETFs attracted large amounts of mostly passive financial investment. Index investments rose from $13 billion in 2003 to $317 billion in 2008 (Masters and White, 2008). Yan et al. (2019) provide aggregate estimates of global commodity-linked investment peaking at more than $400 billion around 2012. Stoll and Whaley (2010) report that index investments made up 24% of open interest in US commodity futures. This development was coined the 'financialization' of commodity markets. ETFs turned commodity markets into an easily accessible asset class. Because commodity futures serve as benchmarks for fragmented spot markets and as a tool for hedging, concerns about the impact on market quality were raised by policymakers and regulators.[2]

Second, market design changed. Starting around mid-2006, commodity

[1] This is also called the "Masters Hypothesis", named after hedge fund manager Michael Masters and his testimony before the US Senate (Masters, 2008; Masters and White, 2008)

[2] See, e.g., U.S. Senate Permanent Subcommittee on Investigations (2006); U.S. Senate Permanent Subcommittee on Investigations (2009).

futures markets saw a technological transformation. For decades, trading activity was concentrated in open-outcry markets where traders interacted physically and used hand signals and their voice to submit orders. The marketplace was an exclusive community with high barriers of entry, where roles were visible by colorful jackets and badges. Traders knew each other and were subject to reputational concerns. Electronic limit order trading existed, but was limited to overnight hours when the pits were closed. Between 2006 and 2008, the major exchanges switched to a hybrid model and allowed side-by-side trading of floor and electronic trading during the main trading hours. Volume gradually shifted to the electronic limit order book and the pits dried out, resulting in most marketplaces closing their commodity pits. For example, CME shut down its commodity trading pits at NYMEX in 2016 due to low trading volume.

In this paper, we investigate how these changes affected market quality, i.e., the duality of liquidity provision and price formation. We use a broad panel of commodity futures that spans 23 years from 1996 to 2018, covering both event periods as well as multiple years before and after. To estimate liquidity and price efficiency, we use 11 years of millisecond-stamped trade and quote data that are available and reliable from 2008 onward. We employ the most suitable proxies for commodity market quality following Lauter and Prokopczuk (2022). For tests that span the entire sample period, we use time-and-sales data to estimate proxies of market quality. After comparing conditional means and trends, we employ a battery of identification strategies to test, if and how, the two changes affected market quality.

Our contribution to the literature is an in-depth study of commodity market quality across regimes. The data consist of a broad cross-section of commodities, is sampled at intraday frequency, and includes both the pre- and post-financialization period. In contrast to most previous studies, we also analyze the effects of commodity futures electronification. Exceptions to

this are Martinez et al. (2011), Raman et al. (2020), and Hu et al. (2020).

First, we perform a simple sample split to assess differences in market quality across regimes as well as index versus non-index commodities. We find that market quality has improved over the sample period, including and especially during the years of financialization and electronification. The improvements also appear to be more pronounced in commodities that are part of a major index. Next, we employ position data on index trading published by the Commodity Futures Trading Commission (CFTC) and run panel regressions to test if the trading activity attributable to speculators or CITs affects market quality. We find that neither predictable nor unpredictable trading activity appears to be convincingly harmful to market quality. Electronic trading appears to coincide with increased price efficiency. We then perform a case study to analyze whether an exogenous but predictable shock in CIT participation impacts market quality. In particular, we study the soybean meal market as this commodity was added to the Bloomberg Commodity Index (BCOM, formerly DJ-UBSCI) in 2013. Using soybean futures as a benchmark, we again find no harmful effect. Lastly, we focus on predictable trades and their effects on market quality during roll days. We argue that event-study approaches suffer from the fact that market quality and volume are not flat around roll days. Thus, we compare market quality during roll days across pre- and post-financialzation regimes. The data suggest that roll activity produces a significant increase in volume which translates to increased liquidity, supporting the 'sunshine trading effect' hypothesis (Admati and Pfleiderer, 1991). Moreover, informational efficiency does not seem to be affected significantly. In total, throughout our tests, we do not find any evidence of harmful effects on market quality.

Our work is related to the literature on empirical market microstructure, especially studies of trader composition and index investing in general. For example, Boehmer and Kelley (2009) show that the informational efficiency

in equity markets is related to both institutional trading and holdings. Israeli et al. (2017) document a decline of liquidity and informational efficiency in increased ETF ownership of stocks. Our work is also related to studies that compare the quality of markets across market designs. In the commodities context, it is related to Shah and Brorsen (2011) who analyze side-by-side trading in KCBT wheat futures. Instead of modeling differences in pit versus electronic trading within a commodity, we model differences between commodities depending on the degree of electronification. It is also related to Raman et al. (2020) who study the electronification of NYMEX WTI futures.

Our work expands the literature on commodity market financialization (Cheng and Xiong, 2014).[3] Haase et al. (2016) provide a meta study on the impact of speculation on commodity futures markets that highlights the controversy. Several studies examine commodity index traders and long-term (several days or weeks) pricing errors (e.g., Singleton, 2014). Others rely on predictable trades like index rolls (e.g., Mou, 2011; Bessembinder et al., 2016) or trades hedging commodity linked notes (Henderson et al., 2015) to study potential price impacts or effects on liquidity. We also model the effects of the Goldman Sachs Commodity Index (GSCI) roll but instead of testing measures during the roll period against that around it, we are able to test the effects between regimes (pre- versus financialization). This is particularly useful because market quality appears to exhibit strong trends around the roll period.

Theoretically, besides no-arbitrage arguments in models that emphasize the role of storage (Kaldor, 1939; Working, 1960), the price formation of commodity futures is driven by the trading motives of market participants. The classical hedging pressure theory of Keynes (1923) explains risk premia in futures prices by the interplay of short producers wishing to hedge and long speculators providing this service in return for a risk premium. The

[3]See Natoli (2021) for a recent literature review.

more recent literature builds on these ideas and describes additional effects of financialization on commodity market quality. One example are models that relate limits to arbitrage to market quality (e.g., Acharya et al., 2013; Singleton, 2014). They explore channels that can explain how financial speculators might lower the informational efficiency of futures prices. Goldstein and Yang (2022) develop a two-period model in which financial traders are informed speculators but also trade to hedge the commodity exposures of other assets in their portfolio. In this model, speculation of financial traders increases price efficiency, while hedging has the opposite effect. Because the former effect only dominates when the share of financial traders is low, their model predicts a negative net effect when the share of financial traders continues to grow beyond a threshold. The effect on liquidity in their model is mainly positive.

The effect of the electronification and the arrival of additional financial traders as a result of this is explored by Raman et al. (2020), who show that informational efficiency in the NYMEX WTI market has improved after the introduction of electronic and open-outcry side-by-side trading on September 5^{th} 2006. Their proprietary CFTC data document that short-term oriented informed institutional traders joined the market due to lower barriers of entry. These provide liquidity, trade in multiple markets and improve informational efficiency measured by intraday pricing errors. Raman et al. (2020) list three channels through which electronification might affect market quality: (1) lower barriers of entry, (2) improved transparency, both of which reduce information asymmetries leading to increased competition, and (3) a simple decrease in order processing costs. Their results suggest that the electronification does not only affect market quality directly, but also indirectly by attracting additional institutional traders. They find that the activity of financial traders is concentrated in short-term contract, so they can use longer-term contracts as a control sample. Our paper, in contrast,

measures effects across commodities and regimes. It also explicitly aims at (uninformed) CITs instead of (informed) financial traders. Martinez et al. (2011) study side-by-side trading of corn, soybeans, and wheat futures. They find that electronic trading exhibits lower transaction costs and is able to incorporate information more quickly. Hu et al. (2020) document a positive impact of algorithmic trading on the market quality of corn, soybean, and live cattle futures.

Studying commodity market quality is not only an interesting avenue per se, since market quality has been shown to affect the real economy. Futures markets are not just mirror images of spot prices because producers rely on futures prices in their decision-making (Sockin and Xiong, 2015; Goldstein and Yang, 2022). Spot markets are often fragmented and futures contracts act as benchmarks. Because they are an input for production decisions, noise in futures prices can be mistaken for fundamental information about future commodity demand, leading to misallocations. Brogaard et al. (2019) study the link between firm profitability and futures price informativeness empirically. They document a decline in the informational efficiency of index commodity futures, leading to reduced operating profitability of companies with high exposure to index commodities.

The existing literature relies on several identification strategies to test the effect of financialization on market quality. Those can be generally classified into easily predictable volume and (largely) unpredictable volume.

Conceptually, the literature tests two opposite effects of easily foreseeable uninformed volume. One is the 'sunshine trading effect' of Admati and Pfleiderer (1991). They argue that pre-announced liquidity demand by traders that are credibly perceived as being uninformed attracts additional liquidity and improves both liquidity and efficient price formation. Opposite to that is the effect of front running or predatory trading. Instead of meeting the liquidity demand with additional supply, other traders can trade ahead of pre-

dictable trades, which leads to a decrease in liquidity and to inefficient pricing (Brunnermeier and Pedersen, 2005). Studies of commodity financialization aim at different predictable trading patterns. A trading pattern that is easily predictable is the activity of index traders mimicking the largely overlapping roll schemes of major diversified commodity indices like the GSCI or BCOM index, or single commodity ETFs like the US Oil Fund (USO). The GSCI and BCOM index consist of the most liquid commodities that are traded in the US or Europe. Their composition is rarely changed. In our sample period, for example, only soybean meal was added to the BCOM index which was announced months ahead. Both indices existed long before the start of the financialization. Thus, index membership can be seen as exogenous and used to identify the financialization treatment group. The evidence of these studies is mixed. Most of them document no significant effect of index roll trades on liquidity (e.g., Shah and Brorsen (2011) for crude oil, Shang et al. (2018) for corn) or a positive effect (Wang et al. (2014) for corn, Bessembinder et al. (2016) for WTI). Mou (2011), on the other hand, documents abnormal returns around GSCI roll days, whereas Stoll and Whaley (2010) do not find such an effect. Yan et al. (2019) take a different approach on predictable trades and study annual GSCI target weight adjustments that come into effect in January. They document temporary abnormal returns with the same sign as the weight change. Yet another approach to studying predictable trades is taken by Henderson et al. (2015), who document abnormal returns on days when commodity linked notes (CLN) are initially priced or their terminal payoff is determined, the days when delta hedges are initialized or unwound in the futures market.

Volume that is less easy to predict might have a different impact on market quality. If liquidity demand hits the market when it is in low supply, effects would be expected to be of larger scale. Liquidity suppliers might not be willing to meet this demand because it is harder to discern whether it

originates from informed traders. Empirical evidence mostly suggests that unpredictable index investment or speculation does not harm market quality (see, e.g., Stoll and Whaley, 2010; Chen and Chang, 2015; Kim, 2015; Bohl et al., 2021). In contrast, Ready and Ready (2022) find price impact estimated from intraday data being related to changes in CFTC SCOT reports.

The remainder of this paper is structured as follows. Section 3.2 describes the sample and our approach to measuring market quality. Section 4.4 presents the results and Section 3.4 concludes.

3.2 Measurement and Data

3.2.1 Data

To study the quality of commodity futures contingent on the market being in a pre- or post-financialization and electronification regime, we aim for a sample period that covers years before and after 2004. In the literature, the year 2004 is commonly considered as a breakpoint indicating the start of the financialization of commodity markets (e.g., Tang and Xiong, 2012; Brogaard et al., 2019). Intraday data in Thomson Reuters Datascope Select (formerly TRTH SIRCA) starts in January 1996. We download Time-and-Sales data (trade prices and time-stamps) for the period January 1996 to December 2018. Additionally, we download millisecond time-stamped trade and quote data from January 2008 to December 2018.[4] We obtain daily (open, high, low, and settlement) price and volume data from Thomson Reuters Datastream.

We include the largest (mostly) US-based commodity futures markets in our analysis. From the energy sector, these are New York Mercantile Exchange (NYMEX) WTI crude oil (CL), heating oil (HO), natural gas (NG), Intercontinental Exchange (ICE) (EU) natural gas (NGLNM), Brent Crude

[4]These data are only reliable from 2008 onwards, since side-by-side trading started around mid-2006 and large parts of the volume had shifted to electronic markets by 2008.

Oil (LCO), and gas oil (LGO). For grains, we use Chicago Board of Trade (CBOT) soybeans (S), corn (C), wheat (W), Kansas City Board of Trade (KCBT) hard red winter wheat (KW), CBOT soybean meal (SM), soybean oil (BO), rough rice (RR), and oats (O). For metals, we consider Commodity Metals Exchange (COMEX) gold (GC), silver (SI), and copper (HG), NYMEX platinum (PL), and palladium (PA). Softs are represented by ICE (US) cotton (CT), sugar No.11 (SB), coffee (KC), cocoa (CC), Chicago Mercantile Exchange (CME) lumber (LB), as well as ICE (US) orange juice (OJ). From the livestock sector, we use CME live cattle (LC), lean hogs (LH), and feeder cattle (FC). We do not always use all commodities in our analysis because some are traded too infrequently during the first half of the sample, which results in patchy estimates of market quality (e.g., PA, PL, and SB). We tabulate the commodities included in our sample in the appendix. We also cannot include futures traded on the London Metals Exchange (LME), because its data is not stored in the Datascope Select database.

For each of the commodities, we use the contract with the highest volume, which is often the front month contract, but can sometimes be the up to fifth-nearest contract, as is the case for soybean meal or soybean oil. To identify the most liquid contract, we use a five-day moving average of volume. Overnight periods are excluded from the sample to avoid low volume periods.

We download data on trader composition like the COT reports, SCOT reports, and index investment data (IID) from the CFTC website. COT data contain weekly (as of each Tuesday) data on the long and short open interest of commercial and non-commercial traders. These reports are available for contracts trading on US exchanges for the entire sample period. The data are, however, of limited use when trying to measure index trading activity, since index traders are present in both the commercial and non-commercial category. Many ETF providers hedge their commodity exposure via swap dealers who are then classified as commercials. Other ETF providers choose

to hedge directly using futures and are classified as non-commercials. Due to this shortcoming, the CFTC started publishing more detailed supplementary reports in 2006. These SCOT reports contain three categories: (1) speculators, (2) hedgers, and (3) commodity index traders, but cover only agricultural commodities. Another shortcoming is that positions are netted internally for each trader. IID is considered to be the best proxy for index investment activity (Irwin and Sanders, 2012). These reports contain non-netted data on all major commodities traded on US exchanges and are available from December 2007 on a quarterly frequency which was increased to monthly in July 2010. The collection and reporting of IID was discontinued by the CFTC after October 2015.

3.2.2 Measuring Market Quality

We follow O'Hara and Ye (2011) and measure market quality with separate proxies for liquidity and price efficiency.

To measure liquidity, we consider using two dimensions: (1) transaction costs captured with the relative effective bid–ask spread, and (2) price impact of market orders. Both are commonly estimated from trade, quote and volume data. However, in the pits, quotes and volumes were not recorded. Side-by-side electronic and pit trading during the main hours started in late 2006 or 2007 depending on the contract and exchange. Therefore, to be able to examine the years before 2008, we resort to proxies for the entire sample period. These proxies are computed from trade prices alone.

Lauter and Prokopczuk (2022) show that liquidity in the form of effective spreads and price impact can effectively be proxied using a volatility-over-volume measure. Their results also demonstrate that market efficiency measures are noisy but can be approximated using 5-minute time-and-sales data. We follow their recommendations and employ the following measures

to capture commodity market quality.

For the time period 2008–2018 in which bid-ask prices were recorded, for each day d and commodity futures contract i, we compute relative effective spreads (EffSprd), price impact as Kyle's λ using regressions of signed root-order imbalance on midquote returns (e.g., Hasbrouck, 2009), absolute deviations of 1-minute to 30-minute variance ratios (Lo and MacKinlay, 1988) from unity (VR), and the volatility of pricing errors (Hasbrouck, 1993) ($\sigma_s^{VAR(5)}$). In samples that include previous years, we resort to proxies computed from time-and-sales data. We estimate the Roll (1984) measure at trade-frequency as a proxy for effective spreads (EffSprd^R) following Easley et al. (2021), a volatility-over-volume measure following Fong et al. (2018) and Kyle and Obizhaeva (2016) (λ^{KO}) to capture price impact, a microstructure-adjusted variance ratio by Smith (1994) (VR^S), and pricing error volatility ($\sigma_s^{MA(1)}$) following Hasbrouck (1993). Note that all measures are inverse measures of market quality, so a high level corresponds to an illiquid and inefficient market. We scale our proxies to their respective benchmarks, so their levels are interpretable. A detailed description of the measures and a brief analysis of their validity can be found in the appendix.

3.3 Empirical Results

Our empirical identification strategy is split into 4 parts. First, we descriptively document an increase in commodity market quality during the financialization period and especially during the electronification regime. We find that this increase is stronger for indexed commodities. Second, we use two different CFTC datasets on aggregate position data of CITs to study the effect of their activity.[5] We find no harmful effects. Third, we use the BCOM index addition of soybean meal (SM) as a predictable but exogenous event.

[5]We provide evidence on the impact of speculative traders in the appendix.

Despite a large increase in long index open interest, market quality does not change. Fourth, we study predictable GSCI roll trades. We find a significant increase in trading volume but no harmful effect on market quality.

3.3.1 Has Commodity Market Quality Changed over Time?

We begin by descriptively studying the time-series of aggregate market quality of indexed commodities. If the influx of index money affects all indexed commodities contemporaneously, we should be able to identify trends in all commodities and thus in cross-sectional aggregates. Aggregating the proxy measures also reduces their noise components.

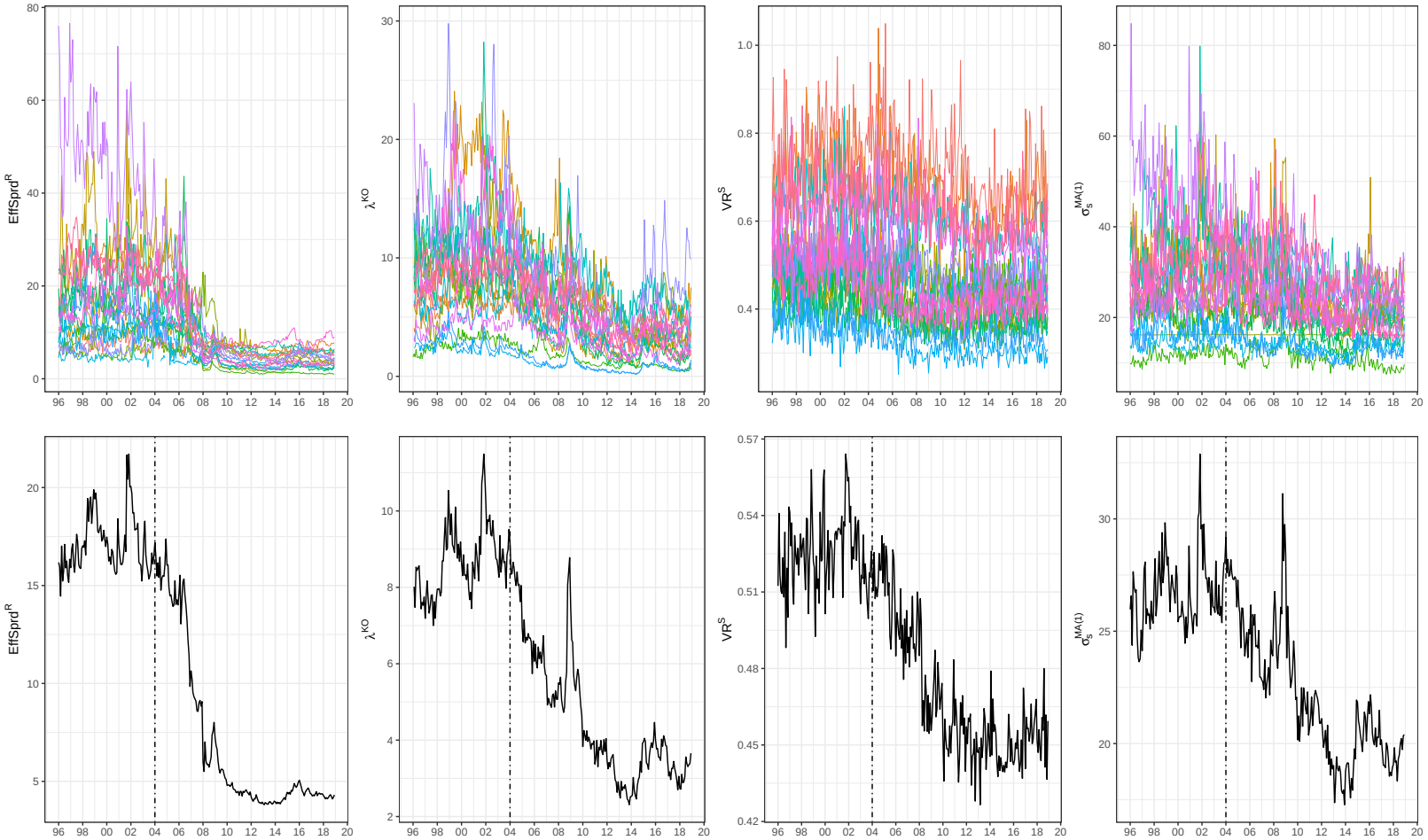

Figure 3.1: **Individual and Aggregate Market Quality**

This figure shows monthly market quality variables (top panels) and aggregated market quality variables that are averaged across commodities (bottom panels). $EffSprd^R$, λ^{KO}, VR^S, and $\sigma_s^{MA(1)}$ are scaled proxies for the relative effective spread, price impact, random walk deviations, and pricing error volatility, respectively. The vertical dashed line indicates January 2004, which is a commonly used breakpoint for the start of the financialization of commodity markets.

First, we aggregate our daily estimates to a monthly frequency by taking simple averages. The top panels in Figure 3.1 depict monthly estimates for our proxy measures. The two left panels show liquidity proxies, i.e., EffSprd^R in bps and λ^{KO} in bps per root million dollar volume. The two right panels show VR^S and $\sigma_s^{MA(1)}$ in % p.a..

We observe that among indexed commodities, there exists considerable cross-sectional heterogeneity. Temporal variation is also visible, as well as some commonality in the variables. For example, a spike in price impact around 2008 is visible when commodity prices peaked. An overall downward trend also seems evident.

The bottom panels in Figure 3.1 show monthly averages across all commodities. Here, trends and spikes are clearly visible. The vertical dashed line indicates January 2004 which is commonly used as the starting point of commodity financialization (e.g., Tang and Xiong, 2012; Brogaard et al., 2019). Up to this breakpoint, all four aggregate inverse measures of market quality seem to plateau and start to decline after around 2003. The overall trend continues until 2014. The period 2004–2014 includes the electronification of commodity futures trading, which occurred between 2006 and 2008. We are not aware of other substantial changes in commodity trading that might be responsible for the trend starting around 2004. Therefore, the financialization might be a plausible candidate for causing improvements in market quality.

Next, we formally test this conjecture for all commodities in a panel setting. To do so, we estimate a model of the form

$$MQ_{i,d} = \mu_i + \gamma_1 FIN_d + \gamma_2 ELE_{i,d} + \epsilon_{i,d}, \tag{3.1}$$

where $MQ_{i,d}$ is a set of (inverse) measures of market quality of commodity i on day d, i.e., EffSprd^R, λ^{KO}, VR^S or $\sigma_s^{MA(1)}$. We capture the financialization effect with a simple dummy FIN_d that takes on a value of one from January

2004 onwards. The dummy variable $ELE_{i,d}$ equals one if the commodity contracts trade electronically during pit trading hours and zero otherwise. With this set up, we can test for a level shift in market quality. Note that $ELE_{i,d}$ is not identical for all commodities. Side-by-side electronic and open-outcry trading on NYMEX started in September 2006, CBOT and KCBT in August 2006, COMEX in December 2006, CME meats in September 2006, ICE Europe in April 2005, and NYBOT (now ICE US) in March 2008 (except for orange juice which started in February 2007). μ_i is a commodity fixed effect and accounts for unobserved cross-sectional heterogeneity in $MQ_{i,d}$. We pool daily data of 20 indexed commodities and estimate the model using OLS. Note that we intentionally do not include any control variables because we aim at a simple first description of the evolution of commodity market quality across commodities and regimes.

We report estimates for γ_1 and γ_2 in Table 3.1. Since all the variables are inverse measures of market quality, a negative level shift is evidence for an improvement of market quality. The results suggest that average effective spreads remained stable after the start of the financialization and were reduced on average by 10.342 bps after the migration of the main trading activity to electronic limit order books. Price impact was lower after the start of the financialization and the electronification of each commodity as well. The results for VR^S and $\sigma_s^{MA(1)}$ are similar. Price efficiency was higher or remained unchanged after 2004 and improved further, reaching a higher level after the staggered electronification of commodity trading.

Comparing levels across regimes as above is most appropriate if the regime shift occurs abruptly instead of gradually. However, the majority of index investors gradually entered commodity markets starting around 2004. Also, once electronic trading became available during pit hours, volume gradually migrated from pits to limit order books. Thus, a piece-wise linear trend model might be more appropriate. We define four different regimes for our

Table 3.1: **Average Market Quality across Regimes**

This table shows estimates of regressions estimated using OLS. The regressions are of the form

$$MQ_{i,d} = \mu_i + \gamma_1 FIN_d + \gamma_2 ELE_{i,d} + \epsilon_{i,d},$$

*where $MQ_{i,d}$ is a set of (inverse) measures of market quality of commodity i on day d, i.e., $EffSprd^R$, λ^{KO}, VR^S or $\sigma_s^{MA(1)}$. We capture the financialization effect with a dummy FIN_d that equals one if day d is past the year 2003. The dummy variable $ELE_{i,d}$ is one if the commodity contracts trade electronically during pit trading hours and zero otherwise. μ_i are commodity fixed effects that account for unobserved cross-sectional heterogeneity in $MQ_{i,d}$. The sample includes daily data of 20 indexed commodities (in GSCI, BCOM or both) and ranges from 1996 to 2018. Standard errors are clustered by day and commodity. t-ratios are in parentheses. ***, **, and * indicate significance at the 1, 5, and 10% level, respectively.*

	Dependent variable:			
	EffSprd^R	λ^{KO}	VR^S	$\sigma_s^{MA(1)}$
	(1)	(2)	(3)	(4)
FIN	-1.590	-1.450^{***}	-0.013^{*}	-0.464
	(-1.424)	(-3.904)	(-2.048)	(-0.646)
ELE	-10.342^{***}	-3.189^{***}	-0.052^{***}	-4.993^{***}
	(-7.102)	(-11.298)	(-7.985)	(-5.959)
Observations	111,182	111,649	111,264	112,445
Adjusted R^2	0.493	0.590	0.262	0.266

sample: (1) the pre-financialization period from January 1996 to December 2003, (2) the financialization period starting in January 2004, (3) the electronification period which is different for each commodity and begins on the day when a specific commodity futures started trading electronically during pit trading hours, and (4) the modern period starting 2 years after the start of the electronification period. In an attempt to separate the effects of the financialization and electronification, we assume that once the electronification started, its effect overshadowed financialization effects. We also assume that it took no more than 2 years for major parts of the volume to migrate from the pits to the electronic limit order books.[6]

[6]If volume migrated within less than 2 years, the estimate of the electronification regime

We estimate regressions of the form

$$MQ_{i,d} = \mu_i + \theta_1 tr_d^{PRE} + \theta_2 tr_d^{FIN} + \theta_3 tr_{i,d}^{ELE} + \theta_4 tr_{i,d}^{MOD} + \epsilon_{i,d}. \tag{3.2}$$

$MQ_{i,d}$ is again a set of (inverse) measures of market quality of commodity i on day d, i.e., EffSprd^R, λ^{KO}, VR^S or $\sigma_s^{MA(1)}$. tr_d^{PRE} is a linear trend starting from 0 on 1996-01-02 and increasing by 1/250 each business day d. tr_d^{FIN} is a piece-wise linear trend (hockey stick function) that is 0 until 2004-01-02 and increases by 1/250 each business day after. tr_d^{ELE} is a piece-wise linear trend that is 0 until the day the commodity futures started trading electronically during pit hours and increases by 1/250 each business day after. Lastly, tr_d^{MOD} is a piece-wise linear trend that is 0 until 2 years after the day the commodity futures started trading electronically during pit hours and increases by 1/250 each business day after. μ_i captures unobserved time-invariant cross-sectional heterogeneity in $MQ_{i,d}$. The resulting piece-wise linear function is continuous by design and admits non-zero slopes in each regime. The parameters $\{\theta_p\}_{p=1:4}$ measure the annual change in $MQ_{i,d}$ during or after the regime. Thus, the estimated annual change in $MQ_{i,d}$ in the n^{th} regime is given by $\sum_{i=1}^{n} \hat{\theta}_p$.

OLS estimates for $\{\theta_p\}_{p=1:4}$ are in Panel A of Table 3.2. To ease the interpretation, the cumulative coefficients are in Panel B of Table 3.2. The latter estimates are regime-conditional annual changes in the respective market quality variable. The results suggest that before 2004 none of the variables exhibits a significantly negative trend. After 2004, the start of the financialization, the commodity-wide trend turns negative. That means that market quality significantly improved between January 2004 and the start of side-by-side trading. During the electronification regime of each commodity, the

will be downward biased. Since 2 years are a rather long period for some commodities, our estimates can be interpreted as a lower bound for the evolution of commodity market quality during the electronification regime.

trend remains negative. For some commodities, however, the 2008/2009 price spike in commodities that was accompanied by a spike in λ^{KO} and $\sigma_s^{MA(1)}$ (Figure 3.1) seems to lead to a reduction in the trend during the electronification regime. In the modern age regime, which we define to start 2 years after the start of the electronification of each commodity, the overall trend remains negative but less steep.

So far, we have considered only indexed commodities. The shift in means and the regime-conditional trends hint at an increase in market quality, but the change could be driven by market forces that we do not consider in our analysis. For example, it could be the case that commodity market quality increased for all commodities, but CITs dampened the increase in market quality of indexed commodities. If this were the case, we would see an even larger increase in market quality of non-indexed commodities. Although our complete sample includes seven non-indexed commodities, three of them (PL, PA, and NGLNM) have no or insufficiently many intraday observations before 2008, which reduces the group of non-indexed commodities to four (LB, O, OJ, and RR).

We begin with a visual inspection of the aggregate variables. First, in order to reduce noise, we average each measure for each commodity over a monthly frequency. Then, we scale each measure for each commodity to have a mean of one before 2004 by dividing it by its pre-2004 mean. Finally, we average each measure across all index or non-index commodities. The resulting averages are shown in Figure 3.2. They hint at a rather larger increase in market quality of indexed versus non-indexed commodities.

Table 3.2: **Trends in Market Quality across Regimes**

This table shows estimates of regressions estimated using OLS. The regressions are of the form

$$MQ_{i,d} = \mu_i + \theta_1 tr_d^{PRE} + \theta_2 tr_d^{FIN} + \theta_3 tr_{i,d}^{ELE} + \theta_4 tr_{i,d}^{MOD} + \epsilon_{i,d}. \tag{3.3}$$

$MQ_{i,d}$ is a set of (inverse) measures of market quality of commodity i on day d, i.e., $EffSprd^R$, λ^{KO}, VR^S, or $\sigma_s^{MA(1)}$. tr_d^{PRE} is a linear trend starting from 0 on 1996-01-02 increasing by $1/250$ each day d. tr_d^{FIN} is a piece-wise linear trend that is 0 until 2004-01-02 and increases by $1/250$ each day after. tr_d^{ELE} is a piece-wise linear trend that is 0 until the day the commodity futures started trading electronically during pit hours and increases by $1/250$ each day after. tr_d^{MOD} is a piece-wise linear trend that is 0 until 2 years after the day the commodity futures started trading electronically during pit hours and increases by $1/250$ each day after. μ_i capture commodity fixed effects. The resulting piece-wise linear function is continuous by design. The sample includes daily data of 20 indexed commodities (in GSCI, BCOM or both) and ranges from 1996 to 2018. Panel A shows the estimates for $\theta_{1:4}$. Standard errors are clustered by day and commodity. t-ratios are in parentheses. Panel B shows cumulative coefficients $\sum_{i=1}^{n} \theta_p$ which is the estimated annual change in $MQ_{i,d}$ in the n^{th} regime. ***, **, and * indicate significance at the 1, 5, and 10% level, respectively.

Panel A: Coefficients

	Dependent variable:			
	EffSprd^R	λ^{KO}	VR^S	$\sigma_s^{MA(1)}$
	(1)	(2)	(3)	(4)
tr^{PRE}	0.015	0.184^{*}	0.0004	0.216^{*}
	(0.079)	(1.811)	(0.389)	(1.742)
tr^{FIN}	-1.422^{**}	-1.782^{***}	-0.010^{***}	-1.567^{***}
	(-2.325)	(-5.318)	(-4.952)	(-3.286)
tr^{ELE}	-2.776^{***}	1.440^{***}	-0.010^{*}	0.892
	(-3.907)	(4.703)	(-2.028)	(1.690)
tr^{MOD}	4.101^{***}	-0.046	0.018^{***}	0.007
	(7.175)	(-0.413)	(5.709)	(0.018)
Observations	111,182	111,649	111,264	112,445
Adjusted R^2	0.504	0.629	0.264	0.274

Panel B: Cumulative Coefficients

	EffSprd^R	λ^{KO}	VR^S	$\sigma_s^{MA(1)}$
tr^{PRE}	0.015	0.184	0.000	0.216
tr^{FIN}	-1.407	-1.598	-0.010	-1.351
tr^{ELE}	-4.183	-0.158	-0.019	-0.460
tr^{MOD}	-0.082	-0.203	-0.001	-0.453

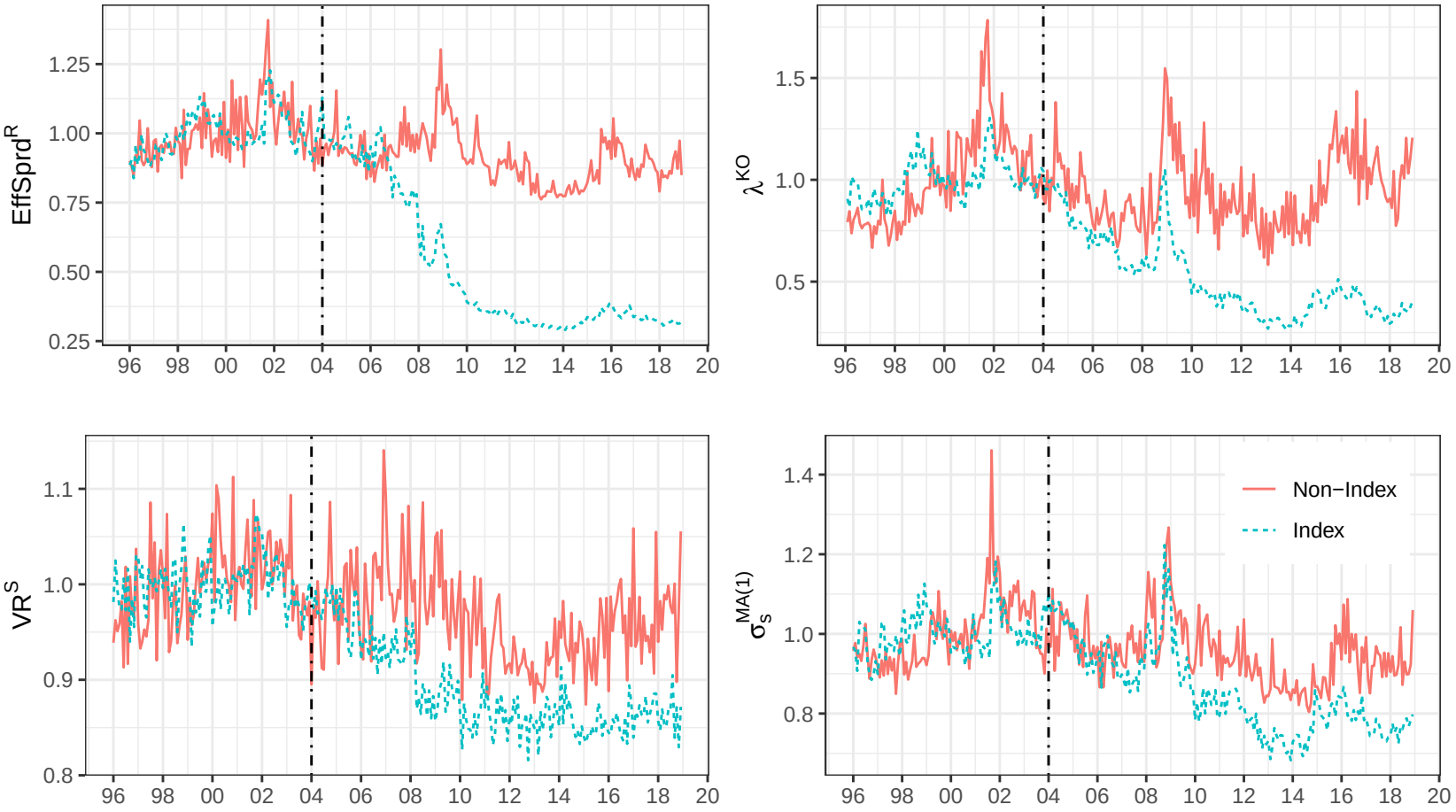

Figure 3.2: **Index (Dashed Blue) vs Non-Index (Solid Red) Aggregate Market Quality**

This figure shows monthly market quality measures averaged across all 20 index (dashed blue) and four non-index (solid red) commodities. Before aggregation, we scale the respective market quality measure of each commodity to a pre-2004 mean of 1 by dividing each series by its pre-2004 mean. $EffSprd^R$ (top left), λ^{KO} (top right), VR^S (bottom left), and $\sigma_s^{MA(1)}$ (bottom right) are scaled proxies for the relative effective spread, price impact, random walk deviations, and pricing error volatility, respectively.

To test this, we adapt and augment the model in Equation (3.1) in the following way

$$MQ_{i,d} = \mu_i + \gamma_1 INDEX_i \times FIN_d + \gamma_2 FIN_d + \gamma_3 ELE_{i,d} + \epsilon_{i,d}, \qquad (3.4)$$

where the variables are defined as previously in Equation (3.1) but we add $INDEX_i$ which is a dummy that is 1 if commodity i is an index commodity (in GSCI or BCOM). The commodity fixed effects μ_i capture cross-sectional heterogeneity in $MQ_{i,d}$ and subsume the $INDEX_i$ dummy outside the interaction term. The sample includes the daily market quality measures of 20 index and four non-index commodities. If the financialization is substantially harmful to market quality, we would expect γ_1 to be positive.

We report OLS estimates in Table 3.3. Two out of four estimates are significantly negative while the other two are not significantly different from zero. These results suggest that the market quality of indexed commodities increased by a larger extent compared to non-index commodities. This is evidence against a harmful effect of index trading on market quality. The size of the estimates suggests that the electronification period appears to coincide with larger increases in market quality than the financialization one.

Overall, this analysis shows that commodity market quality changed substantially during the time period after 2004. Contrary to the hypothesis that the financialization had a negative effect on commodity market quality, we do not find such an effect, but a trend and significant shift in the level that coincides with the financialization and electronification of commodity markets. In the following sections, we follow different identification strategies in order to analyze the effect of speculators and index investors on market quality.

Table 3.3: **Index vs Non-Index Market Quality across Regimes**

This table shows estimates of regressions estimated using OLS. The regressions are of the form

$$MQ_{i,d} = \mu_i + \gamma_1 INDEX_i \times FIN_d + \gamma_2 FIN_d + \gamma_3 ELE_{i,d} + \epsilon_{i,d},$$

*where $MQ_{i,d}$ is a set of (inverse) measures of market quality of commodity i on day d, i.e., $EffSprd^R$, λ^{KO}, VR^S or $\sigma_s^{MA(1)}$. We capture the financialization effect with a dummy FIN_d that equals one if day d is past the year 2003. The dummy variable $ELE_{i,d}$ is one if the commodity contracts trade electronically during pit trading hours and zero otherwise. $INDEX_i$ is a dummy variable that is 1 if commodity i is part of the GSCI, the BCOM index, or both. μ_i capture commodity fixed effects. The sample includes daily data of 20 (4) (non-)indexed commodities and ranges from 1996 to 2018. Standard errors are clustered by day and commodity. t-ratios are in parentheses. ***, **, and * indicate significance at the 1, 5, and 10% level, respectively.*

	Dependent variable:			
	EffSprd^R	λ^{KO}	VR^S	$\sigma_s^{MA(1)}$
	(1)	(2)	(3)	(4)
$FIN \times INDEX$	-8.287^{***}	53.325	-0.027^{**}	-0.317
	(-3.443)	(1.037)	(-2.534)	(-0.126)
FIN	5.431^{***}	-48.533	0.008	-0.441
	(3.383)	(-1.067)	(0.851)	(-0.231)
ELE	-8.928^{***}	-10.780	-0.046^{***}	-4.638^{***}
	(-5.869)	(-1.341)	(-7.705)	(-5.335)
Observations	127,978	133,006	131,487	133,390
Adjusted R^2	0.536	0.709	0.343	0.609

3.3.2 Index Trading and Market Quality

In order to understand the interplay of index traders and market quality, we analyze how market quality is related to their trading activity. We rely on CFTC SCOT data, which consist of weekly reports that contain the aggregated open interest of trader groups classified into commercials (hedgers), non-commercials (speculators), and index traders if their positions exceed the reporting threshold. Open interest in CFTC data is aggregated across all contracts, so roll trades do not lead to changes in the data. This allows changes in open interest to be interpreted as net buying (selling if negative)

volume by the respective group of traders.

To draw conclusions about the impact of CITs on market quality, we rely on the SCOT reports published by the CFTC that explicitly provide information on the net open interest of CITs. In January 2006, the CFTC started publishing SCOT reports for 13 agricultural, livestock, and soft commodity futures. Thus, the sample starts after the common starting point of the financialization but before the electronification period which commenced in mid-2006 for the first US exchanges. CBOT (agriculturals) started trading electronically during pit hours on August 1^{st} 2006, ICE (softs) on February 2^{nd} 2007, and CME livestock on September 5^{th} 2006. This provides an opportunity to study the impact of both financialization and electronification.

We use the ratio of Datascope Select volume to Datastream volume as a proxy for the share of electronic volume relative to total volume. We reduce the sample to only include the seven commodities traded on CME and CBOT, because GLOBEX volume is recorded in Datascope for the entire sample period, but ICE electronic volume was not recorded before April 2008. We pick July 6^{th} 2015 as the end-date for our sample because all CME commodity futures pits were permanently closed after this date.[7]

We use an ARIMA(p,1,q) model estimated for each commodity to decompose relative net CIT open interest into predictable net volume of CITs (absolute *fitted* ARIMA differences) and unpredictable net volume of CITs (absolute *residual* ARIMA differences). We define relative net index open interest (RNI) as

$$RNI = \frac{CIT_L - CIT_S}{OI}, \tag{3.5}$$

where CIT_L (CIT_S) is long (short) open interest of CITs and OI is total open interest. Differences in RNI can be interpreted as relative net trading volume of speculators which we decomopose into predictable and unpredictable

[7] `https://www.cmegroup.com/tools-information/lookups/advisories/ser/SER-7416.html`.

net trading volume. For each commodity, we fit a regression with month-dummies and ARIMA(p,1,q) errors to the RNI time-series where p and q are between 0 and 10. The month-dummies are intended to capture seasonality. Each process includes a linear trend (a constant mean in differences). The functional form is chosen for each commodity based on the minimum Akaike information criterion using the search algorithm of Hyndman and Khandakar (2008). In contrast to equity markets, shorting is as easy as buying in futures markets. Thus, we expect a symmetric effect of net trading on market quality. Therefore, we use fitted (residual) absolute first differences in RNI as proxies for predictable (unpredictable) trading volume by speculators.

We estimate a linear regression model of the form

$$MQ_{i,w} = \mu_i + \gamma_w + \theta_1 Pred.V + \theta_2 Unpred.V + \theta_3 \%GLOBEX + \Theta' X_{i,w} + \epsilon_{i,w}. \tag{3.6}$$

$MQ_{i,w}$ is a set of (inverse) measures of market quality of commodity i in week w, i.e., EffSprd^R, λ^{KO}, VR^S, or $\sigma_s^{MA(1)}$. Pred.V (Unpred.V) is (un-) predictable net trading volume of index traders in %. $\%GLOBEX$ is the share of GLOBEX volume in total volume as a measure of electronification. X is a vector of control variables that are commonly used in empirical market microstructure applications of this kind, i.e., 5-minute intraday volatility, the natural log of dollar volume, the natural log of total (including unclassified) open interest in commodity i, and the inverse settlement price. In order to match frequencies with the SCOT reports, we aggregate daily measures of market quality and controls to a weekly frequency. μ_i (γ_w) capture commodity (year–week) fixed effects. If predictable (unpredictable) index trading or increasing market share of speculators harms market quality, we would expect θ_1 (θ_2) to be significantly positive.

Estimates for Equation (3.6) are reported in Table 3.4. Predictable volume does not appear to affect market quality. This is evidence in favor of

the sunshine trading effect of Admati and Pfleiderer (1991). Unpredictable volume has a large and significant on effective spreads. On average, CITs appear to act as net liquidity suppliers. The migration of volume from pits to electronic markets appears to improve effective spreads slightly. A complete switch from pit to electronic trading, improves effective spreads by 8 bps. This is in line with results by Raman et al. (2020) who argue that the electronification is a change through which more financial traders obtained market access.

Given that the cross-sectional dimension is quite small, with only seven agricultural commodities, we test another CFTC dataset that is considered to be closest to true index trading activity. IID reports were published on a monthly basis between June 2010 and September 2015. By this time, most futures contracts of all considered commodities were traded electronically. The period from 2010 to 2015 allows us to extend the set of dependent variables since we do not have to resort to TAS-based proxies, but can use trade, quote and volume-based measures. We add a non-parametric price impact measure (PrcImpct) which is the volume weighted 5-minute price impact. The measure is based on the signed difference in mid-quote prices with a 5-minute lag after a trade and is calculated as

$$\text{PrcImpct}_k = Q_k \, \frac{M_{k+5min} - M_k}{M_k}, \tag{3.7}$$

where M_{k+5min} is the mid-quote price five minutes after the k^{th} trade of a day and Q_k is a dummy that equals +1 (-1) for a buyer-initiated (seller-initiated) trade. PrcImpct thus measures longer-term price impact compared to λ, which is estimated from 1-minute returns. This could be relevant if a market is less resilient and any transitory price impact does not revert within 1 minute or less. We compute daily estimates by taking volume-weighted averages.

Table 3.4: **Index Trading vs Electronification**

This table shows OLS estimates of the regression

$$MQ_{i,w} = \mu_i + \gamma_w + \theta_1 Pred.V + \theta_2 Unpred.V + \theta_4 \%GLOBEX + \Theta' X_{i,w} + \epsilon_{i,w}.$$

$MQ_{i,w}$ is a set of (inverse) measures of market quality of commodity i in week w, i.e., $EffSprd^R$, λ^{KO}, VR^S, or $\sigma_s^{MA(1)}$. Pred.V (Unpred.V) is (un-) predictable net trading volume by index traders. These are absolute first differences in $ARIMA(p \leq 10, 1, q \leq 10)$ fitted (residual) relative net speculative open interest. $\%GLOBEX$ is the share of GLOBEX volume in total volume as a measure of electronification. X is a vector of control variables, i.e., 5-minute intraday volatility, the natural log of $-volume, the natural log of total reported open interest in commodity i, and the inverse settlement price. μ_i (γ_w) capture commodity (week) fixed effects. The sample includes weekly data of 7 commodities traded on CBOT or CME between January 3^{rd} 2006 and June 30^{th} 2015. Standard errors are double clustered by week and commodity. t-ratios are in parentheses. ***, **, and * indicate significance at the 1, 5, and 10% level, respectively.

	Dependent variable:			
	EffSprd^R	λ^{KO}	VR^S	$\sigma_s^{MA(1)}$
	(1)	(2)	(3)	(4)
Pred.V	11.638	0.879	−0.364	−13.011
	(1.768)	(0.342)	(−1.269)	(−1.123)
Unpred.V	−8.816**	0.743	0.003	8.818
	(−2.550)	(0.905)	(0.018)	(1.049)
%GLOBEX	−0.081**	−0.017*	−0.0001	−0.046
	(−3.155)	(−2.111)	(−0.489)	(−1.494)
Controls	Yes	Yes	Yes	Yes
Observations	3,411	3,417	3,417	3,417
Adjusted R^2	0.729	0.876	0.491	0.670

We also add the time-weighted inside depth (Depth) to the set of market quality measures. This is the number of contracts available at the lowest ask plus the number of contracts offered at the highest bid, weighted by the time these were standing. Then, we multiply it by the contract size, by the price in USD, and divide it by 10^6 to obtain depth in million USD.

We add a proxy for algorithmic trading (AT) activity to the set of control variables by computing

$$AT_{i,d} = \frac{V_{i,d}}{\#\text{Messages}_{i,d}}, \tag{3.8}$$

which is the dollar volume $V_{i,d}$ of commodity i on day d divided by the number of trading messages (trades, order submissions, and order cancellations) following Hendershott et al. (2011) and averaging it in every commodity–month.

We employ the following model to test whether index trading activity affects commodity market quality:

$$MQ_{i,m} = \mu_i + \gamma_m + \theta_1 Pred.V + \theta_2 Unpred.V + \Theta' X_{i,m} + \epsilon_{i,m}. \tag{3.9}$$

$MQ_{i,m}$ is either dollar volume or an (inverse) measure of market quality of commodity i in month m, i.e., EffSprd, λ, PrcImpct, Depth, VR, or $\sigma_s^{VAR(5)}$. Pred.V (Unpred.V) is (un-) predictable net trading volume of index traders. AT is a proxy for algorithmic trading defined in Equation (3.8) which we add to the set of control variables X. All other variables are identical to those in Equation (3.6) but at a monthly frequency. The full sample includes 19 indexed commodities and ranges from June 2010 to October 2015.

Pooled OLS estimation results are reported in Table 3.5. We find no evidence of a harmful effect of trading by CITs on any of the considered market quality measures. Thus, again, the results do not represent convincing

evidence that CITs harm market quality.

Overall, our analysis of aggregate trader position data suggests that it is unlikely that index trader activity is connected to a decrease in market quality. Even if index traders take directional positions unrelated to commodity fundamentals, their orders seem to be met by enough liquidity and informed arbitrageurs, ensuring price efficiency.

Table 3.5: **Index Trading and Market Quality**

This table shows OLS estimates of the regression

$$MQ_{i,m} = \mu_i + \gamma_m + \theta_1 Pred.V + \theta_2 Unpred.V + \Theta' X_{i,m} + \epsilon_{i,m}.$$

*$MQ_{i,m}$ is a set of (inverse) measures of market quality of commodity i in month m, i.e., EffSprd, λ, PrcImpct, Depth, VR, or $\sigma_s^{VAR(5)}$. Pred.V (Unpred.V) is (un-) predictable net trading volume by index traders. These are absolute first differences in ARIMA($p \leq 10, 1, q \leq 10$) fitted (residual) relative net speculative open interest. X is a vector of control variables, i.e., 5-minute intraday volatility, the natural log of $-volume, the natural log of total reported open interest in commodity i, and the inverse settlement price. μ_i (γ_m) capture commodity (year–month) fixed effects. Depth is in thousand USD. The sample includes 19 indexed commodities and ranges from 2010-06 to 2015-10 (65 months). Standard errors are clustered by week and commodity. t-ratios are in parentheses. ***, **, and * indicate significance at the 1, 5, and 10% level, respectively.*

	Dependent variable:					
	EffSprd	λ	PrcImpct	Depth	VR	$\sigma_s^{VAR(5)}$
	(1)	(2)	(3)	(4)	(5)	(6)
Pred.V	0.016	−0.012	0.004	−0.073	−0.004	−0.022
	(1.065)	(−0.746)	(0.282)	(−0.006)	(−1.396)	(−0.371)
Unpred.V	−0.001	0.016	0.003	3.644	−0.002	−0.003
	(−0.122)	(1.266)	(0.355)	(0.531)	(−0.654)	(−0.069)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,207	1,207	1,207	1,207	1,207	1,207
Adjusted R^2	0.887	0.848	0.708	0.794	0.329	0.820

3.3.3 A Case Study of Soybean Meal

In January 2013, CBOT soybean meal (SM) was added to the BCOM index. Previously, soybean meal has never been part of a major commodity index and, thus, commodity index ETF issuers did not have to hedge its price risk in the futures market. The BCOM composition is based on known criteria including liquidity. Changes are known about 2 months in advance before the start of the next year. Index weights for 2013 were announced on October 24, 2012.[8] The addition of SM is visible in CFTC reports, as illustrated in Figure 3.3. IID data exhibit a level shift in long index investment in January 2013.[9] We treat the addition of SM as an exogenous event, which provides an opportunity to test possible effects of (uninformed) index investment on market quality.

First, we compare the means in our market quality variables of SM for 2012 to 2013. We estimate a regression of the form

$$MQ_{SM,d} = \mu + \theta_1 2013_d + \epsilon_d. \tag{3.10}$$

$MQ_{SM,d}$ is either dollar volume or an (inverse) measure of market quality of SM on day d, i.e., EffSprd, λ, PrcImpct, Depth, VR, or $\sigma_s^{VAR(5)}$. Since the sample period starts with the year 2012, we use measures estimated from quote data. 2013_d is a dummy variable that is 1 if the respective date d is in the year 2013 when SM was part of the BCOM index.

[8] `https://www.spglobal.com/spdji/en/documents/index-news-and-announcements/20121024-spdji-ubs-commodity-target-weights.pdf`. In 2012 and 2013, BCOM was named Dow Jones-UBS Commodity Index.

[9] Index investment before 2013 has not been zero, because other (individual) commodity indices exist.

Figure 3.3: **Index Investment in Soybeans (S) and Soybean Meal (SM)**

This figure shows monthly index investment in CBOT Soybeans (S) and Soybean Meal (SM) from CFTC index investment data (IID). The graph shows futures equivalent (futures and options) contracts long in thousands of contracts. SM was added to BCOM in January 2013.

The OLS estimates are reported in Panel A of Table 3.6. They imply that average daily volume remained almost unchanged, while all market quality variables are either insignificantly different from zero or negative. Depth was reduced in 2013, but this is not reflected in increased spreads or price impact. Thus, although long index investment quadrupled in 2013 compared to 2012, market quality does not appear to have undergone a comparable shift.

However, this result might be driven by an overall (negative) trend in the variables that overshadows the effects of index traders. If this was the case, then their influence would unlikely be substantial. Nevertheless, we try to put changes in market quality into perspective. Soybeans (S) have been part of both the GSCI and the BCOM index throughout the depicted sample period. We construct a sample of daily measures of SM and S futures that spans the years 2012 and 2013. Soybean meal is a physical derivative of soybeans, which is 'crushed' into soybean meal and soybean oil. Thus, both commodities' demand and supply dynamics are intertwined. Hedgers and other traders often trade both simultaneously ('crush spread' trading). We deem spillover effects of index trading from SM to S as unlikely since S futures are more liquid than SM futures.[10] In the years 2012 and 2013, the returns of SM and S exhibit a correlation of 89%. The market quality measures of S and SM are also correlated (69% for EffSprd, 84% for λ, 42% for PrcImpct, 74% for Depth, 56% for VR, and 64% for $\sigma_s^{VAR(5)}$).

In our test, we rely on this relation between S and SM. We assume that any demand and supply shock that is likely to affect the market quality of SM futures affects S futures similarly, except for the BCOM index addition in January 2013 which is exclusive to SM. Our goal is to examine if market quality measures of SM are different from what the measures of S would suggest. In this setting, with a two-year sample period of paired commodities, we assume that these variables are parallel in both SM and S so that

[10]The volume of S exceeds that of SM by a factor of more than three in 2012 and 2013.

Table 3.6: **A Case Study of Soybean Meal**

This table shows estimates of regressions estimated using OLS. In Panel A, the regressions are of the form

$$MQ_{SM,d} = \mu + \theta_1 2013_d + \epsilon_d$$

and in Panel B of the form

$$MQ_{i,d} = \mu + \gamma_1 SM_i \times 2013_d + \gamma_2 SM_i + \gamma_3 2013_d + \Theta' X_{i,t} + \epsilon_{i,t}.$$

$MQ_{i,d}$ is either dollar volume, or an (inverse) measure of market quality of commodity i (Soybean meal SM or soybeans S) on day d, i.e., EffSprd, λ, PrcImpct, Depth, VR, or $\sigma_s^{VAR(5)}$. SM is a dummy that is 1 if commodity i is soybean meal SM. 2013 *is a dummy variable that is 1 if the respective date is in the year 2013 when SM was part of the BCOM index. X is a vector of control variables, i.e., 5-minute intraday volatility, the natural log of $-volume, the natural log of total reported open interest in commodity i, and the inverse settlement price. Depth is in million USD. The data are sampled at a daily frequency and spans the years 2012 and 2013. In Panel A, we use Newey and West (1987, 1994) standard errors. In Panel B, standard errors are clustered by day. t-ratios are in parentheses. ***, **, and * indicate significance at the 1, 5, and 10% level, respectively.*

Panel A: Change in SM Market Quality from 2012 to 2013

	Dependent variable:						
	Volume	EffSprd	λ	PrcImpct	Depth	VR	$\sigma_s^{VAR(5)}$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
2013	0.086	0.008	-0.473^{**}	0.074	-0.321^{***}	-0.362^{***}	0.369
	(1.264)	(0.074)	(-2.540)	(0.537)	(-4.160)	(-3.661)	(0.339)
Constant	1.654^{***}	3.393^{***}	3.756^{***}	1.847^{***}	1.052^{***}	0.842^{***}	16.209^{***}
	(32.239)	(44.267)	(28.891)	(21.500)	(14.189)	(8.985)	(22.757)
Observations	495	498	500	500	498	500	498
Adjusted R^2	0.006	-0.002	0.044	-0.001	0.152	0.038	-0.002

Panel B: Difference-in-Differences Approach

	Dependent variable:						
	Volume	EffSprd	λ	PrcImpct	Depth	VR	$\sigma_s^{VAR(5)}$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
$SM \times 2013$	0.338^{**}	0.086^{*}	-0.115	-0.114	0.902^{***}	0.003	0.101
	(2.041)	(1.933)	(-1.480)	(-0.949)	(7.975)	(0.033)	(0.128)
SM	-16.957^{***}	1.277^{***}	0.138	0.752	4.102^{***}	1.371^{***}	4.287
	(-17.026)	(5.986)	(0.334)	(1.212)	(5.849)	(2.649)	(0.945)
2013	-0.386^{**}	0.037	-0.261^{***}	0.267^{***}	-1.212^{***}	-0.355^{***}	0.984
	(-2.324)	(1.251)	(-6.034)	(3.129)	(-9.980)	(-4.456)	(1.321)
Constant	2.980	3.707^{**}	7.995^{***}	-12.537^{***}	-3.907	9.381^{***}	-85.839^{***}
	(0.848)	(2.554)	(2.650)	(-2.849)	(-1.296)	(3.192)	(-2.829)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	993	991	993	992	989	992	992
Adjusted R^2	0.836	0.654	0.687	0.214	0.589	0.086	0.319

any difference can be attributed to SM undergoing a shift in the degree of financialization due to its index addition.

We estimate a regression of the form

$$MQ_{i,d} = \mu + \gamma_1 SM_i \times 2013_d + \gamma_2 SM_i + \gamma_3 2013_d + \Theta' X_{i,t} + \epsilon_{i,t}. \quad (3.11)$$

$MQ_{i,d}$ is either dollar volume or an (inverse) measure of market quality of commodity i (SM or S) on day d, i.e., EffSprd, λ, PrcImpct, Depth, VR, or $\sigma_s^{VAR(5)}$. SM is a dummy that is 1 if commodity i is soybean meal SM. 2013 is a dummy variable that is 1 if the respective date is in the year 2013 when SM was part of the BCOM index. X is a vector of controls, i.e., intraday volatility, log-volume, log-open-interest, inverse price, and algorithmic trading. If the newly arrived index traders harm the market quality of SM futures, we would expect the coefficient of the interaction term γ_1 to be positive.

The estimation results are reported in Panel B of Table 3.6. The coefficients of the interaction term are designed to capture the effect of the index addition. Relative to soybeans, volume, effective spread, and depth increased. The increase in the spread is less than 0.1 bps, which is very small and unlikely to be economically relevant, and only significant at the 10% level. Thus, the market quality of SM relative to that of S does not seem to have suffered from the influx of index money.

Overall, the results of this case study do not show evidence of index trading activity substantially harming market quality, even if the arrival of CITs in a market is highly concentrated.

3.3.4 Market Quality During and Around Index Roll Days

We now turn to studying the possible impact of predictable trades on market quality. One strategy is to focus on measures on and around the roll days of major commodity indices. Bessembinder et al. (2016), for example, compare the (overlapping) USO and GSCI roll periods with the days [-7:-3] before the roll event. This is a valid strategy if the variables tested are considerably constant around the index roll days. In the following, we show that this is not the case for the commodities we consider. However, since our sample contains the pre-financialization period, we are able to compare patterns across regimes. We test if volume or market quality on the index roll days behaves differently after 2004 compared to before.

To do so, we use a sample of 14 commodities that have been constituents of the GSCI from 1996 to 2018. Instead of focusing on the highest-volume (or the front-month) contract, we study the contracts that are explicitly defined in the roll schemes of the GSCI index.[11]

The GSCI rolls 20% of its contracts on each of the five days ranging from the 5^{th} to the 9^{th} business day of a month. Thus, if the index rolls affect market quality, we should see an effect on these days. Since the long and short sides behave differently, we construct two samples. We define days relative to the first roll day, which we index by 0. For those contracts that the index mimicking traders go long, we include the business days [-22:22], i.e., one month before and after the roll period. Because trading in some contracts terminates up to eight days after the first roll day, we include the days [-22:7] for the shorted contracts. In order to eliminate cross-sectional differences and temporal trends, we scale the variables of interest of each individual

[11] `https://www.spglobal.com/spdji/en/documents/methodologies/methodology-sp-gsci.pdf`.

futures contract in the time window [-22:22] or [-22:7] to zero mean and unit variance. We also split each sample (long and short) in 2004 and test averages for equality in the measures during the roll period.

We depict the average scaled volume and market quality measures around the initial roll day in Figure 3.4. The two upper rows show volume and quote-based inverse market quality measures for the time period 2008–2018. We can observe trends in some of the variables around the roll period. As expected, volume exhibits a notable spike on roll days, which is likely to be the activity of index traders. In contrast, such a clear increase is not visible in any of the market quality variables.

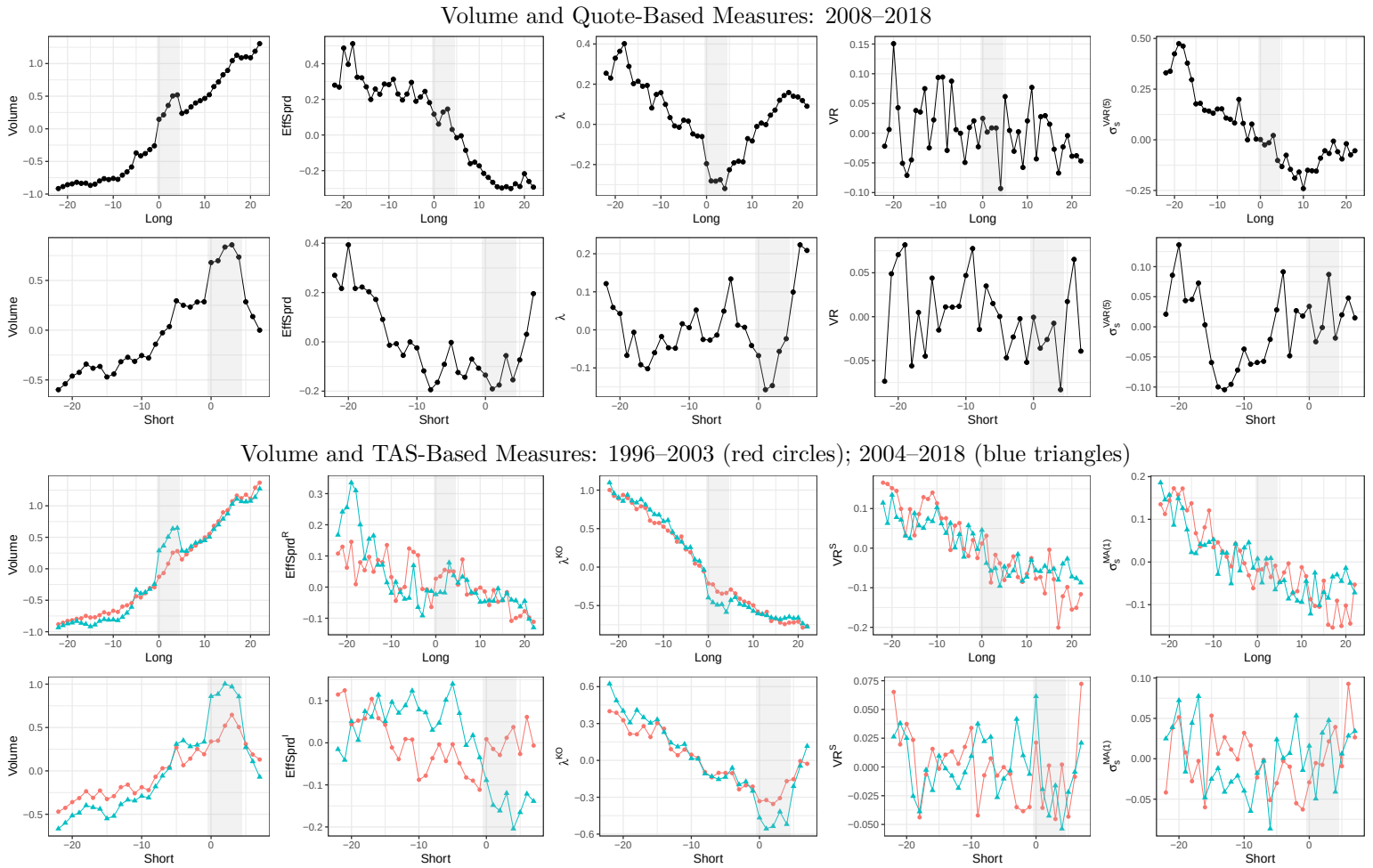

Figure 3.4: **Market Quality During and Around Index Rolls**

This figure shows daily averages of contract-scaled market quality measures around GSCI rolls. $EffSprd^R$, λ^{KO}, VR^S, and $\sigma_s^{MA(1)}$ are proxies for the relative effective spread, price impact, random walk deviations, and pricing error volatility, respectively. We use a sample of 17 commodities that have been constituents of the GSCI from 1996 through 2018. The roll period (0 to 4) is highlighted in gray. We scale each individual contract within the time frame [-22:22] for long and [-22:7] for short trades to zero mean and unit variance.The top (bottom) panels depict the variables corresponding to contracts that are 'bought' ('sold') by index mimicking traders.

Testing the measures in an event study research design as in Bessembinder et al. (2016) would require us to define a pre-event reference period, which might be problematic due to the trending behavior of some variables. Thus, we take a different approach, resorting to TAS-based market quality measures that span the sample period 1996–2018 and split the long and short sample at the start of 2004. The averages of the scaled measures split by regime are depicted in the bottom two rows of Figure 3.4. The overall patterns in TAS-proxies are similar to the measures that use quote data. The spike in volume is visible in post-2003 averages, but not in those of the other half of the sample.

To test if the mean of the scaled variables during the roll period is different for the period before 2004 and after 2003, we use a simple two-sample t-test with standard errors clustered by commodity. Test statistics for a sample of 17 commodities are reported in Panel A of Table 3.7. A high positive (negative) statistic implies that the post-2003 mean is higher (lower) than the pre-2004 mean. In line with the visual inspection in Figure 3.4, the statistics suggest that after 2003 a significantly larger share of volume was concentrated on the GSCI roll days than previously. Effective spreads in short contracts are significantly lower after 2004. The price impact is significantly reduced in both long and short contracts. However, differences in price efficiency measures are all insignificantly different from zero.

As a robustness check, we drop all energy and precious metal commodities from the sample and repeat the analysis with grains, livestock, and softs only. In this sub-sample, the spikes in volume are more pronounced and effects would thus be expected to be stronger. We do this in order to ensure that the differences in VR^S and $\sigma_s^{MA(1)}$ are not diluted by energy and precious metal commodities. Panel B of Table 3.7 shows that the volume hump is significant in both legs, with much higher test-statistics. As in the full sample, most test–statistics suggest that during the roll period, the markets are

Table 3.7: **Volume and Market Quality during the GSCI Roll across Regimes**

This table shows t-test statistics against the null that average volume and market quality during the GSCI roll was the same before and after 2004. We use the [-22:22] (long) or [-22:7] (short) days around the first roll day and scale each individual contract to zero mean and unit variance. Under the null hypothesis, the mean of the standardized measures during days [0:4] is the same for the time-periods 1996–2003 and 2004–2018. Test statistics are calculated using standard errors clustered by commodity. A high positive (negative) test statistic indicates that the post-2003 mean is higher (lower) than before the start of the financialization.

Panel A: Full Sample (17 Commodities)

	Volume	EffSprd^R	λ^{KO}	VR^S	$\sigma_s^{MA(1)}$
Long	4.442	-0.607	-3.406	-1.297	-0.056
Short	3.827	-2.672	-2.842	-0.208	-0.184

Panel B: Without Energy and Precious Metals (10 Commodities)

	Volume	EffSprd^R	λ^{KO}	VR^S	$\sigma_s^{MA(1)}$
Long	15.975	0.585	-3.697	0.944	0.486
Short	9.224	-2.171	-6.019	-1.019	-0.360

more liquid when index traders roll their positions compared to before 2004. This could be seen as evidence for the 'sunshine trading effect' of Admati and Pfleiderer (1991). The price efficiency measures VR^S and $\sigma_s^{MA(1)}$ do not seem to be affected by index roll trades in this sub-sample either. Overall, again, we do not find a harmful effect of index trading activity on commodity market quality.

3.4 Conclusion

Commodity futures markets underwent two substantial changes over the last decades. First, passive long-only index traders became a sizable group that changed the composition of market participants substantially. Second, volume gradually migrated from open-outcry trading pits to electronic limit order books after exchanges introduced side-by-side trading. It is natural to ask how these developments have impacted on market quality. We es-

timate liquidity and intraday informational efficiency measures in a sample that spans from 1996 to 2018 to study their long-term trends. After the start of the financialization in 2004 and during the electronification of commodity markets, both illiquidity and price inefficiency decreased in levels and their trends were negative. To study if and how index traders affect commodity market quality, we combine our market quality measures with different public CFTC reports on aggregate holdings data. We do not find a substantially harmful effect of index investor participation on liquidity and intraday price efficiency. A case study of soybean meal which had seen a sudden shift in index trader open interest affirms this result. We also compare the market quality around the index roll days across regimes. Overall, our results imply that, while electronification had a positive impact, index trading does not seem to be harmful to the quality of commodity futures markets measured via liquidity and intraday price efficiency.

B Appendix

This appendix lists the commodities in our sample in Table B.1, provides details on the employed measures of commodity market quality and their validity, and reports results for speculative trading.

B.1 Measuring Market Quality

In the following, we present the measures and their respective proxies.

Effective Spread (2008–2018) We compute the relative effective spread as of the k^{th} transaction of a day as

$$\text{EffSprd}_k = 2Q_k \, \frac{P_k - M_k}{M_k}, \tag{B.1}$$

where Q is a binary variable that is +1 for a buyer-initiated trade and -1 for a seller-initiated trade flagged using the Lee and Ready (1991) algorithm, P is the trade price, and M is the prevailing mid-quote price. Chakrabarty et al. (2015) show that the Lee–Ready algorithm has a higher accuracy than the tick rule or bulk volume classification in classifying trades. We compute the effective spread for each transaction and take the volume-weighted average to obtain a daily estimate. Since a high effective spread indicates high transaction costs and thus low liquidity, it is an inverse measure of liquidity.

Effective Spread Proxy: Roll's Measure (1996–2018) As a proxy for the effective spread, we estimate the model proposed by Roll (1984). Under the assumption that the fair price follows a random walk and that the bid–ask spread consists of order processing costs alone, he shows that the relative effective spread can be estimated via first-order co-variances of returns. We estimate it on a tick-by-tick frequency for every day in our sample and follow

Table B.1: **Commodity Futures Considered**

This table gives an overview of the commodities we consider in our analysis. NYMEX = New York Mercantile Exchange, ICE = Intercontinental Exchange, CBOT = Chicago Board of Trade, KCBT = Kansas City Board of Trade, COMEX = Commodity Exchange, CME = Chicago Mercantile Exchange. CME, NYMEX, CBOT, KCBT, and COMEX are all part of CME Group.

Sector	Exchange	Commodity	Ticker
Energy			
	NYMEX	WTI Crude Oil	CL
	NYMEX	Heating Oil	HO
	NYMEX	Natural Gas	NG
	ICE (EU)	Natural Gas	NGLNM
	ICE (EU)	Brent Crude Oil	LCO
	ICE (EU)	Gas Oil	LGO
Grains			
	CBOT	Soybeans	S
	CBOT	Corn	C
	CBOT	Wheat	W
	KCBT	Hard Red Winter Wheat	KW
	CBOT	Soybean Meal	SM
	CBOT	Soybean Oil	BO
	CBOT	Rough Rice	RR
	CBOT	Oats	O
Metals			
	COMEX	Gold	GC
	COMEX	Silver	SI
	COMEX	Copper	HG
	NYMEX	Platinum	PL
	NYMEX	Palladium	PA
Softs			
	ICE (US)	Cotton	CT
	ICE (US)	Sugar 11	SB
	ICE (US)	Coffee	KC
	ICE (US)	Cocoa	CC
	ICE (US)	Orange Juice	OJ
	CME	Lumber	LB
Livestock			
	CME	Live Cattle	LC
	CME	Lean Hogs	LH
	CME	Feeder Cattle	FC

Easley et al. (2021) who correct for possible positive autocovariances by taking the absolute value. The Roll measure we employ is given by

$$\text{EffSprd}_{i,d}^{R} = 2\sqrt{|Cov[\Delta p_{i,d,k}, \Delta p_{i,d,k-1}]|}, \tag{B.2}$$

where $p_{i,d,k} = ln(P_{i,d,k})$ is the k^{th} log price of commodity futures i on day d and $\Delta p_k = p_k - p_{k-1}$ are tick-by-tick log returns. The superscript R indicates that the effective spread is the Roll proxy.

Price Impact (2008–2018) We estimate price impact as the slope coefficient (λ) of a linear regression of 1-minute mid-quote log-returns on signed root-dollar-volume order imbalance. We estimate λ using a regression of the form

$$\Delta ln(M_{i,d,t}) = \alpha_{i,d} + \lambda_{i,d}\ sign(OIB_{i,d,t})\sqrt{|OIB_{i,d,t}|} + \epsilon_{i,d,t}, \tag{B.3}$$

where t denotes 1-minute intervals, $\Delta ln(M_{i,d,t})$ is the mid-quote return calculated from the prevailing mid-quote price at the end of the 1-minute interval, and OIB_t is the aggregate signed volume (order imbalance; buyer- minus seller-initiated trades) within the time-interval t measured in USD. Dollar volume is the number of contracts traded times the futures price in USD times the contract size. The theoretical results of Kyle (1985) imply a linear relationship between order imbalance and returns. In empirical studies, however, a concave functional form with signed root-volume is preferred (e.g., Hasbrouck, 2009; Collin-Dufresne and Fos, 2015). A high value indicates a high price-impact, which makes it an inverse measure of liquidity.

Price Impact Proxy: Kyle and Obizhaeva Measure (1996–2018) Kyle and Obizhaeva (2016) suggest a measure of illiquidity that they derive from their market microstructure invariance hypotheses. The measure is

similar to the popular Amihud (2002) measure and given by

$$\lambda_{i,d}^{KO} = \left(\frac{\sigma_{i,d}^2}{V_{i,d}/CPI_d} \right)^{\frac{1}{3}}, \tag{B.4}$$

where σ is the annualized realized volatility of 5-minute trade returns, V is US dollar volume which we deflate by CPI to January 2000 USD, following Fong et al. (2018). Dollar volume is the number of contracts traded times the futures price in USD times the contract size.

In the empirical microstructure literature, tests of market efficiency commonly are against the null hypothesis that a market is (weak-form) efficient. We measure price efficiency using two standard measures: (1) variance ratios and (2) pricing error volatility.

Variance Ratio (2008–2018) Variance ratios as a weak-form price efficiency test against the null hypothesis of a random walk were first proposed by Lo and MacKinlay (1988). We compute a variance ratio of 1-minute to 30-minute mid-quote log-returns. That is, we compute

$$V(30) = \frac{\sigma^2(mr_{30,t})}{30\sigma^2(mr_{1,t})}, \tag{B.5}$$

where $mr_{30,t}$ refers to overlapping mid-quote log-returns aggregated to a 30-minute frequency, while $mr_{1,t}$ are mid-quote returns at a 1-minute frequency. We compute this variance ratio for every day d and every commodity i in the subsample. Deviations from unity are commonly used as a measure for the degree of inefficiency of a price process. Thus, we transform the variance ratio to a scale where a higher value indicates a higher degree of inefficiency by calculating

$$VR_{i,d} = |1 - V(30)_{i,d}|. \tag{B.6}$$

A high VR indicates deviation from a random walk. It is thus an inverse measure of price efficiency.

Variance Ratio Proxy: Microstructure-Adjusted Variance Ratio (1996–2018) For the estimation of variance ratios over the whole sample, we rely on 5-minute end-of-interval trade price returns. During such short time intervals, a large part of the variance comes from trade prices hitting standing ask and buy orders – in limit order markets as well as in pits. To limit the influence of this bid–ask bounce, we compute Smith (1994) variance ratios. His modified variance ratio accounts for bid–ask-induced effects in recorded prices by incorporating the model of Blume and Stambaugh (1983). He provides a closed-form solutions for a bid–ask-adjusted variance ratio. The adjusted variance ratio is

$$V_k^* = \frac{\frac{1}{T}\sum_{t=1}^{T}(ln(P_t) - ln(P_{t-k}) - k\mu)}{km_2} - 1, \tag{B.7}$$

where P_t is a trade price in the 5-minute time interval t,

$$\mu = \frac{1}{T}\sum_{t=1}^{T} ln(P_t) - ln(P_{t-1}), \tag{B.8}$$

$$m_2 = \frac{1}{T}\sum_{t=1}^{T}(ln(P_t) - ln(P_{t-1}) - \mu)^2 - 2\sigma_\delta^2, \tag{B.9}$$

$$\sigma_\delta^2 = \frac{\frac{j}{T}\sum_{t=1}^{T}(ln(P_t) - ln(P_{t-1}) - \mu) - \frac{1}{T}\sum_{t=1}^{T}(ln(P_t) - ln(P_{t-j}) - j\mu)^2}{2j - 2}. \tag{B.10}$$

We set the lag parameters to j = 2 and k = 4. This way, we keep the data requirements low. If the underlying price process follows a random walk, then $V_k^* = 1$. We capture deviations from unity by transforming V_k^* using

$$VR_{i,d}^S = |1 - V_{i,d}^*|. \tag{B.11}$$

The superscript S indicates that VR was calculated from trade prices using the variance ratio by Smith (1994).

Pricing Error Volatility (2008–2018) As a second measure, we estimate the lower bound of the pricing error volatility σ_s using a vector autoregressive process (VAR) of returns and order imbalance (buyer-initiated minus seller-initiated trades). Hasbrouck (1993) suggests decomposing the log-trade price process p_t into a random-walk rw_t and a stationary component s_t

$$p_t = rw_t + s_t. \tag{B.12}$$

We estimate the volatility of s_t , σ_s, using vector-autoregressive regressions with 5 lags, VAR(5), using 1-minute returns. We invert the VAR(5) to a vector moving average (VMA) process and truncate the VMA parameters at 11 lags. After imposing the Beveridge and Nelson (1981) restriction, we obtain a lower bound estimate for σ_s which we denote $\sigma_s^{VAR(5)}$. In the literature, σ_s is used as a measure for price efficiency (e.g., Hendershott and Moulton, 2011; Rösch et al., 2017) but also as a measure of price impact (e.g., Collin-Dufresne and Fos, 2015). Hasbrouck (1993) refers to it as a measure of market quality. In contrast to variance ratios, pricing errors are related to the semi-strong form of price efficiency (Boehmer and Kelley, 2009).

Pricing Error Volatility Proxy (1996–2018) With only trade prices at hand, we estimate a proxy for the lower bound of the pricing error volatility σ_s using a moving average process of order one, MA(1), as proposed by Hasbrouck (1993) using trade prices sampled at a 5-minute frequency. We estimate an MA(1) model of the form

$$r_t = \epsilon_t - a\epsilon_{t-1}. \tag{B.13}$$

After imposing the Beveridge and Nelson (1981) restriction, a lower bound for the pricing error variance can be estimated by

$$\sigma_s^{MA(1)} = \sqrt{a^2\sigma_\epsilon^2}. \tag{B.14}$$

The superscript indicates that σ_s is estimated assuming an MA(1) process. A large σ_s indicates large deviations from the efficient price, making it a measure of inefficiency.

To remove outliers, we first calculate the mean and standard deviation of each measure (for every commodity) after trimming the top and bottom 2.5%. Then, we remove those data points from the sample whose standardized absolute value exceeds 10.

We test for a unit–root using Augmented Dickey-Fuller (Said and Dickey, 1984) tests for each of the eight measures and each commodity. For few time-series, the null cannot be rejected.[12] We do not difference our time-series of measures, because (1) we believe that valuable information would be lost, (2) theoretically, return variances, price impact, and relative spreads can be assumed to be finite which puts the empirical test results into perspective, and (3) to be in line with the majority of the previous literature.

We multiply EffSprd by 10^4, so it can be interpreted in bps relative to the futures price. λ is also multiplied by $10^4 \times \sqrt{10^6}$ such that its unit of measurement is the relative price change in bps induced by a root million dollars of trading volume. VR is the absolute deviation from unity. We annualize volatilities like $\sigma_s^{VAR(5)}$ estimated from, e.g., 1-minute data by computing $\sigma_{annualized} = \sigma_{1min}\sqrt{60 \cdot 24 \cdot 250}$ and use it in %.

To assess the effect sizes on our proxies in the same manner as for our trade-, quote- and volume-based measures (benchmarks), we scale them using

[12] Out of 25 commodities the following number of ADF-tests have $p > 0.05$: EffSprd 6, EffSprd^R 1, λ 2, λ^{KO} 0, VR 0, VR^S 0, $\sigma_s^{VAR(5)}$ 0, and $\sigma_s^{MA(1)}$ 0.

linear regressions commodity-by-commodity for each measure–proxy pair. We estimate regressions of the form

$$\text{Benchmark}_{i,d} = \theta_{0,i} + \theta_{1,i}\text{Proxy}_{i,d} + \epsilon_{i,d} \tag{B.15}$$

for each commodity i with daily proxy and benchmarks estimates using Huber (1964) M-estimation (with $k = 1.345$) which gives less weight to extreme values than simple OLS. Then, we compute a scaled version of each proxy for each commodity by

$$\text{Scaled Proxy}_{i,d} = \hat{\theta}_{0,i} + \hat{\theta}_{1,i}\text{Proxy}_{i,d}, \tag{B.16}$$

where the parameters are the commodity-specific estimates from Equation (B.15). In our analysis, we use the scaled version of each proxy.

To assess how proxies are related to their benchmarks, we compute pairwise correlations for each commodity and average the correlations across all commodities. Panel A of Table B.2 shows the average correlation coefficients. We can observe relatively high correlations for the liquidity measures. The efficiency measures also show considerable degrees of correlation, albeit slightly lower. This is likely due to high levels of noise in those measures.

B.2 Proxy Validity

Before we study commodity market quality, we further study the validity of the proxies we employ, since one might be concerned that the identified proxies might be correlated in the post-financialization sample but not before. A drastic change in the composition of market participants might affect the measurement of our proxies in a way that does not allow us to compare them across the two regimes. To address this concern, we compute the correlation of our measures conditional on the commodity being in an index (GSCI or

Table B.2: **Proxy–Benchmark Correlations**

This table reports (average) correlation coefficients of benchmark measure X and proxy measure Y (labeled X/Y). Panel A shows the results for a sample of 28 commodities (All). In Panel B, we report proxy–benchmark correlations for soybean meal (SM). Before 2013, SM was not part of any major commodity index. From 2013 onwards, it was part of the BCOM index. The top (bottom) row shows correlations from 2008–2012 (2013–2018). In Panel C, we report average proxy–benchmark correlations of all other commodities in our sample. 20 are part of the GSCI or the BCOM index while 7 are non-index commodities (NGLNM, O, OJ, RR, LB, PA, and PL). The sample consists of daily estimates between January 2008 and December 2018.

Panel A: All

EffSprd/EffSprd^R	λ/λ^{KO}	VR/VR^S	$\sigma_s^{VAR(5)}/\sigma_s^{MA(1)}$
0.644	0.756	0.300	0.402

Panel B: Soybean Meal (SM)

Index	EffSprd/EffSprd^R	λ/λ^{KO}	VR/VR^S	$\sigma_s^{VAR(5)}/\sigma_s^{MA(1)}$
Yes	0.480	0.593	0.291	0.291
No	0.813	0.607	0.265	0.265

Panel C: All but SM

Index	EffSprd/EffSprd^R	λ/λ^{KO}	VR/VR^S	$\sigma_s^{VAR(5)}/\sigma_s^{MA(1)}$
Yes	0.642	0.781	0.318	0.381
No	0.635	0.694	0.251	0.464

BCOM) or not. If proxy–benchmark correlations are unaffected by index membership, then this can be interpreted as evidence that our proxies are valid before the financialization started.

First, we study proxy–benchmark correlations of a single commodity across regimes. Soybean meal (SM) was not part of a major index before 2013 but was added to the BCOM index in January 2013. This allows us to compare correlations of proxy–benchmark measures within a commodity. We report the correlations of our proxies and benchmarks in Panel B of Table B.2. Most estimates are in a similar range. Only the correlation of EffSprd and EffSprd^R is higher before SM became an index–commodity. All correlations are significantly different from zero at the 1% level.

In a second robustness check, we compare the correlations of proxy and benchmark measures using a cross-section of commodities that either have

or have not been in a major index throughout the 11-year sample period.[13] We compute correlations for all 27 commodities and report the average correlation for each proxy-benchmark pair and for each group in Panel C of Table B.2. The estimates of index and non-index commodities are of similar size. Individual correlations of proxy–benchmark pairs of all commodities are significantly different from zero at the 1% level (not reported).

Overall, the employed proxies appear to be robust to the 'degree of financialization' of a commodity proxied by index–membership. Thus, we expect these proxies to be able to capture market quality in the pre-financialization regime.

B.3 Speculative Trading and Market Quality

For (un-) predictable net trading activity of speculators, we rely on the same procedure as for CITs. First, we compute relative net open interest of speculators, and then model it for each commodity using an ARIMA(p,1,q) process. Fitted (residual) absolute first differences are then proxies for (un-) predictable net speculative trading volume.

We aim to capture possible effects in a regression of the form

$$MQ_{i,w} = \mu_i + \gamma_w + \theta_1 Pred.V + +\theta_2 Unpred.V + \Theta' X_{i,w} + \epsilon_{i,w}. \qquad \text{(B.17)}$$

$MQ_{i,w}$ is a set of (inverse) measures of market quality of commodity i in week w, i.e., EffSprd^R, λ^{KO}, VR^S, or $\sigma_s^{MA(1)}$. Pred.V (Unpred.V) is (un-) predictable net trading volume of speculators. X is a vector of control variables that are commonly used in empirical market microstructure applications of this kind, i.e., 5-minute intraday volatility, the natural log of dollar volume,

[13] For this analysis we use 20 index commodities, i.e., BO, C, CC, CL, CT, FC, GC, HG, HO, KC, KW, LC, LCO, LGO, LH, NG, S, SB, SI, and W; as well as 7 non-index commodities, i.e., LB, NGLNM, O, OJ, PA, PL, and RR.

the natural log of total (including unclassified) open interest in commodity i, and the inverse settlement price. In order to match frequencies with the COT reports, we aggregate daily measures of market quality and controls to a weekly frequency. μ_i (γ_w) capture commodity (year–week) fixed effects. If predictable (unpredictable) speculative trading or increasing market share of speculators harms market quality, we would expect θ_1 (θ_2) to be significantly positive.

The OLS estimates are reported in Table B.3. The coefficient that relates $Unpred.V$ to λ^{KO} ($\sigma_s^{MA(1)}$) is significantly negative at the 5% (10%) level. A one percentage point change in $Unpred.V$ thus appears to coincide with reduced price impact and smaller variance ratio deviations, although the coefficients are relatively small. However, the contemporaneous relationship could be due to reverse-causality. Speculators might concentrate their trading activity on periods of high market quality. Furthermore, the design of CFTC COT data allows, to some degree, conclusions about the interplay of speculation and commodity market quality. However, they do not allow conclusions about a possible effect of index trading. This is because commodity index traders (CITs) are represented both in the commercial and non-commercial group. Some ETFs hedge their exposure using futures, so they are classified as non-commercials. Other issuers rely on swap dealers for hedging, which in turn are classified as commercials.

Table B.3: **Speculative Trading and Market Quality**

This table shows OLS estimates of the regression

$$MQ_{i,w} = \mu_i + \gamma_w + \theta_1 Pred.V + \theta_2 Unpred.V + \Theta' X_{i,w} + \epsilon_{i,w}.$$

$MQ_{i,w}$ is a set of (inverse) measures of market quality of commodity i in week w, i.e., $EffSprd^R$, λ^{KO}, VR^S, or $\sigma_s^{MA(1)}$. Pred.V (Unpred.V) is (un-) predictable net trading volume by speculators. These are absolute first differences in $ARIMA(p \leq 10, 1, q \leq 10)$ fitted (residual) relative net speculative open interest. X is a vector of control variables, i.e., 5-minute intraday volatility, the natural log of $-volume, the natural log of total reported open interest in commodity i, and the inverse settlement price. μ_i (γ_w) capture commodity (week) fixed effects. The sample includes weekly data of 18 indexed commodities (in GSCI, BCOM or both) for which CFTC COT data are available for a sufficiently long time period. The sample ranges from 1996 to 2018. Standard errors are clustered by week and commodity. t-ratios are in parentheses. ***, **, and * indicate significance at the 1, 5, and 10% level, respectively.

	Dependent variable:			
	EffSprd^R	λ^{KO}	VR^S	$\sigma_s^{MA(1)}$
	(1)	(2)	(3)	(4)
Pred.V	−0.022	0.001	0.00003	−0.002
	(−1.376)	(0.249)	(0.149)	(−0.104)
Unpred.V	−0.040	-0.014^{**}	−0.0004	-0.031^{*}
	(−1.254)	(−2.160)	(−1.626)	(−1.809)
Controls	Yes	Yes	Yes	Yes
Observations	20,353	20,397	20,449	20,452
Adjusted R^2	0.667	0.871	0.579	0.665

Chapter 4

Market Quality, Index Trading and Arbitrage Opportunities in Commodity Markets: High-Frequency Evidence from the Options Market*

4.1 Introduction

Commodity markets underwent drastic changes over the last decades. After 2004, large institutional financial investors like hedge funds, but also passive long-only index investors, entered the markets. During the years after around mid-2006, trading volume migrated from open-outcry to electronic limit order books. In 2008, in the midst of the global financial crisis, commodity prices collectively spiked. Financial investors with no physical interest in the com-

*This chapter is based on the Working Paper "Market Quality, Index Trading and Arbitrage Opportunities in Commodity Markets: High-Frequency Evidence from the Options Market" by Tobias Lauter and Marcel Prokopczuk, 2023.

modities themselves were accused of driving prices beyond their fair values and harming the process of price formation.[1] Consequently, this has been controversially discussed and tested in the literature.[2]

In this paper, we explore the role of ETF-related trading in New York Mercantile Exchange (NYMEX) West Texas Intermediate (WTI) sweet crude oil futures and options. We focus on the information content conveyed in these trades and how they relate to market quality and especially arbitrage opportunities. Besides uninformed commodity index traders (CITs), ETFs could also be used by informed investors attempting to hide (Eglite et al., 2023, find that informed investors trade equity ETFs to conceil insider trading). ETFs are arguably easier to trade: their shares offer a highler granularity than an exposure to the price of 1,000 barrels. Trading ETF shares also does not require opening and maintaining a margin account. Both WTI futures and large ETFs are very liquid with low spreads which in turn allows arbitrageurs to eliminate price differences between ETFs and futures quickly. Since modern electronic markets are fast and the effects of ETF-related trading might be short lived, we measure frictions and mispricings at high frequencies.

To do so, we use a new approach that has, to the best of our knowledge, not been used in the context of commodity financialization: the measurement of market quality using high-frequency options data. In a frictionless complete market, option contracts are redundant assets. This feature greatly aids the detection of those frictions. In contrast to futures contracts, where risk premia arising from hedging pressure and unobservable cost-of-carry variables like storage costs and convenience yield influence futures prices, in options markets, the no-arbitrage relationship is much clearer to analyze. Measuring market quality via options data has the following advantages:

[1] This is also called the "Masters Hypothesis", named after hedge fund manager Michael Masters and his testimony before the US Senate (Masters, 2008; Masters and White, 2008)

[2] For a recent literature overview, see Natoli (2021).

(1) Put–call-parity (PCP) is a simple but powerful no-arbitrage relation that allows us to measure market quality (almost) without having to select a possibly misspecified model for the price process that is compared to another also possibly misspecified model as a benchmark (such as a random walk). Thus, we expect the resulting measures to be more robust and less noisy. Futures prices are tied to spot prices via no-arbitrage in the form of cost-of-carry (theory of storage of Working (1933) and Kaldor (1939)). However, this is relatively difficult to measure, e.g., due to storage costs and the implicit and unobservable convenience yield. Complementarily, the hedging pressure theory of Keynes (1923) gives rise to a risk premium that long speculators demand from short producers. With so many unobservable determinants, deviations from the fair price are difficult to discern in futures markets alone. This is why their efficiency has to be measured via return predictability in the cross-section (see, e.g., (Bakshi et al., 2019)) or time-series (e.g., variance ratios (Lo and MacKinlay, 1988), or pricing errors (Hasbrouck, 1993)). In all these cases, mispricing can only be defined relative to a benchmark model, which is often a linear factor model of the returns in the cross-section or a random walk price process in the time-series. In contrast to spot or futures markets, options have the advantage that their price arises from a clearer no-arbitrage relationship because they are written on the futures price (another financial asset), and not the physical spot price. In a complete market without frictions, they are redundant in the way that their payoffs can be replicated from other financial assets. The most simple replication is using PCP and replicating one option with a portfolio involving another. Alternatively, a single option's payoff can be replicated by trading the underlying (futures contracts) and the risk-free asset. Futures options, the most liquid options contract form for commodities, have the advantage that they are not a function of any payments during the lifetime of the contract. Futures do

not pay dividends, and have no storage costs or convenience yields.[3]

(2) We expect variation in inefficiency and illiquidity of futures contracts to be reflected in an amplified way in option contracts, which should increase the signal-to-noise ratio further. Option market making is more complicated compared to futures markets so that we expect any variation in futures market quality to be more pronounced and thus be measurable with more ease. This means that slight variations in inventory risk or adverse selection risk that might remain undetected in futures-based measures are expected to be amplified and thus more visible in options-based measures. These advantages make commodity futures options an ideal research subject for studies of market frictions that has been overlooked in the debate on commodity financialization. We aim to fill this gap.

(3) The measurement only requires contemporaneous prices, which allows a highly frequent measurement and detection of very short-lived inefficiencies.

We contribute to the literature by studying the role of ETF-related trading activity at the 5-minute frequency. We employ a simple approach to identify ETF-related trading in the order flow of WTI futures. Our option- and bid-ask-based measures are almost model-free and yield high-frequency pictures of frictions and mispricings. This allows a closer look at the role of ETF-related trading that is not possible using daily or weekly data and aggregate volume. First, we sign the volume of large ETFs and WTI futures at the tick frequency and compute 5-minute order imbalances for each. Using a simple linear regression, we decompose futures volume into ETF-related (fitted) and residual imbalances. We compute a PCP-based measure of market quality using the bid and ask prices of futures, calls, and puts to relate it in a panel regression to decomposed imbalances. Our findings suggest that ETF-

[3]One complication arises due to the fact that most commodity futures options are of the American type. However, since no dividends are paid, early exercise is rarely optimal (Back et al., 2013).

related trading coincides with larger PCP deviations than non-ETF-related trading. At the same time, however, it is also less variable, which makes inventory risk explanations less plausible. In order to gauge the information content transmitted through those trades, we relate order imbalances to past and subsequent returns. Slight positive futures returns following ETF-related trading activity hint at private information contained in those trades. Lastly, we relate order imbalances to the occurrence and size of arbitrage opportunities to test if those profits stem from arbitrage trades. Our results imply that this is not the case. From the evidence we conclude that ETF-related trading likely carries private information that gets permanently transmitted into prices, coming at the cost of adverse selection risk for market makers.

We add to a stream of literature that investigates the influence of specific groups on the pricing process in financial markets. In equity markets, for example, both institutional ownership and trading appears to be associated with more efficient pricing (Boehmer and Kelley, 2009) while ETF ownership is associated with less informative prices (Israeli et al., 2017). In the context of commodity markets, the influence of ETF or index investors is highly controversial. Bessembinder et al. (2012) test predatory versus sunshine trading in predictable ETF roll trades and find evidence of the latter. Ready and Ready (2022) on the other hand, find modest but temporary price impacts of index trade flows. They rely on weekly aggregate CFTC trader positions data and issuances of commodity-linked notes. Our approach is more granular, as our entire analysis is performed at the 5-minute frequency.

The main advantages of intraday data over daily data are the following: (1) Mispricings, i.e., arbitrage opportunities, are likely to be short-lived, and an expansion of the sample size by increasing the sampling frequency represents a natural approach to facilitate detection. (2) CME Group in-the-money options are settled by the exchange in accordance with PCP. Thus, measures that rely on PCP are unable to detect mispricings in daily settlement data.

(3) Intraday options data enable us to incorporate the option bid–ask spread into our analysis. (4) Asynchronicity issues are less severe. When either the option or the underlying futures contract is not traded while the other is, stale prices can imply mispricings that are not exploitable (Battalio and Schultz, 2006). The choice of the sampling frequency (5 minutes) is a trade-off between stale prices and dropping observations because they are too far apart. For the PCP-base measures, at least one price update in a call, the complimentary put, and the underlying futures have to occur in a sampling interval of 5 minutes. Finally, price-staleness is also an aspect of market quality. In a liquid and efficient futures and options market, arbitrageurs ensure that prices only remain stale for very short periods of time. In order to exclude sparsely traded options, we only retain at-the-money ($0.95 > \frac{F}{X} > 1.05$, where F is the futures price and X is the option strike) options with a time-to-maturity of 10 days to 100 days.

Our paper is also related to studies of option market liquidity and efficiency. In theory and the empirical literature, the main channels are adverse selection risk, hedging costs, and inventory risk (for single-stock options see, e.g., Christoffersen et al., 2018). Hedging costs are especially important if the option gamma is high which requires frequent rebalancing to achieve delta-neutrality. Thus, higher effective spreads of the underlying directly result in expensive hedging. Overall, equity option market quality appears to be mainly driven by the liquidity of the underlying stock and option market maker capacities. For single-stock options, Muravyev (2016) shows that inventory risk is a major concern for those option market makers. This is especially true in net supply markets in which short-selling is costly. Shorting is very easy in WTI futures markets and thus less relevant in our setting. Our measurement approach is similar to (Cremers and Weinbaum, 2010) who study PCP deviations and subsequent expected returns of the underlying stocks. Their findings point at option mispricings as a consequence of

adverse selection risk. Our design aims at the role of ETF-related traders in commodity markets.

The remainder of the paper is structured as follows. Section 4.2 describes the sample, Section 4.3 the PCP-based measures and a simple regression that we use to decompose order imbalances, Section 4.4 presents our results, and Section 4.5 concludes.

4.2 Data

We download 5-minute intraday futures and options data from Refinitiv Datascope Select (RDS, formerly Thomson Reuters Tick History, TRTH). Our sample spans from January 2008 to October 2021. Quotes and volume are only recorded for electronic trading for both options and futures. Option volume is only reliably available after November 6, 2009, which is why we drop data before 2010 for analyses that include option volume. We download tick trade and quote data (including inside depth) for the first 5 New York Mercantile Exchange Western Texas Intermediate Sweet Crude Oil (NYMEX WTI) futures contracts (Reuters symbol CL) spanning January 2008 to October 2021. We also download tick-by-tick trade and quote data of two ETFs for the time period January 2008 to October 2021. The first, United States Oil fund (USO), offers exposure to WTI.[4] At the end of 2022, it had total net assets of $2.3 Billion. The second, iShares S&P GSCI Commodity-Indexed Trust ETF (GSG) tracks the S&P Goldman Sachs Commodity Index (GSCI), the most popular broad commodity index.[5] At the end of 2022, it had total

[4] `https://www.uscfinvestments.com/uso`. USO was initiated in April 2006. Until April 17, 2020, it held front-month futures only. Nowadays, it holds the first 7 contracts. Since May 2020, it is rolled over the course of 10 days instead of 4. USO managers may trade any energy futures (or option) contract on NYMEX, ICE or other exchanges, but their benchmark is the performance of NYMEX front-month futures. On April 29 2020, USO underwent a 1 for 8 reverse split.

[5] `https://www.ishares.com/us/products/239757/ishares-sp-gsci-commodityindexed-trust-fund`

net assets of $1.3 Billion. With about 20%, WTI has the largest weight in the GSCI. Additionally, we obtain 5-minute trade and quote (plus inside depth) data of American-style at-the-money (ATM; ±5%) NYMEX WTI futures options for the sample period January 2010 to October 2021. For the risk-free rate, we use overnight, 1 week, 1, 2, 3, 6, 12 month LIBOR offered rates by the ICE Benchmark Administration (IBA) downloaded from Refinitiv Datastream and interpolate it using cubic splines.

We focus on the NYMEX WTI oil market since this is the largest and most liquid commodity market. Trading activity has been high during our sample period which ensures that the sample includes sufficiently many 5-minute intervals in which both the option and the underlying futures contract have been traded so that the resulting sample has few gaps.

We filter the futures data as follows. Since volume is still concentrated to the former pit trading hours 9 AM to 2:30 PM, we restrict our sample to these hours. We remove implausible negative prices from the sample.[6] Then, we remove outliers. We define an outlier as a 100%-deviation of the bid ask or trade price from the daily settlement price or the daily median price of all bid, ask, last, and settlement prices of that contract–day. We remove volume entries that exceed the daily volume and additionally truncate it at the top 0.1%. Options data are filtered by eliminating non-positive spreads and one-sided or non-positive quotes.

Some of the market quality measures we employ require the estimation of implied volatilities (IV). To do so, we use the Black (1976) model for calls and the Barone-Adesi and Whaley (1987) model for puts[7], and the secant method for optimization.[8]

[6]That includes the sample period during which WTI futures prices turned negative. We do this, because our market quality measures cannot handle negative prices since they include a transformation of option prices to implied volatilities.

[7]See also Trolle and Schwartz (2009) for a justification for this approach.

[8]The main analysis is performed in R. Due to the size of the data and the complexity of the computations, we write parts of our code in C++ (which can be integrated seamlessly

4.3 Methodology

4.3.1 Measuring Market Quality from Options Data

Our approaches to measuring market quality build on put–call–parity (PCP) for American-style futures options

$$p - EEP + Fe^{-rT} = Xe^{-rT} + c, \tag{4.1}$$

where p (c) is the price of an American-style futures put (call) option, T is their common remaining time to maturity, X is their common strike, r is the T-period-risk-free rate, EEP is the early exercise premium of the put, and F is the futures price. In a frictionless complete market, PCP dictates a no-arbitrage relation that holds irrespective of the process governing the underlying's price. An option pricing model is only required to compute the early-exercise premium because most traded commodity options are American-style, and PCP only holds for European options. The fact that the underlying is a futures contract and thus does not pay dividends makes early exercise rarely optimal (Back et al., 2013). Early exercise depends on the risk-free rate which was stable and low during the sampling period of our analysis. The estimated average daily EEP in our sample is mostly below 0.1 ct and always below 1 ct. In the following, we describe measures all of which build on PCP in Equation (4.1). All measures are driven by both price efficiency and liquidity aspects of the markets and thus capture market quality.

Implied Volatility Deviations

We follow an idea of Amin et al. (2004) that is also used by Cremers and Weinbaum (2010) and Rösch et al. (2017). PCP suggests that the IVs of

in R using the Rcpp package) and run it on parallel on a 32-core machine.

put–call option pairs should be identical. We compute the absolute difference between the IVs of puts and calls with matching underlying, maturity, and strike.

Implied Volatility Deviations using Bid and Ask Prices (IVDBA) With bid–ask quotes, this is

$$IVDBA = |IV_C^{Ask} - IV_P^{Bid}| + |IV_C^{Bid} - IV_P^{Ask}| \tag{4.2}$$

where IV_C^{Ask} (IV_C^{Bid}) is computed from call ask (bid) quotes and futures ask (bid) quotes while IV_{Put}^{Ask} (IV_{Put}^{Bid}) is computed from put ask (bid) quotes and futures bid (ask) quotes.

Synthetic Futures Differences and Implied Arbitrage Profits

With our second measure, we follow Battalio and Schultz (2006) who study the price efficiency of American single-stock options by computing option-implied synthetic prices of the underlying and comparing them to observed prices. We solve Equation (4.1) for F and use observed option prices to compute a synthetic futures price implied by PCP. Then, we compare the synthetic futures price to the observed one and compute absolute differences as a measure of market quality.

In detail, the synthetic underlying futures prices is

$$F_{Syn} = e^{rT}\left(c - (p - EEP)\right) + X \tag{4.3}$$

where c and p are observed American-style call and put futures option prices written on the same underlying with an identical time to maturity T as a fraction of a year and the same strike X and r is the annualized T-period risk-free interest rate. EEP is the early exercise premium. We estimate it by

taking the average IV of all futures–day calls and transform it to put option prices using the result by Black (1976). The EEP is then the difference this estimated European and the observed American put option price.

With bid and ask quotes available, the relationship in Equation (4.3) becomes

$$F_{Syn}^{Ask} = e^{rT}\left(c^{Ask} - \left(p^{Bid} - EEP\right)\right) + X \tag{4.4}$$

$$F_{Syn}^{Bid} = e^{rT}\left(c^{Bid} - \left(p^{Ask} - EEP\right)\right) + X \tag{4.5}$$

for bid and ask prices, respectively (Battalio and Schultz, 2006). There exists an arbitrage opportunity if $F_{Syn}^{Ask} < F^{Bid}$ or $F_{Syn}^{Bid} > F^{Ask}$, i.e., if buying (selling) the underlying synthetically at the ask (bid) while simultaneously selling (buying) it at the bid (ask) yields an instantaneous risk-free profit.

Implied Arbitrage Profit (IAP) From bid-ask prices and inside depth of options and futures, we are then able to compute a rough estimate of arbitrage profits available. In detail, we estimate implied arbitrage profits as

$$IAP = min(size_c^{Ask}, size_p^{Bid}, size_F^{Bid}) \times max(F^{Bid} - F_{Syn}^{Ask}, 0) \tag{4.6}$$

$$+ min(size_c^{Bid}, size_p^{Ask}, size_F^{Ask}) \times max(F_{Syn}^{Bid} - F^{Ask}, 0), \tag{4.7}$$

where $size$ is the bid or ask size (inside depth) of the respective futures or option contract. For actual dollar profits, the profit has to be multiplied by the contract size of the futures. Of course, IAP does not perfectly measure arbitrage profits because we use bid-ask prices and sizes at the end of 5-minute periods and only inside depth which does not allow walking the book. We estimate these measures for every 5-minute interval in our sample. Recall that in a complete frictionless market IVDBA should be zero, so it is an inverse measure of market quality. To limit the impact of extreme values,

we winsorize the top 0.1% of $IVDBA$. Since IAP contains many zeros, we winsorize the top 0.1% of all positive IAP. We also winsorize the top (non-negative variables) or bottom and top 0.1% of explanatory variables that exhibit extreme values.

4.3.2 ETF-Related Futures Order Imbalances

In this section, we describe how we decompose WTI futures order imbalances into ETF-related and residual order imbalances using a simple regression. We use tick-by-tick quotes and trades of WTI futures, GSG, and USO to compute 5-minute signed volume. Futures and ETFs are linked through creation-redemption and (index) arbitrage mechanisms. Both ETFs, USO and GSG, and WTI futures are highly liquid instruments and any profitable arbitrage opportunity can be assumed to be traded away within 5 minutes – the frequency of our sample. Order imbalance in one market transmits through price impact, where we expect similar order imbalances to occur. The conventional functional form of price impact in empirical microstructure that links order imbalance to returns is the square–root (Hasbrouck and Saar, 2013; Collin-Dufresne and Fos, 2015). Thus, we assume a root-root link between ETF and futures order imbalances. The square–root transformation also lowers the influence of very high order imbalances. Lee–Ready-signed volume ($s.Vol$) is aggressive buy minus aggressive sell volume in the number of shares or contracts.[9] We define root order imbalance as $OIB = \pm\sqrt{|s.Vol|} = sign(s.Vol)\sqrt{|s.Vol|}$ and explore the relations of futures and ETF OIB.

We estimate several models whose coefficients are presented in Table 4.1. Model (1) is a logit regression and Model (2) an OLS regression, whose estimates imply that absolute futures OIB is related to trading in futures contracts. When ETFs are aggressively traded in one direction, futures are

[9]Using dollar volume instead would induce a spurious relation since the prices of futures and ETFs tracking them are closely related.

more likely to be traded in that same 5-minute interval and are also more likely to be traded aggressively in one direction. This shows that trading activity in futures and ETFs is generally contemporaneously correlated. In Model (3), we test if and how signed OIB in ETFs and futures are related. The estimates and large t-stats indicate that on average, buy or selling pressure in futures and both ETFs occur simultaneously. This could be due to OIB-induced price pressure transmitting from one market to the other via arbitrage activity, or simply because new information is simultaneously incorporated in all markets. Since the intercept in Model (3) is only significantly different from zero at the 10% level, we drop it in Model (4) and use this specification to decompose futures OIB into fitted (ETF-related) and residual (non-ETF-related) OIB. In this case, zero OIB in ETFs also results in zero ETF-related OIB in futures. ETF-related imbalance accounts for 2.8% of the variation in futures OIB.

We extend Model (3) by adding 12 leads and lags (±60 minutes) of GSG and USO OIB. The sample is restricted to the main trading hours but lagged and leading imbalances are also sampled from before 9 AM and after 2:30 PM. This serves as a test of possible delayed transmissions between markets but also as a robustness check. We present OLS point estimates with 2 times the standard error bars for all 22 slope coefficients in Figure 4.1. The estimates show that the relationship is only significant for the contemporaneous terms. This makes us confident that our simple regression is in fact picking up ETF-related trading activity in futures. By ETF-related trading, we understand trades in ETFs that are subsequently hedged using futures contracts or trades in futures that are subsequently mimicked in ETF shares. Same-information trades in both futures and ETFs is also captured by this approach.

Table 4.1: **Order Imbalances in WTI Futures and ETFs**

*This table shows the relationship between absolute signed volume in large commodity ETFs and WTI futures. GSG is an ETF tracking the S&P GSCI, and USO offers exposure to WTI futures only. Sample: 5-minute frequency, Jan 2008–Dec 2021, 5 shortest-maturity CL futures contracts, 9 AM to 2:30 PM. Standard errors of OLS models are clustered by futures contract. t-stats are in parentheses. ***, **, and * indicate significance at the 1%, 5%, and 10% level, respectively.*

	Dependent variable:			
	$1_{s.Vol_F \neq 0}$	$\sqrt{\lvert s.Vol_F \rvert}$	$\pm\sqrt{\lvert s.Vol_F \rvert}$	
	Logit	OLS	OLS	OLS
	(1)	(2)	(3)	(4)
$\sqrt{\lvert s.Vol_{GSG} \rvert}$	0.058***	0.043***		
	(35.923)	(17.146)		
$\sqrt{\lvert s.Vol_{USO} \rvert}$	0.003***	0.024***		
	(22.895)	(15.497)		
$\pm\sqrt{\lvert s.Vol_{GSG} \rvert}$			0.013***	0.013***
			(9.931)	(9.934)
$\pm\sqrt{\lvert s.Vol_{USO} \rvert}$			0.016***	0.016***
			(18.089)	(18.092)
Constant	3.480***	6.365***	0.038*	
	(528.016)	(36.179)	(1.666)	
Observations	1,118,925	1,118,925	1,118,925	1,118,925
R^2		0.167	0.028	0.028
Adjusted R^2		0.167	0.028	0.028
Log Likelihood	-121,368.000			

4.4 Empirical Results

We study the determinants of IVDBA by conducting a regression analysis. The model includes absolute futures order imbalances as our variable of interest. We estimate two regressions. The second uses absolute order imbalances disaggregated into ETF- and non-ETF-related using the approach described in Section 4.3.2. Theoretically, we expect a positive relationship between $|OIB_F|$ and IVDBA if the futures market is not deep and resilient enough, so price impact leads to increased PCP violations. This might happen if liquidity suppliers perceive the aggressive trader to be informed (e.g., in the sense of Glosten and Milgrom (1985)) so they adjust spreads and the price impact is persistent and does not reverse within the 5-minute interval. We

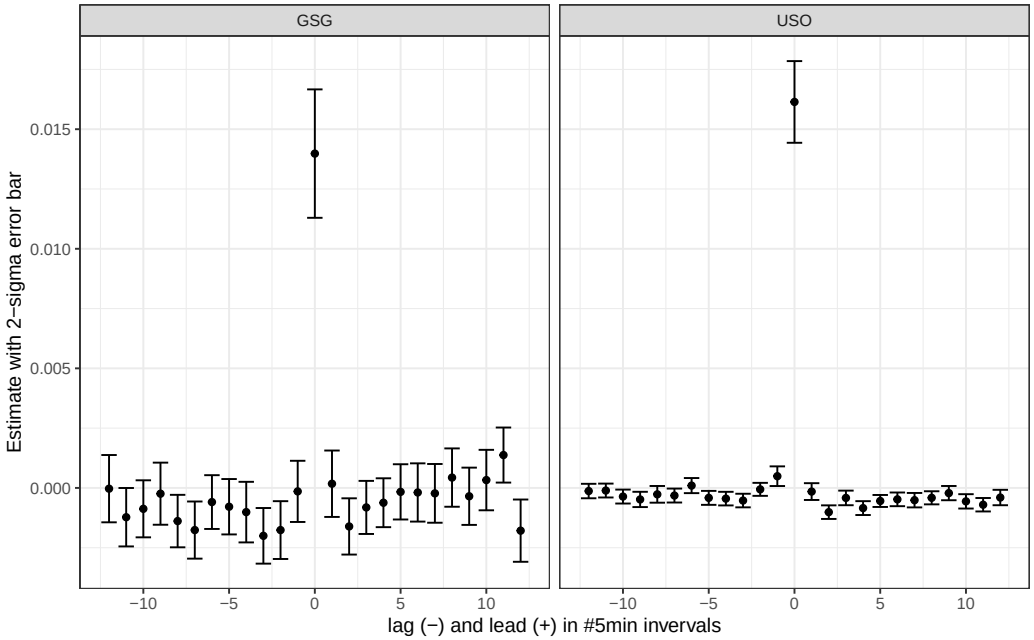

Figure 4.1: **Lead and Lagged Signed ETF Volume**

This figure shows point estimates and error bars indicating 2 standard errors. The model is

$$\left(\pm\sqrt{|s.Vol_F|}_{i,t}\right) = \gamma_0 + \sum_{l=-12}^{12} \theta_l \left(\pm\sqrt{|s.Vol_{GSG}|}_{i,t-l}\right) + \sum_{l=-12}^{12} \phi_l \left(\pm\sqrt{|s.Vol_{USO}|}_{i,t-l}\right).$$

Standard errors are clustered by futures contract.

include further regressors that have been shown to capture option liquidity along option contract and time-of-day fixed effects (FE) to control for unobserved heterogeneity throughout the average day and across option contracts. Those regressors are the following. $\Gamma_C \times \sigma_F$ proxies hedging costs (Engle and Neri, 2010). Γ_C is the Black (1976) option-greek gamma of the call, which is collinear with that of the put. σ_F is the 5-minute standard deviation of futures returns on that day. $\mathcal{V}_C \times BAS_F$ proxies inventory rebalancing costs (Leland, 1980; Boyd et al., 2018; Christoffersen et al., 2018). $\mathcal{V}_C$ is the Black (1976) option-greek vega, which is collinear with that of the put. $-\frac{\Delta_C}{\Delta_P}$ is a ratio of the Black (1976) option-greeks delta which captures initial hedging

costs. Deltas of calls and puts are closely related (correlation of around 0.8 in our sample), so we use a ratio to capture variation in both. Time-to-maturity effects are captured by Γ_C and moneyness by deltas. To further account for potentially temporary unobserved time or option contract effects we follow Petersen (2009) and double cluster the standard errors by option contract and time.

The regression results are presented in Table 4.2. In Model (1), we can see that large order imbalances in the underlying futures coincide with increased PCP deviations. On average, an order imbalance of plus or minus one futures contract is associated with an about 0.8 bps higher total difference in implied volatilites between put and call bid and ask prices. This is to be expected because when large orders walk the book, the spread widens which is reflected in an increase in IVDBA. Also, proxies for hedging costs and inventory rebalancing costs are positively associated with IVDBA. If option market makers have to adjust their hedges frequently and at a higher cost, this leads to inefficient option quotes. The positive coefficient of the delta ratio (put deltas are always non-positive) also implies that increased initial hedging costs result in inefficiencies. Positive coefficients of option volume are also expected, which is why the negative one for puts in Model (1) is surprising.

In Model (2), we use absolute root order imbalances split into ETF-related and non-ETF-related. The estimates suggest that large ETF-related order imbalances are associated with increased IVDBA while the effect is substantially lower for non-ETF-related trading. On average, an ETF-related aggressive net buy or sell order imbalance in futures coincides with a combined 3 bps higher than average put and call bid and ask implied volatility deviation. This result could be driven by ETF-related volume being more directional and thus more frequently walking the book. The sample standard deviations of $|OIB_F(ETF)|$ is 1.77 while that of $|OIB_F(resid)|$ is 7.74 which makes this

Table 4.2: **Implied Volatility Deviations in Bid-Ask Prices**

This table presents OLS estimates of models for market quality in the form of bid-ask implied volatility deviations (IVDBA) in basis points. $|OIB_F|$ *is absolute square-root order imbalance in the underlying futures contract.* (ETF) *and* $(resid)$ *indicate ETF-related and non-ETF-related order imbalance as described in Section 4.3.2.* F*,* C*, and* P *indicate futures, calls, and puts, respectively.* Γ*,* $\mathcal{V}$*,* Δ *are Black (1976) option-greeks.* σ_F *is the 5-minute return standard deviation of the underlying futures during between 9 AM and 2:30 PM of that day.* BAS_F *is quoted log futures bid-ask spread.* $\sqrt{Vol}$ *is square-root of volume. Sample: 5-minute frequency, Jan 2010–Oct 2021,* $\pm 10 - 100$ *day 5% ATM WTI options with their underlying futures and USO and GSG, 9 AM to 2:30 PM. All models include option contract and 5-minute interval fixed effects. Standard errors are clustered by option contract and time. t-stats are in parentheses.* ****,* ***, and* * *indicate significance at the 1%, 5%, and 10% level, respectively.*

	Dependent variable:	
	IVDBA	
	(bps)	(bps)
	(1)	(2)
$\lvert OIB_F \rvert$	0.799^{***}	
	(18.434)	
$\lvert OIB_F(ETF) \rvert$		3.267^{***}
		(9.799)
$\lvert OIB_F(resid) \rvert$		1.028^{***}
		(15.571)
$\Gamma_C \times \sigma_F$	3.851^{***}	4.728^{***}
	(6.276)	(7.502)
$\mathcal{V}_C \times BAS_F$	2.375^{***}	2.212^{***}
	(15.623)	(14.973)
$-\frac{\Delta_C}{\Delta_P}$	35.293^{***}	36.044^{***}
	(28.299)	(28.720)
$\sqrt{Vol_C}$	0.917^{***}	1.438^{***}
	(7.093)	(11.074)
$\sqrt{Vol_P}$	-1.530^{***}	-1.032^{***}
	(-13.203)	(-9.246)
Observations	8,028,646	8,028,646
R^2	0.108	0.107
Adjusted R^2	0.107	0.106

less plausible. Another possible explanation is that ETF-related trading is on average based on private information.

We test this by relating ETF- and non-ETF order imbalance to leading and lagged futures returns. To do so, we regress order imbalances on 5 to 30-minute lagged returns, contemporaneous returns and on 5 to 30-minute leading returns. Note that we do not use absolute but signed root–volume in this specification. We also exclude the May 2020 WTI futures contract from the sample because its prices turned negative. OLS estimates are presented in Table 4.3. Estimates of returns lagged by 10 minutes or more suggest that both ETF and non-ETF trades tend to trade aggressively against past price trends (reversal-style). Positive non-ETF-related trading however appears to coincide with lagged 5-minute short-term price increases (momentum-style). Both order imbalances are positively associated with contemporaneous returns. This could be a result of price impact or faster return-chasing (momentum-style) within below 5 minutes. More interesting are subsequent returns, as they are indicative of possible short-term private information as a motivation for aggressive order submission. ETF-related order imbalances appear to be weakly related to positive future returns. Returns 10, 20, and 25 minutes after a large ETF-related OIB are in the same direction with t-ratios of 1.956, 1.714, and 2.329, respectively. This could be interpreted as evidence for short term private information in ETF-related trading which manifests as permanent price impact. If price impact was large but transitory due to ETF-related volume being uninformed we would expect returns to turn negative after some time, but our estimates are positive or insignificant. We also extend leads and lags to ± 60 minutes and obtain similar results (not tabulated).

Arbitrage Opportunities Positive subsequent returns of ETF-related trading could also stem from the exploitation of arbitrage opportunities. In the

Table 4.3: **Futures Returns and (Non-) ETF Order Imbalances**

This table presents OLS estimates. OIB_F is the square-root order imbalance in a futures contract. (ETF) and $(resid)$ indicate ETF-related and non-ETF-related order imbalance as described in Section 4.3.2. r_t are 5-minute futures mid-quote log-returns. Sample: 5-minute frequency, Jan 2010–Oct 2021, first 5 WTI futures and USO and GSG, 9 AM to 2:30 PM. All models include futures contract and 5-minute interval fixed effects. Standard errors are clustered by futures contract. t-stats are in parentheses. ***, **, and * indicate significance at the 1%, 5%, and 10% level, respectively.

	Dependent variable:	
	$OIB_F(ETF)$	$OIB_F(resid)$
	(1)	(2)
$r_{t-30min}$	-6.905^{***}	-15.322^{***}
	(-3.630)	(-2.812)
$r_{t-25min}$	-7.785^{***}	-21.337^{***}
	(-3.427)	(-3.521)
$r_{t-20min}$	-3.793^{***}	-32.016^{***}
	(-5.311)	(-7.122)
$r_{t-15min}$	-2.916^{**}	-21.122^{***}
	(-2.339)	(-2.777)
$r_{t-10min}$	-3.243^{***}	-17.554^{**}
	(-2.918)	(-2.285)
r_{t-5min}	3.356	81.086^{***}
	(0.998)	(8.129)
r_t	67.316^{***}	775.020^{***}
	(13.015)	(8.763)
r_{t+5min}	2.778	2.916
	(1.373)	(0.265)
$r_{t+10min}$	3.912^{*}	-7.680
	(1.956)	(-1.582)
$r_{t+15min}$	2.238	-1.916
	(1.241)	(-0.631)
$r_{t+20min}$	1.236^{*}	-0.901
	(1.714)	(-0.309)
$r_{t+25min}$	1.982^{**}	-1.850
	(2.329)	(-0.478)
$r_{t+30min}$	0.384	-1.371
	(0.425)	(-0.293)
Observations	1,109,878	1,109,878
R^2	0.014	0.046
Adjusted R^2	0.014	0.046

following, we test this alternative explanation. In our setting, an arbitrage opportunity exists if $F_{Syn}^{Ask} < F^{Bid}$ or $F_{Syn}^{Bid} > F^{Ask}$. Details are described in Section 4.3. We define a dummy variable 1_{Arb} that is 1 if either case occurs and 0 else. In our sample, we record $1,732$ arbitrage opportunities. 23% and 34% of those were in 2010 and 2020, respectively. Moderately many occurred in 2018 (11%) and 2019 (12%) while other years saw few. 575 of them were in low-strike June 2020 option contracts. Excess supply in the entire market and especially at the bottleneck in Cushing distorted prices at the start of the Corona pandemic leading to extremely low prices. Observations with negative futures prices are excluded from our sample because standard option pricing methods break down.

We model the likelihood of an arbitrage opportunity occurring using a logit regression. Again, we use absolute (ETF and non-ETF) order imbalances as our main regressands of interest. The regressions include option contract and time-of-day fixed effects. Regression estimates are presented in Table 4.4. The results suggest that arbitrage opportunities are more likely to occur when absolute order imbalances are large. However, we do not find differences between the effects of ETF-related and non-ETF-related imbalances.

We also investigate the size of the available arbitrage profits conditional on an opportunity being present. To do so, we reduce our sample to those instances with an arbitrage opportunity and model the available implied arbitrage profit. The sample size is thus reduced to $1,732$ observations. Results are presented in Table 4.5. High absolute imbalances are associated with reduced arbitrage profits according to Model (1). On average, a one contract higher order imbalance coincides with 3.4 cents lower arbitrage profits. Thus, as expected, futures traders appear to trade against market inefficiencies. Our previous results have shown that non-ETF-related trading is more short momentum-oriented. Nevertheless, our analysis does not provide any

Table 4.4: **Arbitrage Opportunities**

This table presents logit estimates of arbitrage opportunities. 1_{Arb} *is a dummy that is 1 if there is an arbitrage opportunity after spreads.* OIB_F *is the square-root order imbalance in a futures contract.* (ETF) *and* $(resid)$ *indicate ETF-related and non-ETF-related order imbalance as described in Section 4.3.2.* F*,* C*, and* P *indicate futures, calls, and puts, respectively.* Γ*,* $\mathcal{V}$*,* Δ *are Black (1976) option-greeks.* σ_F *is the 5-minute return standard deviation of the underlying futures during between 9 AM and 2:30 PM of that day.* BAS_F *is quoted log futures bid-ask spread.* $\sqrt{Vol}$ *is square-root of volume. Sample: 5-minute frequency, Jan 2010–Oct 2021, +- 10–100 day 5% ATM WTI options with their underlying first 5 futures and USO and GSG, 9 AM to 2:30 PM. All models include futures contract and 5-minute interval fixed effects. t-stats are in parentheses.* ****,* ***, and* * *indicate significance at the 1%, 5%, and 10% level, respectively.*

	1_{Arb} (1)	1_{Arb} (2)
$\lvert OIB_F \rvert$	0.060^{***}	
	(16.990)	
$\lvert OIB_F(ETF) \rvert$		0.041^{***}
		(3.083)
$\lvert OIB_F(resid) \rvert$		0.051^{***}
		(13.923)
$\Gamma \times \sigma_F$	0.022^{***}	0.022^{***}
	(7.163)	(6.849)
$\mathcal{V}_C \times BAS_F$	-0.026^{***}	-0.027^{***}
	(-6.871)	(-7.177)
$-\frac{\Delta_C}{\Delta_P}$	0.090^{***}	0.092^{***}
	(2.502)	(2.585)
$\sqrt{Vol_C}$	0.071^{***}	0.074^{***}
	(5.049)	(5.355)
$\sqrt{Vol_P}$	0.039^{***}	0.044^{***}
	(2.945)	(3.377)

evidence that either group's activity is associated with higher or lower contemporaneous arbitrage profits.

4.4.1 Discussion

We present evidence that ETF-related trading is coinciding with larger PCP deviations than non-ETF-related trading. Lower variation in the former makes larger directional trades less likely as a causal explanation. Slightly positive subsequent returns hint private information contained in ETF-related

Table 4.5: **Arbitrage Profits**

*This table presents OLS estimates of arbitrage profits conditional on an opportunity being present. IAP is implied arbitrage profit which is the available dollar arbitrage considering spreads and inside depth. OIB_F is the square-root order imbalance in a futures contract. (ETF) and $(resid)$ indicate ETF-related and non-ETF-related order imbalance as described in Section 4.3.2. F, C, and P indicate futures, calls, and puts, respectively. Γ, $\mathcal{V}$, Δ are Black (1976) option-greeks. σ_F is the 5-minute return standard deviation of the underlying futures during between 9 AM and 2:30 PM of that day. BAS_F is quoted log futures bid-ask spread. $\sqrt{Vol}$ is square-root of volume. Sample: 5-minute frequency, Jan 2010–Oct 2021, $\pm 10-100$ day 5% ATM WTI options with their underlying first 5 futures and USO and GSG, 9 AM to 2:30 PM. We exclude all instances without an arbitrage opportunity being present. All models include option contract and 5-minute interval fixed effects and standard errors are clustered by option contract and time. t-stats are in parentheses. ***, **, and * indicate significance at the 1%, 5%, and 10% level, respectively.*

	Dependent variable:	
	IAP	
	(USD)	(USD)
	(1)	(2)
$\lvert OIB_F \rvert$	-0.034^{**}	
	(-2.130)	
$\lvert OIB_F(ETF) \rvert$		-0.006
		(-0.126)
$\lvert OIB_F(resid) \rvert$		-0.012
		(-0.428)
$\Gamma_C \times \sigma_F$	-0.0003	-0.0003
	(-0.257)	(-0.250)
$\mathcal{V}_C \times BAS_F$	0.029	0.031
	(1.270)	(1.366)
$-\frac{\Delta_C}{\Delta_P}$	-0.010	-0.007
	(-0.043)	(-0.030)
$\sqrt{Vol_C}$	-0.088	-0.098
	(-0.995)	(-1.075)
$\sqrt{Vol_P}$	0.061	0.057
	(0.584)	(0.521)
Observations	1,732	1,732
R^2	0.697	0.696
Adjusted R^2	0.447	0.446

trades that is permanently incorporated into futures prices. Our analysis of arbitrage opportunities and conditional arbitrage profits suggests that ex-

ploitable arbitrage opportunities are more likely to emerge when imbalances occur but show no differences between the two. In line with the findings of Ready and Ready (2022), we find that ETF-related trading appears to coincide with price impact. This price impact is likely permanent on average. Besides moving prices, spread adjustments due to adverse selection risk of market makers is another plausible channel. Our analysis of arbitrage opportunities suggests that the average short-term profits of ETF-related trading does not stem from put–call–futures arbitrage activity. Among the determinants of efficient option prices, i.e., order processing costs, hedging costs (Engle and Neri, 2010), inventory risk (Muravyev, 2016), and adverse selection risk, our results highlight the importance of the latter one. Another possible channel for higher PCP deviations of informed ETF trades is that market makers learn about the share of informed traders and widen spreads after observing one-directional imbalances (Aliyev et al., 2022). Due to their lower liquidity compared to their underlying, option prices only appear to react to large and more permanent changes in the prices of the underlying. In the context of commodity financialization, our results are opposite to the notion of the Master's Hypothesis that ETF-related trading is uninformed.

4.5 Conclusion

In this paper, we study the impact of ETF-related trading on put–call-parity deviations in WTI futures markets. Our estimates suggest that it is associated with higher inefficiencies in the futures–call–put triangle than other trading activity but slightly positive subsequent expected returns imply that it is rather informed volume. Overall, our results hint at informed trading of financial investors improving the informativeness of prices in modern commodity markets while coming at the cost of adverse selection risk to market makers.

C Appendix

C.1 Robustness: Correlations of Order Imbalances with Futures Returns

As a robustness check to the results in Section 4.3, we compute simple unconditional pairwise correlations of leading and lagged returns with (ETF-related) order imbalance. The estimates (first two columns) and their respective p-values (last two columns) are presented in Table C.1. The results are similar. ETF-related order imbalance is positive (negatively) correlated with future (past) returns.

Table C.1: **Correlations of Leading/Lagged Returns with (ETF-related) Trading**

This table presents correlation estimates in the first two and their corresponding p-values in the second two columns. OIB_F is the square-root order imbalance in a futures contract. (ETF) and $(resid)$ indicate ETF-related and non-ETF-related order imbalance as described in Section 4.3.2. r_t are 5-minute futures mid-quote log-returns. Sample: 5-minute frequency, Jan 2010–Oct 2021, first 5 WTI futures and USO and GSG, 9 AM to 2:30 PM.

	$OIB_F(ETF)$	$OIB_F(resid)$	$p(OIB_F(ETF))$	$p(OIB_F(resid))$
$r_{t-30min}$	-0.0109	-0.0035	0.0000	0.0002
$r_{t-25min}$	-0.0131	-0.0065	0.0000	0.0000
$r_{t-20min}$	-0.0060	-0.0080	0.0000	0.0000
$r_{t-15min}$	-0.0041	-0.0038	0.0000	0.0001
$r_{t-10min}$	-0.0059	-0.0054	0.0000	0.0000
r_{t-5min}	0.0026	0.0173	0.0058	0.0000
r_t	0.1084	0.2109	0.0000	0.0000
r_{t+5min}	0.0016	-0.0043	0.0881	0.0000
$r_{t+10min}$	0.0058	-0.0023	0.0000	0.0163
$r_{t+15min}$	0.0040	0.0013	0.0000	0.1823
$r_{t+20min}$	0.0018	-0.0001	0.0560	0.9502
$r_{t+25min}$	0.0025	-0.0014	0.0081	0.1544
$r_{t+30min}$	0.0005	-0.0000	0.5805	0.9635

C.2 Average Implied Volatility Deviations over Time

In order to compare the evolution of commodity market quality over time with our results from Chapter 3, we compute absolute implied volatility deviations

$$IVD = |IV_C - IV_P|. \tag{C.1}$$

That is, the difference in implied volatilities computed from futures and options trade prices observed at the end of each 5-minute interval in which all assets were traded at least once. We depict daily averages of the natural logarithm of IVD in Figure C.1. In line with the findings from Chapter 3, market quality has improved during the financialization and the electronification.

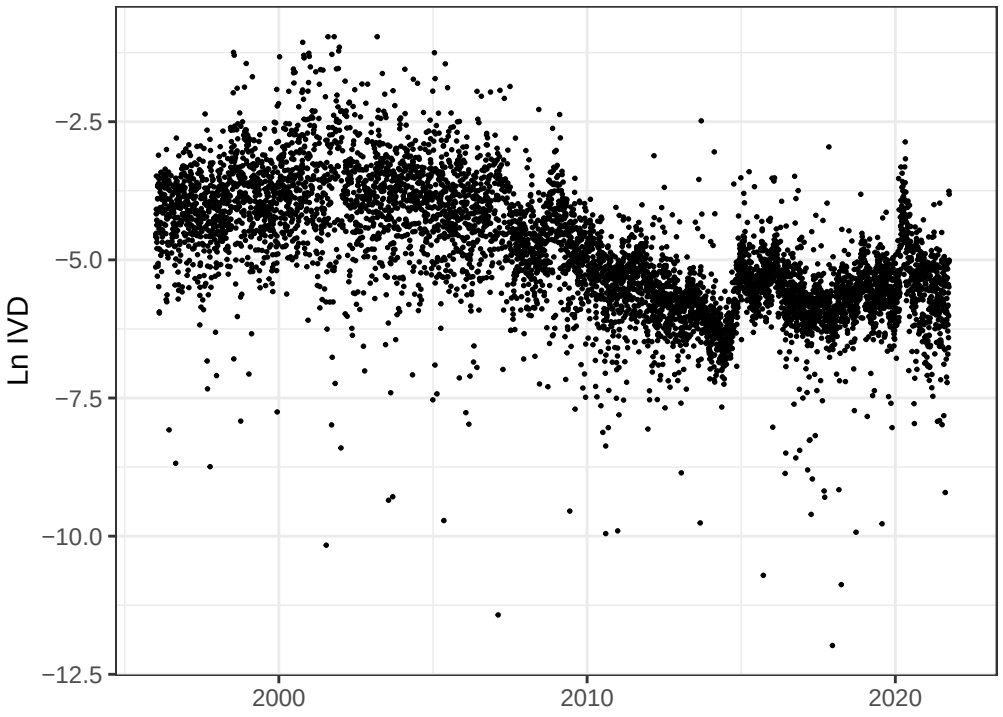

Figure C.1: **Implied Volatility Deviations over Time**

This figure shows daily averages of the natural logarithm of implied volatility deviations computed from trade prices.

Chapter 5

Conclusions and Further Research

5.1 Summary and Conclusion

This thesis investigates the measurement and drivers of commodity market quality. Chapter 2 conducts a horserace to identify the best approach to measuring commodity futures liquidity and price efficiency. It provides guidance for researchers facing the decision, which proxy measure and which sampling frequency to employ. We find that volatility-over-volume measures (Kyle and Obizhaeva, 2016; Fong et al., 2018) are best at capturing (aggregate) time-series and cross-sectional variation in liquidity and also being superior to the commonly used Amihud (2002) ratio. While proxies computed from daily data exhibit high correlations with both bid–ask spread and price impact benchmarks, price efficiency requires intraday data in order to reduce the noise to an acceptable level.

Chapter 3 studies commodity market quality before and after both the start of the financialization and the introduction of side-by-side open-outcry and electronic trading. First, we study its evolution across regimes and then

relate it to both predictable and unpredictable index trader activity by using aggregate open interest data, an exogenous shift of index open interest in soybean meal, and predictable roll trades. We find that commodity market quality increased after the start of the financialization and continued to improve when volume shifted to electronic markets. Neither predictable nor unpredictable index trading appear to be harmful, but coincide with improved market quality.

Chapter 4 studies inefficiencies in West Texas Intermediate sweet crude oil options and futures to unveil the role of ETF-related trading in commodity markets. Using put–call–parity allows us to obtain almost model-free high-frequency measures of market quality. We find that implied volatility differences are increased when ETF-related absolute order imbalances are high. Our subsequent analysis for drivers of this effect hints at adverse selection risk being the main channel. Inventory risk and arbitrage activity seem to be less plausible explanations.

The results presented in this thesis have implications for academics, market participants and regulators. The liquidity of commodity derivatives is important for the producing sector to hedge price risks in the desired quantity at any time. Efficient pricing of benchmarks for goods that make up a substantial part of the (poorer) peoples' consumption basket is of central interest to regulators. This thesis provides insights into measurement, evolution, and drivers of commodity market quality. Overall, our findings highlight the beneficial impact of increased competition through easier market access and transparency of the trading process. The overall quality of commodity markets appears to benefit from them being more integrated into financial markets with players being active across markets and having a variety of investment vehicles available including ETFs which can be used not only for passive portfolio diversification but also informed speculation. Extreme commodity prices are thus likely driven by physical demand and supply shocks

and therefore present in fundamental prices. If regulators aim at limiting these, they should focus on diversifying supply chains, encouraging reasonable levels of inventory, and fostering more sustainable production especially in farming.

5.2 Suggestions for Further Research

Related to the measurement and analysis of market quality of commodity markets (in the context of commodity financialization) but also other asset classes, the following questions and avenues for future research emerge.

During the measurement of price efficiency of commodity derivatives but also of other asset classes, the sampling frequency is a decision made by the researcher. However, it is not arbitrary and its influence would be worth exploring. At what frequency are asset prices unpredictable? Chordia et al. (2005) perform an empirical study of equity markets and find a lag between 5 and 60 minutes, but their latest data are now almost 20 years old. Instead of measuring return predictability at a certain frequency the way it is standard in modern empirical market microstructure, it would be more direct to measure the speed at which information incorporated in prices (price adjustment). Hillmer and Yu (1979) and Damodaran (1993) propose such measures, but they are based on returns alone. A new measure could include not only returns, but also lagged order imbalances. In Chapter 2, a combination of proxies did not prove to be useful. In equity markets, cross-sectional differences are likely to be more pronounced, which could make proxy combinations more effective.

Rösch et al. (2017) find a systematic price efficiency component in equity markets using principal component analysis. It appears to be correlated with funding liquidity (proxied by the TED spread), algorithmic trading activity, and hedge fund size. Their analysis could be extended to other asset classes

including commodities. The extent to which a systematic price efficiency factor exists for commodities and its relation to those of other markets could be interesting to analyze in order to understand the degree of integration of commodity markets into financial markets.

Analyzing the activity of commodity index traders in detail at a high frequency is hindered by the aggregation level of positions data that are made public by the U.S. Commodities Futures Trading Commission (CFTC). The non-public Large Trader Reporting System (LTRS) contains daily positions data disaggregated by contract maturity. For intraday microstructure analyses, tools are required to infer the trading activity of financial traders and especially index traders. A combination of tick-level ETF trades (as used in Chapter 4) and (LTRS) CFTC data might result in higher-quality proxies of index-related trading activity. Ideally, future work on the impact of financial (index) traders on commodity prices should source data from exchanges like the Commodity Mercantile Exchange (CME) or the Intercontinental Exchange (ICE) that include both aggressor flags and trader IDs, which allow the researcher to assign signed volume to particular groups.

Generally, the microstructure of commodity markets has many open questions. For example, what is the role of liquidity in the pricing of futures along the futures curve or option moneyness? Or, how do market makers handle inventory risk when the assets have a limited life (in contrast to stocks where inventory can persist for months (Hasbrouck and Sofianos, 1993; Subrahmanyam, 2008))? Muravyev (2016), for example, suggests using multiple exchanges to separate asymmetric information risk from inventory risk in trading the same asset. WTI futures are traded on CME and ICE and could be studied using his methodology. What is the role of ETFs and options in price discovery? Further research is also required that explores how the liquidity provision of hedgers, speculators and other investors is linked to time-varying risk premia or funding liquidity constraints.

Bibliography

Abdi, F., and A. Ranaldo, 2017, A simple estimation of bid–ask spreads from daily close, high, and low prices, *Review of Financial Studies* 30, 4437–4480.

Acharya, V. V., L. A. Lochstoer, and T. Ramadorai, 2013, Limits to arbitrage and hedging: Evidence from commodity markets, *Journal of Financial Economics* 109, 441–465.

Acharya, V. V., and L. H. Pedersen, 2005, Asset pricing with liquidity risk, *Journal of Financial Economics* 77, 375–410.

Admati, A. R., and P. Pfleiderer, 1991, Sunshine trading and financial market equilibrium, *Review of Financial Studies* 4, 443–481.

Akerlof, G. A., 1970, The market for "Lemons": Quality uncertainty and the market mechanism, Technical Report 3.

Aliyev, N., X.-Z. He, and T. J. Putnins, 2022, Learning about adverse selection in markets, Working Paper.

Amihud, Y., 2002, Illiquidity and stock returns: Cross-section and time-series effects, *Journal of Financial Markets* 5, 31–56.

Amin, K., J. D. Coval, and H. N. Seyhun, 2004, Index option prices and stock market momentum, *Journal of Business* 77, 835–873.

Back, J., M. Prokopczuk, and M. Rudolf, 2013, Seasonality and the valuation of commodity options, *Journal of Banking & Finance* 37, 273–290.

Bagehot, W., 1971, The only game in town, *Financial Analysts Journal* 27, 12–14.

Bakshi, G., X. Gao, and A. G. Rossi, 2019, Understanding the sources of risk underlying the cross section of commodity returns, *Management Science* 65, 619–641.

Barone-Adesi, G., and R. E. Whaley, 1987, Efficient analytic approximation of American option values, *Journal of Finance* 42, 301–320.

Bates, J. M., and C. W. J. Granger, 1969, The combination of forecasts, *Journal of the Operational Research Society* 20, 451–468.

Battalio, R., and P. Schultz, 2006, Options and the bubble, *Journal of Finance* 61, 2071–2102.

Bessembinder, H., A. Carrion, L. Tuttle, and K. Venkataraman, 2012, Predatory or sunshine trading? Evidence from crude oil ETF rolls, Working Paper.

Bessembinder, H., A. Carrion, L. Tuttle, and K. Venkataraman, 2016, Liquidity, resiliency and market quality around predictable trades: Theory and evidence, *Journal of Financial Economics* 121, 142–166.

Beveridge, S., and C. R. Nelson, 1981, A new approach to the decomposition of economic time series into permanent and transitory components with particular attention to the measurement of the 'business cycle', *Journal of Monetary Economics* 7, 151–174.

Black, F., 1976, The pricing of commodity contracts, *Journal of Financial Economics* 3, 167–179.

Blume, M. E., and R. F. Stambaugh, 1983, Biases in computed returns: An application to the size effect, *Journal of Financial Economics* 12, 387–404.

Boehmer, E., K. Fong, and J. Wu, 2021, Algorithmic trading and market quality: International evidence, *Journal of Financial and Quantitative Analysis* 56, 2659–2688.

Boehmer, E., and E. K. Kelley, 2009, Institutional investors and the informational efficiency of prices, *Review of Financial Studies* 22, 3563–3594.

Bohl, M., A. Pütz, and C. Sulewski, 2021, Speculation and the informational efficiency of commodity futures markets, *Journal of Commodity Markets* 23.

Boyd, N. E., J. H. Harris, and B. Li, 2018, An update on speculation and financialization in commodity markets, *Journal of Commodity Markets* 10, 91–104.

Breiman, L., 2001, Random forests, *Machine Learning* 45, 5–32.

Brennan, M. J., and A. Subrahmanyam, 1996, Market microstructure and asset pricing: On the compensation for illiquidity in stock returns, *Journal of Financial Economics* 41, 441–464.

Brogaard, J., M. C. Ringgenberg, and D. Sovich, 2019, The economic impact of index investing, *Review of Financial Studies* 32, 3461–3499.

Brunnermeier, M. K., and L. H. Pedersen, 2005, Predatory trading, *Journal of Finance* 60, 1825–1864.

Chakrabarty, B., R. Pascual, and A. Shkilko, 2015, Evaluating trade classification algorithms: Bulk volume classification versus the tick rule and the Lee–Ready algorithm, *Journal of Financial Markets* 25, 52–79.

Chen, Y. L., and Y. K. Chang, 2015, Investor structure and the informational efficiency of commodity futures prices, *International Review of Financial Analysis* 42, 358–367.

Cheng, I. H., and W. Xiong, 2014, Financialization of commodity markets, *Annual Review of Financial Economics* 6, 419–441.

Chordia, T., R. Roll, and A. Subrahmanyam, 2000, Commonality in liquidity, *Journal of Financial Economics* 56, 3–28.

Chordia, T., R. Roll, and A. Subrahmanyam, 2005, Evidence on the speed of convergence to market efficiency, *Journal of Financial Economics* 76, 271–292.

Chordia, T., R. Roll, and A. Subrahmanyam, 2008, Liquidity and market efficiency, *Journal of Financial Economics* 87, 249–268.

Christoffersen, P., R. Goyenko, K. Jacobs, and M. Karoui, 2018, Illiquidity premia in the equity options market, *Review of Financial Studies* 31, 811–851.

Clemen, R. T., 1989, Combining forecasts: A review and annotated bibliography, *International Journal of Forecasting* 5, 559–583.

Cleveland, W. S., E. Grosse, and W. M. Shyu, 1993, Local regression models, in T. J. Hastie, ed., *Statistical Methods in S*, 309–376 (Chapman & Hall, New York).

Collin-Dufresne, P., and V. Fos, 2015, Do prices reveal the presence of informed trading?, *Journal of Finance* 70, 1555–1582.

Corwin, S., and P. Schultz, 2012, A simple way to estimate bid-ask spreads from daily high and low prices, *Journal of Finance* 67, 719–760.

Cremers, M., and D. Weinbaum, 2010, Deviations from put-call parity and stock return predictability, *Journal of Financial and Quantitative Analysis* 45, 335–367.

Damodaran, A., 1993, A simple measure of price adjustment coefficients, *The Journal of Finance* 48, 387–400.

Daskalaki, C., A. Kostakis, and G. Skiadopoulos, 2014, Are there common factors in individual commodity futures returns?, *Journal of Banking & Finance* 40, 346–363.

Daskalaki, C., and G. Skiadopoulos, 2011, Should investors include commodities in their portfolios after all? New evidence, *Journal of Banking & Finance* 35, 2606–2626.

Daskalaki, C., and G. Skiadopoulos, 2016, The effects of margin changes on commodity futures markets, *Journal of Financial Stability* 22, 129–152.

Daskalaki, C., G. Skiadopoulos, and N. Topaloglou, 2017, Diversification benefits of commodities: A stochastic dominance efficiency approach, *Journal of Empirical Finance* 44, 250–269.

Easley, D., M. L. de Prado, M. O'Hara, and Z. Zhang, 2021, Microstructure in the machine age, *Review of Financial Studies* 34, 3316–3363.

Eglite, E., D. Staermans, V. Patel, and T. J. Putnins, 2023, Using ETFs to conceal insider trading, Working Paper.

Engle, R., and B. Neri, 2010, The impact of hedging costs on the bid and ask spread in the options market, Working Paper.

Fama, E. F., and J. D. MacBeth, 1973, Risk, return, and equilibrium: empirical tests, *Journal of Political Economy* 81, 607–636.

Fernandez-Perez, A., A. M. Fuertes, and J. Miffre, 2019, A comprehensive appraisal of style-integration methods, *Journal of Banking & Finance* 105, 134–150.

Fong, K. Y. L., C. W. Holden, and O. Tobek, 2018, Are volatility over volume liquidity proxies useful for global or US research?, Working Paper.

Fong, K. Y., C. W. Holden, and C. A. Trzcinka, 2017, What are the best liquidity proxies for global research?, *Review of Finance* 21, 1355–1401.

Glosten, L. R., and P. R. Milgrom, 1985, Bid, ask and transaction prices in a specialist market with heterogeneously informed traders, *Journal of Financial Economics* 14, 71–100.

Goldstein, I., and L. Yang, 2022, Commodity financialization and information transmission, *The Journal of Finance* 77, 2613–2667.

Goyenko, R. Y., C. W. Holden, and C. A. Trzcinka, 2009, Do liquidity measures measure liquidity?, *Journal of Financial Economics* 92, 153–181.

Haase, M., Y. Seiler Zimmermann, and H. Zimmermann, 2016, The impact of speculation on commodity futures markets–A review of the findings of 100 empirical studies, *Journal of Commodity Markets* 3, 1–15.

Hasbrouck, J., 1993, Assessing the quality of a security market: A new approach to transaction-cost measurement, *Review of Financial Studies* 6, 191–212.

Hasbrouck, J., 2004, Liquidity in the futures pits: inferring market dynamics from incomplete data, *Journal of Financial and Quantitative Analysis* 39, 305–326.

Hasbrouck, J., 2009, Trading costs and returns for U.S. equities: Estimating effective costs from daily data, *Journal of Finance* 64, 1445–1477.

Hasbrouck, J., and G. Saar, 2013, Low-latency trading, *Journal of Financial Markets* 16, 646–679.

Hasbrouck, J., and G. Sofianos, 1993, The trades of market makers: An empirical analysis of NYSE specialists, *The Journal of Finance* 48, 1565–1593.

Hendershott, T., C. M. Jones, and A. J. Menkveld, 2011, Does algorithmic trading improve liquidity?, *Journal of Finance* 66, 1–33.

Hendershott, T., and P. C. Moulton, 2011, Automation, speed, and stock market quality: The NYSE's hybrid, *Journal of Financial Markets* 14, 568–604.

Henderson, B. J., N. D. Pearson, and L. Wang, 2015, New evidence on the financialization of commodity markets, *Review of Financial Studies* 28, 1285–1311.

Hillmer, S. C., and P. L. Yu, 1979, The market speed of adjustment to new information, *Journal of Financial Economics* 7, 321–345.

Holden, C. W., 2009, New low-frequency spread measures, *Journal of Financial Markets* 12, 778–813.

Hsiao, C., and S. K. Wan, 2014, Is there an optimal forecast combination?, *Journal of Econometrics* 178, 294–309.

Hu, Z., S. Teresa, and P. Garcia, 2020, Algorithmic quoting, trading, and market quality in agricultural commodity futures markets, *Applied Economics* 52, 6277–6291.

Huber, P. J., 1964, Robust estimation of a location parameter, *Annals of Mathematical Statistics* 73–101.

Hyndman, R. J., and Y. Khandakar, 2008, Automatic time series forecasting: The forecast package for R, *Journal of Statistical Software* 27, 1–22.

Irwin, S. H., and D. R. Sanders, 2012, Testing the masters hypothesis in commodity futures markets, *Energy Economics* 34, 256–269.

Israeli, D., C. M. Lee, and S. A. Sridharan, 2017, Is there a dark side to exchange traded funds? An information perspective, *Review of Accounting Studies* 22, 1048–1083.

Kaldor, N., 1939, Speculation and economic stability, *Review of Economic Studies* 7, 1–27.

Keynes, J. M., 1923, Some aspects of commodity markets, *Manchester Guardian Commerical: European Reconstruction Series* 13, 784–786.

Kim, A., 2015, Does futures speculation destabilize commodity markets?, *Journal of Futures Markets* 35, 696–714.

Kyle, A. S., 1985, Continuous auctions and insider trading, *Econometrica* 53, 1315.

Kyle, A. S., and A. A. Obizhaeva, 2016, Market microstructure invariance: Empirical hypotheses, *Econometrica* 84, 1345–1404.

Lauter, T., and M. Prokopczuk, 2022, Measuring commodity market quality, *Journal of Banking & Finance* 145, 106658.

Lee, C. M. C., and M. J. Ready, 1991, Inferring trade direction from intraday data, *Journal of Finance* 46, 733–746.

Leland, H. E., 1980, Who should buy portfolio insurance?, *The Journal of Finance* 35, 581–594.

Lesmond, D. A., J. P. Ogden, and C. A. Trzcinka, 1999, A new estimate of transaction costs, *Review of Financial Studies* 12, 1113–1141.

Lo, A. W., and A. C. MacKinlay, 1988, Stock market prices do not follow random walks: Evidence from a simple specification test, *Review of Financial Studies* 1, 41–66.

Lou, X., and T. Shu, 2017, Price Impact or trading volume: Why is the Amihud (2002) measure priced?, *Review of Financial Studies* 30, 4481–4520.

Marshall, B. R., N. H. Nguyen, and N. Visaltanachoti, 2012, Commodity liquidity measurement and transaction costs, *Review of Financial Studies* 25, 599–638.

Marshall, B. R., N. H. Nguyen, and N. Visaltanachoti, 2013, Liquidity commonality in commodities, *Journal of Banking and Finance* 37, 11–20.

Martinez, V., P. Gupta, Y. Tse, and J. Kittiakarasakun, 2011, Electronic versus open outcry trading in agricultural commodities futures markets, *Review of Financial Economics* 20, 28–36.

Masters, M., 2008, Testimony before the committee on agriculture, nutrition and forestry. United States Senate.

Masters, M. W., and A. K. White, 2008, The accidental Hunt brothers: How institutional investors are driving up food and energy prices, Technical report.

Mincer, J. A., and V. Zarnowitz, 1969, The evaluation of economic forecasts, in Jacob A Mincer, ed., *Economic forecasts and expectations: Analysis of forecasting behavior and performance*, 14–20 (NBER, New York, NY).

Mou, Y., 2011, Limits to arbitrage and commodity index investment: Front-running the Goldman roll, Working Paper.

Muravyev, D., 2016, Order flow and expected option returns, *Journal of Finance* 71, 673–708.

Natoli, F., 2021, Analyzing the structural transformation of commodity markets: Financialization revisited, *Journal of Economic Surveys* 35, 488–511.

Newbold, P., and C. W. J. Granger, 1974, Experience with forecasting univariate time series and the combination of forecasts, *Journal of the Royal Statistical Society. Series A (General)* 137, 131.

Newey, W. K., and K. D. West, 1987, A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix, *Econometrica* 55, 703–708.

Newey, W. K., and K. D. West, 1994, Automatic lag selection in covariance matrix estimation, *Review of Economic Studies* 61, 631–653.

O'Hara, M., and M. Ye, 2011, Is market fragmentation harming market quality?, *Journal of Financial Economics* 100, 459–474.

Parkinson, M., 1980, The extreme value method for estimating the variance of the rate of return, *Journal of Business* 53, 61.

Pástor, L., and R. F. Stambaugh, 2003, Liquidity risk and expected stock returns, *Journal of Political Economy* 111, 642–685.

Petersen, M. A., 2009, Estimating standard errors in finance panel data sets: Comparing approaches, *Review of Financial Studies* 22, 435–480.

Raman, V., M. A. Robe, and P. K. Yadav, 2020, The third dimension of financialization: Intraday institutional financial traders and commodity market quality, Working Paper, CFTC, January.

Ready, M., and R. Ready, 2022, Order flows and financial investor impacts in commodity futures markets, *Review of Financial Studies* .

Roll, R., 1984, A simple implicit measure of the effective bid–ask spread in an efficient market, *Journal of Finance* 39, 1127–1139.

Rösch, D. M., A. Subrahmanyam, and M. A. Van Dijk, 2017, The dynamics of market efficiency, *Review of Financial Studies* 30, 1151–1187.

Sadka, R., 2006, Momentum and post-earnings-announcement drift anomalies: The role of liquidity risk, *Journal of Financial Economics* 80, 309–349.

Said, S. E., and D. A. Dickey, 1984, Testing for unit roots in autoregressive-moving average models of unknown order, *Biometrika* 71, 599–607.

Schestag, R., P. Schuster, and M. Uhrig-Homburg, 2016, Measuring liquidity in bond markets, *Review of Financial Studies* 29, 1170–1219.

Shah, S., and B. W. Brorsen, 2011, Electronic vs. open outcry: Side-by-side trading of KCBT wheat futures, *Journal of Agricultural and Resource Economics* 36, 48–62.

Shang, Q., M. Mallory, and P. Garcia, 2018, The components of the bid–ask spread: Evidence from the corn futures market, *Agricultural Economics* 49, 381–393.

Sheshkin, D. J., 2004, *Handbook of parametric and nonparametric statistical procedures*, third edition (Chapman & Hall/CRC, Boca Raton, FL).

Singleton, K. J., 2014, Investor flows and the 2008 boom/bust in oil prices, *Management Science* 60, 300–318.

Smith, T., 1994, Econometrics of financial models and market microstructure effects, *Journal of Financial and Quantitative Analysis* 29, 519–540.

Sockin, M., and W. Xiong, 2015, Informational frictions and commodity markets, *Journal of Finance* 70, 2063–2098.

Stock, J. H., and M. W. Watson, 2004, Combination forecasts of output growth in a seven-country data set, *Journal of Forecasting* 23, 405–430.

Stoll, H. R., 1978, The supply of dealer services in securities markets, *The Journal of Finance* 33, 1133–1151.

Stoll, H. R., and R. E. Whaley, 2010, Commodity index investing and commodity futures prices, *Journal of Applied Finance* 20, 7–46.

Subrahmanyam, A., 2008, Lagged order flows and returns: A longer-term perspective, *Quarterly Review of Economics and Finance* 48, 623–640.

Szymanowska, M., F. De Roon, T. Nijman, and R. Van Den Goorbergh, 2014, An anatomy of commodity futures risk premia, *Journal of Finance* 69, 453–482.

Tang, K., and W. Xiong, 2012, Index investment and the financialization of commodities, *Financial Analysts Journal* 68, 54–74.

Tran, V. L., and T. Leirvik, 2019, A simple but powerful measure of market efficiency, *Finance Research Letters* 29, 141–151.

Trolle, A. B., and E. S. Schwartz, 2009, Unspanned stochastic volatility and the pricing of commodity derivatives, *Review of Financial Studies* 22, 4423–4461.

U.S. Senate Permanent Subcommittee on Investigations, 2006, The role of market speculation in rising oil and gas prices: A need to put the cop back on the beat, Technical report, Committee on Homeland Security and Governmental Affairs.

U.S. Senate Permanent Subcommittee on Investigations, 2009, Excessive speculation in the wheat market: Majority and minory staff report, Technical report, Committee on Homeland Security and Governmental Affairs.

Wang, X., P. Garcia, and S. H. Irwin, 2014, The behavior of bid–ask spreads in the electronically-traded corn futures market, *American Journal of Agricultural Economics* 96, 557–577.

Working, H., 1933, Price relations between July and September wheat futures at Chicago since 1885, *Wheat Studies* 9, 187–240.

Working, H., 1960, Speculation on hedging markets, *Food Research Institute Studies* 1, 185–220.

Yan, L., S. H. Irwin, and D. R. Sanders, 2019, Is the supply curve for commodity futures contracts upward sloping?, Working Paper.

Yang, F., 2013, Investment shocks and the commodity basis spread, *Journal of Financial Economics* 110, 164–184.

Zou, G. Y., 2007, Toward using confidence intervals to compare correlations, *Psychological Methods* 12, 399–413.