The impact of speculation on commodity prices: A Meta-Granger analysis[☆]

Thomas Wimmer[a,b,*], Jerome Geyer-Klingeberg[a], Marie Hütter[a], Florian Schmid[a], Andreas Rathgeber[a]

[a] *Professorship for Finance & Resource Management, Institute of Materials Resource Management, University of Augsburg, Germany*
[b] *MEAG MUNICH ERGO AssetManagement GmbH, Munich, Germany*

ARTICLE INFO

JEL classification:
G13
Q02
C58
C83

Keywords:
Speculation
Commodity prices
Non-commercial trader
Meta-analysis
Granger causality

ABSTRACT

This paper uses Meta-Granger analysis to explain and summarize the mixed results in the literature on the impact of financial speculation on commodity prices. The sample covers 2106 manually collected *p*-values from Granger causality (GC) tests reported in 54 prior studies. Our results show that the heterogeneity in previous findings can be largely explained by the commodity type under examination, the sample period of the data, the measurement of the focus variables (return, volatility, or spread), and the inclusion of control variables in the GC model. Even after accounting for 23 observable differences in study and test design, our results indicate that studies published in higher ranked journals present significantly less evidence for speculation to drive commodity prices. Moreover, we use the Meta-Granger results to predict 'best choice' models considering preferred model setups. The results reveal that the hypothesis of Granger non-causality between speculation and commodity prices cannot be rejected at standard significance levels when assuming a best choice study design and various variations of it. We conclude that either there is no genuine overall speculation effect in agricultural, energy and metal markets, or the research design of the frequently applied GC testing is not powerful enough to detect those effects.

1. Introduction

The market environment of commodity trading has undergone substantial changes over the last decades. Often referred to as *financialization*, commodities have become an increasingly attractive asset class for investors. Financial speculation, which is amplified by the emerging popularity of index related financial products, is often associated with the increased trading activity in commodity

☆ We would like to thank Etienne Borocco, Stephan Bruns, Chris Doucouliagos, Ayla Kayhan, Philipp Klein, Delphine Lautier, Hermann Locarek-Junge, Klaus Moeltner, Shinichi Nakagawa, Tom Stanley, Ke Tang, and Sheridan Titman for their helpful comments and suggestions. We also thank participants at the 26th Annual Meeting of the German Finance Association in Essen, Germany (September 2019), Association of Environmental and Resource Economists Annual Summer Conference 2019 in Incline Village, USA (May 2019), Commodity and Energy Markets Association Annual Meeting in Pittsburgh, USA (June 2019), European Conference on Data Analysis in Bayreuth, Germany (March 2019), Meta-Analysis of Economics Research Network (MAER-Net) Colloquium in Sydney, Australia (October 2018), as well as seminar participants at the University of Augsburg and the Université Paris-Dauphine. All remaining errors are ours. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect those of MEAG MUNICH ERGO AssetManagement GmbH, its subsidiaries, or affiliate companies.

* Corresponding author. Universitätsstraße 2, 86159, Augsburg, Germany.
E-mail address: thomas.wimmer@mrm.uni-augsburg.de (T. Wimmer).

futures markets (Tang and Xiong, 2012). This rise of speculation is frequently claimed to be a driver for surging commodity prices and volatilities, especially in the 2000s, sparking an ongoing public debate (e.g., Greely and Currie, 2008; Masters and White, 2009; U.S. Senate, 2009).

From a theoretical perspective, the influence of speculation on commodity prices can be seen through the lens of two major transmission mechanisms: risk sharing and information discovery (Cheng and Xiong, 2014). According to Keynes (1923), Hicks (1939), and Hirshleifer (1988), futures markets are characterized by hedging pressure as commercial hedgers are usually net short. Speculators acting as counterparty via long positions accommodate this pressure, leading to more efficient risk sharing. On the contrary, price distortions might arise if speculators engage in selloffs when seeking risk reductions in their portfolios. Considering information discovery (Grossman and Stiglitz, 1980; Hellwig, 1980), informational frictions hamper an efficient price discovery process in commodity markets. On the one hand, well-informed speculators might enhance this process by providing information regarding supply and demand through their trading activity. On the other hand, speculators might drive prices away from fundamental values due to imperfect information and heterogeneous beliefs (Singleton, 2014).

The empirical literature picks up the ambiguity in the public debate and academic theory by analyzing pricing mechanisms of commodities. The most common methodological approach is the analysis of the relation between non-commercial traders' or index traders' open interest and commodity futures prices using methods of Granger causality (GC) testing.[1] Looking at the findings of this literature stream, we see that results are rather diverse. One strand finds no or limited evidence for GC from speculation to commodity prices (among others, Alquist and Gervais, 2013; Büyükşahin and Harris, 2011; Sanders et al., 2004). In contrast, other authors detect substantial evidence that speculators Granger-cause changes in commodity prices (among others, Bohl et al., 2018; Mayer, 2012; Obadi and Korecek, 2018). A third group of studies finds evidence for speculation effects depending on the model and data characteristics of the GC test (among others, Ciner, 2002; Fujihara and Mougou, 1997; Huchet and Fam, 2016).

The previous GC studies often vary in terms of their study design, especially their sample composition (data period, frequency and data source), the configuration of empirical testing (variations of the GC model and inclusion of control variables), and the diverging measurement of financial speculation and commodity market behavior. Hence, it is challenging to directly compare the previous evidence without accounting for this heterogeneity.

Driven by the wide range of literature and its inconclusive outcomes, several review articles aggregate the existing research record to find out what we really know about the impact of speculation on commodities. Boyd et al. (2018), Grosche (2014), Haase et al. (2016), Shutes and Meijerink (2012), and Will et al. (2016) conduct literature reviews of articles on commodity (index) speculation. As an overall result, these reviews document mixed evidence for speculation to raise commodity prices or amplify its volatility. The paper closest to this study is Haase et al. (2016). The authors apply a vote counting approach to summarize the distribution and apparent disagreement among 100 studies on the effects of financial speculation. They use an integer scale from -2 to $+2$ to categorize the studies' results. This scale refers to the direction and strength of the impact of speculation on commodity markets. In addition, they apply this categorization to subgroups of results depending on the examined speculation measure, the response variable, paper quality, as well as the commodity type. The authors find that within their sample of 100 studies, the evidence is equally distributed among weakening effects (-2 and -1), no impact (0), and reinforcing effects ($+1$ and $+2$).[2] For specific subsamples, they report more conclusive results. For example, weakening effects dominate in the case of direct speculation measures. However, the vote counting procedure is widely criticized (Borenstein et al., 2009; Friedman, 2001; Hedges and Olkin, 1985; Mann, 1994; Stanley and Doucouliagos, 2012). First, it ignores the sample size of the primary studies and, thus, also the fact that the probability of finding a significant relation increases with the number of observations, i.e., when applying vote counting procedures, small-sample studies receive the same weight as large-sample studies. Second, the subsampling approach divides vote counts into separate categories. However, the examined categories – like paper quality or commodity type – might influence the collected primary studies' findings simultaneously. To avoid spurious and biased aggregation through omitted-variable bias, regression-based meta-methods are widely preferred to the univariate approach (Stanley and Doucouliagos, 2012). Third, the vote counting by Haase et al. (2016) collects one finding per study,[3] albeit empirical articles typically report a wide range of test results for different commodity types, lag structures, and time periods. Thus, condensing a study outcome into a single result neglects valuable information about the distribution of findings and the drivers of heterogeneity.

The Meta-GC analysis presented in this paper extends the existing reviews by aggregating 2106 reported *p*-values from GC tests on the relation between speculation and commodity markets reported in 54 empirical studies. Through this approach, we address three main research questions:

RQ1: Is there an overall effect of financial speculation on commodity prices?
RQ2: Is the literature contaminated by publication bias or overfitting?
RQ3: Which observable differences in study design explain variation across studies?

[1] It should be noted that studies also apply other methods beyond GC testing. However, they cannot be considered in this Meta-Granger analysis, as meta-analysis inherently requires the empirical results gathered from the primary studies to be comparable. The prerequisite of homogenous results is not fulfilled when pooling results obtained from fundamentally different methods. This can be seen as a limitation of meta-analysis as compared to conventional literature reviews.

[2] A further group contains papers where the direction of the impact could not be determined.

[3] In total, the authors collect 122 findings out of 100 studies. The literature search conducted in our study shows that there are 2106 test results from 54 GC studies that meet the criteria to apply meta-analysis.

By collecting and aggregating statistical data from previous studies, meta-analysis provides a powerful method to improve our understanding of why reported study results are so diverse and, thus, explains sources of heterogeneity. The Meta-GC analysis contributes to the literature in several ways: (i) We provide a statistical integration of previous GC tests. By applying the panel GC method according to Dumitrescu and Hurlin (2012), we explore differences between individual commodities. (ii) We apply the meta-analysis method for GC testing by Bruns and Stern (2019) to investigate the presence and impact of publication selection bias and overfitting via lag selection. (iii) We use meta-regression to explicitly test the joint impact of 23 aspects of study design and quantitatively discover heterogeneity drivers such as the commodity type, measurement differences, methodological characteristics, as well as journal quality of the prior studies. (iv) Based on the results from meta-regression analysis, we define a best choice model with preferred study and test characteristics.

The remainder of the article is structured as follows. Section 2 describes the construction of the meta-data set. Section 3 presents the methodology of the Meta-GC analysis. Section 4 explains the sources of heterogeneity among primary study results. Section 5 presents and discusses the empirical findings. Section 6 concludes.

2. Literature search and data construction

The literature search process and the subsequent meta-analysis are in line with the reporting guidelines for meta-regression research by Stanley et al. (2013) and Havranek et al. (2020). To find the sample of relevant studies, a systematic database search,[4] a forward search with Google Scholar's cited-by-option, as well as a backward search of the reference lists of all previously identified studies was conducted. The search strategy is based on relevant keywords and frequent synonyms for 'speculation' and 'commodity markets'. The last study was added in April 2018.

To construct a homogenous set of primary research studies, we apply three selection criteria: (i) The study must report results from GC tests examining a measure of financial speculation and a measure of commodity prices. GC testing was found to be the most prevalent empirical method and ensures the comparability of collected statistics. (ii) The associated F-statistics, χ^2-statistics, or p-values of the GC tests must be reported in each study. (iii) The number of lags used for GC testing must be available. This information is required to analyze the impact of overfitting on the reported GC results. If the empirical test statistics and lag orders are unreported, we requested the missing information from the study authors or re-calculated them from other reported statistics. (iv) As there is no observable measure that perfectly identifies speculation, studies must rely on proxies. We consider studies using speculation measures based on the following primary data:

- Non-commercial trader as well as commodity index trader positions[5]
- Flows into commodity funds and individual fund positions
- Overall open interest and trading volume in futures markets

Studies that do not explicitly define their measure are excluded from our analysis. Filtering the results from the literature search against the selection criteria left us with a total of 54 primary studies. The full reference list of studies is available in Appendix A. The difference in the sample size as compared to the previous review by Haase et al. (2016), who analyze 100 studies, can be explained as follows. First, we extend the literature search by the more recent period (April 2015 – April 2018). Second, applying the above-mentioned selection criteria substantially reduces our sample. Many of the studies in the sample by Haase et al. (2016) had to be excluded as they either do not (fully) report results from GC tests or apply a fundamentally different measure for speculative activity (e.g., stock returns).

Out of the final sample of studies, we manually extract GC test results in the form of p-values and the respective test statistics (mostly F or χ^2). If possible, missing values are re-calculated from other reported statistics. As studies usually report multiple results, e.g., for different commodities, lags, or method choices, we collect all reported findings per study.[6] The final sample covers 2106 p-values from 54 studies. This discrepancy between the number of observations and number of studies arises from the various 'sub-studies', because most papers present results for different commodities, sample periods, models, and robustness tests. In the subsequent meta-regression models, we explicitly account for the within-study dependencies arising from multiple estimates collected from the same article.

3. Methodology

Meta-regression analysis is a form of meta-analysis designed to analyze empirical research in economics and business (Stanley, 2001, 2007). It covers statistical methods to condense information from a sample of studies and provides insights into why empirical outcomes vary or even contradict on a certain phenomenon. Empirical research studies typically exhibit large variation in terms of the analyzed

[4] We searched in ABI/Inform Complete, EconLit, Google Scholar, ScienceDirect, and the Social Science Research Network.

[5] Although index investment is not unambiguously regarded as speculation (Stoll and Whaley, 2011), we follow Haase et al. (2016) and include those positions in our analysis. Unreported Meta-GC results have shown no significant differences in effects compared to non-commercial trader positions.

[6] In this regard, we follow established practice in meta-regression analysis (among others, Feld et al., 2013; Hang et al., 2020; Kysucky and Norden, 2016; Rusnak et al., 2013). This procedure maximizes the amount of retained information, avoids subjectively selecting individual estimates, and retains the natural variability within primary studies occurring from differences in model design and subsample selection.

time periods, sample composition, applied methods, and model specification. This heterogeneity creates demand for reviews examining the sources of this variation and identifying the overall effect implied by the literature. Unlike narrative reviews, meta-analysis allows testing for such moderating effects across papers, to reconcile conflicting evidence, and to draw the big picture of piecemeal findings. Meta-methods have been frequently applied in finance research (van Ewijk and Groot, 2012; Feld et al., 2013; Rahim et al., 2014; Kysucky and Norden, 2016; Geyer-Klingeberg et al., 2018a, 2019, 2020; Hang et al., 2018).

Meta-analysis is especially valuable for the speculation-commodity price nexus, as there are many studies based on a similar setup (GC testing) reporting divergent results. Thus, specific study characteristics might play an essential role for driving outcomes. Moreover, differences in study quality, like journal ranking, can only be examined on a meta-level and not in a single primary study.

3.1. GC testing in primary studies

We analyze p-values from GC tests (Granger, 1969), where financial speculation Granger-causes commodity prices when past information on speculation improves the prediction of prices compared to a prediction solely based on price information of the past. Originally, the test employs an autoregressive distributed lag (ADL) model:

$$y_t = \phi + \sum_{j=1}^{p} \lambda_j y_{t-j} + \sum_{k=1}^{q} \theta_k x_{t-k} + \varepsilon_t, \tag{1}$$

where in our context y_t refers to a measure of the commodity price, x_t refers to a measure of speculative activity, and (p, q) refers to the lag structure. The relevant test statistic is an F or χ^2 test, with $H_0 : \theta_k = 0 \ \forall k = 1, ..., q,$ testing for Granger non-causality. If the null hypothesis is rejected, speculative activity is said to Granger-cause commodity prices. Owed to the nature of GC testing, which focuses on the presence of a lead-lag relationship instead of the size or direction of the relation, we treat p-values from F or χ^2 tests reported in prior studies as effect sizes in our meta-analysis. Although GC might only indicate (but does not imply) causality, we refer to it as analysis of speculation effects on commodity prices.

Eq. (1) illustrates the basic GC model. Its interpretability and explanatory power are often criticized in the literature due to the sensitivity of results to the proper specification of the time series, omitted variables, non-linearity, and other econometric challenges (Grosche, 2014). Many studies in our sample address those issues, e.g., by applying advanced GC models. We control for and quantify the impact of a wide range of GC model and data specifications in the subsequent meta-regression analysis.

3.2. Meta-GC analysis I: Panel Granger approach

Several approaches exist for the combination of p-values in order to test overall effects across studies (Hartung, 1999; Hou, 2005; Dumitrescu and Hurlin, 2012; van Aert et al., 2016; Chien, 2018; Heard and Rubin-Delanchy, 2018). As we focus on GC tests, we combine the prior results by applying the panel GC approach proposed by Dumitrescu and Hurlin (2012). To determine speculation effects for individual commodities, we construct one panel per commodity. Therefore, a panel contains M test statistics from all prior studies that analyzed the same commodity. The null hypothesis of no significant coefficients in any of the M tests within the panel can be tested against the alternative hypothesis of at least one significant coefficient in the panel.

Assuming that the residuals are independently distributed across individuals, we use the panel GC test as a combining test in the sense of a meta-analysis. This enables us to calculate the average Wald statistic per commodity according to:

$$\overline{W}_M = \frac{1}{M} \sum_{i=1}^{M} w_i W_{i,T_i} \quad \text{with} \quad w_i = \frac{\frac{1}{J_i}}{\sum_{i=1}^{M} \frac{1}{J_i}}, \tag{2}$$

where W_{i,T_i} is the χ^2-distributed Wald statistic of the primary GC test i with sample size T_i. M refers to the total number of GC tests per commodity. To account for the assumption of independence across individuals, we choose the weights w_i of a test statistic as inverse of the number of estimates J_i reported in the corresponding primary study.

In the case of an unbalanced panel with a unit specific lag order, we are able to normalize the statistics and generate a random variable which asymptotically converges to standard normal distribution if $M \to \infty$ and $T_i > 5 + 2K_i$, where T_i is the length of the time series applied to estimate the Wald statistic i and K_i is the lag order in the estimation procedure of the Wald statistic i. The corresponding standard normal distributed test statistic Z_M^{Hnc} of the panel GC test is calculated as follows[7]:

$$Z_M^{\text{Hnc}} = \sqrt{M} \ \frac{\overline{W}_M - \sum_{i=1}^{M} w_i K_i \frac{T_i - 2K_i - 1}{T_i - 2K_i - 3}}{\sqrt{\sum_{i=1}^{M} 2 w_i K_i \frac{(T_i - 2K_i - 1)^2 (T_i - K_i - 3)}{(T_i - 2K_i - 3)^2 (T_i - 2K_i - 5)}}} \ . \tag{3}$$

We use the average Wald statistic across the individual GC tests, $\overline{W}_M$, to determine if at least one test in the primary literature or none test has found speculation effects for a specific commodity. Z_M^{Hnc} is the corresponding test statistic to determine the statistical significance of the panel GC.

[7] This equation is equivalent to formula (33) in Dumitrescu and Hurlin (2012) with the exception that we additionally include weights w_i.

3.3. Meta-GC analysis II: Publication bias and overfitting test

Reported tests in primary studies can be subject to so called p-hacking (Head et al., 2015; Brodeur et al., 2016; Ioannidis et al., 2017). Accordingly, authors might – consciously or unconsciously – present only a subset of estimated model specifications that deliver a p-value of an F or χ^2 test below the common threshold levels of statistical significance. Consequently, a substantial number of model specifications might remain unreported (Simonsohn et al., 2014b), resulting in publication bias and uncertainty about genuine effects. In the context of GC testing, p-hacking might be caused by sampling errors and overfitting (Bruns and Stern, 2019). The former arises, for example, from changing the time period of the data until a p-value below a certain level of statistical significance is detected (Bruns, 2017). Similarly, overfitting might arise by increasing the lag length of speculative activity. The consequence is reporting of false positive findings, most notably in small samples.

Bruns and Stern (2019) present a meta-regression model testing for genuine GC while controlling for p-hacking based on sampling errors and overfitting. Since both, the absence and the presence of speculation effects are prevalent results in the existing literature, suggesting that both are widely accepted findings for publication, we test for p-hacking in both directions. Hence, we do not only investigate potential bias in significant, but also in non-significant test results. Therefore, we extend the model by Bruns and Stern (2019) by additionally testing for the genuine lack of GC besides testing for the genuine presence of GC:

$$\tilde{p}_{\mathrm{ij}} = \alpha + \beta_1 \sqrt{\mathrm{df}_{\mathrm{ij}}} + \beta_2 \mathrm{lags}_{\mathrm{ij}} + \varepsilon_{\mathrm{ij}}, \tag{4}$$

$$\text{with } \tilde{p}_{\mathrm{ij}} = \Phi^{-1}\left[p_{\mathrm{ij}}\right],$$

where we use the inverse cumulative distribution function of the standard normal distribution, Φ^{-1}, to convert the p-value p_{ij} from GC test i reported in study j to probit-transformed p-values $\tilde{p}_{ij}$. For this part of the analysis, we employ Eq. (4) separately for two groups of p-values: $\tilde{p}_{ij,p\geq 0.05}$ and $\tilde{p}_{ij,p<0.05}$, referring to test results denying GC ($p_{ij} \geq 0.05$) and indicating GC ($p_{ij} < 0.05$).

As Bruns and Stern (2019) demonstrate, in the case of reported p-values lower than 0.05, p-values decrease with higher degrees of freedom if genuine GC is present. However, if prior studies report significant findings when there is no genuine effect, p-values should be unrelated to df_{ij} or even increase when df_{ij} rises. Since selective publication of results from various model specifications might not only lead to biased false positive findings, we additionally test for false negatives by investigating the relationship between p-values larger than 0.05 and the degrees of freedom in the corresponding model specifications. In the case of a genuine lack of GC, p-values should increase when df_{ij} rises. In contrast, if non-significant test results are biased due to false negatives, we expect $\beta_1 \leq 0$.

Additionally, we include the lag lengths of primary models to control for overfitting. Especially in small samples, overfitting might even occur if information criteria are used (Bruns and Stern, 2019). If lags of the speculative activity measure ($lags_{ij}$) are overfitted, test results might over-reject the null of no GC. Analogously, if less lags are included than required, tests might be biased towards failing to reject the null hypotheses. Therefore, by including the lag length of the speculation measure in the publication bias test, we control for overfitting in primary tests that reject Granger non-causality and control for underfitting in primary tests that fail to reject Granger non-causality.

3.4. Meta-GC analysis III: Heterogeneity analysis

While still controlling for publication bias and overfitting, we include moderating variables that potentially drive the variation in primary test results, leading to the following extended Meta-GC model:

$$\tilde{p}_{\mathrm{ij}} = \alpha + \beta_1 \sqrt{\mathrm{df}_{\mathrm{ij}}} + \beta_2 \mathrm{lags}_{\mathrm{ij}} + \sum_{l=1}^{L} \gamma_l Z_{l,\mathrm{ij}} + \varepsilon_{\mathrm{ij}} \tag{5}$$

where $Z_{l,ij}$ is the variable referring to moderator l of GC test i reported in study j. In the following section, we present details about primary tests' characteristics that we modeled through moderating variables $Z_{l,ij}$. The Meta-GC coefficients γ_l indicate the change in p-values from GC tests of the sample of studies collected for the meta-analysis due to changes in the moderating variables.

For the estimation of the Meta-GC models in Eqs. (4) and (5), we use the following econometric specifications to account for heteroscedasticity and the unbalanced panel data structure:

Weighted regression. It is an established approach in meta-regression research to use weighted least squares (WLS) to obtain efficient estimates and to reduce the apparent heteroscedasticity in effect sizes that occurs as each study or even each GC test is based on a different sample composition (Stanley and Doucouliagos, 2012). We apply three alternative weighting schemes in this study: (i) The inverse of the number of estimates reported in each primary study. This weight accounts for the unbalancedness of the meta-data set as some studies report many estimates and others only a few. It assigns equal weights to studies independently of the number of reported estimates. This approach follows recent meta-studies by Geyer-Klingeberg et al. (2018b), Havranek and Irsova (2017), or Zigraiova and Havranek (2016). (ii) The degrees of freedom of the primary study data sample. This implies that more precise and thus more reliable statistical estimates (those with more observations, i.e., lower standard errors) receive a larger weight in the regression. Degrees of freedoms are directly observable from the primary studies or re-calculated by the number of reported observations and the number of estimated coefficients in the GC models. (iii) The inverse of the variance in the GC model. The inverse variance is the common weighting in meta-regression research (Stanley et al., 2010). However, as authors do not report estimates but p-values from GC tests, the variance is unreported in primary studies. We apply the variance approximation by Dumitrescu and Hurlin (2012) to estimate the variance of the GC test i in study j:

$$\mathrm{Var}_{\mathrm{ij}} = 2K_{\mathrm{ij}}\frac{\left(T_{\mathrm{ij}} - 2K_{\mathrm{ij}} - 1\right)^2\left(T_{\mathrm{ij}} - K_{\mathrm{ij}} - 3\right)}{\left(T_{\mathrm{ij}} - 2K_{\mathrm{ij}} - 3\right)^2\left(T_{\mathrm{ij}} - 2K_{\mathrm{ij}} - 5\right)} \tag{6}$$

where T_{ij} are the degrees of freedom and K_{ij} is the number of lags used in the primary GC test.

Clustering of standard errors. As described in Section 2, we collect all results from GC testing to maximize data availability and to avoid biases arising from subjective selection of specific outcomes. With 54 studies and 2106 collected p-values, we have on average 39 observations per study. By sampling multiple estimates per study, dependency on the study level occurs. Moreover, several study authors published more than one paper in the field (e.g., Aulerich, Brunetti or Gilbert) leading to between-study dependency. To control for the issue of non-independence, we adopt robust standard errors in the Meta-GC model with clusters at both the level of the individual studies and the study authors. This approach follows the two-way error clustering by Cameron et al. (2011).

4. Sources of heterogeneity

Eq. (5) incorporates covariates to control for the sources of heterogeneity that potentially influence empirical results at the primary study level. We use explanatory variables as presented in Table 1 to explain disparate findings in the literature, as well as to derive 'best choice models' in order to reveal overall effects from speculation on commodity markets.

Sample and data characteristics. We include the start date of the data sets used for the GC tests in the primary studies to measure time effects in our sample. As the sample data in the primary studies starts at different dates during a year, we account for the exact date and not just the sample year. Especially since the early 2000s, commodity market conditions have changed due to financialization (Tang and Xiong, 2012). Consequently, we expect a higher probability for the presence of speculation effects in more recent primary samples.

Table 1
Description and summary statistics of primary GC tests' characteristics.

Variable	Description	Mean	Std. Dev.
Sample and data characteristics			
Sample start date	Start date of primary study data used for GC test	08/17/2002	2074 days
Sample start date after 2004	= 1 if start date of primary study data is after January 01, 2004, 0 otherwise	0.64	0.48
Degrees of freedom	Square root of the difference between sample size and number of included covariates	21.73	11.82
Daily	= 1 if daily data is examined	0.33	0.47
Weekly*	= 1 if weekly data is examined	0.62	0.49
Monthly	= 1 if monthly data is examined	0.05	0.22
Commodity class			
Metals*	= 1 if metals are examined	0.34	0.47
Energy	= 1 if energy commodities are examined	0.17	0.38
Agriculturals	= 1 if agriculturals only or a basket of commodities incl. agriculturals is examined	0.49	0.50
Measurement of the response variable			
Return	= 1 if the effect of speculation on price returns is examined	0.61	0.49
Volatility*	= 1 if the effect of speculation on price volatilities is examined	0.37	0.48
Spread	= 1 if the effect of speculation on price spreads is examined	0.02	0.13
Econometric specification			
#Lags	Number of lagged values of the independent variable	2.07	5.77
VAR/VEC	= 1 if a VAR/VEC model is examined, 0 if an ADL model is examined	0.40	0.49
Controls	= 1 if control variables are examined, 0 otherwise	0.09	0.29
Multivariate GC	= 1 if a multivariate GC test is applied, 0 otherwise	0.05	0.21
Contemporaneous	= 1 if lag 0 of the independent variable is included, 0 otherwise	0.02	0.13
Non-parametric	= 1 if the test is non-parametric, 0 otherwise	0.01	0.08
Publication characteristics			
Higher ranked journal	= 1 if the Scimago journal rank (SJR) in the publication year is in the first ranking quartile of the categories Economics/Econometrics or Finance, 0 otherwise.	0.13	0.34
Author affiliation			
Academic authors only	= 1 if all authors only hold academic positions and have not declared any external funding, 0 otherwise	0.08	0.27
Market liquidity			
Large futures market	= 1 if a highly traded commodity is examined, 0 otherwise	0.77	0.42
Major exchange	= 1 if data from CBOT, CME, ICE, LME or NYMEX is examined, 0 otherwise	0.91	0.28
Measurement of speculative activity			
First differences	= 1 if the change in speculative activity is examined, 0 otherwise	0.36	0.48
Position and flow data	= 1 if speculative measures are based on trader positions or fund flows, 0 if speculative measures are based on general open interest/volumes	0.74	0.44

Notes: This table presents the definitions and summary statistics of the variables measuring data-related and methodological heterogeneity across studies. All variables are manually collected from primary studies. (*) marks the omitted base category in the meta-regression analysis.

As a further control variable for time effects, we define a dummy variable that is equal to one for GC tests with data sets starting after January 01, 2004. The years 2004/2005 are commonly regarded as the beginning of the financialization of commodity markets (Büyükşahin et al., 2008; Hamilton and Wu, 2014; Tang and Xiong, 2012). Furthermore, we also control for the data frequency, as speculation effects might be transitory and only measurable in shorter frequencies. Degrees of freedom are included to control for publication bias as outlined in the previous section.

Commodity class. The primary studies analyze 46 different commodities in total. We group them into three main categories: metals, energy, and agricultural commodities. As GC tests that analyze a basket of commodities all include agricultural commodities, we assign them to the agriculturals group.

Measurement of the response variable. Primary tests greatly differ regarding the dependent variable that captures the commodity price. We implement a set of dummy variables to disentangle speculation effects on various forms and transformations of commodity prices: returns, volatilities,[8] and term spreads for different maturities.

Econometric specification. Next to the number of lags that might lead to overfitting or underfitting, we analyze the influence of the model type on the primary studies' results. We differentiate between classical GC tests that use an autoregressive distributed lag (ADL) model and GC tests embedded in a vector-autoregressive regression (VAR) or vector-error correction (VEC) framework that accounts for the potential bi-directionality of the relationship between speculation and commodity prices. Furthermore, while traditional GC tests establish causality in the mean, modified versions of the test are able to detect causality in, e.g., quantiles or higher moments of the response variable (Bell et al., 1996). Such non-linear GC and non-parametric test types might produce results that are different from traditional tests. Moreover, we analyze whether primary results change if the GC test is extended by controls or the contemporaneous value of the speculation measure. Finally, we also control if primary studies apply multivariate GC tests, i.e., if a primary study investigates multiple traders' positions simultaneously in the same model (e.g. of commercial and non-commercial traders).

Publication characteristics. We consider if a study is published in an important journal. Differences in publication quality are captured by the Scimago Journal Rank (SJR) of the publication outlet (González-Pereira et al., 2010). If the journal is ranked in the top quartile of the categories 'Economics and Econometrics' or 'Finance' in the publication year, the dummy variable takes the value 1, otherwise it is 0. Working papers and book chapters are coded as 0.

Author affiliation. The professional background might have an influence on the research outcome selected for publication. The process of selecting speculation proxies, designing the study, as well as reporting of results is not immune to bias (even unconsciously). Such a bias might arise from the institutional affiliation or financial support (Lexchin, 2012; Krimsky, 2013). We differentiate between studies where all authors are affiliated with an academic institution and have not declared any funding and the remaining studies where at least one author has declared an external affiliation or funding.

Liquidity. We define two variables referring to the market liquidity of individual commodities examined in a primary study. First, we consider how actively the analyzed commodity is traded. Every commodity with an average trading volume of more than 30,000 futures contracts per day (between 2004 and 2013) is classified as having a large futures market. 30,000 contracts refer to the mean of the average trading volume across all analyzed commodities that are traded in futures contracts at the Chicago Mercantile Exchange (CME), the Chicago Board of Trade (CBOT), and the Intercontinental Exchange (ICE). The exact classification is reported in Appendix B. Second, we code the respective commodity exchange the individual primary study refers to. We differ between important trading hubs (CBOT, CME, ICE, London Metal Exchange, and New York Mercantile Exchange) and less important exchanges.

Definition and measurement of speculative activity. Heterogeneous empirical findings might stem from dissimilar concepts of financial speculation. We define a dummy variable that is 1 if a study uses measures like Working's (1960) T-index or other measures that are based on positioning (of non-commercial and index traders) or fund flows.[9] The dummy variable is 0 if primary studies use general open interest and trading volume data from futures markets. The latter group of studies uses, e.g., the ratio of trading volume to open interest as proxy for speculative dominance in the market. In our opinion, position and fund flows data (that at least tries to distinguish between speculative and non-speculative activity) should deliver more reliable results than general open interest and volume data that cannot separate effects from commercial trading at all. Furthermore, we analyze how the speculative variable enters primary econometric models. We introduce another dummy variable to account for the fact that variables measured in levels should exhibit more vulnerability to spurious findings as compared to first differenced variables.

5. Empirical results

5.1. Meta-GC analysis I: Panel Granger approach

To synthesize the sample of p-values, we apply panel GC testing as described in Eq. (2). This test relies on the primary tests' Wald statistics with an asymptotic χ^2-distribution.[10] The results of the panel GC test are presented in Table 2. The rows of Table 2 refer to the

[8] Moreover, we distinguished between conditional, historical, and implied volatility. As we could not detect any significant differences in the primary studies' results, we model those variations jointly by one dummy variable.

[9] Based on unreported empirical meta-findings, different notions as to which types of traders count as speculators do not lead to significantly different results, such that we do not further differentiate trader classifications.

[10] If a primary study applies a F-test instead, we use the F-statistic to calculate the corresponding χ^2-statistic. In the case that only the p-value and the type of the test statistic was given (F or χ^2), we re-calculated the χ^2-statistics. We excluded 15 tests from two studies, because their testing strategy is neither a χ^2-test nor an F-test.

different commodities examined in the primary studies. Columns 2 to 3 report the findings for the full data set, columns 4 to 5 show only those results for GC tests that analyze returns or spreads, and columns 6 to 7 present results for GC tests that analyze volatilities.

There are two main findings. First, the major commodity type that shows non-significant results are metals (lead, tin, and zinc in the overall sample). In addition, we find relatively low $\overline{W}$ statistics for gold, copper, palladium, platinum, and silver, although we aggregate more than 100 observations for each of those metals. This is a first hint that for metals, primary tests reject the null hypothesis of no GC less often than for other commodities. Second, we see that the $\overline{W}$ statistics calculated from tests that analyze returns or spreads ($\overline{W} = 6.41$) are substantially lower than those based on tests that analyze volatilities ($\overline{W} = 10.95$). To conclude, for most commodities the panel GC based on primary tests rejects the null of Granger non-causality. However, we must not interpret this finding as evidence for overall speculation effects, because the rejection of the null only indicates that there is at least one test in the primary literature that detects GC for a specific commodity. Therefore, we rather interpret the apparent differences between the average Wald statistics driven by the commodity classes and the focus variable (return/spread or volatility) as a hint that heterogeneity in primary tests is driving the disparate results. We explicitly model and account for multiple sources of heterogeneity related to study design, data and publication characteristics in the next two subsections.

5.2. *Meta-GC analysis II: Publication bias and overfitting test*

Before quantitatively testing for publication bias, we investigate the graphical distribution of the 2106 collected *p*-values in our sample. Fig. 1(a) depicts the distribution of all *p*-values included in our meta-data set and reveals that results are right-skewed. However, only 362 *p*-values, which is equivalent to around 17% of the full sample, are below the 5% threshold of statistical significance.

To further investigate significant findings, Fig. 1(b) visualizes the distribution of the subsample of *p*-values between 0.00 and 0.05. Likewise, the distribution is skewed to the right with a median of 0.016. While the distribution of *p*-values larger than 0.005 is almost uniformly distributed, we observe a dominance of smaller *p*-values. In general, *p*-hacking might be present if *p*-values just below the common threshold of statistical significance are overrepresented in the literature (Simonsohn et al., 2014a; van Aert et al., 2016). The apparent pattern in our sample, especially the absence of a noticeable hump just below 0.05, suggests that significant results do not stem from *p*-hacking.

To explicitly test for publication bias and overfitting, we employ the Meta-GC model from Eq. (4). As publication bias and overfitting might be present in both directions (towards significant and non-significant results), we split the sample for this part of the analysis in two samples based on the threshold of $p = 0.05$. Results are reported in Table 3. The baseline meta-regression (Model II) inversely weights probit-transformed *p*-values by the number of estimates per study to give equal weight to each primary study. Model II is our preferred setup, as we have a largely unbalanced panel data set with some studies reporting just one *p*-value and other studies with over 100 GC tests. Furthermore, we apply the baseline model setup to subsamples for agricultural commodities (Model V), studies published in journals with an SJR in the top quartile (Model VI) or GC tests based on more recent data, i.e., with a sample start date after January 01, 2004 (Model VII).

Panel A covers tests with the result that speculation Granger-causes commodity prices. The regression coefficient for the degrees of freedom ($\sqrt{df}$) is always negative in the full sample models (Models I to IV). As expected, in the presence of genuine GC, primary findings become more significant when degrees of freedom increase. This result suggests that *p*-values below 0.05 do not suffer from publication bias. Nevertheless, as our estimates are not consistently significant, we cannot fully rule out publication bias in the full sample. These findings are robust if we apply no weighting of primary *p*-values (Model I), a weighting scheme based on the inverse of the degrees of freedom (Model III) and the inverse of the approximated standard errors (Model IV).

When applying the meta-test to subsamples (Models V to VII), we find mixed evidence. On the one hand, tests that analyze agricultural commodities and tests that use data sets after January 01, 2004 show no indication of publication bias. On the other hand, considering the relatively small subset of *p*-values below 0.05 from GC tests published in higher ranked journals (Model VI), those *p*-values rise with the degrees of freedom, indicating a slight bias by false positives in this subgroup. However, this result is only weakly significant. Additionally, Table 3 shows that this potential bias only applies to 23 tests, while 258 tests in primary studies that are published in higher ranked journals find no empirical evidence that speculation Granger-causes commodity prices.

Panel B analyzes *p*-values from tests that find no speculation effects. If there is a genuine lack of GC, *p*-values from primary studies should increase with an increase in the degrees of freedom. However, our analysis mostly reveals a negative relationship for the sample with *p*-values above 0.05. Nevertheless, we do not treat these findings as strong evidence for publication bias as the coefficients of $\sqrt{df}$ are only significantly negative in Model IV. Furthermore, controlling for overfitting or underfitting has no major effect, as coefficients of the lag lengths are mostly non-significant and near zero in both panels. Overall, we may conclude that significant results in the literature do not suffer from publication bias. For non-significant primary results doubts about publication bias remain, but we do not see strong evidence either.

5.3. *Meta-GC analysis III: Heterogeneity analysis*

As we find that publication bias and overfitting do not explain the large variation in primary results, we continue with the analysis of the observed heterogeneity arising from study and test characteristics. Histograms of several subgroups of GC tests in the primary literature are depicted in Fig. 2.

Fig. 2 shows that primary studies find more significant GC from speculation to commodity markets when analyzing agricultural commodities in contrast to metals. Furthermore, *p*-values are higher in studies published in higher ranked journals, as compared to

Table 2
Panel GC test.

Commodity	Full sample		Returns and spreads		Volatility	
	$\overline{W}$	No. of obs.	$\overline{W}$	No. of obs.	$\overline{W}$	No. of obs.
Panel A: Agriculturals						
Barley	1.68***	11	–	–	1.68***	11
Castor seed	7.32***	1	–	–	7.32***	1
Chickpeas	46.23***	2	–	–	46.23***	2
Cocoa	4.91***	36	5.13***	23	3.84***	13
Coffee	6.50***	44	6.91***	31	1.83***	13
Corn	6.20***	154	6.25***	104	6.04***	50
Cotton	4.80***	44	3.04***	23	7.32***	21
Feeder cattle	8.67***	35	10.16***	22	2.92***	13
Guar seed	32.95***	5	–	–	32.95***	5
Lean hogs	9.64***	39	12.12***	22	4.76***	17
Live cattle	6.39***	52	6.82***	35	4.84***	17
Maize	5.70***	30	4.65***	4	5.90***	26
Mix	2.01***	17	2.01***	17	–	–
Mustard seed	6.74***	10	–	–	6.74***	10
Oat	14.91***	2	14.91***	2	–	–
Orange juice	0.18	1	0.18	1	–	–
Palm oil	38.58***	2	–	–	38.58***	2
Pepper	17.72***	17	–	–	17.72***	17
Potato	0.47	2	–	–	0.47	2
Rapseed oil	7.56***	2	–	–	7.56***	2
Rice	10.29***	16	14.53***	3	6.59***	13
Rubber	18.24***	1	18.24***	1	–	–
Soybean flour	15.80***	8	15.19***	6	17.79***	2
Soybean oil	12.81***	49	9.82***	32	22.30***	17
Soybeans	7.38***	158	6.41***	96	9.98***	62
Sugar	6.92***	41	4.03***	22	11.32***	19
Turmeric	8.42***	1	–	–	8.42***	1
Wheat (CBOT)	3.00***	145	2.38***	95	7.15***	50
Wheat (KCBT)	5.58***	68	6.00***	51	3.61***	17
Wheat (other exch.)	17.12***	8	17.10***	6	17.16***	2
Panel B: Energy						
Crude oil	8.55***	324	9.29***	304	5.01***	20
Gasoline	0.30	2	0.30	2	–	–
Heating oil	2.49***	5	2.49***	5	–	–
Natural gas	3.85***	44	3.23***	28	5.36***	16
Panel C: Metals						
Aluminum	5.81***	3	5.81***	3	–	–
Copper	4.42***	158	4.93***	86	1.73***	72
Gold	4.68***	150	5.13***	78	2.06***	72
Lead	0.58	1	0.58	1	–	–
Nickel	3.58***	3	3.58***	3	–	–
Palladium	1.80***	108	1.97***	54	1.62***	54
Platinum	3.97***	146	4.41***	74	1.90***	72
Silver	1.67***	144	1.16***	72	2.18***	72
Tin	1.09	1	1.09	1	–	–
Zinc	0.47	1	0.47	1	–	–
All	7.60***	2091	6.41***	1308	10.95***	783

Notes: This table reports the results of Eq. (2): $\overline{W}_M = \frac{1}{M}\sum_{i=1}^{M} w_i W_{i,T_i}$ *with* $w_i = \frac{\frac{1}{J_i}}{\sum \frac{1}{J_i}}$, where $W_{i,T}$ are the M effect sizes (Wald statistics) of the primary studies and w_i are the weights of the studies. We choose the weights as inverse of the number of estimates J_i. "Returns and spreads" refers to tests examining returns or spreads and "Volatility" refers to tests examining volatility. $^*p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

unpublished studies or studies published in lower ranked journals. GC tests applied to data sets starting after January 01, 2004 produce smaller p-values than tests using samples with a sample start date before January 01, 2004.

We quantitatively analyze the heterogeneity using the Meta-GC model in Eq. (4). For this part of the analysis, we again examine the full data set without splitting it in significant and non-significant results. Hence, we assume that the test results reported in the primary studies are driven by the same heterogeneity factors. The Meta-GC results are presented in Table 4. Models II to VII are estimated by weighted least squares using different weighting schemes. As discussed before, Model II is our preferred model as it accounts for the

(a) (b)

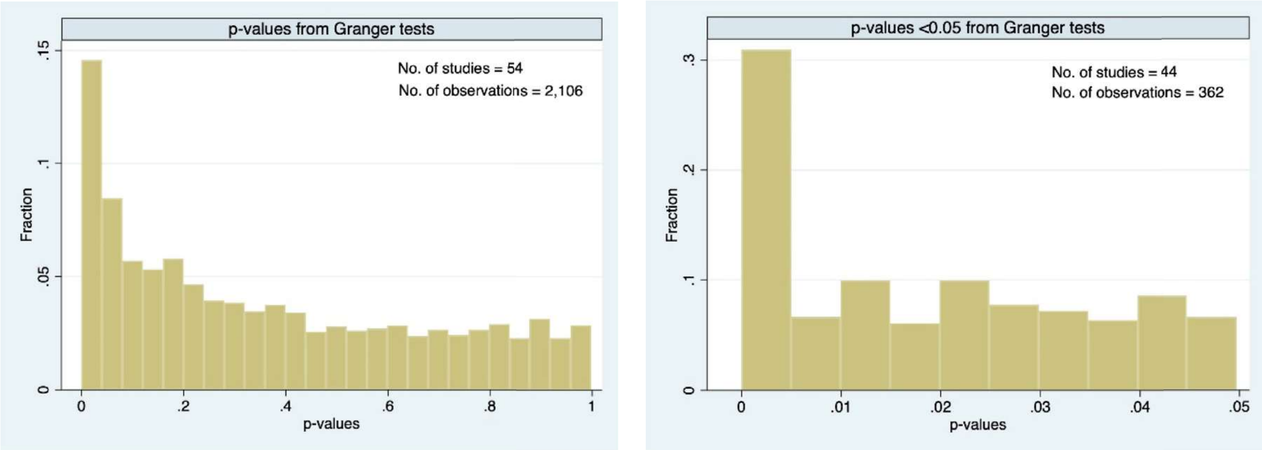

Fig. 1. Distribution of *p*-values from GC tests. *Notes:* The figures show the histogram of *p*-values from GC tests reported in primary studies: (a) the full sample and (b) the subsample of *p*-values between 0.00 and 0.05.

Table 3
Analysis of publication selection and overfitting bias.

Weighting	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)
	Full sample	Full sample (Baseline)	Full sample	Full sample	Agriculturals only	Higher ranked journals only	Sample start date after 2004
	–	$[1/J_i]$	$[\sqrt{df_{ij}}]$	$[1/\widetilde{SE_{ij}^2}]$	$[1/J_i]$	$[1/J_i]$	$[1/J_i]$
Panel A: Primary studies rejecting Granger non-causality (dependent variable: $\tilde{p}_{ij,p<0.05}$)							
Constant	−1.971***	−2.256***	−1.825***	−1.978***	−1.823***	−6.147**	−1.751***
	(-19.27)	(-5.13)	(-13.23)	(-20.55)	(-9.57)	(-2.57)	(-2.77)
$\sqrt{df}$	−0.017***	−0.021*	−0.022***	−0.012**	−0.033***	0.112*	−0.039**
	(-2.92)	(-1.73)	(-3.49)	(-2.49)	(-3.24)	(1.65)	(-2.33)
#Lags	−0.005	0.002	−0.008	−0.028*	−0.003	0.061*	−0.101
	(-1.09)	(0.36)	(-1.14)	(-1.73)	(-0.62)	(1.84)	(-1.35)
#Obs.	362	362	362	362	237	23	261
#Studies	44	44	44	44	31	4	36
Panel B: Primary studies not rejecting Granger non-causality (dependent variable: $\tilde{p}_{ij,p\geq 0.05}$)							
Constant	−0.032	−0.165	−0.069	0.004	−0.089	−0.451	0.016
	(-0.36)	(-1.36)	(-0.70)	(0.06)	(-0.76)	(-1.53)	(0.07)
$\sqrt{df}$	−0.005*	−0.004	−0.004	−0.006**	−0.009	0.015	−0.006
	(-1.77)	(-0.66)	(-1.55)	(-1.96)	(-1.62)	(1.19)	(-0.59)
#Lags	−0.006	−0.002	−0.006	−0.004	−0.003	−0.038	−0.055**
	(-1.23)	(-0.55)	(-1.57)	(-0.21)	(-0.98)	(-1.44)	(-2.22)
#Obs.	1744	1744	1744	1744	788	258	1088
#Studies	53	53	53	53	42	4	36

Notes: This table reports the results of Eq. (4): $\tilde{p}_{ij} = \alpha + \beta_1\sqrt{df_{ij}} + \beta_2 lags_{ij} + \varepsilon_{ij}$, where $\tilde{p}_{ij}$ is the probit-transformed *p*-value and df_{ij} the degrees of freedom of the corresponding GC test *i* of study *j*. The variable $lags_{ij}$ refers to the lag length of the speculation activity measure x_{ij}. Models II and V-VII are estimated by weighted least squares using the inverse of the number of estimates as weights. Model I is unweighted while Model III and IV use the square root of the degrees of freedom respective the inverse of the squared approximated standard errors as weights. The *t*-statistics of the regression coefficients reported in parentheses are based on standard errors adjusted for within-study and across-study correlation. *$p < 0.1$, **$p < 0.05$, ***$p < 0.01$.

diverging number of test results reported per study. Model VI is a reduced form of Model II where only significant variables are kept.[11] Model VII is a meta-probit model with a dummy as dependent variable that is 0 if $p < 0.05$, and 1 otherwise. This serves as a robustness test that narrows the analysis on two categories: *p*-values below and above a standard significance threshold. Although loosing valuable information about the size of individual *p*-values, the meta-probit approach has been applied in several previous meta-studies (among

[11] We consider all variables from Model II that are significant at least at the 5% level.

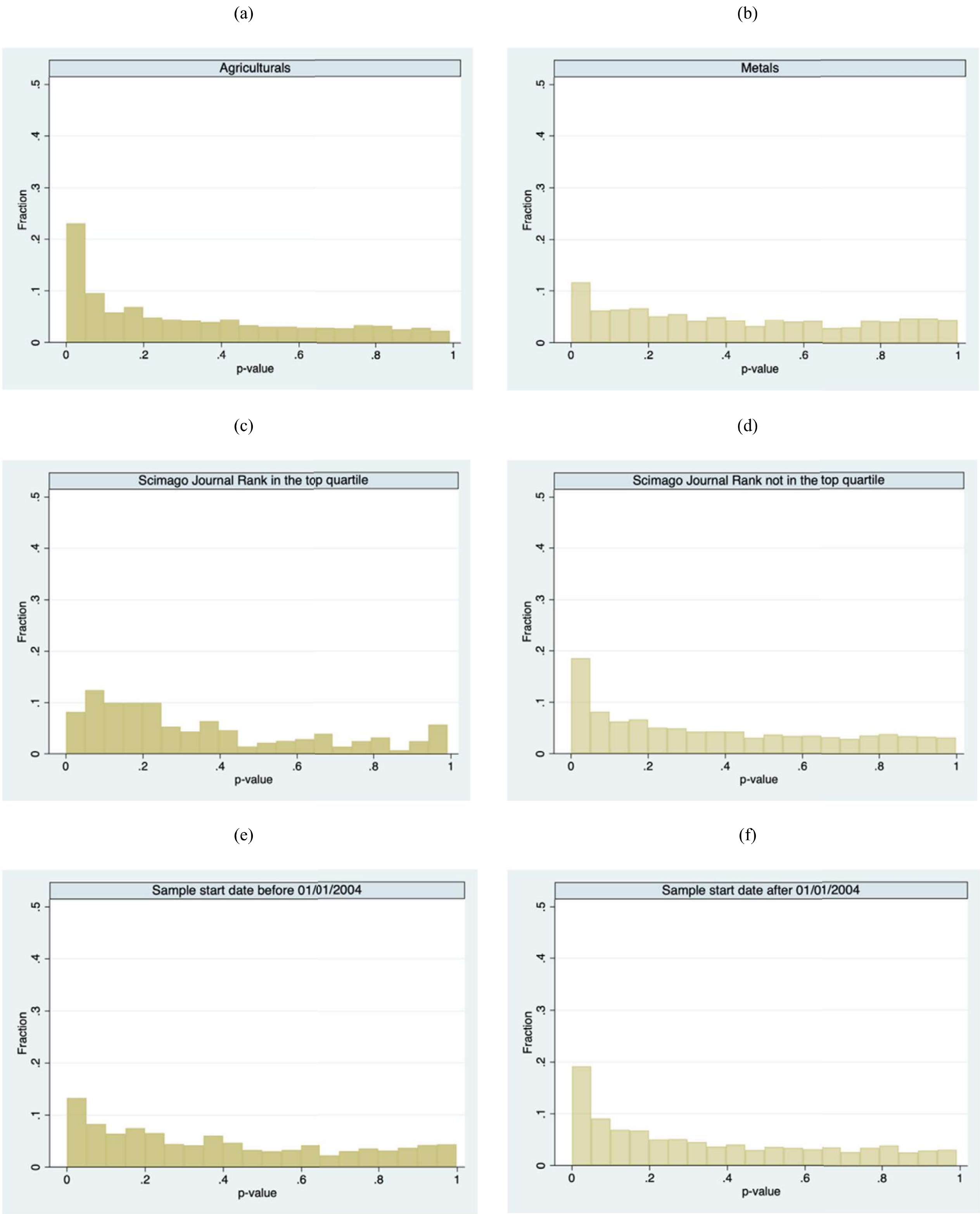

Fig. 2. Distribution of *p*-values from subsamples of GC tests. *Notes:* The figures show the histogram of *p*-values from subsamples of GC tests reported in primary studies: (a) GC tests that examine agricultural commodities and (b) GC tests that examine metals, (c) studies published in journals with SJR in the top quartile, (d) studies published in lower ranked or unranked journals and unpublished studies, (e) GC tests with sample start date before January 01, 2004, (f) GC tests with sample start date equal to or after January 01, 2004.

others, Horváthová, 2010; Koetse et al., 2009; Kysucky and Norden, 2016). For Models I to VI, we report the *t*-statistics based on standard errors clustered on study as well as on author level, the standard errors in Model VII are only clustered on study-level. The

Table 4
Analysis of heterogeneity.

	Linear						Probit
Model	(I)	(II) (Baseline)	(III)	(IV)	(V)	(VI) (Reduced)	(VII)
Weighting	Unweighted	$[1/J_i]$	$[1/J_i]$	$[\sqrt{df_{ij}}]$	$[1/\widehat{SE}^2_{ij}]$	$[1/J_i]$	$[1/J_i]$
Constant	−0.408*	0.262	0.576	−0.424	−0.455**	0.212	2.194***
	(-1.67)	(0.55)	(1.14)	(-1.43)	(-2.08)	(0.48)	(3.48)
$\sqrt{df}$	−0.022***	−0.065***	−0.064***	−0.022**	−0.012	−0.058***	−0.066***
	(-2.58)	(-4.69)	(-4.38)	(-1.96)	(-1.50)	(-3.79)	(-3.26)
#Lags	−0.003	−0.003	−0.004	−0.002	−0.012		−0.006
	(-1.17)	(-0.89)	(-1.00)	(-0.98)	(-0.66)		(-1.27)
Sample start date	−0.089**	−0.223**		−0.028	−0.080**	−0.217**	−0.443***
	(-2.08)	(-2.44)		(-0.36)	(-2.36)	(-2.45)	(-3.16)
Sample start date after 2004			−0.579**				
			(-2.52)				
Daily	0.300**	1.188***	1.222***	0.385**	0.027	1.100***	0.950**
	(2.26)	(3.26)	(3.11)	(2.42)	(0.19)	(2.83)	(2.07)
Monthly	0.039	0.015	0.020	0.307	0.029	0.080	−0.280
	(0.19)	(0.05)	(0.07)	(1.11)	(0.15)	(0.34)	(-0.64)
Weekly				Omitted base group			
Agriculturals	−0.583***	−0.717***	−0.636**	−0.801***	−0.384***	−0.749**	−0.394
	(-4.24)	(-2.80)	(-2.10)	(-4.21)	(-2.83)	(-2.48)	(-1.22)
Energy	−0.679***	−1.072***	−1.009***	−0.871***	−0.406**	−1.111***	−0.663*
	(-2.99)	(-3.15)	(-2.76)	(-3.19)	(-2.19)	(-2.81)	(-1.75)
Metals				Omitted base group			
Return	0.378***	0.592***	0.652***	0.457***	0.374***	0.544***	0.377**
	(4.99)	(2.67)	(2.92)	(4.28)	(4.60)	(2.77)	(2.08)
Spread	0.563***	0.917***	1.053***	0.548***	0.583***	0.798***	0.790
	(3.57)	(3.62)	(4.22)	(3.36)	(3.40)	(3.16)	(1.56)
Volatility				Omitted base group			
VAR/VEC	0.030	0.139	0.130	0.048	−0.130		0.216
	(0.30)	(0.83)	(0.75)	(0.42)	(-1.19)		(1.28)
ADL				Omitted base group			
Controls	0.507***	0.912***	1.149***	0.544***	0.482***	0.856***	0.822**
	(4.54)	(2.59)	(3.17)	(3.66)	(4.84)	(3.62)	(2.22)
Multivariate GC	−0.203	−0.369	−0.376	−0.193	−0.321		−0.441
	(-0.87)	(-1.03)	(-1.08)	(-0.72)	(-0.96)		(-1.62)
Contemporaneous	−0.448	−0.751	−0.796	−0.135	−0.362		−0.162
	(-1.07)	(-0.95)	(-1.04)	(-0.39)	(-1.05)		(-0.37)
Non-parametric	0.146	0.022	−0.173	0.203	0.045		0.780***
	(1.53)	(0.10)	(-0.72)	(1.52)	(0.39)		(3.30)
Higher ranked journal	0.414	0.991**	0.885**	0.393	0.296	0.904***	0.775**
	(1.41)	(2.38)	(2.02)	(1.34)	(1.12)	(2.58)	(2.05)
Academic authors only	0.360	0.155	0.058	0.382	0.125		0.015
	(1.49)	(0.42)	(0.16)	(1.38)	(0.58)		(0.04)
Large futures market	−0.019	0.279*	0.255*	−0.073	−0.049		0.168
	(-0.39)	(1.80)	(1.71)	(-1.03)	(-1.05)		(1.03)
Major exchange	0.179	−0.305	−0.211	0.196	0.141		−0.881***
	(1.03)	(-1.53)	(-1.05)	(1.24)	(0.86)		(-2.86)
First differences	0.097	0.338**	0.367**	0.077	0.124	0.369**	0.307
	(0.99)	(2.00)	(2.05)	(0.62)	(1.04)	(2.12)	(1.54)
Position and flow data	0.127	0.089	−0.055	0.273*	0.093		0.339
	(1.48)	(0.29)	(-0.21)	(1.91)	(1.54)		(1.09)
#Obs.	2106	2106	2106	2106	2106	2106	2106
#Studies	54	54	54	54	54	54	54

Notes: This table reports the results of Eq. (5): $\tilde{p}_{ij} = \alpha + \beta_1 \sqrt{df_{ij}} + \beta_2 lags_{ij} + \sum_{l=1}^{L} \gamma_l Z_{l,ij} + \varepsilon_{ij}$, where $\tilde{p}_{ij}$ is the probit-transformed p-value and df_{ij} the degrees of freedom of the corresponding GC test i of study j. The variable $lags_{ij}$ refers to the lag length of the speculation activity measure x_{ij}. $Z_{l,ij}$ is the variable referring to moderator l of GC test i of study j, with $l \in$ (*Average time*, *Daily*, …, *Position and flow data*). Model II, III, VI and VII are estimated by weighted least squares using the inverse of the number of estimates as weights. Model I is an unweighted regression, while Model IV and V use the square root of the degrees of freedom and the inverse of the squared approximated standard errors as weights. Model VI is a reduced model including only the most significant variables. Model VII serves as a robustness check, where we group p-values into a dummy variable of non-significant and significant values with a threshold of 0.05. The t-statistics of the regression coefficients reported in parentheses are based on standard errors adjusted for within-study and across-study correlation. $^{*}p < 0.1, ^{**}p < 0.05, ^{***}p < 0.01$.

Meta-GC coefficients reported in Table 4 indicate the impact of the moderator variables on the *p*-values collected from the primary studies' GC tests, i.e., the coefficients measure changes or differences in the collected *p*-values. To assess the impact of speculation we derive absolute *p*-values conditional on specific values of the moderator variables in section 5.4.

Coefficients of the degrees of freedom ($\sqrt{df}$) are significantly negative, i.e., larger sample sizes lead to lower probit-transformed *p*-values in the literature. However, in the light of the discussion in Section 5.2, the interpretation of $\sqrt{df}$ as a test for publication bias is different for *p*-values below than for *p*-values above 0.05. Since the outcome is the result of simultaneously testing for two directions of publication biases, it cannot be regarded as evidence for or against *p*-hacking. Consistent with our previous findings, lag selection of the speculation measure (*#Lags*) does not influence GC test results on speculation effects.

In line with Mayer et al. (2017), we find that GC tests observed from studies with more recent samples are more likely to discover a price impact of speculation. We define the *Sample start date* as *z*-score of the sample start date of the primary GC test. The size of the time effect becomes more visible if we compare, e.g., a test with a hypothetical sample start date on September 02, 1994 to a test with a sample start date on February 23, 2011. As these exemplary samples are 1.5 standard deviations below, respective above, the mean start date of August 17, 2002, the regression coefficients must be multiplied by 3 to obtain the difference in resulting probit-transformed *p*-values. Our results imply that the probability of speculation effects increased over time. This supports the view that increased trading activity of financial investors beginning around 2004 lead to the financialization of commodity markets (Tang and Xiong, 2012). When using the dummy variable *Sample start date after 2004* instead of a continuous time variable (Model III), we find significantly lower *p*-values in GC tests with samples starting after January 01, 2004. However, one must bear in mind that the coefficient only measures the relative difference in *p*-values between samples after and before January 01, 2004. We shed light on the *p*-values' absolute levels to assess the presence of speculation effects with best choice models in section 5.4.

Furthermore, we observe statistical differences if the primary test is based on time series data with daily frequency. The results for the variable *Daily* indicate higher *p*-values as compared to models using weekly data (omitted base group). However, the sign of the regression coefficient for daily frequencies seems counterintuitive. As our base group is weekly frequency, we would expect lower *p*-values for higher frequencies: assuming that speculators are often short-term oriented (Froot et al., 1992; Kang et al., 2019), price effects might be more present in shorter frequencies. For the analyses of monthly data sets, we do not observe significant differences in the likelihood of speculation leading prices as compared to weekly tests.

The results suggest significant differences between commodity types. Primary tests that analyze speculation effects on *Agriculturals* find significantly lower *p*-values than for *Metals*. Hence, the literature detects more often speculation effects for agriculturals than for metals. The same holds for *Energy* commodities, such as natural gas, crude or processed oil, which exhibit also significantly lower *p*-values from GC tests as compared to metals. Thus, we may conclude that agricultural and energy commodity prices are more driven by speculation than metal prices. This result is in line with the panel GC results (Table 2). There, we also find non-significance for several industrial metals, which indicates that not a single test rejected Granger non-causality. We suppose that LME traded metals (like lead, nickel, tin and zinc) are less in the focus of financial investors resulting in less vulnerability to non-commercial trading activities. In contrast, precious metals have been traditionally traded by financial investors already before the 2000s, e.g., as save haven investment (Baur and Lucey, 2010). Hence, we expect that the impact of financialization on precious metal markets, which is the dominant metals group in primary studies, is less pronounced, leading to less significant results.

In addition, our results also suggest that GC effects from speculation to commodity prices are less likely to be found in price returns (*Returns*) than in price volatilities. The difference between returns and volatilities as dependent variable in the primary tests is significant and relatively large across all models. Hence, it is more likely to detect GC from speculation to volatility than to returns. The GC testing strategy allows only to determine if volatility is changing, but not whether volatility is increasing or decreasing with higher trading activity. However, evidence from market microstructure research shows that increased liquidity, as a result of increased trading activity, reduces the bid-ask spread and lastly reduces volatility (Hasbrouck, 2007; O'Hara, 2008). Hence, speculation might not significantly change returns, but might significantly decrease volatility as a consequence of increased liquidity. The finding of larger effects of speculation on volatility is in line with Brunetti and Büyükşahin (2009), Brunetti et al. (2016), Mayer et al. (2017), Prokopczuk et al. (2014) and Sanders and Irwin (2011), that also detect a reduction in volatility in some markets while mostly no effects on returns.

Although rarely, some primary studies apply models with term spreads (*Spread*) as dependent variable to analyze speculation effects (among others, Alquist and Gervais, 2013; Irwin et al., 2011; Naderian and Javan, 2017). We observe that using spreads is associated with higher, less significant *p*-values as compared to using volatilities as a response variable in GC tests. This finding emphasizes the previous result of a lower probability of speculation effects on returns than on volatilities, as term spreads are also based on prices. However, following the ‘preferred habitat hypothesis’ in bond markets (Modigliani and Sutch, 1966), investors prefer futures with specific maturities. If this also holds for commodity speculators, term spreads should increase, and one might have expected lower *p*-values for term spread effects in the primary literature.

Regarding the econometric specification of the primary study tests, we differentiate between ADL and VAR (or VEC) models. In contrast to our expectations, vector regressions (*VAR/VEC*) that account for bi-directionality of the relationship between speculation and commodity prices do not result in significantly different *p*-values than standard ADL models.

Prior studies that use control variables in their GC framework end up with less significant results. All five models show highly significant and large coefficients for *Controls*. Consequently, adding control variables, e.g., market fundamentals (Shanker, 2017), to the GC test leads to less findings of GC from speculation to prices. This can be reasoned by the fact that omitting relevant variables leads to biased estimates of the regression coefficients followed by biased *t*-statistics and *p*-values.

Some studies analyze multiple trader positions at once in the same model. However, the meta-estimates for the *Multivariate GC* variable do not imply that there is a systematic difference between those multivariate models and regular GC tests examining just one speculation measure. Occasionally, studies also incorporate a lag of zero of the speculation measures in their model setups (among others, Gilbert and Pfuderer, 2014; Irwin and Sanders, 2012; Irwin et al., 2011). These studies aim to capture the contemporaneous relationship between speculation and commodity prices, as market microstructure theory suggests that trading impacts are instantaneous (O'Hara, 2008). Our results reveal that adding the speculation variable at time t (*Contemporaneous*) indeed reduces p-values, yet not significantly. Furthermore, the application of non-parametric frameworks that examine non-linear relationships mostly show no significant differences to the standard linear GC test. Only Model VI reveals a significant positive relation for the variable *Non-parametric*. However, the probit approach is less granular than the Meta-Granger models (Models I to V), and therefore, less reliable.

Another driver of variation within our meta-sample is given by the publication outlet of the primary study. Articles published in journals with an SJR in the top quartile are less likely to present significant p-values as compared to lower ranked or unranked journal publications and unpublished studies. This is indicated by the positive coefficients of the variable *Higher ranked journal*. However, the estimates are not consistently significant across all models. As we observe significance at the 5% level in our preferred model (Model II) and the reduced model (Model V), we see this result as indicator that a GC test that is published in a higher ranked journal is associated with a substantially larger p-value on average. The results for this variable are considerable because the significance is present even though we control for several other variables, which characterize an article in a journal of high quality. Albeit having a large data set with adequate treatment, using controls and applying more advanced methods, the higher ranked journal variable shows significant results. Consequently, there must be additional factors in higher ranked journals that drive reported results. Since the publication process attempts to ensure the compliance of academic research with certain standards in conducting, reporting and interpreting scientific outcome, we regard papers published in top journals as more trustworthy. Looking at the peer review process in top journals, the reviewers and editors might ensure a more rigorous process to guarantee that data are treated with care, the right set of controls are added, or methods are applied in an adequate manner as compared to, e.g., to unpublished studies without peer review. These factors cannot be measured by the other Meta-GC variables and, thus, might be reflected in our results for higher ranked journals. Another explanation might stem from the fact that the first studies in higher ranked journals mostly report no GC effects from speculation to prices. Consequently, as this result was once established, it might be favorable for authors to not disagree in GC tests in order to increase the probability of getting accepted by those journals. However, we consider the reviewer's quality hypothesis as more reliable than the hypothesis of asserting established results.

Furthermore, we include a variable to analyze the effect of research independence. If study authors are exclusively affiliated to academic institutions and receive no external funding, reported p-values are higher according to the results for the variable *Academic authors only*. An external affiliation or funding might be associated with pressure to publish significant findings to fulfil potential expectations of a third party. However, we cannot provide evidence for this hypothesis as our coefficients are non-significant throughout all meta-models.

Surprisingly, we find that overall liquidity in the respective commodity futures market has no influence on speculation effects. Liquidity is estimated by two dummy variables, *Large futures market* and *Major exchange*, which both exhibit no significant effects on reported p-values in primary studies.

We observe some significant results for first differenced speculation measures. Reported p-values are higher for *First differences*, supporting our expectations that differencing should remove non-stationarity and thus reduces the probability of spurious findings. However, we have to note that this effect is only significant in our main Model II, Model III and the reduced Model VI.

Lastly, the choice of the speculation measure seems to have no significant influence according to the variable *Position and flow data*. Using non-commercial trader or CIT position data or flows into commodity funds does not lead to different results than using speculation measures based on open interest and trading volume. On the one hand, this finding is surprising as the latter measures cannot explicitly separate out commercial trading activities. On the other hand, this could indicate that the categorization of trader groups or market participants in general is challenging and as a consequence results do not significantly differ.

5.4. Best choice model

The results of the previous section provide explanations for the heterogeneity in primary studies. Nevertheless, it is challenging to estimate a mean implied p-value and to derive an ‘average statement’ about the presence of speculation effects. This is because the p-value is conditional on study and test characteristics as shown in the previous results for the heterogeneity analysis. To derive practical implications, we follow Stanley and Doucouliagos (2012) and create a ‘best choice’ model by inserting best practice values for the explanatory variables. We then predict the probit-retransformed p-values based on the regression coefficients from our reduced baseline model in the meta-regression (Model V in Table 4).

We define the ‘best practice’ case as follows. In general, a large sample size and hence many degrees of freedom are desirable for any empirical analysis. However, as we estimated the coefficient of $\sqrt{df}$ based on our collected sample of primary tests, we insert the mean of the primary studies' square root of degrees of freedom as best practice value. The best practice value refers to a weekly test with over 600 degrees of freedom implying a sample period of nearly 12 years. We follow the majority of empirical studies and assume weekly models due to primary data availability of most studies. For *Sample start date* we consider three hypothetical dates to visualize differences over time. The two samples starting on January 01, 1971 and on January 01, 2015 represent the earliest and latest sample start dates of the primary studies' tests included in our meta-analysis. Additionally, we include January 01, 2004 to explicitly cover a sample that starts at the beginning of the financialization of commodity markets in 2004. For the commodity class, price type (return and volatility) and journal quality, we report results conditional on varying choices. Ideally, a GC test should include controls to prevent omitted variable bias. However, traditional GC analysis comprises only two variables, and only few studies implement additional variables (Aulerich et al., 2010, 2014; Etienne et al., 2017; Shanker, 2017). Therefore, we offer results for both models, with and without control variables. We apply a speculative measure in first differences to prevent spurious findings. This choice is also supported by tests in the primary literature that mostly cannot reject non-stationarity of position data (Malliaris and Urrutia, 1998; Brunetti and Büyükşahin, 2009; Irwin and Sanders, 2012; Sehgal et al., 2012). The results for the best practice estimates are reported in Table 5.

The predictions reported in Table 5 can be interpreted as the mean *p*-value given a study with best practice research design after controlling for publication selection and lag selection bias. Panel A shows that tests including control variables cannot reject Granger non-causality in any model setup. This implies that aggregating the literature about GC and using the integrated knowledge from an analysis of 2106 primary tests delivers a very clear conclusion of no speculation effects on commodity prices in the mean. Even after considering various study and test characteristics, speculation in metal, energy and agricultural markets does not Granger-cause commodity returns or volatilities when control variables are included in the GC model. Panel B demonstrates that this unambiguous finding largely holds for a (less preferred) model setup without control variables as well. Only in the case of energy and agricultural commodities, when analyzing the two more recent samples (starting in 2004 and 2015) and if the study is published in a lower ranked journal, speculation effects on volatility at the 5% significance level can be detected. Additionally, in the same specific setup, the best choice model predicts that speculation significantly Granger-causes returns in samples with a start date in 2015. Hence, only six – less preferred, due to the lack of controls and lower publication quality – model setups out of 72 deliver significant GC results.

Overall, we do not find evidence that speculation genuinely Granger-causes commodity price returns or volatilities. This result implies either that speculation genuinely does not drive overall commodity prices or that primary studies' and test designs have low power to identify speculation effects. Issues might especially arise from data availability. Even daily data for the speculative measure – which we found to be associated with higher *p*-values than weekly data – might have an unsuitable data frequency to systematically detect speculation effects. Furthermore, the variety of applied speculation measures shows that there is no single method that perfectly identifies speculative activity. Consequently, identification of effects from speculation remains challenging, as discussed in previous literature (Büyükşahin and Harris, 2011; Alquist and Gervais, 2013). Another explanation why Granger causality tests cannot provide empirical evidence for speculation effects might stem from the failure to account for multiple roles and motives of non-commercial traders (Cheng et al., 2015; Kang et al., 2019). According to Cheng et al. (2015), prices might decrease when speculators increase their positions to accommodate the hedging pressure of commercial traders, whereas an increase in positions of speculators for other trading purposes might increase prices – possibly resulting in no overall effect on average.

Table 5
Best choice model.

Publication quality	Sample start date	Price	Metals	Energy	Agriculturals
Panel A: p-values of model setups with control variables					
Journal with an SJR in the top quartile	1971	Return	0.996	0.939	0.972
		Volatility	0.983	0.842	0.914
	2004	Return	0.919	0.613	0.742
		Volatility	0.804	0.399	0.542
	2015	Return	0.836	0.447	0.590
		Volatility	0.668	0.249	0.376
Lower ranked journal or unpublished	1971	Return	0.960	0.740	0.843
		Volatility	0.887	0.540	0.678
	2004	Return	0.690	0.269	0.400
		Volatility	0.480	0.123	0.212
	2015	Return	0.529	0.150	0.250
		Volatility	0.319	0.057*	0.111
Panel B: p-values of model setups without control variables					
Journal with an SJR in the top quartile	1971	Return	0.964	0.756	0.854
		Volatility	0.896	0.559	0.695
	2004	Return	0.706	0.285	0.418
		Volatility	0.499	0.133	0.227
	2015	Return	0.548	0.161	0.265
		Volatility	0.336	0.063*	0.121
Lower ranked journal or unpublished	1971	Return	0.816	0.416	0.560
		Volatility	0.639	0.225	0.347
	2004	Return	0.359	0.070*	0.133
		Volatility	0.183	0.022**	0.049**
	2015	Return	0.217	0.029**	0.063**
		Volatility	0.092*	0.007***	0.019**

Notes: This table presents hypothetical *p*-values of our best choice models in different setups depending on publication quality, sample start, commodity price measurement, as well as commodity type. The sample start dates refer to hypothetical samples that start on January 01 of the respective year and have a length of the primary studies' median sample length of a single GC test of 3.99 years. 'Return' and 'Volatility' refer to the dependent variable used in the primary studies' GC tests.

6. Conclusion

In this study, we apply Meta-Granger analysis to systematically aggregate and compare 54 primary studies reporting 2106 GC test results for the impact of financial speculation on commodity prices. The analysis is conducted in four steps.

First, we apply panel GC methods to aggregate the reported primary study tests separately for each commodity type. The results suggest that there is heterogeneity among the test results driven by the commodity class under examination and by the measurement of commodity prices (returns/spreads vs. volatilities).

Second, we explore if selective reporting of (non-)significant test results or over-/underfitting via lag selection are present in the literature. When applying the Meta-GC model by Bruns and Stern (2019), we do not detect any indication for publication bias towards significance. For the subsample of non-significant results, we find no clear evidence for publication bias either, although this finding is not as strong as for the subsample of significant results. Unlike many other research fields, publication bias seems to be no notable issue in this literature strand. Moreover, we do not find strong support for the presence of lag selection bias.

Third, we use a Meta-GC model with moderator variables capturing various primary study characteristics and methodological differences of the test design in order to explain the disparate findings. Our results show that GC from speculation is more present in agricultural and energy than in metal markets. Speculative activity is more likely to Granger-cause volatilities than returns or term spreads. Next, we find empirical evidence supporting the financialization hypothesis, as tests that use more recent samples are more likely to reject Granger non-causality from speculation to commodity prices. Moreover, we advocate enriching tests with control variables in order to reduce omitted variable bias that leads to an overrejection of Granger non-causality. Furthermore, one of the largest drivers of variation within our meta-sample is the journal quality of the primary study. Journal articles with an SJR in the top quartile are less likely to present significant *p*-values compared to articles in lower ranked or unranked journals and unpublished studies. Thus, even after controlling for more than twenty other factors, with some even already capturing the primary study's quality, the journal impact plays a substantial role. This finding could exclusively be uncovered by meta-analysis methods as they allow a 'macro' perspective on the previous literature and its sources of heterogeneity.

Fourth, we compute best choice models derived from the heterogeneity analysis. We observe that the hypothesis of Granger non-causality between speculation and commodity returns or volatilities cannot be rejected at standard significance levels. Therefore, when aggregating the entire literature of 2106 test results and assuming a best practice study design, our Meta-GC results indicate that there are no overall effects from speculation in three commodity markets, namely agriculturals, metals and energy. The results are robust against alternative definitions of best choice models. Hence, we might conclude from this meta-study that either there is genuinely no overall effect, or the research design of GC testing is not powerful enough to detect those effects.

Declaration of competing interest

Besides the PhD studies, the first author works at MEAG Munich Ergo AssetManagement GmbH, an asset management company investing in a broad range of asset classes. The PhD studies as well as this research project are not funded by MEAG Munich Ergo AssetManagement GmbH. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect those of MEAG Munich Ergo AssetManagement GmbH, its subsidiaries or affiliate companies.

Appendix A. Studies included in the meta-analysis

Algieri, B., 2016. Conditional price volatility, speculation, and excessive speculation in commodity markets: sheep or shepherd behaviour? International Review of Applied Economics 30 (2), 210–237.

Alquist, R., Gervais, O., 2013. The Role of Financial Speculation in Driving the Price of Crude Oil. The Energy Journal 34 (3), 35–54.

Antonakakis, N., Chang, T., Cunado, J., Gupta, R., 2018. The relationship between commodity markets and commodity mutual funds: A wavelet-based analysis. Finance Research Letters 24, 1–9.

Aulerich, N.M., Irwin, S.H., Garcia, P., 2010. The Price Impact of Index Funds in Commodity Futures Markets: Evidence from the CFTC's Daily Large Trader Reporting System. Working Paper at the University of Illinois.

Aulerich, N.M., Irwin, S.H., Garcia, P., 2014. Bubbles, Food Prices, and Speculation: Evidence from the CFTC's Daily Large Trader Data Files. In: Chavas, J.-P., Hummels, D., Wright, B.D. (Eds.): The Economics of Food Price Volatility. University of Chicago Press, Chicago.

Babalos, V., Balcilar, M., 2017. Does institutional trading drive commodities prices away from their fundamentals: Evidence from a nonparametric causality-in-quantiles test. Finance Research Letters 21, 126–131.

Bohl, M.T., Siklos, P.L., Wellenreuther, C., 2018. Speculative activity and returns volatility of Chinese agricultural commodity futures. Journal of Asian Economics 54, 69–91.

Borin, A., Di Nino, V., 2012. The Role of Financial Investments in Agricultural Commodity Derivatives Markets. Working Paper at the Bank of Italy.

Bos, J.W.B., van der Molen, M., 2012. A Bitter Brew? Futures Speculation and Commodity Prices. Working Paper at Maastricht University School of Business and Economics.

Brunetti, C., Büyükşahin, B., 2009. Is Speculation Destabilizing? Working Paper at the Johns Hopkins University.

Brunetti, C., Büyükşahin, B., Harris, J.H., 2016. Speculators, Prices, and Market Volatility. Journal of Financial and Quantitative Analysis 51 (5), 1545–1574.

Bu, H., 2011. Price dynamics and speculators in crude oil futures market. Systems Engineering Procedia 2, 114–121.

Büyükşahin, B., Harris, J.H., 2011. Do speculators drive crude oil futures prices? The Energy Journal 32 (2), 167–202.
Capelle-Blancard, G., Coulibaly, D., 2011. Index Trading and Agricultural Commodity Prices: A Panel Granger Causality Analysis. International Economics 126-127, 51–71.
Chakraborty, R., Das, R., 2013. Dynamic Relationship Between Futures Trading and Spot Price Volatility: Evidence from Indian Commodity Market. Journal of Applied Finance 19 (4), 5–19.
Ciner, C., 2002. Information content of volume: An investigation of Tokyo commodity futures markets. Pacific-Basin Finance Journal 10 (2), 201–215.
Coleman, L., Dark, J., 2012. Economic significance of non-hedger investment in commodity markets. Working Paper at the University of Melbourne.
Ding, H., Kim, H.-G., Park, S.Y., 2014. Do net positions in the futures market cause spot prices of crude oil? Economic Modelling 41, 177–190.
Ederer, S., Heumesser, C., Staritz, C., 2013. The role of fundamentals and financialisation in recent commodity price developments: An empirical analysis for wheat, coffee, cotton, and oil. Working paper at the Austrian Foundation for Development Research.
Etienne, X.L., Irwin, S.H., Garcia, P., 2017. New evidence that index traders did not drive bubbles in grain futures markets. Journal of Agricultural and Resource Economics 42, 45–67.
Fagan, S., Gencay, R., 2008. Liquidity-induced dynamics in futures markets. Working Paper at the Simon Fraser University.
Fujihara, R.A., Mougou, M., 1997. An examination of linear and nonlinear causal relationships between price variability and volume in petroleum futures markets. Journal of Futures Markets 17 (4), 385–416.
Gilbert, C.L., 2009. Speculative influences on commodity futures prices 2006-2008. Working Paper at the University of Trento.
Gilbert, C.L., 2010. Commodity speculation and commodity investment. Food and agriculture organization of the United Nations, Rome.
Gilbert, C.L., 2010. How to Understand High Food Prices. Journal of Agricultural Economics 61 (2), 398–425.
Gilbert, C.L., Pfuderer, S., 2012. Index funds do impact agricultural prices. Working Paper at the University of Trento.
Gilbert, C.L., Pfuderer, S., 2014. The Role of Index Trading in Price Formation in the Grains and Oilseeds Markets. Journal of Agricultural Economics 65 (2), 303–322.
Gupta, C.P., Sehgal, S., Wadhwa, S., 2018. Agricultural Commodity Trading: Is it Destabilizing Spot Markets? The Journal for Decision Makers 43 (1), 47–57.
Haase, M., Seiler Zimmermann, Y., Zimmermann, H., 2018. Permanent and transitory price shocks in commodity futures markets and their relation to speculation. Empirical Economics 32 (2), 167.
Huchet, N., Fam, P.G., 2016. The role of speculation in international futures markets on commodity prices. Research in International Business and Finance 37, 49–65.
Irwin, S.H., Garcia, P., Good, D.L., Kunda, E.L., 2011. Spreads and Non-Convergence in Chicago Board of Trade Corn, Soybean, and Wheat Futures: Are Index Funds to Blame? Applied Economic Perspectives and Policy 33 (1), 116–142.
Irwin, S.H., Sanders, D.R., 2010. The Impact of Index and Swap Funds on Commodity Futures Markets. OECD Food, Agriculture and Fisheries Papers No. 27.
Irwin, S.H., Sanders, D.R., 2012. Testing the Masters Hypothesis in commodity futures markets. Energy Economics 34 (1), 256–269.
Malliaris, A.G., Urrutia, J.L., 1998. Volume and price relationships: hypotheses and testing for agricultural futures. Journal of Futures Markets 18 (1), 53–72.
Mayer, H., Rathgeber, A., Wanner, M., 2017. Financialization of metal markets: Does futures trading influence spot prices and volatility? Resources Policy 53, 300–316.
Mayer, J., 2012. The Growing Financialisation of Commodity Markets: Divergences between Index Investors and Money Managers. Journal of Development Studies 48 (6), 751–767.
Merino, A., Albacete, R., 2010. Econometric modelling for short-term oil price forecasting. OPEC Energy Review 34 (1), 25–41.
Naderian, M.A., Javan, A., 2017. Distortionary effect of trading activity in NYMEX crude oil futures market: post crisis. OPEC Energy Review 41 (1), 23–44.
Obadi, S.M., Korecek, M., 2018. The Crude Oil Price and Speculations: Investigation Using Granger Causality Test. International Journal of Energy Economics and Policy 8 (3), 275–282.
Often, E.M., Wisen, C.H., 2013. Disaggregated commitment of Traders Data and prospective Price Effects. Journal of Applied Business Research 29 (5), 1381–1400.
Prokopczuk, M., Symeonidis, L., Verlaat, T., 2014. Rising and Volatile Food Prices: Are Index Fund Investors to Blame? Working Paper at the Zeppelin University.
Sanders, D.R., Boris, K., Manfredo, M., 2004. Hedgers, funds, and small speculators in the energy futures markets: an analysis of the CFTC's Commitments of Traders reports. Energy Economics 26 (3), 425–445.
Sanders, D.R., Irwin, S.H., 2011. New evidence on the impact of index funds in US grain futures markets. Canadian Journal of Agricultural Economics 59 (4), 519–532.
Sanders, D.R., Irwin, S.H., 2011. The Impact of Index Funds in Commodity Futures Markets: A Systems Approach. The Journal of Alternative Investments 14 (1), 40–49.
Sanders, D.R., Irwin, S.H., 2014. Energy futures prices and commodity index investment: New evidence from firm-level position data. Energy Economics 46, 57–68.
Sanders, D.R., Irwin, S.H., Merrin, R.P., 2009. Smart money: The forecasting ability of CFTC large traders in agricultural futures markets. Journal of Agricultural and Resource Economics, 276–296.

Sassi, M., Werner, H.A., 2013. Non-commercial actors and the recent futures prices of wheat. Economia e Diritto Agroalimentare 18 (3), 309–330.
Sehgal, S., Rajput, N., Dua, R.K., 2012. Futures trading and spot market volatility: evidence from Indian commodity markets. Asian Journal of Finance and Accounting 4 (2), 199–217.
Shanker, L., 2017. New indices of adequate and excess speculation and their relationship with volatility in the crude oil futures market. Journal of Commodity Markets 5, 18–35.
Shanmugam, V., Armah, P., 2012. Role of speculators in agricultural commodity price spikes during 2006-2011. Academy of Accounting and Financial Studies Journal 16, 97.
Sharma, D.K., Malhotra, M., 2015. Impact of futures trading on volatility of spot market-a case of guar seed. Agricultural Finance Review 75 (3), 416–431.
Sharma, T., 2016. The Impact of Future Trading on Volatility in Agriculture Commodity: A Case of Pepper. Journal of Financial Risk Management 13 (4), 47.
Stoll, H.R., Whaley, R.E., 2011. Commodity Index Investing: Speculation or Diversification? The Journal of Alternative Investments 14 (1), 50–60.
Yang, J., Balyeat, R.B., Leatham, D.J., 2005. Futures Trading Activity and Commodity Cash Price Volatility. Journal of Business Finance and Accounting 32 (1-2), 297–323.

Appendix B. Commodity market size

Large futures market	Small futures market	Large futures market	Small futures market	Large futures market	Small futures market
Agricultural		*Energy*		*Metals*	
Coffee	Barley	Brent oil		Aluminum	Lead
Corn	Castor seed	Crude oil		Copper	Nickel
Cotton	Chickpeas	Ethanol		Gold	Palladium
Lean hogs	Chili	Gasoline		Silver	Platinum
Live cattle	Cocoa	Heating oil			Tin
Maize	Cumin	Natural gas			Zinc
Mix	Feeder cattle				
Soybean oil	Guar seed				
Soybeans	Lumber				
Sugar	Mint oil				
Wheat (CBOT)	Mustard seed				
	Oat				
	Orange				
	Orange juice				
	Palm oil				
	Pepper				
	Pork				
	Potato				
	Rapseed oil				
	Rice				
	Rubber				
	Soybean flour				
	Sunflower oil				
	Turmeric				
	Wheat (KCBT)				
	Wheat (other exchanges)				

Notes: A commodity with a higher average trading volume than 30,000 contracts per day – during the years 2004–2013 – is classified as part of a large futures market. We use futures data from the Chicago Mercantile Exchange (CME), Chicago Board of Trade (CBOT) and the Intercontinental Exchange (ICE). Additionally, we use data from the Kansas City Board of Trade (KCBT) in the case of wheat futures.

References

Alquist, R., Gervais, O., 2013. The role of financial speculation in driving the price of crude oil. Energy J. 34, 35–54.
Aulerich, N.M., Irwin, S.H., Garcia, P., 2010. The Price Impact of Index Funds in Commodity Futures Markets: Evidence from the CFTC's Daily Large Trader Reporting System. Working Paper at the University of Illinois.
Aulerich, N.M., Irwin, S.H., Garcia, P., 2014. Bubbles, food prices, and speculation: evidence from the CFTC's daily large trader data files. In: Chavas, J.-P., Hummels, D., Wright, B.D. (Eds.), The Economics of Food Price Volatility. University of Chicago Press, Chicago, Illinois.
Baur, D.G., Lucey, B.M., 2010. Is gold a hedge or a safe haven? An analysis of stocks, bonds and gold. Financ. Rev. 45, 217–229.
Bell, D., Kay, J., Malley, J., 1996. A non-parametric approach to non-linear causality testing. Econ. Lett. 51, 7–18.
Bohl, M.T., Siklos, P.L., Wellenreuther, C., 2018. Speculative activity and returns volatility of Chinese agricultural commodity futures. J. Asian Econ. 54, 69–91.
Borenstein, M., Hedges, L.V., Higgins, J.P.T., Rothstein, H.R., 2009. Introduction to Meta-Analysis. John Wiley & Sons, Chichester.
Boyd, N.E., Harris, J.H., Li, B., 2018. An update on speculation and financialization in commodity markets. J. Commod. Mark. 10, 91–104.
Brodeur, A., Lé, M., Sangnier, M., Zylberberg, Y., 2016. Star wars: the empirics strike back. Am. Econ. J. Appl. Econ. 8, 1–32.

Brunetti, C., Büyükşahin, B., 2009. Is Speculation Destabilizing? Working Paper at the Johns Hopkins University.
Brunetti, C., Büyükşahin, B., Harris, J.H., 2016. Speculators, prices, and market volatility. J. Financ. Quant. Anal. 51, 1545–1574.
Bruns, S.B., 2017. Meta-regression models and observational research. Oxf. Bull. Econ. Stat. 79, 637–653.
Bruns, S.B., Stern, D.I., 2019. Lag length selection and p-hacking in Granger causality testing: prevalence and performance of meta-regression models. Empir. Econ. 56, 797–830.
Büyükşahin, B., Haigh, M.S., Harris, J.H., Overdahl, J.A., Robe, M.A., 2008. Fundamentals, Trader Activity and Derivative Pricing. U.S. Commodity Futures Trading Commission, Washington, DC. Technical report.
Büyükşahin, B., Harris, J.H., 2011. Do speculators drive crude oil futures prices? Energy J. 32, 167–202.
Cameron, A.C., Gelbach, J.B., Miller, D.L., 2011. Robust inference with multiway clustering. J. Bus. Econ. Stat. 29, 238–249.
Cheng, I.-H., Kirilenko, A., Xiong, W., 2015. Convective risk flows in commodity futures markets. Rev. Finance 19, 1733–1781.
Cheng, I.-H., Xiong, W., 2014. Financialization of commodity markets. Ann. Rev. Financ. Econ. 6, 419–441.
Chien, L.-C., 2018. A method for combining p-values in meta-analysis by gamma distributions. J. Appl. Stat. 46, 247–261.
Ciner, C., 2002. Information content of volume: an investigation of Tokyo commodity futures markets. Pac. Basin Finance J. 10, 201–215.
Dumitrescu, E.-I., Hurlin, C., 2012. Testing for Granger non-causality in heterogeneous panels. Econ. Modell. 29, 1450–1460.
Etienne, X.L., Irwin, S.H., Garcia, P., 2017. New evidence that index traders did not drive bubbles in grain futures markets. J. Agric. Resour. Econ. 42, 45–67.
Feld, L.P., Heckemeyer, J.H., Overesch, M., 2013. Capital structure choice and company taxation: a meta-study. J. Bank. Finance 37, 2850–2866.
Friedman, L., 2001. Why vote-count reviews don't count. Biol. Psychiatr. 49, 161–162.
Froot, K.A., Scharfstein, D.S., Stein, J.C., 1992. Herd on the street: informational inefficiencies in a market with short-term speculation. J. Finance 47, 1461–1484.
Fujihara, R.A., Mougou, M., 1997. An examination of linear and nonlinear causal relationships between price variability and volume in petroleum futures markets. J. Futures Mark. 71, 385–416.
Geyer-Klingeberg, J., Hang, M., Rathgeber, A., 2020. Meta-analysis in financial economics: opportunities, challenges, and contemporary applications. Int. Rev. Financ. Anal. 17, 101524.
Geyer-Klingeberg, J., Hang, M., Rathgeber, A., Stöckl, S., Walter, M., 2018a. What do we really know about corporate hedging? A meta-analytical study. Bus. Res. 11, 1–31.
Geyer-Klingeberg, J., Hang, M., Rathgeber, A.W., 2019. What drives financial hedging? A meta-regression analysis of corporate hedging determinants. Int. Rev. Financ. Anal. 61, 203–221.
Geyer-Klingeberg, J., Hang, M., Walter, M., Rathgeber, A., 2018b. Do stock markets react to soccer games? A meta-regression analysis. Appl. Econ. 50, 2171–2189.
Gilbert, C.L., Pfuderer, S., 2014. The role of index trading in price formation in the grains and oilseeds markets. J. Agric. Econ. 65, 303–322.
González-Pereira, B., Guerrero-Bote, V.P., Félix, Moya-Anegón, 2010. A new approach to the metric of journals' scientific prestige: the SJR indicator. J. Inf. 4, 379–391.
Granger, C.W.J., 1969. Investigating causal relations by econometric models and cross-spectral methods. Econometrica 37, 424–438.
Greely, D., Currie, J., 2008. Speculators, Index Investors, and Commodity Prices. Goldman Sachs Commodities Research.
Grosche, S.-C., 2014. What does Granger causality prove? A critical examination of the interpretation of Granger causality results on price effects of index trading in agricultural commodity markets. J. Agric. Econ. 65, 279–302.
Grossman, S.J., Stiglitz, J.E., 1980. On the impossibility of informationally efficient markets. Am. Econ. Rev. 70, 393–408.
Haase, M., Seiler Zimmermann, Y., Zimmermann, H., 2016. The impact of speculation on commodity futures markets – a review of the findings of 100 empirical studies. J. Commod. Mark. 3, 1–15.
Hamilton, J.D., Wu, J.C., 2014. Risk premia in crude oil futures prices. J. Int. Money Finance 42, 9–37.
Hang, M., Geyer-Klingeberg, J., Rathgeber, A., 2020. The Speed of Adjustment towards Target Capital Structure - A Meta-Regression Analysis. Working Paper at the University of Augsburg.
Hang, M., Geyer-Klingeberg, J., Rathgeber, A.W., Stöckl, S., 2018. Measurement matters—a meta-study of the determinants of corporate capital structure. Q. Rev. Econ. Finance 68, 211–225.
Hartung, J., 1999. A note on combining dependent tests of significance. Biom. J. 41, 849–855.
Hasbrouck, J., 2007. Empirical Market Microstructure: the Institutions, Economics and Econometrics of Securities Trading. Oxford University Press, Oxford, New York.
Havranek, T., Irsova, Z., 2017. Do borders really slash trade? A meta-analysis. IMF Econ. Rev. 65, 365–396.
Havranek, T., Stanley, T.D., Doucouliagos, H., Bom, P., Geyer-Klingeberg, J., Iwasaki, I., Reed, R.W., Rost, K., van Aert, R.C.M., 2020. Reporting guidelines for meta-analysis in economics. J. Econ. Surv. 34 (3), 469–475.
Head, M.L., Holman, L., Lanfear, R., Kahn, A.T., Jennions, M.D., 2015. The extent and consequences of p-hacking in science. PLoS Biol. 13, e1002106.
Heard, N.A., Rubin-Delanchy, P., 2018. Choosing between methods of combining p-values. Biometrika 105, 239–246.
Hedges, L.V., Olkin, I., 1985. Statistical Methods for Meta-Analysis. Academic Press, San Diego, California.
Hellwig, M.F., 1980. On the aggregation of information in competitive markets. J. Econ. Theor. 22, 477–498.
Hicks, J.R., 1939. The foundations of welfare economics. Econ. J. 49, 696–712.
Hirshleifer, D., 1988. Residual risk, trading costs, and commodity futures risk premia. Rev. Financ. Stud. 1, 173–193.
Horváthová, E., 2010. Does environmental performance affect financial performance? A meta-analysis. Ecol. Econ. 70, 52–59.
Hou, C.-D., 2005. A simple approximation for the distribution of the weighted combination of non-independent or independent probabilities. Stat. Probab. Lett. 73, 179–187.
Huchet, N., Fam, P.G., 2016. The role of speculation in international futures markets on commodity prices. Res. Int. Bus. Finance 37, 49–65.
Ioannidis, J.P.A., Stanley, T.D., Doucouliagos, H., 2017. The power of bias in economics research. Econ. J. 127, F236–F265.
Irwin, S.H., Garcia, P., Good, D.L., Kunda, E.L., 2011. Spreads and non-convergence in Chicago board of trade corn, soybean, and wheat futures: are index funds to blame? Appl. Econ. Perspect. Pol. 33, 116–142.
Irwin, S.H., Sanders, D.R., 2012. Testing the Masters Hypothesis in commodity futures markets. Energy Econ. 34, 256–269.
Kang, W., Rouwenorst, K.G., Tang, K., 2019. A tale of two premiums: the role of hedgers and speculators in commodity futures markets. J. Finance 75 (1), 377–417.
Keynes, J.M., 1923. A Tract on Monetary Reform. Macmillan, London.
Koetse, M.J., Groot, H.L.F. de, Florax, R.J., 2009. A meta-analysis of the investment-uncertainty relationship. South. Econ. J. 76, 283–306.
Krimsky, S., 2013. Do financial conflicts of interest bias research? Sci. Technol. Hum. Val. 38, 566–587.
Kysucky, V., Norden, L., 2016. The benefits of relationship lending in a cross-country context: a meta-analysis. Manag. Sci. 62, 90–110.
Lexchin, J., 2012. Sponsorship bias in clinical research. Int. J. Risk Saf. Med. 24, 233–242.
Malliaris, A.G., Urrutia, J.L., 1998. Volume and price relationships: hypotheses and testing for agricultural futures. J. Futures Mark. 18, 53–72.
Mann, C.C., 1994. Can meta-analysis make policy? Science 266, 960–962.
Masters, M., White, A., 2009. The 2008 Commodities Bubble: Assessing the Damage to the United States and its Citizens. Masters Capital Management and White Knight Research and Trading.
Mayer, H., Rathgeber, A., Wanner, M., 2017. Financialization of metal markets: does futures trading influence spot prices and volatility? Resour. Pol. 53, 300–316.
Mayer, J., 2012. The growing financialisation of commodity markets: divergences between index investors and money managers. J. Dev. Stud. 48, 751–767.
Modigliani, F., Sutch, R., 1966. Innovations in interest rate policy. Am. Econ. Rev. 56, 178–197.
Naderian, M.A., Javan, A., 2017. Distortionary effect of trading activity in NYMEX crude oil futures market: post crisis. OPEC Energy Rev. 41, 23–44.
Obadi, S.M., Korecek, M., 2018. The crude oil price and speculations: investigation using Granger causality test. Int. J. Energy Econ. Pol. 8, 275–282.
O'Hara, M., 2008. Market Microstructure Theory. Blackwell, Malden, Massachusetts.
Prokopczuk, M., Symeonidis, L., Verlaat, T., 2014. Rising and Volatile Food Prices: Are Index Fund Investors to Blame? Working Paper at the. Zeppelin University.
Rahim, N.A., Goodacre, A., Veld, C., 2014. Wealth effects of convertible-bond and warrant-bond offerings: a meta-analysis. Eur. J. Finance 20, 380–398.
Rusnak, M., Havranek, T., Horvath, R., 2013. How to solve the price puzzle? A meta-analysis. J. Money Credit Bank. 45, 37–70.

Sanders, D.R., Boris, K., Manfredo, M., 2004. Hedgers, funds, and small speculators in the energy futures markets: an analysis of the CFTC's Commitments of Traders reports. Energy Econ. 26, 425–445.
Sanders, D.R., Irwin, S.H., 2011. The impact of index funds in commodity futures markets: a systems approach. J. Altern. Investments 14, 40–49.
Sehgal, S., Rajput, N., Dua, R.K., 2012. Futures trading and spot market volatility: evidence from Indian commodity markets. Asian J. Finance Account. 4, 199–217.
Shanker, L., 2017. New indices of adequate and excess speculation and their relationship with volatility in the crude oil futures market. J. Commod. Mark. 5, 18–35.
Shutes, K., Meijerink, G.W., 2012. Food Prices and Agricultural Futures Markets: A Literature Review. Wageningen School of Social Sciences. WASS Working PAPER No. 3.
Simonsohn, U., Nelson, L.D., Simmons, J.P., 2014a. p-Curve and effect size: correcting for publication bias using only significant results. Perspect. Psychol. Sci. 9, 666–681.
Simonsohn, U., Nelson, L.D., Simmons, J.P., 2014b. P-curve: a key to the file-drawer. J. Exp. Psychol. Gen. 143, 534–547.
Singleton, K.J., 2014. Investor flows and the 2008 boom/bust in oil prices. Manag. Sci. 60, 300–318.
Stanley, T.D., 2001. Wheat from Chaff: meta-analysis as quantitative literature review. J. Econ. Perspect. 15, 131–150.
Stanley, T.D., 2007. Meta-regression methods for detecting and estimating empirical effects in the presence of publication selection. Oxf. Bull. Econ. Stat. 70, 103–127.
Stanley, T.D., Doucouliagos, H., 2012. Meta-regression Analysis in Economics and Business. Routledge, London.
Stanley, T.D., Doucouliagos, H., Giles, M., Heckemeyer, J.H., Johnston, R.J., Laroche, P., Nelson, J.P., Paldam, M., Poot, J., Pugh, G., Rosenberger, R.S., Rost, K., 2013. Meta-Analysis of economics research reporting guidelines. J. Econ. Surv. 27, 390–394.
Stanley, T.D., Jarrell, S.B., Doucouliagos, H., 2010. Could it Be better to discard 90% of the data? A statistical paradox. Am. Statistician 64, 70–77.
Stoll, H.R., Whaley, R.E., 2011. Commodity index investing: speculation or diversification? J. Altern. Investments 14, 50–60.
Tang, K., Xiong, W., 2012. Index investment and financialization of commodities. Financ. Anal. J. 68, 54–74.
U.S. Senate, 2009. Excessive speculation in the wheat market. Majority and minority staff report. Perm. Subcommittee Invest. 24, 107–108.
van Aert, R.C.M., Wicherts, J.M., van Assen, M.A.L.M., 2016. Conducting meta-analyses based on p values: reservations and recommendations for applying p-uniform and p-curve. Perspect. Psychol. Sci. 11, 713–729.
van Ewijk, C., Groot, H.L.F. de, 2012. A meta-analysis of the equity premium. J. Empir. Finance 19, 819–830.
Will, M.G., Prehn, S., Pies, I., Glauben, T., 2016. Is financial speculation with agricultural commodities harmful or helpful? A literature review of current empirical research. J. Altern. Investments 18, 84–102.
Zigraiova, D., Havranek, T., 2016. Bank competition and financial stability: much ado about nothing? J. Econ. Surv. 30, 944–981.