ISSN No. 2454 - 1427

CDE
July 2026

# Tokenized Gold Under Stress: Divergent Transaction Costs in Thin Gold Markets During Crypto-to-Gold Rotations

**Reetika Garg**
Email: rgarg@econdse.org
Department of Economics,
Delhi School of Economics

**Srishti Jaiswal**
Email: srishtijaiswal@econdse.org
Department of Economics,
Delhi School of Economics

**Working Paper No. 366**

Centre for Development Economics
Delhi School of Economics
Delhi- 110007

# Tokenized Gold Under Stress:

## Divergent Transaction Costs in Thin Gold Markets During Crypto-to-Gold Rotations

**Reetika Garg***
**Srishti Jaiswal†**

## Abstract

We examine whether tokenized gold remains cost-effective to trade during crypto market sell-offs, when investors are most likely to shift from Bitcoin to the relative safety of gold. A safe haven that becomes expensive to transact in a crisis, is only a partial safe haven, yet we know little about how these instruments behave under stress. Using 1,574 daily observations (September 2019-December 2025) on tokenized gold (PAX Gold, PAXG; Tether Gold, XAUT), a deep gold ETF (GLD), and junior gold miners (GDXJ), we set a tokenized thin market against a conventional but equally thin one to separate what is specific to tokenization from what merely reflects a small market. Our dependent variable is the daily high-low range, which we treat throughout as a proxy for the cost of transacting under stress.

Our primary result comes from the difference-in-differences (DiD) specification that absorbs a full set of 1,574 date fixed effects and so removes the common gold-price, volatility, and information shock that hits every gold vehicle on a given day. On Bitcoin to gold rotation days, the two thin vehicles move in opposite relative directions: the cost of transacting tokenized gold rises against GLD while for the junior-miner it falls.

To understand why the spreads diverge, we estimate them jointly in a seemingly-unrelated-regression system. We find that the one force that formally separates the two is equity-market stress (VIX): the miner spread depends heavily on the VIX while the token spread does not. We read this as relative insulation of tokenized gold from equity fear, not immunity, because the ETF and the token co-move with the VIX by similar amounts and their difference nets out. Every other stressor, bond stress included, moves the two spreads alike.

An event study around 31 candidate Bitcoin to gold rotation episodes then validates the timing and levels: both thin market vehicles (token and miner), but not the deep market GLD, deteriorate; the raw token and miner magnitudes look alike only because they share the same gold-market shock, and the resemblance dissolves once that shock is removed which is evident from DiD results.

**Keywords:** tokenized gold; gold-backed cryptocurrencies; transaction costs; thin markets; crypto-to-gold rotation; difference-in-differences; PAXG; XAUT; GDXJ.

**JEL classification:** G12; G14; G15; G23.

* Assistant Professor, Department of Economics, Delhi School of Economics, University of Delhi, Delhi-110007. Email: rgarg@econdse.org

† Ph.D. Scholar, Department of Economics, Delhi School of Economics, University of Delhi, Delhi-110007. Email: srishtijaiswal@econdse.org

## 1. Introduction

Physical gold has served as a safe haven across centuries of crises, and exchange-traded funds such as GLD have made the rotation into gold operationally frictionless. An investor can buy deep gold exposure in a millisecond during a panic at a near-zero spread. Tokenized gold promises the same exposure with blockchain settlement. Paxos' PAX Gold (PAXG) and Tether's Tether Gold (XAUT), the two leading gold-backed cryptocurrencies (GBCs), are each redeemable claims on vaulted London Bullion Market association (LBMA) good-delivery bars, and together they account for 89% of the tokenized-gold market share. But gold backed cryptocurrencies (GBCs) do not trade on commodity exchanges. They trade on the same order books, venues, and market-maker networks as Bitcoin. That raises a natural question. When a Bitcoin-driven stress episode is what prompts an investor to rotate into gold, does tokenized gold stay as cost-effective to trade as deep market gold ETF (GLD), or does its thin, crypto-native microstructure reassert itself at the very moment the investor needs to transact?

We take up that question with a daily panel of four gold-linked vehicles and three complementary identification strategies. Alongside the two tokens and GLD we include the VanEck Junior Gold Miners ETF (GDXJ), an exchange-traded fund that tracks the performance of small-cap and mid-cap companies involved in gold mining, a thin but conventional, non-tokenized gold-linked vehicle. Setting a tokenized thin market against a non-tokenized thin market is what lets us separate anything specific to tokenization from generic market thinness. Our proxy for trading conditions is the daily high-low range, and we read it not as a clean gauge of illiquidity but as a proxy for the cost of transacting under stress, a quantity that, in a thin market, folds together bid-ask cost, depth, and idiosyncratic price volatility, which is difficult to pull apart using daily data for all the four vehicles considered in this study.

### 1.1 Main findings

Our results are derived from one daily panel but three deliberately different estimators, i.e. a date-fixed-effects difference-in-differences (DiD), a jointly estimated seemingly-unrelated-regression (SUR) system, and an event study around crypto-to-gold rotation episodes, each of which nets out the shocks common to all gold vehicles in a different way.

- From the date-fixed-effects difference-in-differences approach we find that on an average day, when people are dumping Bitcoin and buying gold, PAXG's intraday price swings become wider than GLD's do, so the token gets relatively more expensive to trade at that moment, while the junior-miner gap (GDXJ vs GLD) shrinks. The results also indicate that both tokens (Tether gold XAUT and Paxos gold PAXG) behave identically, indicating something structural about tokenised gold rather than one issuer's quirk.

- From the seemingly-unrelated-regression system results, we find that when stock-market fear spikes (VIX rises), junior miners get much more expensive to trade relative to GLD. This is because junior minors trade on equity exchange and their market-makers are equity desks that pull back when volatility hits. Tokenized gold barely feels it, not because it is immune, but because PAXG and GLD both widen by roughly the same amount when the VIX rises, so their difference stays flat. Equity market stress is the only thing formally separating how the two spreads behave (widens miner spread 9-11 times more than token spread). Every other stressor i.e. bond stress, dollar strength, policy uncertainty, and even the gold return, moves the two spreads by statistically equal amounts. In the case of bond stress, the coefficients are virtually identical, and the confidence interval on their difference is estimated to be zero.

- In the event study, events are identified as rotation from Bitcoin-to- gold on days of joint Bitcoin weakness and gold strength. Around each of the 31 rotation episodes, both tokens (PAXG and XAUT) get harder and more expensive to trade while GLD stays flat. But junior miners spike by a similar raw magnitude, which at first looks like all three thin vehicles are fragile for the same reason. But combining this finding with the results from DiD this similarity is misleading. Tokens and junior miners, both, are simply riding the same surge in gold-market activity, and once you strip that shared component out, the two diverge sharply in opposite directions, revealing completely different mechanisms. Junior miners are equity instruments, and the people making their markets are equity desks. On a crypto-specific rotation day, when Bitcoin is sliding while stocks hold steady, those desks have no reason to flinch, and the miner spread actually tightens. Tokenized gold sits in a different place entirely. It trades on the same crypto venues as Bitcoin, so the same wave of selling that pushes investors toward gold lands directly on the token's market-makers, and PAXG's spread against GLD widens. One external shock, two very different mechanisms- one settles, the other seizes up.

- Finally, the COVID crash of early 2020 hands us something close to a natural experiment. It was the most frightening stretch for equity markets in the entire sample, with the VIX touching 82, yet it was not a Bitcoin-to-gold rotation, because this time Bitcoin and gold fell together rather than money fleeing the one for the other. That accident of timing is what we want, since it lets us watch what tokenized gold does when equity fear is at its worst but the crypto-rotation trigger is switched off. The answer is that it does almost nothing. The token's spread against GLD barely moves, its sensitivity to fear essentially zero, even as the junior-miner spread blows out to more than four times its usual response. If the token's fragility were simply a reaction to market panic in general, this is the precise moment it should have surfaced, and it did not. What widens the token, then, is not any abstract or a systemic fear but the specific stress of a crypto sell-off rotating into gold, the very mechanism the rest of the paper sets out to isolate.

### 1.2 Literature and Research Contribution

Our paper sits at the junction of three research streams that have, until now, largely run on separate tracks. The safe-haven role of gold, the safe-haven debate around Bitcoin, and the small but growing empirical work on gold-backed cryptocurrencies. Gold's status as a safe haven, an asset that holds or gains value just when riskier assets fall, is among the most durable findings in asset pricing (Baur and Lucey, 2010; Baur and McDermott, 2010; Reboredo, 2013; Beckmann, Berger and Czudaj, 2015). Bitcoin was widely promoted as "digital gold," and some studies find hedge or safe-haven properties for it and for related crypto assets in particular settings (Bouri et al., 2017; Bouri, Shahzad and Roubaud, 2020; Baur and Hoang, 2021), while others conclude that it is too volatile and too correlated with risk assets to substitute for gold's stability (Klein, Thu and Walther, 2018). Tokenized gold sits at the meeting point of these two literatures- it promises gold's price behaviour with the settlement convenience of a crypto token, which makes the natural question whether it inherits gold's stability or crypto's fragility once markets come under stress.

The empirical literature on gold-backed cryptocurrencies (GBCs) has concentrated on two questions, and neither is the one we ask. The first is how faithfully these tokens track the price of physical gold. Ashfaq, Pfeifer, Gürpinar and Gulum (2026) assess the tracking accuracy of PAXG and XAUT against spot gold and find strong long-run alignment but short-term deviations that they attribute to continuous crypto-market trading, liquidity fragmentation, and issuer-specific design, with XAUT tracking spot gold more closely than PAXG. The second and larger strand asks whether GBCs inherit gold's safe-haven and hedging properties, and answers it with returns-based volatility and connectedness models. Jalan, Matkovskyy and Yarovaya (2021) study five gold-backed stablecoins through the COVID-19 pandemic and find that their volatility remained comparable to Bitcoin's rather than to gold's, that they did not display gold's safe-haven potential, and that the gold market, not Bitcoin, was the dominant source of their volatility. Wang, Ma and Wu (2020) reach a related conclusion for gold-pegged stablecoins, which are less effective than gold as safe havens but still useful for limiting extreme losses, while Wasiuzzaman and Haji Abdul Rahman (2021) fit an ARMA-GARCH model to PAX Gold and document a rise in volatility during the COVID-19 bear market that is economically visible but statistically insignificant. Mnif, Salhi, Trabelsi and Jarboui (2022) examine market efficiency and herding in GBCs, and Aloui, Ben Hamida and Yarovaya (2021) contrast Islamic and non-Islamic gold-backed tokens and find the former less exposed to geopolitical risk. Closest to our stress framing, Hoque, Billah, Alam and Tiwari (2024) test XAUT and PAXG, alongside gold and Bitcoin, as hedges and safe havens against categorical and regional financial stress. Across this literature the object of study is the return series and its second moments; the cost of transacting these tokens, and how that cost behaves in a crisis, is left unexamined.

That omission matters, because a safe haven is only useful to the extent that an investor can actually move into it cheaply when the crisis arrives. Liquidity is state-dependent and tends to evaporate in the episodes when it is most needed, as funding and market liquidity reinforce one another (Brunnermeier and Pedersen, 2009) and effective spreads and trading activity shift with market conditions (Amihud, 2002; Chordia, Roll and Subrahmanyam, 2001). For crypto assets specifically, the microstructure is fragmented across venues and market-makers, arbitrage is impeded (Makarov and Schoar, 2020), and transaction costs are non-trivial (Dyhrberg, Foley and Svec, 2018). Measuring these trading conditions for tokenized gold is constrained by data. Intraday bid-ask and depth series do not exist for PAXG and XAUT over a multi-year sample, which forces reliance on low-frequency proxies built from daily open, high, low, and close prices (Parkinson, 1980; Chung and Zhang, 2014). Tellingly, the one GBC study that computes a liquidity measure at all, Jalan, Matkovskyy and Yarovaya (2021), uses the same daily high-low range we adopt, but only descriptively, plotting it token by token without a stress-conditional or cross-vehicle design. No study places a tokenized thin market beside a conventional thin one to ask what, if anything, tokenization adds to generic market thinness.

Three gaps follow. First, the tokenized-gold literature is a literature about returns and price-tracking, not about the cost of transacting under stress; whether these tokens stay cost-effective to trade during a crypto-to-gold rotation is simply unknown. Second, no study separates what is specific to tokenization from what merely reflects a thin market, because none sets the tokenized vehicle against a non-tokenized but equally thin one. Third, the evidence rarely extends beyond a single token or a single issuer, so it cannot say whether a finding is a property of the tokenized-gold category or an artifact of one issuer's plumbing. This paper takes up all three.

Our contribution is a result about channels, not levels. Two thin gold-linked vehicles share a vulnerability under stress but transmit it through different mechanisms, and once the common shock is stripped away, they move in opposite directions relative to their benchmark. Relative to the tokenized-gold literature, which has mostly studied price-tracking accuracy, we provide the first evidence on stress-conditional trading conditions for tokenized gold, an explicit contrast between a tokenized and a conventional thin market, and a cross-issuer replication across two independently governed tokens. We are equally careful about what the evidence does not establish. It does not show that tokenized gold possesses a proven crypto stress channel of its own, and it does not cleanly separate illiquidity from idiosyncratic volatility. What it does establish is relative insulation of the token from equity fear and a divergence in the cost of transacting under stress that holds up across specifications.

Section 2 describes data and measurement, Section 3 reports descriptive analysis and diagnostics, and documents the mid-2020 level shift as a pre-estimation diagnostic. Section 4 presents the primary result based on the difference-in-differences with date-fixed-effects. Section 5 asks why the two spreads diverge and answers with the jointly estimated system. Section 6 presents the event study as validation of the timing and levels. Section 7 collects robustness tests and confronts the volatility-conflation concern head-on. Sections 8 and 9 discuss and conclude; limitations are stated in Section 8, after the contribution is established.

## 2. Data and Measurement

### 2.1 Vehicles and sample

We obtain daily Open, High, Low, Close and Volume (OHLCV) data for PAXG-USD, XAUT-USD, GLD, and GDXJ from yfinance, avoiding source-splicing. PAXG runs from its 26 September 2019 launch through December 2025; XAUT begins 7 February 2020; GLD and GDXJ provide continuous coverage. The common-weekday intersection yields 1,574 observations. LBMA PM-fix data, used only for episode detection, come from the LBMA.

### 2.2 The high-low range: what it measures, and what it does not

Our cross-vehicle measure of trading conditions is the daily high-low range,

$$HLR_t = (\text{High} - \text{Low}) / [(\text{High} + \text{Low})/2],$$

the one proxy computable for all four vehicles. We use it as a low-frequency proxy in the tradition of daily high/low estimators (Chung and Zhang, 2014;), but we are careful about its interpretation, because this is the single most consequential measurement point in the paper. Up to scaling, the high-low range is the Parkinson (1980) volatility estimator. It blends a volatility/information component with a liquidity component, and in a thin market the two are often inseparable in practice. An order that cannot be absorbed without moving the price shows up as a wider range, and so does a burst of idiosyncratic volatility that has nothing to do with the cost of trading.

For that reason we do not treat the high-low range as an absolute baseline for illiquidity. We read it instead as a proxy for the cost of transacting under stress, the effective cost an investor bears when trying to move size, whether that cost arrives as a wider bid-ask spread, as thinner depth, or as idiosyncratic price volatility. All three raise the cost of execution in the same direction, and for a trader they are close substitutes. This framing is deliberate. It lets us make a defensible claim about relative trading conditions while conceding, up front, that we cannot fully disentangle the liquidity and volatility components with daily data. Our response to the conflation is design rather than assertion: the difference-in-differences system in Section 4 absorbs a full set of date fixed effects that remove the *common* volatility, gold-price, and information shock hitting every gold vehicle on a given day, leaving the cross-vehicle differential as the estimand. What that design cannot remove is *idiosyncratic* volatility specific to one vehicle; we return to this limit explicitly in Section 7.1.

### 2.3 The tokenization and hierarchy spreads

Two pre-specified spreads organize the analysis:

$$TLS_t = \mathrm{HLR}_{\mathrm{PAXG},t} - \mathrm{HLR}_{\mathrm{GLD},t} \text{ (tokenization spread)},$$

$$LHS_t = \mathrm{HLR}_{\mathrm{GDXJ},t} - \mathrm{HLR}_{\mathrm{GLD},t} \text{ (liquidity-hierarchy spread or miner spread)},$$

with an analogous TLS for XAUT. In normal times TLS averages 0.0073, PAXG's intraday range runs about 80% wider than GLD's, while LHS averages 0.0212, roughly three times larger. Both spreads difference out the common gold-market component embedded in GLD. The design logic is a nested comparison. If tokenized gold simply inherited a generic thin-market fragility, TLS and LHS would respond to the same stressors; if the tokenized wrapper matters, they would separate. Section 5 shows they separate on one axis.

### 2.4 Stress variables and controls

Four exogenous stress proxies enter the regressions, the CBOE Volatility Index VIX (equity-market fear), the change in the 2-year Treasury yield (ΔUS2Y, bond stress), the broad-dollar change (ΔDXY), and the Economic Policy Uncertainty index (EPU; Baker, Bloom and Davis, 2016), each of them z-scored on a rolling 252-day window. Controls include the LBMA gold return and the Bitcoin return, which enters only the token equation because GDXJ and GLD have no crypto-exchange exposure. Autoregressive lags absorb the strong persistence of the spreads.

### 2.5 Candidate crypto-to-gold rotation episodes

We identify candidate rotation episodes i.e. the days on which Bitcoin weakens while gold strengthens, following the detection procedure of Wang and Xiao (2025), with thresholds calibrated on the pre-PAXG window (January 2016-August 2019) so that detection is out of sample. A day qualifies when the Bitcoin return falls below its calibration 20th percentile (-2.0%), the LBMA gold return rises above its 80th percentile (+0.54%), and each also departs from its 30-day rolling mean by more than one rolling standard deviation. The joint condition is what gives these days their economic interest. On the same session Bitcoin sells off and gold rallies, a pattern consistent with investors reaching for a safe haven rather than simply de-risking across the board. The price pattern is what we observe directly, and while it is suggestive of capital rotating out of Bitcoin and into gold, our data cannot confirm that such a flow actually took place, so we treat the rotation reading as a plausible interpretation rather than an established fact. On that understanding we label the resulting set flight-to-safety (FTS) events and use that term throughout. We use "flight to safety" in a deliberately narrow sense i.e. a move specifically from crypto into gold, rather than the broader market usage of a general shift into safe assets across many risk markets.

The procedure yields 31 FTS events, each coincident with an identifiable market episode (see Table A1 in Appendix). These days anchor both the difference-in-differences interaction (Section 4) and the event study (Section 6). By construction the March 2020 COVID-19 shock is not an FTS event, as Bitcoin and gold fell together that month, so the Bitcoin-down/gold-up condition failed. We exploit that exclusion as a natural placebo in Sections 5, a period of acute stress in which the crypto-to-gold rotation did not occur lets us check that our results track the rotation itself and not stress in general.

## 3. Descriptive Analysis and Diagnostics

### 3.1 The liquidity hierarchy

Table 1 reports summary statistics. The central descriptive fact is a stable hierarchy in trading conditions. Mean HLR rises monotonically from GLD (0.0092), XAUT (0.0146), and PAXG (0.0166) to GDXJ (0.0304). PAXG's average intraday range is roughly 1.8 times GLD's; GDXJ's is 3.3 times. Our analysis concerns how they move under stress, not their level.

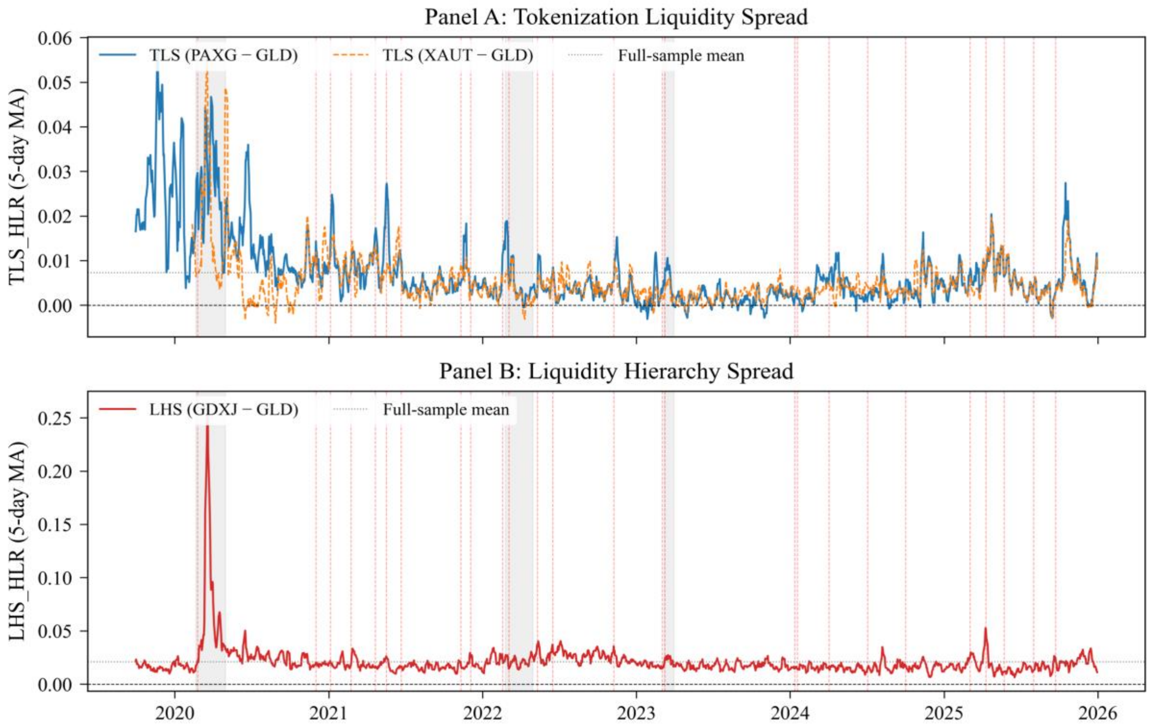


***Figure 1.*** *Tokenization spread (Panel A) and miner spread /hierarchy spread (Panel B), 5-day moving average. Vertical dashed lines mark rotation episodes; the grey band highlights COVID. The compression of the tokenization spread after mid-2020 is the durable level shift documented in Section 3.3 (Table 2).*

**Table 1. Descriptive statistics (cross-vehicle weekday panel, N = 1,574)**

| Variable | Mean | Median | Std | p01 | p99 | Skew |
|---|---|---|---|---|---|---|
| PAXG_HLR | 0.01655 | 0.01289 | 0.01230 | 0.00436 | 0.06743 | 2.88 |
| XAUT_HLR | 0.01457 | 0.01220 | 0.01090 | 0.00462 | 0.05078 | 6.24 |
| GLD_HLR | 0.00923 | 0.00792 | 0.00513 | 0.00318 | 0.02958 | 2.58 |
| GDXJ_HLR | 0.03041 | 0.02594 | 0.02229 | 0.01132 | 0.08969 | 8.33 |
| TLS_HLR (PAXG-GLD) | 0.00732 | 0.00430 | 0.01080 | -0.00491 | 0.05616 | 2.91 |
| LHS_HLR (GDXJ-GLD) | 0.02118 | 0.01780 | 0.01937 | 0.00416 | 0.07256 | 8.97 |
| TLS_XAUT (XAUT-GLD) | 0.00518 | 0.00355 | 0.00892 | -0.00530 | 0.03016 | 7.21 |
| VIX_z | -0.086 | -0.395 | 1.281 | -1.789 | 4.796 | 2.03 |
| ΔUS2Y_z | 0.049 | 0.066 | 1.090 | -3.267 | 3.370 | 0.18 |

*Note.* All series reject normality (Jarque-Bera $p < 0.001$), which motivates Newey-West inference.

### 3.2 Tokenization is not hierarchy: distinctness, persistence and stationarity

The two spreads share only about 12% of their variance (correlation 0.342), so they are separate objects rather than two labels for one thing, and there is no collinearity worth worrying about (maximum VIF 1.19). Both are highly persistent, first-order autocorrelation of 0.51 for TLS and

0.72 for LHS , which is why the later specifications carry autoregressive terms and lean on Newey-West HAC standard errors rather than treating the errors as clean.

Before working with the spreads we establish the behaviour of the underlying range panel. Our four vehicles are all claims on the same metal, so their intraday ranges should move together. A gold-price shock, a macro release, or a volatility spike reaches every one of them on the same day. That co-movement is something to test, not to assume away. Pesaran's cross-sectional dependence test on the panel of high-low ranges ($HLR_{it}$) returns a CD statistic of 54.933 ($p < 0.01$), so the null of cross-sectional independence is rejected decisively. This is the result one would expect given the shared gold factor, and it also settles the choice of tool. A first-generation panel unit-root test would be biased in the presence of that dependence, so we use Pesaran's cross-sectionally augmented CIPS test instead. The CIPS statistic for $HLR_{it}$ is -6.190, comfortably beyond the 1% critical band, so we reject the unit-root null and treat the range panel as stationary, I(0).

The series-by-series picture agrees. Lee-Strazicich (2003) unit-root tests, which allow for an endogenously dated structural break, place every series, both spreads and the macro controls, in the I(0) category (Table 2), which is what licenses the SUR framework used later. For the tokenization spread the test also flags a break we did not impose, dated to 22 July 2020. Because that date carries more than a stationarity check, we take it up in Section 3.3.

The only significant post-2020 break is in VIX (Aug 2024, the Japan carry-trade unwind), which we carry as the D_VIX_shock control in the miner spread equation in section 5. All series are I(0), so OLS/SUR with Newey-West HAC is appropriate.

**Table 2. Lee-Strazicich minimum-LM unit-root tests with structural break**

| Variable | LM t-stat | I(0)? | Break date | Break coeff | Sig. |
|---|---|---|---|---|---|
| TLS_HLR | -6.924 | Yes (1%) | 2020-07-22 | -0.016 | 10% |
| LHS_HLR | -7.048 | Yes (1%) | 2020-07-09 | -0.014 | No |
| VIX_z | -5.530 | Yes (1%) | 2024-08-30 | +1.865 | 1% |
| ΔUS2Y_z | -7.554 | Yes (1%) | 2022-05-18 | -1.066 | No |
| EPU_z | -4.431 | Yes (1%) | 2020-07-27 | -0.327 | No |
| BTC_Ret | -3.600 | Yes (5%) | 2020-06-19 | -0.004 | No |

*Note.* Crash model, max 22 lags, trimming $\pi = 0.10$; 1%/5% critical values -3.798/-3.230.

### 3.3 A durable level shift: the mid-2020 maturation break

Before the stress-conditional analysis we document, and then control for, a single durable feature of the tokenization spread.

A unit-root test that ignores breaks invites the spurious inference that the Lee-Strazicich procedure is designed to avoid, since it dates any break endogenously rather than at a point we choose. Here it locates one break in the tokenization spread, at 22 July 2020 (Table 2). We read this as a pre-estimation diagnostic, a one-time level shift that has to be absorbed so it does not leak into the stress-conditional estimates.

Two things about the timing are worth putting on record. The break lands about five weeks before PAXG's main exchange listing on Binance (26 August 2020), and it falls within a broader period when the token was visibly maturing: more venues, deeper order books, a growing roster of market-makers. It also coincides with the day the U.S. Office of the Comptroller of the Currency issued Interpretive Letter No. 1170 (22 July 2020), which confirmed that national banks may hold cryptocurrency in custody for their customers. A trust company like Paxos, already chartered and supervised in New York, stood to benefit from that clarification, so it is at least coherent that the token's perceived custody and counterparty risk were repriced somewhat around that date.

These two events play different roles. The OCC letter is a regulatory signal of the sort that can move the baseline expectations attached to a regulated, asset-backed token, while the Binance listing five weeks later supplies the mechanical channel: the depth and volume that let larger orders go through without pushing the spread around. The listing left the token's legal standing untouched, and by the time it happened, any repricing the July news set off had already had a month to filter through. Read that way, the exchange listing scaled and confirmed a regime that was mostly already in place, rather than bringing it about.

The tokenization spread drops once, and durably, in mid-2020, so we anchor a maturation dummy to that break. Its coefficient is roughly −0.008 in the TLS_HLR specification of Section 5, which is worth pausing on: since the dependent variable is the PAXG-minus-GLD range, a negative shift means the spread narrowed for good, by about 0.8 percentage points of intraday range — larger than the mean spread itself of 0.0073. (This is the coefficient from the controlled

Section 5 specification; the −0.016 reported in the Lee–Strazicich test of Table 2 is the raw structural-break estimate from a different model, so the two are expected to differ.) The same narrowing shows up for XAUT, which tells us this is a feature of tokenized gold and not of one issuer. That is why we keep the dummy in every stress-conditional specification, so that the maturation cannot pass itself off as a stress response.

## 4. The Primary Result: Divergent Transaction Costs in a Difference-In-Differences Design with Date Fixed Effects

The question is on days when Bitcoin sells off while gold rallies (our candidate flight-to-safety (FTS) episodes) what happens to the cost of transacting each thin gold vehicle relative to the deep ETF? The identification problem is equally simple to state. Rotation days are gold-appreciation days, and gold-appreciation days carry a mechanical widening of the intraday range for every gold vehicle. Any honest test has to strip out that common component before comparing vehicles. The date-fixed-effects difference-in-differences does exactly that, and it does so without asking us to specify the stress-response function.

As defined in Section 2.5, we use "flight-to-safety (FTS)" and "rotation" as labels for this Bitcoin-down/gold-up price pattern; the underlying capital flow out of Bitcoin and into gold is a plausible explanation that we do not test explicitly.

Stacking the vehicle-day observations, we regress vehicle HLR on vehicle indicators interacted with the Flight to safety event.

*Model 1 (Day 0):*

$$HLR_{i,t} = \alpha_i + \gamma_t + \beta_3(\mathrm{PAXG}_i \times \mathrm{FTS}_{\mathrm{Day}\ 0,t}) + \beta_4(\mathrm{XAUT}_i \times \mathrm{FTS}_{\mathrm{Day}\ 0,t}) + \beta_5(\mathrm{GDXJ}_i \times \mathrm{FTS}_{\mathrm{Day}\ 0,t}) + \varepsilon_{i,t}$$

*Model 2 ([0, +2]):*

$$HLR_{i,t} = \alpha_i + \gamma_t + \beta_3(\mathrm{PAXG}_i \times \mathrm{FTS}_{[0,+2],t}) + \beta_4(\mathrm{XAUT}_i \times \mathrm{FTS}_{[0,+2],t}) + \beta_5(\mathrm{GDXJ}_i \times \mathrm{FTS}_{[0,+2],t}) + \varepsilon_{i,t}$$

GLD is the omitted reference and serves as benchmark.

The estimator is a fixed-effects difference-in-differences (DiD), absorbing vehicle fixed effects ($\alpha_i$) and date fixed effects ($\gamma_t$)[‡]. The identification however rests on the 1,574 date fixed effects ($\gamma_t$) which absorb every daily shock shared across vehicles, the VIX level, the gold return, the Bitcoin return, the dollar, policy uncertainty, leaving each interaction as a clean difference-in-differences relative to GLD. The common volatility, gold-price, and information component that the high-low range would otherwise attribute to trading cost is what the date fixed effects remove. Standard errors are clustered by date (1,574 clusters); entity clustering is infeasible with four vehicles. Hence, we use the term date-fixed-effects DiD throughout, to emphasise the identification mechanism.

The primary specification uses a $[0,+2]$event window, defined as the event day and the subsequent two trading days, consistent with the event-study design. Unlike announcement-based event studies, candidate flight-to-safety episodes are identified from daily Bitcoin and gold return realizations rather than from an information release. They represent the emergence of a broader market-stress episode rather than a one-day shock. A short post-event window allows sufficient time for the market adjustment to be reflected in the HLR-based transaction-cost measure while limiting contamination from unrelated information that becomes increasingly likely over longer event windows.

The coefficient $\beta_3$ and $\beta_4$ have a clean DiD interpretation:

$\beta_3$ = *(PAXG_FTS - GLD_FTS) - (PAXG_normal - GLD_normal)*

$\beta_4$ = *(XAUT_FTS - GLD_FTS) - (XAUT_normal - GLD_normal)*

This is the excess widening of PAXG /XAUT over GLD during FTS days, relative to their usual relationship. Positive $\beta_3$ means PAXG deteriorates more than GLD during FTS. The date fixed effects $\gamma_t$ ensures that any common daily shock (macro, gold price, sentiment) affecting all vehicles equally is absorbed. What remains is the pure cross-vehicle differential.

Table 3 reports the results, and three findings come out of it. The first is that PAXG and GDXJ move in opposite relative directions. PAXG's interaction is +0.0016, so the cost of transacting tokenized gold rises relative to GLD, whereas GDXJ's is −0.0018, meaning the miner spread falls; the Wald test rejects equality of the two ($p = 0.038$). This sign reversal is the finding the

[‡] This is not the staggered-treatment case that the recent two-way FE literature (Goodman-Bacon; de Chaisemartin–D'Haultfœuille) warns about.

rest of the paper sets out to explain and to validate. The second is that PAXG and XAUT come out statistically identical, at +0.0016 apiece with a Wald p of 0.975,  a cross-issuer replication that holds up under the strictest macro control we have and carries across different issuers, vaults, jurisdictions, and blockchains. The third concerns significance: the individual interactions are only marginal (PAXG p = 0.111, GDXJ p = 0.168), yet the difference between them is estimated precisely. This is a property of the design rather than a weakness of it. The estimation error the two share cancels when we take the difference, so the contrast is pinned down much more tightly than either coefficient on its own.

The absence of a significant differential on the event day alone (Wald p =0.980) suggests that the treatment effect is not fully realized on day 0 but becomes apparent over the subsequent trading sessions, consistent with the primary $[0, +2]$ specification.

**Table 3.  Pooled panel DiD with 1,574 date fixed effects**

**Panel A. Vehicle × FTS interactions**

| | Model 1 (Day 0) | | Model 2 ([0, +2]) | | |
|---|---|---|---|---|---|
| **Term** | **Coeff** | **p-value** | **Coeff** | **p-value** | **Reading** |
| PAXG × FTS | +0.0013 | 0.365 | +0.0016 | 0.111 | widens vs GLD |
| XAUT × FTS | +0.0035 | 0.015 | +0.0016 | 0.042 | widens vs GLD |
| GDXJ × FTS | +0.0014 | 0.515 | -0.0018 | 0.168 | narrows vs GLD |

**Panel B. Wald tests of equality Model 2 ([0, +2])**

| **Hypothesis** | **Wald -stat** | **p-value** | **Reading** |
|---|---|---|---|
| $H_0$: PAXG = GDXJ | 4.327 | 0.038 | opposite directions |
| $H_0$: PAXG = XAUT | 0.001 | 0.975 | cross-issuer identical |

*Note.* Standard errors clustered by date (1,574 clusters). Wald tests use the joint panel covariance.

Two implications deserve emphasis before we ask why the divergence arises. First, since the individual PAXG and GDXJ interactions are each only marginal, we do not claim that "cost of transacting tokenized gold increases under stress" in an absolute sense on rotation days. We claim that the *relative* cost of transacting tokenized gold rises and the *relative* cost of transacting miner falls, so the two diverge. Secondly, the divergence reproduces across two independently governed tokens (PAXG as well as XAUT), on the same days, to the same magnitude, is difficult to attribute to any single issuer's plumbing and points instead to a feature of the tokenized-gold category.

## 5. Mechanism behind Why the Two Spreads Diverge?

Section 4 establishes what happens; this section asks why. We use a jointly estimated system as mechanism analysis, an attempt to identify the economic force that makes a crypto-specific rotation push the two thin-market spreads in opposite relative directions once the common daily shock is gone.

### 5.1 A Jointly Estimated Seemingly-Unrelated-Regression System

We estimate the TLS and LHS equations together as a seemingly-unrelated-regression system (Zellner, 1962) with HAC-robust covariance. Joint estimation is essential because comparing p-values from two separate OLS equations is not a valid test of whether a coefficient truly differs across them. A regressor can be "significant" in one equation and "insignificant" in the other while the underlying coefficients are statistically identical (Gelman and Stern, 2006). The SUR joint covariance permits a proper Wald test of cross-equation equality:

$$H_0\text{:}\ \beta_{\mathrm{TLS}} = \beta_{\mathrm{LHS}}.$$

The regressors included in the two equations are not symmetric, because each equation carries the shocks that have an economic channel into that particular spread, rather than a common list imposed for both equations. The token spread (TLS) is a crypto-native variable, so it should absorb crypto-systematic risk directly; we include the Bitcoin return as the natural proxy for that risk, alongside the July-2020 maturation dummy, which marks the deepening of tokenized-gold liquidity that shifts the level of the tokenization spread. The miner spread (LHS) has no such channel: GDXJ is an equity, so a Bitcoin regressor there would contribute noise rather than signal, and the maturation break (a tokenized-gold event) has no reason to move it. What the miner equation does carry is the August-2024 VIX regime shift, an equity-market event that belongs in the equation whose spread is equity-sensitive.

### 5.2 The one confirmed differential: relative insulation from equity fear

The system turns up one differential that clears formal testing, and it is equity-market fear. The VIX moves the miner spread strongly while leaving the tokenization spread indistinguishable from zero i.e. a gap of roughly eleven-fold in the estimated coefficients, and the Wald test rejects equality. There is nothing borderline about it; the difference is significant at the 1% level

**Table 4. SUR system: TLS and LHS with cross-equation Wald tests**

| Variable | TLS coeff | SUR p | LHS coeff | SUR p | Wald W | Wald p |
|---|---|---|---|---|---|---|
| **VIX_z** | **+0.0002** | **0.155** | **+0.0022** | **<0.001** | **8.870** | **0.0029** |
| ΔUS2Y_z | -0.0005 | 0.022 | -0.0005 | 0.045 | 0.000 | 0.998 |
| ΔDXY_z | -0.0002 | 0.511 | +0.0003 | 0.243 | 1.651 | 0.199 |
| EPU_z | -0.0000 | 0.939 | -0.0002 | 0.495 | 0.179 | 0.672 |
| Gold_Ret | -0.0554 | 0.029 | -0.1152 | 0.144 | 0.682 | 0.409 |
| BTC_Ret | -0.0033 | 0.655 | excl. | n/a | n/a | n/a |
| D_maturation | -0.0078 | <0.001 | excl. | n/a | n/a | n/a |
| D_VIX_shock | excl. | n/a | -0.0013 | 0.013 | n/a | n/a |

*Note.* Estimated jointly via SUR (Zellner, 1962), HAC-robust (Newey-West, bw = 7); Overall system $R^2 = 0.5453$**,** cross-equation residual correlation 0.0840. Wald tests of $H_0$: $\beta_{TLS} = \beta_{LHS}$ use the full joint covariance; Constant and AR suppressed. AR(1 to 5) lags were included in both equations to account for persistence and serial correlation. N = 1,526

This implies that the spread of GDXJ over GLD is strongly sensitive to equity-market fear, whereas the spread of PAXG over GLD exhibits no statistically detectable response. This is relative insulation, not immunity. Showing that the tokenization spread is insulated from equity fear does not, on its own, demonstrate that some crypto-specific force stresses it instead; that is a stronger claim and lies beyond what our data can establish.

What the data supports is that the tokenization spread barely responds to the VIX, and the reason is instructive. Vehicle-level regressions (Tables 5A ,5B) show the insulation is a netting-out, not an absence of response. VIX affects both PAXG and GLD significantly, but the estimated responses are of similar magnitude (0.0008 versus 0.0007). As a result, the common equity-fear component largely cancels out in the spread, leaving no statistically detectable differential response in the TLS equation.

**Table 5A. Vehicle-Level Evidence behind TLS**

| Variable | PAXG_HLR | GLD_HLR | Difference (≈TLS) | Why TLS result |
|---|---|---|---|---|
| **VIX_z** | +0.0008 (p<0.001) *** | +0.0007 (p<0.001) *** | +0.0001 | Both respond equally → TLS insig |
| **dUS2Y_z** | -0.0007 (p=0.005) *** | -0.0003 (p=0.025) ** | -0.0004 | PAXG 2.3× larger → TLS sig |
| **dDXY_z** | +0.0001 (p=0.880) | +0.0003 (p=0.155) | -0.0002 | Neither sig individually → TLS insig |
| **EPU_z** | -0.0001 (p=0.637) | +0.0000 (p=0.687) | -0.0001 | Neither sig individually → TLS insig |
| **Gold_Ret** | -0.0702 (p=0.081)* | -0.0187 (p=0.324) | -0.0515 | PAXG 3.7× more sensitive → TLS marginal |
| **BTC_Ret** | -0.0069 (p=0.538) | -0.0020 (p=0.567) | -0.0049 | Neither sig individually → TLS insig |

Note: ***, **, * indicates significance at 1%, 5% and 10 % respectively

VIX is highly significant for both PAXG (0.0008, $p < 0.001$) and GLD (0.0007, $p < 0.001$). PAXG's response is only about 14% larger than GLD's, and this modest differential is not statistically distinguishable. By contrast, PAXG's response to Treasury-yield changes (-0.0007) is more than 130% larger than GLD's (-0.0003), allowing the differential to emerge as a significant predictor of the tokenization spread.

In contrast, the miner spread exhibits the opposite mechanism. Vehicle-level regressions (Table 5B) show that both GDXJ and GLD respond significantly to equity-market fear, and GDXJ's estimated VIX coefficient (0.0027) is approximately 3.7 times larger than GLD's (0.0007). Thus, unlike the tokenization spread, the differential survives in the spread itself, producing the strong and statistically significant VIX effect observed in the LHS regression.

**Table 5B. Vehicle-Level Evidence behind LHS**

| Variable | GDXJ_HLR | GLD_HLR | Difference (≈LHS) | Why LHS result |
|---|---|---|---|---|
| **VIX_z** | +0.0027 (p<0.001) *** | +0.0007 (p<0.001) *** | +0.0020 | GDXJ ≈3.7× more sensitive → LHS significant |
| **dUS2Y_z** | -0.0007 (p=0.055) * | -0.0003 (p=0.024) ** | -0.0004 | GDXJ ≈2.3× more sensitive → LHS significant |
| **dDXY_z** | +0.0006 (p=0.278) | +0.0003 (p=0.134) | +0.0003 | Neither sig individually → LHS insig |
| **EPU_z** | -0.0004 (p=0.260) | -0.0001 (p=0.577) | -0.0003 | Neither sig individually → LHS insig |
| **Gold_Ret** | -0.1239 (p=0.113) | -0.0193 (p=0.312) | -0.1045 | GDXJ ≈6.4× more sensitive → LHS insignificant |

Note: ***, **, * indicates significance at 1%, 5% and 10 % respectively

**How this explains the primary result.** In Section 4, the difference-in-differences results demonstrated that the two thin vehicles do not simply react by different amounts on rotation days; they move in opposite directions, i.e., the token spread widens (meaning tokenized gold becomes relatively more expensive to trade than GLD) while the hierarchy spread narrows (meaning the junior-miner gap against GLD contracts).

The SUR results explain this divergence. The only channel that formally separates the two spreads is equity-market fear. Flight-to-safety episodes are characterised by falling Bitcoin prices and rising gold prices, not by a common increase in every source of financial stress. The vehicle-level regressions show why the VIX differential appears only in the hierarchy spread. PAXG and GLD respond to equity-market fear by similar amounts, so the common VIX component is largely removed when the tokenization spread is formed. GDXJ, however, responds much more strongly than GLD, so the VIX differential survives in the hierarchy spread. The same shock generates opposite movements in the two thin-market spreads because the underlying vehicles transmit that shock differently.

Taken together, SUR identifies the differential channel, the vehicle-level regressions explain why that differential survives (or disappears) in the spread construction, and the difference-in-differences estimates show the divergence those channels produce during flight-to-safety episodes.

### 5.3 What is not a differential channel: the discipline of the joint test

Not every stress variable separates the two thin-market spreads. Bond stress (ΔUS2Y) is statistically significant in both the TLS and LHS equations (Table 5A,5B), but the joint Wald test shows that the estimated coefficients are statistically indistinguishable:

$$\boldsymbol{\beta_{TLS} - \beta_{LHS} = -0.000001}, \text{Wald } \boldsymbol{p = 0.998}.$$

Accordingly, we cannot conclude that Treasury-yield shocks affect the tokenization spread differently from the hierarchy spread. The same conclusion applies to ΔDXY and EPU, both fail the cross-equation equality test.

One coefficient deserves clarification. The gold return coefficient is individually significant only in the TLS equation, but not in the LHS equation. If we interpret it carelessly, we may conclude "significant for the token spread, insignificant for the miner spread" looks like evidence of a second, token-specific channel operating alongside the VIX. But significance is a statement about each coefficient relative to zero, not about the two coefficients relative to each other, and the two are not the same test. This is the Gelman-Stern point which says that the difference between significant and not-significant is not itself statistically significant. The joint test of $\beta_{\mathrm{TLS}} = \beta_{\mathrm{LHS}}$ dissolves the apparent contrast. The coefficients are not statistically different. Gold-price appreciation should be interpreted as a common market factor.

The VIX remains the only variable for which the equality restriction is rejected, and hence the only confirmed differential channel between the two spreads.

### 5.4 COVID as a natural experiment: Equity fear with crypto-to-gold rotation trigger switched off

The COVID subsample runs from 20 February to 30 April 2020, and it gives us something close to a natural experiment: the VIX peaked at 82 during it, the highest reading anywhere in our sample. We re-estimate the TLS and LHS SUR system over this window using an AR(1) specification and no structural-break dummies, since both breaks fall outside it (Table 6). What makes the window so useful is a quirk of timing. It brought the sharpest equity-fear shock in the whole sample, but it was not a crypto-to-gold rotation, because Bitcoin and gold fell together this time rather than Bitcoin sliding while gold climbed. That leaves us with pure, extreme equity fear and the rotation trigger switched off, which separates the two forces that our ordinary data may leave confounded.

**Table 6. COVID-subsample SUR: TLS and LHS with cross-equation Wald tests**

| Variable | TLS coeff | SUR p | LHS coeff | SUR p | Wald W | Wald p |
|---|---|---|---|---|---|---|
| **VIX_z** | **-0.0004** | **0.681** | **+0.0095** | **<0.001** | **36.004** | **<0.001** |
| ΔUS2Y_z | -0.0056 | 0.004 | +0.0074 | 0.070 | 5.790 | 0.016 |
| ΔDXY_z | -0.0007 | 0.560 | -0.0007 | 0.629 | 0.000 | 0.996 |
| EPU_z | +0.0056 | <0.001 | +0.0021 | 0.565 | 0.696 | 0.404 |
| Gold_Ret | +0.0351 | 0.765 | -0.8452 | 0.035 | 4.808 | 0.028 |
| BTC_Ret | -0.0852 | <0.001 | excl. | n/a | n/a | n/a |

*Note.* Estimated jointly via SUR (Zellner, 1962), HAC-robust (Newey-West, bw = 7); COVID window 20 February-30 April 2020. AR(1) only; structural-break dummies are omitted because both breaks fall outside this window. Overall system $R^2 = 0.706$. Constant and AR suppressed. N = 48.

The two spreads part ways here. On the miner side, the VIX coefficient climbs to +0.0095, which is 4.3 times its full-sample value; the worst of the fear produced the widest miner spreads, which is what you would expect from a vehicle so closely tied to equities. The token side does the opposite, or rather does nothing: its coefficient is −0.0004 ($p = 0.681$), meaning the largest fear shock in our whole sample left PAXG's spread against GLD basically where it was. One coefficient being significant and the other not is not, by itself, evidence that the two differ, a point we flag in Section 5.3, so we test the gap head-on. The cross-equation Wald test rejects equality of the two VIX coefficients ($W = 36.0$, $p < 0.001$), and the separation is much wider here than in the full sample ($W = 8.87$). With only 48 observations behind it, though, we treat this as support for the story rather than proof of it.

This is what the COVID window lets us see. If the token's calm were just a muted version of a general panic response, then COVID, with fear at its peak and no rotation to trigger the crypto channel, is precisely when the token spread should have widened, and it did not, even as the miner spread widened hard. So the fragility we find in the token is not something broad market stress switches on; it seems to need the specific strain that a crypto-to-gold rotation puts on the token's own infrastructure.

## 6. Validation: The Event Study

Our three research designs are best read as answers to a nested set of questions. The date-fixed-effects difference-in-differences carries the primary result. Once the common daily shock is swept out, the token and miner spreads do not simply widen by different amounts around rotation episodes; they move in opposite directions, with the token growing more expensive against GLD as the miner gap tightens. The seemingly-unrelated-regression system then explains that split, isolating one force that formally separates the two spreads: the token's relative insulation from equity-market fear. The event study that follows is the junior partner of the three by design. It does not carry the identification, only corroborates it.

What it contributes is a check in a different unit of measurement. The DiD reasons in differences: it defines each thin vehicle's move net of GLD, so the shared gold-market shock that hits every vehicle on a given day is removed by construction and never enters the estimate. The event study works the other way, tracking each vehicle's own abnormal trading cost, its high-low range measured against its pre-event baseline, without netting it against GLD. It sees the gross move rather than the net one. That single difference is what reconciles a result that would otherwise look contradictory. In these raw, un-netted terms the token and the miner deteriorate by almost the same absolute amount, which seems hard to square with the DiD's finding that their responses differ. The answer is that both are being lifted by the same broad gold-market appreciation: the event study captures that lift in full, while the DiD strips it away and exposes the divergence underneath. Taken on its own, then, the event study confirms that both thin vehicles, the token and the miner, see their trading cost rise around rotation episodes, whereas the deep-market vehicle, GLD, does not. Taken alongside the DiD, it explains how the token and miner can look so alike on the surface and still be driven by different machinery beneath.

### 6.1 Measuring abnormal trading cost: the CAL statistic

For each episode we compute abnormal trading cost as HLR minus its mean over a [-65, -6] estimation window, cumulate it over the event window, and test the resulting cumulative abnormal liquidity cost (CAL) with a sign test, a Wilcoxon signed-rank test (primary), and a block bootstrap (1,000 replications).

The intuition is straightforward. We want to know whether a vehicle became unusually expensive to trade during a rotation episode, so we first need a picture of what "normal" looks like for that vehicle. The [-65, -6] window supplies it: the average trading cost over the roughly three months leading up to the episode, stopping a week before so the run-up itself doesn't contaminate the baseline. Subtracting that baseline from the observed HLR gives *abnormal* cost i.e. how far trading conditions departed from the vehicle's own recent norm, which matters because a naturally wide-spread vehicle and a naturally tight one can't be compared in raw levels but can be compared in departures from their own baselines. We then cumulate this abnormal cost across the event window rather than reading a single day, because a rotation's effect on liquidity cost can build and persist over several sessions; summing it captures the total increase in trading cost instead of one possibly-noisy snapshot. That cumulative quantity is the CAL.

We use three tests to answer one question reliably which is that around rotation episodes, does trading cost widen systematically, or could the widening we see just be noise? All three are testing the same null hypothesis (that the episode has no effect, so abnormal cost is centered on zero).

We run three tests rather than one because trading-cost data can be noisy, and each test guards against a different weakness. The cost changes we study are skewed and heavy-tailed, which makes a standard t-test unsafe, so we lean on tests that don't assume normality. The Wilcoxon signed-rank test does most of the work: it asks whether abnormal costs come out consistently positive across episodes, weighing both the direction and the size of each move. The sign test is a plainer backup that ignores size and simply counts how many episodes widened, so no handful of large moves can carry the result on its own; when it agrees with Wilcoxon, we know the effect is spread across most episodes rather than concentrated in a few. The block bootstrap deals with something else again. Daily spreads are correlated from one day to the next, and tests that overlook this tend to overstate significance, so the bootstrap resamples the data in connected blocks, preserving that day-to-day structure and returning a more honest p-value. Between them, the three cover the bases: Wilcoxon establishes the effect through magnitudes, the sign test shows it isn't the work of outliers, and the bootstrap shows it survives once serial correlation is accounted for. A widening that clears all three is one we can trust.

Our primary sample is the full set of 31 crypto-to-gold rotation episodes identified in Section 2.5; for robustness we also use an independent subsample of 17 episodes spaced at least 60 trading days apart.

### 6.2 Both thin markets deteriorate; the deep ETF does not

For each crypto-to-gold rotation episode we obtain one number per vehicle, which is its cumulative abnormal cost, or CAL. We take the vehicle's high-low range over the [0,+2] window, subtract what that range normally is over the pre-event baseline, and sum across the three days. A positive CAL means that the vehicle is more expensive to trade during the rotation event, relative to its own recent norm. The question this section answers is simple, asked once for each of the 31 episodes, did trading cost rise, and by how much?

For PAXG the answer is yes, and it holds up from every direction we test it (Table 7). The mean CAL is 0.0123 and the median 0.0116, and 71% of the events produce positive CAL for PAXG. Since the mean and median are close to each other, it means that the positive CAL results for PAXG are not driven by outliers.

We then use three tests which interrogate those 31 numbers in different ways, and reporting all three is not redundancy but triangulation, since each test guards against a weakness the others share. The sign test ($p = 0.015$) is the bluntest of the three: it discards magnitudes and counts only how many episodes came out positive. Under the null that a rotation is as likely to tighten the spread as to widen it, we would expect roughly half of the 31 events to show a positive CAL. What we find instead is that the widening runs across most of them rather than being manufactured by a few, and that is what gives us breadth. The Wilcoxon signed-rank test ($p < 0.001$) uses both direction and size, asking whether the large moves are systematically the widenings; its answer tells us the effect is not just broad but weighted toward the bigger episodes, and it reaches that conclusion without assuming the data are normal, which matters because they are skewed and fat-tailed enough that a t-test would not be safe. The bootstrap ($p =$

0.009) then takes on the problem the other two set aside: daily ranges are autocorrelated, and treating the episodes as fully independent would overstate significance, so by resampling in contiguous blocks it keeps that day-to-day dependence intact and still finds that a CAL as large as 0.0123 essentially never turns up by chance. It is the agreement of all three that makes PAXG's widening credible.

The other vehicles sharpen the contrast. XAUT, a different token from a different issuer trading on a different exchange, reproduces the pattern on its own: a mean CAL of 0.0091, again positive, with 69% of events producing a positive CAL and a Wilcoxon p below 0.001. It behaves like an independent second draw rather than an echo of PAXG, and the results line up all the same.

GLD barely moves (with a positive mean CAL of 0.0042) which indicates that the result is significant, and the little it does is the mechanical range-widening any gold vehicle inherits on a big gold-price day, about a third of PAXG's; its borderline p-value is why we read it as little evidence of deterioration rather than a clean effect.

GDXJ is the awkward case as its CAL is 0.0091 against PAXG's 0.0123, and a paired Wilcoxon on the episode-by-episode PAXG-GDXJ difference returns $p = 0.1954$, meaning the gap between the two thin vehicles is indistinguishable from zero. Why that similarity is misleading is the whole point of the next subsection. For now the cleaner figure is the differential CAL, which nets GLD out episode by episode and isolates a genuinely token-specific excess of 0.0080 for PAXG ($p = 0.004$) and 0.0059 for XAUT ($p = 0.008$).

Similar results are obtained using an independent-subsample of 17 events. Results are directionally identical but weaker by construction and reported in the Appendix in Table A2.

**Table 7. Event study, full sample (N = 31), window [0,+2]**

| Vehicle | Mean CAL | Median | Positive CAL % | Sign p | Wilcoxon p | Boot p |
|---|---|---|---|---|---|---|
| **PAXG** | **0.0123** | **0.0116** | **71.0%** | **0.015**** | **<0.001***** | **0.009***** |
| **XAUT** | **0.0091** | **0.0078** | **69.0%** | **0.031**** | **<0.001***** | **0.002***** |
| GLD | 0.0042 | 0.0016 | 67.7% | 0.035** | 0.094* | 0.010** |
| GDXJ | 0.0091 | 0.0063 | 61.3% | 0.141 | 0.073* | 0.013** |
| ΔCAL: PAXG-GLD | 0.0080 | 0.0093 | 80.6% | 0.004*** | 0.004*** | 0.047** |
| ΔCAL: XAUT-GLD | 0.0059 | 0.0056 | 75.9% | 0.041** | 0.008*** | 0.034** |
| PAXG – GDXJ | 0.0032 | 0.0099 | 67.7% | 0.071* | 0.1954 | 0.550 |

*Note.* ***, **, * indicates significance at 1%, 5% and 10 % respectively.

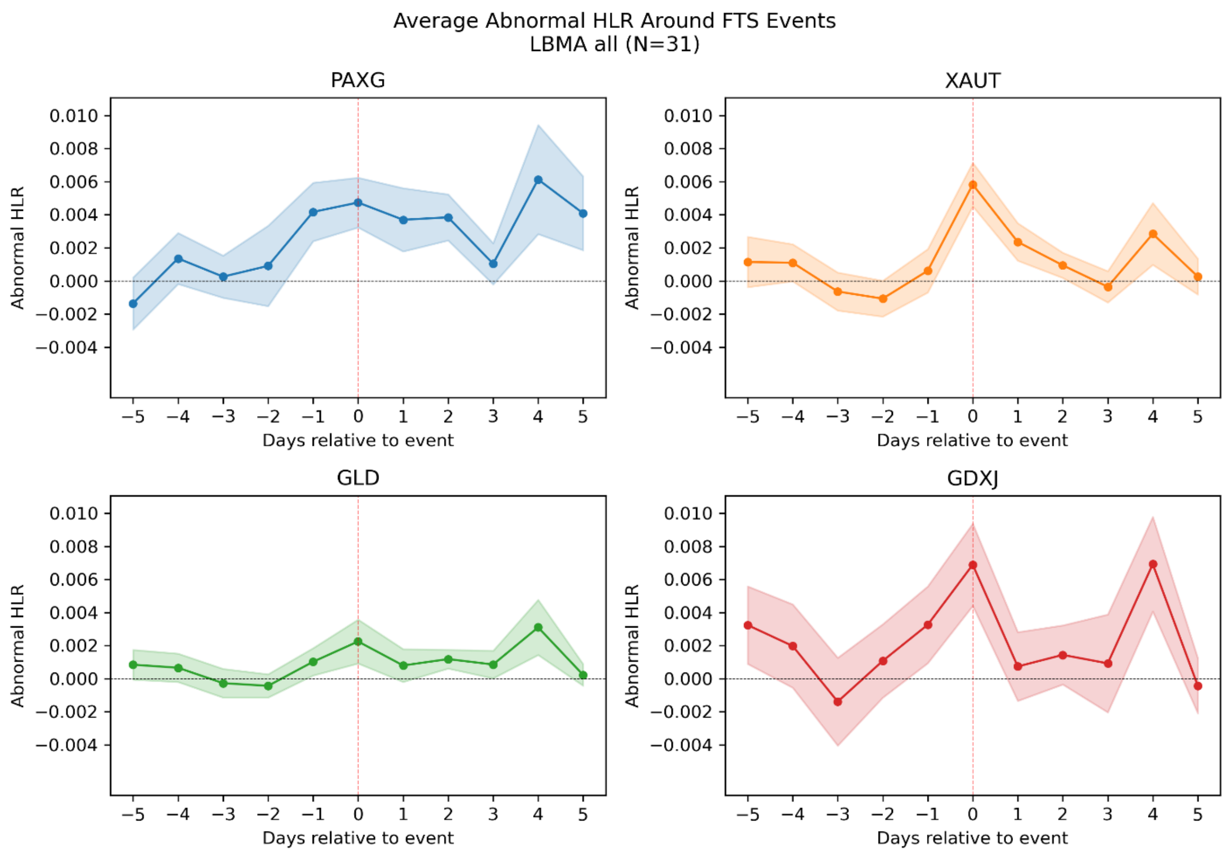


**Figure 2.** *Average abnormal HLR around 31 candidate rotation episodes (±1 SE).*

The event-time paths make the level result visual and reveal its timing (Figure 2, N=31). PAXG sits at or below zero in the pre-event days, meaning slightly cheaper to trade than its trailing baseline, then steps up sharply at t = 0 and stays elevated through the post-event window, the [0,+2] concentration the tests pick up. XAUT shows a cleaner spike: flat and even negative before the event, a pronounced jump at t = 0, partial reversion afterward, independently reproducing PAXG's onset with a different issuer and exchange. GLD is the benchmark; its abnormal HLR hugs zero across the whole window, never separating from baseline. GDXJ does spike at t = 0, but with markedly wider ±1-SE bands, a signal that its deterioration is noisier and more heterogeneous across episodes than the tokens'.

Three features carry the interpretation. Timing: for both tokens the widening is an onset effect that materializes at t = 0 and persists a day or two, consistent with a rotation-driven order-book shock rather than pre-event drift or post-event leakage; the pre-event descent into negative territory also rules out that the tokens were already deteriorating going in. Benchmark contrast: placing GLD's flat path beside the tokens' step-ups reinforces the core level fact that thin gold-linked vehicles deteriorate around rotations while deep-market gold does not. Robustness of shape: the independent subsample of 17 events (Figure A1 in the Appendix) sharpens rather than softens the pattern, confirming the effect is not an artifact of overlapping event windows.

### 6.3 Reading the event study as validation of the DiD

Why should PAXG and GDXJ deteriorate by such similar absolute amounts if, as Sections 4 and 5 argue, they are structurally different? Answering that is what turns the event study into a validation of the DiD instead of a rival to it. The event study measures each vehicle against its own baseline, so both absolute CALs carry the common gold-market component inside them. On a rotation day GLD itself widens as gold-demand trading surges, and every thin vehicle's own-baseline CAL picks up that same widening. The date-fixed-effects DiD strips the component out, and with it gone the two vehicles part ways: PAXG widens relative to GLD (+0.0016) while GDXJ narrows (−0.0018). The event study records the gross outcome; the DiD isolates the net one. Read together, the raw similarity turns out to be not a shared mechanism at all but the mark of a shared shock, and the two designs tell a single story: the event study shows that the token's cost genuinely rises around episodes and GLD's does not, and the DiD shows that, once the common shock is removed, the token and the miner move in opposite directions.

## 7. Robustness

### 7.1 The volatility-conflation concern, honestly

The central measurement objection is that the high-low range is, up to scaling, the Parkinson (1980) volatility estimator, so it captures volatility as well as trading cost. We take this seriously, and we are careful not to claim more than our design supports.

Start with what the date effects do and do not remove. The 1,574 date fixed effects in Section 4 absorb any shock that is *common* across the four vehicles on a given day, the market-wide gold move, the aggregate volatility spike, the shared information event. What survives in the interaction coefficients is a genuinely cross-vehicle differential. This differencing is powerful against *common* volatility. It is not a defense against *idiosyncratic* volatility. If the token's own price becomes more volatile on rotation days for reasons unrelated to the cost of trading, a token-specific information shock, say, that would also show up in the token's HLR and would not be absorbed by the date effects. Daily high-low data cannot separate an idiosyncratic volatility shock to PAXG from a genuine deterioration in PAXG's trading conditions. An earlier and looser version of this defense argued that because GLD's range does not rise on gold-up days, the token's rise must be liquidity rather than volatility. That inference does not go through. GLD's flat path rules out the *common* gold-appreciation volatility that all vehicles share; it says nothing about idiosyncratic volatility specific to the token.

This is why we frame the dependent variable as a proxy for the cost of transacting under stress rather than as illiquidity. The claim we defend is deliberately the narrower one: on rotation days the cost of transacting tokenized gold rises relative to a deep ETF, whether that cost arrives as a wider spread, as thinner depth, or as idiosyncratic price volatility. For an investor trying to move out of Bitcoin and into gold, all three raise the effective cost of execution in the same direction, and it is that effective cost, not a decomposition of its microstructural sources, that the high-low range captures.

The clean separation of trading cost from idiosyncratic volatility requires intraday bid-ask or depth data, which do not exist for these tokens over our sample, and we flag it as the natural next step rather than a box we have already checked.

### 7.2 The panel difference-in-differences across event windows

Section 4 reports the divergence between the tokenization and hierarchy spread over the [0,+2] window, and the natural question is whether that result is an artifact of the window we happened to pick. It is not. Estimating the same vehicle-and-date fixed-effects panel over three horizons - the event day alone, [0,+2], and [0,+5] - shows the divergence is a dynamic that emerges as the rotation plays out, not a property of any single window. Table 8 collects all three.

On the event day itself, PAXG and GDXJ are almost indistinguishable and both widen slightly against GLD (+0.0013 and +0.0014), and the Wald test for equality of their spreads finds nothing to separate them ($p = 0.980$). Extending the window across the two sessions that follow changes the picture completely. PAXG's excess widening firms up to +0.0016 while GDXJ's coefficient crosses zero and turns negative (-0.0018), and the same Wald test now rejects equality at the 5% level ($p = 0.038$). Pushing out to [0,+5] leaves that pattern intact - PAXG at +0.0016, GDXJ deeper into negative territory at -0.0021 - so once the separation appears it holds rather than fades. The two tokens give similar results - their spreads are marginally apart on the event day (Wald $p = 0.111$) but statistically identical over [0, +2] ($p = 0.975$), the cross-issuer replication reasserting itself at the window the mechanism actually concerns.

Two things follow. First, the [0,+2] window was fixed in advance by the event-time analysis of Section 6, not chosen after inspecting these panel estimates, and the divergence is absent at day-0, present at [0,+2], and stable at [0,+5]. A sign that is missing at one horizon, appears at the next, and persists at the third. Thus, there is no event window in which the estimated ordering reverses. The divergence is absent at day-0, emerges over [0, +2], and is preserved over [0, +5].

Second, is that the divergence takes a day or two to surface. The event day captures mostly the raw, simultaneous shock, in which the common gold-market component dominates and lifts every vehicle's range together - which is why PAXG and GDXJ look alike at day-0. Only once that common component works through, and the vehicle-specific responses take over, that the two diverge. The token widens against GLD, while the miner - whose stressor is equity fear rather than crypto rotation -fails to match GLD's own gold-driven widening and narrows relative to it. The window progression is the gross-versus-net logic of Section 6.3 unfolding in time. At day-0 the gross common shock swamps the difference; by [0, +2] the net divergence has emerged.

**Table 8. Panel DiD across event windows: vehicle × FTS interactions relative to GLD**

| | Model 1 (Day 0) | | Model 2 ([0, +2]) | | Model 3 ([0, +5]) | |
|---|---|---|---|---|---|---|
| **Term** | **Coeff** | **p-value** | **Coeff** | **p-value** | **Coeff** | **p-value** |
| PAXG × FTS | +0.0013 | 0.365 | +0.0016 | 0.111 | +0.0016 | 0.089 |
| XAUT × FTS | +0.0035 | 0.015 | +0.0016 | 0.042 | +0.0005 | 0.333 |
| GDXJ × FTS | +0.0014 | 0.515 | -0.0018 | 0.168 | -0.0021 | 0.052 |

**Wald tests of equality**

| | Model 1 (Day 0) | | Model 2 ([0, +2]) | | Model 3 ([0, +5]) | |
|---|---|---|---|---|---|---|
| **Hypothesis** | **Wald-stat** | **p-value** | **Wald =stat** | **p-value** | **Wald-stat** | **p-value** |
| $H_0$: PAXG = GDXJ | 0.000 | 0.980 | 4.327 | 0.038 | 8.489 | 0.003 |
| $H_0$: PAXG = XAUT | 2.537 | 0.111 | 0.001 | 0.975 | 1.550 | 0.213 |

*Notes.* Each column is a separate estimation of $HLR_{it} = \alpha_i + \gamma_t + \beta\cdot(\text{vehicle}_i \times FTS_t) + \varepsilon_{it}$, with vehicle effects and 1,574 date effects throughout; GLD is the omitted reference, so each coefficient is that vehicle's excess widening relative to GLD on FTS days. Standard errors are clustered by date (1,574 clusters); entity clustering is infeasible with four vehicles.

### 7.3 Ruling out the alternatives in event study: noise, gold, and one token

Three alternative explanations could reproduce the Table 7 result without any token-specific mechanism, first, if PAXG is simply a noisy series, second, that we are picking up gold volatility rather than trading cost, and third that the finding is peculiar to a single token. We address each in turn.

The first concern is that the abnormal widening reflects nothing more than the ordinary variability of a thin trading series, so that almost any set of dates would look similar. To check this, we re-ran the full event-study procedure on 500 randomly drawn sets of 31 non-event dates from the eligible sample, building a placebo distribution for each vehicle (Table 9A). The two tokens stand well clear of theirs. PAXG's observed CAL of 0.0123 sits far above its placebo mean of −0.0022 and exceeds 98.6% of the random draws (empirical p = 0.014), and XAUT's 0.0091 exceeds 99.6% of its own draws (p = 0.004). The benchmark and the miner behave differently. GLD's event-window CAL falls within the range its random dates routinely produce (p = 0.140), which is just what we would expect of a vehicle that barely reacts, and GDXJ's does the same (p = 0.348), the latter because the miner's own-baseline range is so volatile that a CAL the size of its event figure is unremarkable against random dates; its placebo standard deviation, 0.0213, is roughly four times the tokens' and swamps the signal.

This is the split we would hope to see. For the tokens, the placebo pins the widening to the rotation episodes themselves, rather than to the event-study machinery or to a fortunate choice of dates. GLD's null result fits its job as the benchmark that does not move, and GDXJ's null follows from how noisy the miner's raw range is to begin with. One caveat on how this is to be interpreted. The test establishes that the tokens' event-window costs belong to the episodes, not that tokenization itself is what drives them. That stronger claim rests on the differential CALs and on the DiD and SUR results, and not on the placebo.

**Table 9A. Placebo validation based on 500 random event sets**

| Vehicle | Placebo Mean CAL | Placebo SD | Actual CAL | Exceeds Placebo Draws | Empirical p |
|---|---|---|---|---|---|
| PAXG | -0.0022 | 0.0059 | 0.0123 | 98.6% | 0.014 |
| XAUT | -0.0080 | 0.0058 | 0.0091 | 99.6% | 0.004 |
| GLD | 0.0008 | 0.0032 | 0.0042 | 86.0% | 0.140 |
| GDXJ | 0.0021 | 0.0213 | 0.0091 | 65.2% | 0.348 |

*Note***.** Placebo results are based on 500 random samples of 31 non-event dates drawn without replacement from the eligible trading days. For each draw, the complete event-study procedure was repeated using the same estimation window and event window as the main analysis. Placebo Mean CAL and Placebo SD denote the mean and standard deviation across the 500 placebo experiments. Actual CAL is the observed abnormal cost during the identified LBMA flight-to-safety events. Exceeds Placebo Draws reports the percentage of placebo experiments with CAL values below the observed event-study estimate. The empirical p-value equals the proportion of placebo draws at least as large as the observed CAL (minimum attainable $p = 1/(500+1) \approx 0.002$).

The second is gold volatility. A large gold move mechanically widens every vehicle’s range, so the token effect might simply be tracking gold. To test this, within each event window we strip out the part of each vehicle’s range that co-moves with the gold return and re-run the event study (Table 9B). If gold volatility were driving the result, removing it would shrink the token effect; instead the adjusted token CAL is, if anything, slightly larger than the raw figure (PAXG rises from 0.0123 to 0.0129, XAUT from 0.0091 to 0.0101), and both tokens remain significant on the sign, Wilcoxon, and bootstrap tests alike. GLD and GDXJ keep small positive adjusted CALs, but these are flagged only by the mean-driven bootstrap (sign $p = 0.14$ and $0.24$); the tokens are the only vehicles significant across all three tests. Because the gold component is common to all four vehicles, removing it strips out what they share and leaves the token-specific part exposed, a purely shared gold effect would not survive the adjustment, and the tokens’ effect does.

**Table 9B. Gold-return-adjusted event study (N = 31, window [0,+2])**

| Vehicle | Adj CAL | Median | Positive CAL % | Sign p | Wilcoxon p | Boot p |
|---|---|---|---|---|---|---|
| **PAXG** | **0.0129** | **0.0114** | **74.2%** | **0.005*** ** | **<0.001*** ** | **0.007*** ** |
| **XAUT** | **0.0101** | **0.0071** | **75.9%** | **0.004*** ** | **<0.001*** ** | **0.001*** ** |
| GLD | 0.0043 | 0.0017 | 61.3% | 0.141 | 0.079* | 0.004*** |
| GDXJ | 0.0088 | 0.0026 | 58.1% | 0.237 | 0.082* | 0.016** |

*Note.* Each vehicle's HLR is orthogonalized to the contemporaneous gold return within the event window before the abnormal-cost sum is formed; N = 31 (XAUT = 29). Stars mark the significance of each test statistic: *** p<0.01, ** p<0.05, * p<0.10. PAXG and XAUT are significant on all three tests; GLD and GDXJ only on the mean-driven bootstrap.

The third is that the result is specific to one token. PAXG and XAUT behave alike on every dimension that matters (Table 9C), the size of the deterioration (0.0123 versus 0.0091), its direction, its significance on all three tests, and its survival of the gold adjustment, even though they come from different issuers (Paxos versus Tether), trade on different primary exchanges (Coinbase versus Bitfinex), and hold their gold in vaults in different countries. A result driven by Paxos's setup or Coinbase's trading conditions would not reproduce in a Tether token on Bitfinex; this one does, which points to a shared feature of tokenized gold rather than an idiosyncrasy of a single vehicle. Together the three checks remove the "noisy asset," "just gold," and "single token" explanations in turn.

**Table 9C. Cross-issuer replication: PAXG and XAUT (N = 31, window [0,+2])**

| Vehicle | CAL [0,+2] | Gold-adj CAL | Positive CAL % | Sign p | Wilcoxon p | Boot p |
|---|---|---|---|---|---|---|
| PAXG | 0.0123 | 0.0129 | 71.0% | 0.015** | <0.001*** | 0.009*** |
| XAUT | 0.0091 | 0.0101 | 69.0% | 0.031** | <0.001*** | 0.002*** |

*Note*. CAL and its three tests are the raw [0,+2] event-study figures; the gold-adjusted CAL is from Table 9B. N = 31 for PAXG and 29 for XAUT. *** p<0.01, ** p<0.05, * p<0.10.

### 7.4 Robustness of the Mechanism to the Maturation and VIX Regime Dummies

The SUR system of Section 5 carries two dummy variables- the July-2020 maturation dummy in the tokenization-spread equation and the August-2024 VIX-regime dummy in the miner-spread equation, each absorbing a one-time level shift documented in Section 3.3. We ask whether the headline result, the VIX channel that separates the two spreads, is in any way an artifact of how those breaks are handled? It is not because the dummies shift the intercept of a spread, whereas the VIX differential is a matter of slope, of how each spread responds to equity fear. Removing a level control cannot manufacture or dissolve a slope difference, so the separation should survive, and it does.

We re-estimate the system three ways, dropping the maturation dummy, dropping the VIX-regime dummy, and dropping both, and track the only confirmed differential, the cross-equation VIX restriction. Dropping the maturation dummy leaves the picture essentially unchanged: the bond-stress coefficient remains numerically identical (−0.0005), while its p-value shifts only modestly from 0.0218 to 0.0374. The VIX channel separation is likewise unaffected (Wald $W = 11.10$, $p = 0.0009$; LHS/TLS coefficient ratio ≈ 7.36×). Dropping the VIX-regime dummy from the miner equation also leaves the differential intact (Wald $W = 7.75$, $p = 0.0054$; ratio ≈ 8.82×), as does dropping both together (Wald $W = 9.85$, $p = 0.0017$; ratio ≈ 7.12×). Against the full-sample benchmark of Wald $W= 8.87$ ($p = 0.0029$) and an LHS/TLS VIX coefficient ratio of approximately 9.14×, these changes are quantitatively modest and do not alter the substantive inference.

The maturation break marks the point at which tokenized-gold liquidity deepened and the spread settled onto a permanently lower level; the VIX-regime break marks a discrete shift in the miner spread's baseline. Both are features of where the spreads sit, not of how they react when equity fear rises. Absorbing them sharpens the estimates by removing variance the model would otherwise have to explain, but the stress-conditional response that identifies the mechanism, lives in the interaction between the spread and the VIX, which those level controls never touch. That the VIX differential is unchanged whether or not the dummies are present is not a lucky robustness result but what one would expect if the mechanism is genuine, and it rules out the alternative reading in which the break controls, rather than the equity-fear channel, are quietly doing the work.

**Table 10. SUR robustness: the VIX differential with and without the break dummies**

| Specification | VIX Wald W | P | LHS/TLS ratio |
|---|---|---|---|
| Full (both dummies) | 8.87 | 0.0029 | 9.14× |
| Drop maturation dummy | 11.10 | 0.0009 | 7.36× |
| Drop VIX-regime dummy | 7.75 | 0.0054 | 8.82× |
| Drop both | 9.85 | 0.0017 | 7.12× |

*Note. The TLS/LHS SUR of Table 4, re-estimated after dropping the July-2020 maturation dummy, the August-2024 VIX-regime dummy, and both. W and p are the cross-equation Wald test of equality of the two VIX coefficients ($H_0$: $\beta TLS = \beta LHS$); the ratio is the LHS VIX coefficient divided by the TLS VIX coefficient. N = 1,526 in every specification; HAC-robust (Newey–West) kernel covariance.*

## 8. Discussion

### 8.1 One primary result, one mechanism, one validation

The three designs answer nested questions and are mutually consistent. The date-fixed-effects DiD, our primary result, shows that once the common daily shock is absorbed, the two thin markets diverge in opposite relative directions (Wald p = 0.038) implying that tokenized gold's transaction cost rises against GLD while the junior-miner spread falls. The SUR helps identify the mechanism i.e., the two spreads have different stress-response functions, and equity-market fear is the one force that formally separates them, driving the miner spread 9-11 times more than the token spread (Wald p = 0.0029). The event study validates that both thin vehicles deteriorate in absolute terms while GLD does not, and the raw token-miner similarity is a shared-shock artifact that the DiD nets out. Thin-market status is the necessary condition; the stress channel and the relative dynamics are what distinguish tokenized gold from a conventional thin gold vehicle.

An analogy sharpens the distinction. The junior-miner spread behaves like an equity product, when the VIX rises, its cost against GLD widens. Tokenized gold sits on crypto-market foundations, it is strained when Bitcoin sells off. A rotation episode stresses both, which is why their event-study magnitudes look alike; but the underlying force, and therefore the remedy, differs. The cross-issuer replication (PAXG ≡ XAUT) says the crypto-native foundation is a property of the tokenized-gold category, not of one issuer's plumbing, so remedies should target the category, exchange listings, dedicated market-maker programs, native derivatives, rather than any single token.

### 8.2 Economic significance and an investor's takeaway

The magnitude matters. The maturation dummy enters the Section 5 specification with a coefficient of about −0.008, and because the high-low range is measured as a fraction of price, that figure reads directly as 0.8 percentage points of intraday range, so the break lowered the calm-time tokenization spread by roughly that much. For an investor moving out of Bitcoin and into gold during a bout of crypto stress, GLD gives the cleanest execution; a tokenized-gold token stays tradable but costs meaningfully more to transact, because the same Bitcoin selling that prompts the rotation also weighs on the token's market-makers, who sit on the very same venues. GDXJ's fragility works differently, since it is driven by equity fear, so a rotation that is purely crypto-driven and arrives without accompanying equity stress may leave the miner steadier than its raw event-study CAL would suggest.

### 8.3 Limitations

Four limitations bound the claims, and we would rather state them plainly than bury them. The first is measurement. The high-low range is a low-frequency proxy, and it folds trading cost together with idiosyncratic volatility; the date-fixed-effects design and the gold-return adjustment take care of the common-volatility part, but neither can isolate trading cost on its own in the time-series regressions, and cleaner intraday measures simply do not exist for these tokens over our sample. The second is the cross-section. The investable tokenized-gold universe with enough history is basically PAXG and XAUT, so our cross-issuer replication runs across two issuers, which is a genuine step up from a single-token study but still a thin panel. The third concerns what the episodes are. They are Bitcoin-down, gold-up co-movements consistent with a rotation, not verified capital flows, so we make no causal claim that tokenization is what produces the widening. The fourth is the benchmark. GLD is an imperfect counterfactual, an ETF with a microstructure of its own, and here we lean on the DiD's date effects and the contrast with GDXJ to hold the comparison steady. None of the four overturns the primary result; each simply marks the point where progress has to come from a better dataset rather than a better argument.

## 9. Conclusion

Tokenized gold promises the safety of physical gold with the settlement convenience of a crypto token. But the literature had mostly asked whether these instruments track the gold price or inherit gold's safe-haven properties, these are questions about returns, not about whether an investor can actually use them when a crisis hits. A safe haven that becomes expensive to transact in the moment it's needed is only a partial safe haven. That question had simply not been asked.

Our answer is that tokenized gold does become more costly to trade when investors are most likely to rotate out of Bitcoin and into gold. But the more interesting finding is ‘how’ this happens. A conventional thin gold vehicle-junior miners-deteriorates too, but through a completely different mechanism. The miner spread widens when equity markets panic; the token spread widens when crypto markets sell off. Both are fragile, but for different reasons, and the divergence only becomes visible once you strip away the common gold-market shock that hits every gold-linked vehicle on a given day.

That divergence is what makes the result worth taking seriously. If tokenized gold simply inherited generic thin-market fragility, it would behave like any other thinly traded gold vehicle, but our finding is that it does not. The token's stress channel is not equity-driven, but crypto native. And the fact that two independently issued tokens PAXG and XAUT, from different issuers, on different exchanges, produce identical results tells us this is a feature of the category, not a result that holds only for a specific issuer only.

There is a broader point here about how we think about safe havens. An asset's price behavior is only part of the story; its microstructure, how it actually trades when markets are under stress, matters just as much. Tokenized gold may offer gold-like returns with crypto-like settlement, but it also inherits crypto-like fragility in its trading infrastructure at the wrong moment. Investors who expect GLD-like execution should adjust their expectations accordingly.

We are careful about what we have not established. The high-low range is a proxy, not a definitive measure of liquidity: it folds together spread, depth, and idiosyncratic volatility, and daily data cannot pull them apart. Cleaner intraday measures do not exist for these tokens over our sample, and we flag that as the natural next step. Nor can we prove that capital actually moved from Bitcoin into gold on the episodes we identify; what we observe is a price pattern consistent with that reading, not the flow itself.

What we can say is that tokenized gold's relative transaction costs rise under stress. The effect is economically material, holds up across specifications, and replicates across issuers. And it is not a reaction to generic market fear - COVID showed that much - but to the particular strain of a crypto sell-off rotating into gold. For an investor, that is the practical takeaway. For the literature, it is a reminder that safe-haven status rests not only on what an asset is, but on how it trades when trading matters most.

## References

Aloui, C., Ben Hamida, H., Yarovaya, L., 2021. Are Islamic gold-backed cryptocurrencies different? Finance Research Letters 39, 101615.

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

Ashfaq, M., Pfeifer, M., Gürpinar, T., Gulum, M.A., 2026. Tokenized gold in crypto markets: tracking accuracy and portfolio performance. FinTech 5 (1), 19.

Baker, S.R., Bloom, N., Davis, S.J., 2016. Measuring economic policy uncertainty. Quarterly Journal of Economics 131, 1593-1636.

Baur, D.G., Hoang, L.T., 2021. A crypto safe haven against Bitcoin. Finance Research Letters 38, 101431.

Baur, D.G., Lucey, B.M., 2010. Is gold a hedge or a safe haven? An analysis of stocks, bonds and gold. Financial Review 45, 217-229.

Baur, D.G., McDermott, T.K., 2010. Is gold a safe haven? International evidence. Journal of Banking and Finance 34, 1886-1898.

Beckmann, J., Berger, T., Czudaj, R., 2015. Does gold act as a hedge or a safe haven for stocks? A smooth transition approach. Economic Modelling 48, 16-24.

Bouri, E., Molnár, P., Azzi, G., Roubaud, D., Hagfors, L.I., 2017. On the hedge and safe haven properties of Bitcoin. Finance Research Letters 20, 192-198.

Bouri, E., Shahzad, S.J.H., Roubaud, D., 2020. Cryptocurrencies as hedges and safe-havens for US equity sectors. Quarterly Review of Economics and Finance 75, 294-307.

Brunnermeier, M.K., Pedersen, L.H., 2009. Market liquidity and funding liquidity. Review of Financial Studies 22, 2201-2238.

Chordia, T., Roll, R., Subrahmanyam, A., 2001. Market liquidity and trading activity. Journal of Finance 56, 501-530.

Chung, K.H., Zhang, H., 2014. A simple approximation of intraday spreads using daily data. Journal of Financial Markets 17, 94-120.

Dyhrberg, A.H., Foley, S., Svec, J., 2018. How investible is Bitcoin? Analyzing the liquidity and transaction costs of Bitcoin markets. Economics Letters 171, 140-143.

Gelman, A., Stern, H., 2006. The difference between “significant” and “not significant” is not itself statistically significant. The American Statistician 60, 328-331.

Hoque, M.E., Billah, M., Alam, M.R., Tiwari, A.K., 2024. Gold-backed cryptocurrencies: a hedging tool against categorical and regional financial stress. Global Finance Journal 60, 100964.

Jalan, A., Matkovskyy, R., Yarovaya, L., 2021. “Shiny” crypto assets: a systemic look at gold-backed cryptocurrencies during the COVID-19 pandemic. International Review of Financial Analysis 78, 101958.

Klein, T., Thu, H.P., Walther, T., 2018. Bitcoin is not the new gold. International Review of Financial Analysis 59, 105-116.

Lee, J., Strazicich, M.C., 2003. Minimum Lagrange multiplier unit root test with two structural breaks. Review of Economics and Statistics 85, 1082-1089.

Makarov, I., Schoar, A., 2020. Trading and arbitrage in cryptocurrency markets. Journal of Financial Economics 135, 293-319.

Mnif, E., Salhi, B., Trabelsi, L., Jarboui, A., 2022. Efficiency and herding analysis in gold-backed cryptocurrencies. Heliyon 8 (12), e11982.

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

Parkinson, M., 1980. The extreme value method for estimating the variance of the rate of return. Journal of Business 53, 61-65.

Pesaran, M. H. (2007). A Simple Panel Unit Root Test in the Presence of Cross Section Dependence. *Journal of Applied Econometrics*, 22(2), 265-312

Reboredo, J.C., 2013. Is gold a safe haven or a hedge for the US dollar? Implications for risk management. Journal of Banking and Finance 37, 2665-2676.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. Transactions of the American Mathematical Society, 54(3), 426–482.

Wang, G.-J., Ma, X.-Y., Wu, H.-Y., (2020). Are stablecoins truly diversifiers, hedges, or safe havens against traditional cryptocurrencies as their name suggests? Research in International Business and Finance 54, 101225.

Wang, Yuhan, and Di Xiao. Novel symbolic detection for flight-to-safety in Bitcoin and investigation of information flow dynamics alongside multiple markets. Physica A: Statistical Mechanics and its Applications 660 (2025): 130358.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. Biometrics Bulletin, **1**(6), 80–83.

Wasiuzzaman, S., Haji Abdul Rahman, H.S.W., 2021. Performance of gold-backed cryptocurrencies during the COVID-19 crisis. Finance Research Letters 43, 101958.

Zellner, A., 1962. An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. Journal of the American Statistical Association 57, 348-368.

## Appendix

**Table A1 : List of 31 Bitcoin – to – Gold Rotation Events**

| Event | Date | BTC Return (%) | LBMA Gold Return (%) | Verified Market Context |
|---|---|---|---|---|
| **1** | 19 Feb 2020 | -5.14 | +0.90 | Early COVID-19 market uncertainty |
| **2** | 24 Feb 2020 | -2.80 | +1.71 | Global market sell-off following the spread of COVID-19 outside China |
| **3** | 01 Dec 2020 | -4.28 | +2.70 | _____ |
| **4** | 22 Feb 2021 | -5.97 | +1.18 | Broad cryptocurrency market correction |
| **5** | 15 Mar 2021 | -5.90 | +1.10 | Rising U.S. Treasury yields ahead of the March FOMC meeting |
| **6** | 21 Apr 2021 | -4.65 | +1.14 | Broad cryptocurrency market correction following Bitcoin's April record high |
| **7** | 04 May 2021 | -7.00 | +1.69 | Continued cryptocurrency market correction |
| **8** | 19 May 2021 | -14.81 | +1.16 | Renewed Chinese cryptocurrency regulatory restrictions |
| **9** | 21 Sep 2021 | -5.15 | +0.95 | China Evergrande debt crisis |
| **10** | 10 Nov 2021 | -3.00 | +1.74 | Higher-than-expected U.S. CPI inflation release |
| **11** | 05 Jan 2022 | -5.21 | +0.82 | Federal Reserve tightening expectations after FOMC minutes |
| **12** | 17 Feb 2022 | -8.11 | +1.64 | Escalating Russia-Ukraine tensions |
| **13** | 05 May 2022 | -8.19 | +1.53 | Market reaction to Federal Reserve policy decision |
| **14** | 09 Nov 2022 | -15.49 | +2.16 | FTX collapse |
| **15** | 09 Mar 2023 | -6.44 | +0.83 | Silicon Valley Bank crisis |
| **16** | 14 Nov 2023 | -2.68 | +1.94 | Lower-than-expected U.S. CPI inflation data |
| **17** | 21 Nov 2023 | -4.54 | +1.91 | __________ |
| **18** | 12 Jan 2024 | -7.88 | +1.30 | U.S. spot Bitcoin ETF launch |
| **19** | 29 Feb 2024 | -2.11 | +0.76 | Heightened cryptocurrency market volatility during the ETF adoption period |
| **20** | 05 Mar 2024 | -6.86 | +1.72 | Heightened volatility following Bitcoin's new all-time high |
| **21** | 02 Apr 2024 | -6.30 | +2.24 | Record gold prices amid geopolitical tensions |

| | | | | |
|---|---|---|---|---|
| **22** | 12 Apr 2024 | -4.17 | +2.35 | Middle East geopolitical tensions |
| **23** | 10 May 2024 | -3.65 | +1.99 | ______ |
| **24** | 03 Jul 2024 | -3.04 | +1.26 | Mt. Gox creditor repayment concerns |
| **25** | 01 Oct 2024 | -4.02 | +1.42 | Iran-Israel military escalation |
| **26** | 09 Dec 2024 | -3.83 | +1.30 | Correction following Bitcoin's US$100,000 milestone |
| **27** | 03 Mar 2025 | -9.08 | +1.62 | Renewed U.S. tariff announcement |
| **28** | 10 Apr 2025 | -3.64 | +2.18 | U.S. tariff policy and trade tensions |
| **29** | 12 Jun 2025 | -2.57 | +1.84 | Escalating Israel-Iran conflict |
| **30** | 01 Aug 2025 | -2.13 | +1.44 | U.S. reciprocal tariff announcement |
| **31** | 22 Sep 2025 | -2.24 | +1.53 | Federal Reserve rate-cut expectations |

Note: 31 algorithmically identified crypto-to-gold rotation or flight-to-safety (FTS) episodes used in the analysis, defined by concurrent negative Bitcoin returns and positive gold returns. The final column provides the dominant contemporaneous macroeconomic, geopolitical, or cryptocurrency-related market development identified from contemporaneous news sources to contextualize each episode.

**Table A2. Event study, independent subsample (N = 17), window [0,+2]**

| Vehicle | Mean CAL | Median | Positive CAL % | Sign p | Wilcoxon p | Boot p |
|---|---|---|---|---|---|---|
| **PAXG** | **0.0103** | **0.0112** | **70.6%** | **0.072*** | **0.002***** | **0.053*** |
| **XAUT** | **0.0076** | **0.0084** | **68.8%** | **0.105** | **0.006***** | **0.048**** |
| GLD | 0.0014 | 0.0013 | 70.6% | 0.072* | 0.165 | 0.214 |
| GDXJ | 0.0099 | 0.0073 | 76.5% | 0.025** | 0.066* | 0.033** |
| ΔCAL: PAXG−GLD | 0.0089 | 0.0093 | 76.5% | 0.025** | 0.002*** | 0.087* |
| ΔCAL: XAUT−GLD | 0.0062 | 0.0057 | 75.0% | 0.038** | 0.013** | 0.094* |

*Note. The independent subsample requires ≥60 trading days between episodes, ruling out overlapping estimation windows. PAXG deterioration is confirmed (Wilcoxon p = 0.002); GLD remains insignificant (Wilcoxon p = 0.165). The token-specific excess (ΔCAL PAXG−GLD) is significant (Wilcoxon p = 0.002). Results are directionally identical to the full sample (Table 7) but with reduced power from the smaller sample. Stars: *** p<0.01, ** p<0.05, * p<0.10.*

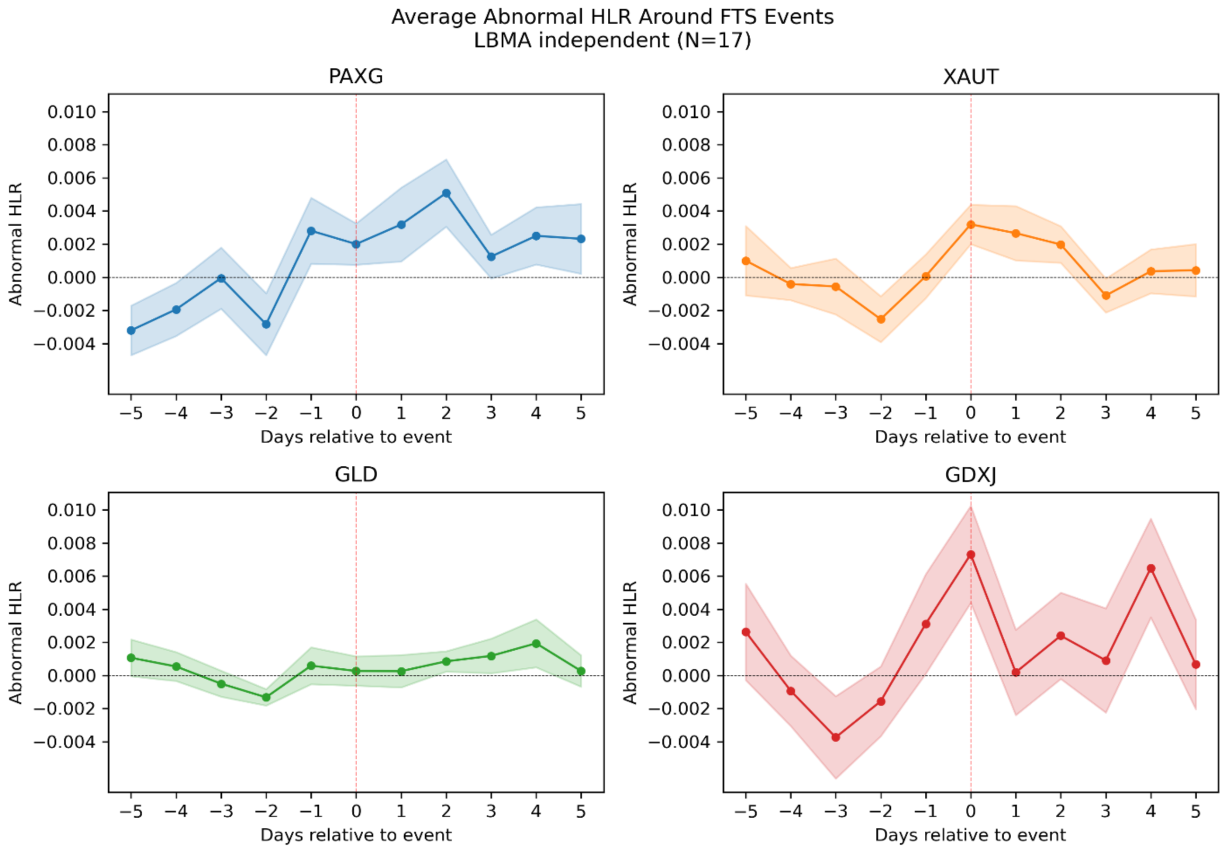


**Figure A1.** Average abnormal HLR around 17 independent episodes (±1 SE). The independent subsample is cleaner: PAXG transitions to positive post-event abnormal cost peaking at t = +2, while GLD stays near zero throughout**.**