Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Beyond Benford's Law: A Rotation Benchmark for First-Digit Screening in Small Bank Panels

Dien Noviany Rahmatika

diennovi@upstegal.ac.id

Universitas Pancasakti Tegal

Dewi Amaliah Nafiati

Universitas Pancasakti Tegal

Foster Mesreh

Universitas Pancasakti Tegal

Research Article

Keywords: Benford's Law, rotation benchmark, fraud hexagon, first-digit screening, bank financial reporting, Indonesian banks

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11518726/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Beyond Benford's Law: A Rotation Benchmark for First-Digit Screening in Small Bank Panels

Dien Noviany Rahmatika[1]*, Dewi Amaliah Nafiati[2], Foster Mesreh[3]
[1] Universitas Pancasakti Tegal, Department of Accounting, Tegal City, Central Java, Indonesia
E-mail: diennovi@upstegal.ac.id
ORCID: 0000-0002-2012-8525

[2] Universitas Pancasakti Tegal, Department of Economics Education, Tegal City, Central Java, Indonesia
E-mail: dewiamaliah@upstegal.ac.id
ORCID: 0000-0002-0278-1248

[3] Universitas Pancasakti Tegal, Department of Management, Tegal City, Central Java, Indonesia
E-mail: mesrehfoster5@gmail.com
ORCID: 0009-0003-8648-2564

*Corresponding author: Dien Noviany Rahmatika, diennovi@upstegal.ac.id

Abstract

Fixed Benford conformity bands classify every unit as nonconforming when each observation rests on a few figures, and ratios built from the tested amounts can be associated with a digit statistic through arithmetic alone. We propose a panel-based benchmark that builds a bank-year index of first-digit deviation from five reported amounts and judges conformity and regression coefficients against a rotation reference, which preserves each bank's ratios and growth path and randomises its scale. In simulations without departures, the benchmark flagged 3.7% to 3.9% of units at the bank level and 5.6% to 7.2% pooled, but its power against altered leading digits was at most 0.07 for a single bank and 0.26 pooled. We illustrate it on twelve Indonesian banks with four fraud hexagon proxies. The pooled deviation lay at the 67th percentile of the benchmark, and one of 13 units exceeded its 95th percentile. In fixed-effects models the benchmark reproduced about 17% of the leverage coefficient, and the remainder could not be distinguished from zero; the other proxies were indistinguishable from zero, with minimum detectable effects 1.7 to 3.6 times the estimates, so the illustration shows the limits of such a panel. The extracted amounts do not satisfy the balance identity within 1% of assets in 33 of 72 bank-years, and the proxies are unvalidated, so results are exploratory. Restricting the panel to the 39 bank-years that reconcile within 0.1% left these conclusions unchanged. Digit-based studies using ratios built from tested figures should benchmark associations against arithmetic dependence.

Keywords: Benford's Law; rotation benchmark; fraud hexagon; first-digit screening; bank financial reporting; Indonesian banks

JEL classification: C23; G21; M41; M42

1. Introduction

1.1 Background and research gap

The figures that banks report inform investors, regulators, creditors and auditors about performance, risk and prudential soundness. Irregular reporting need not appear as an explicit misstatement; it may surface only as an unusual numerical pattern, and supervisors and auditors therefore look for inexpensive screening devices that indicate where deeper examination is warranted. We treat detection in this sense, as the identification of early signals that direct examination and not as the classification of a report as fraudulent (Shahana et al. 2023).

Two established bodies of work bear on this task, and they have rarely been combined. The first is Benford's Law, which describes the expected distribution of leading digits in many naturally occurring numerical datasets; accountants use it to screen data that merit closer scrutiny (Durtschi et al. 2004), and it remains an active field, with recent extensions to financial risk (Cerqueti et al. 2024) and banking (Grammatikos and Papanikolaou 2021). A deviation does not identify its own cause, however: it may reflect accounting conventions (Nguyen and Duong 2022), the number of figures tested (Barney and Schulzke 2016) or the nature of the data, as well as manipulation. The second is the fraud hexagon, which interprets fraud as shaped by pressure, opportunity, rationalisation, capability, arrogance and collusion, and which has been applied to Indonesian banking (Achmad et al. 2023).

Whether hexagon proxies are associated with numerical anomalies identified directly from reported figures is not known, because the banking studies reviewed in Section 2.2 use other outcomes. Two features of bank data make the question hard to answer. First, an annual report supplies only a handful of comparable aggregate amounts per bank-year, and the mean absolute deviation (MAD) of first digits depends on the number of figures tested (Barney and Schulzke

2016), so with five amounts even independent Benford draws lie above the conventional conformity bands (Section 3.4). Second, the amounts of a bank-year are linked by the balance identity and a common scale, and a ratio such as leverage, the natural proxy for pressure, is built from the figures whose digits are tested, so some association could arise from arithmetic alone. Conformity and within-bank associations must therefore be judged against the number of figures tested and the dependence among them.

1.2 Framework, research questions and hypotheses

The study joins the two bodies of work through a bank-year index, BENFORD_MAD, the MAD of the five amounts of a bank-year. The reasoning is that if hexagon conditions make manipulation more likely, and manipulation disturbs the leading-digit pattern in a way the index can register, a bank should show a larger departure in years of stronger conditions than in its other years (Section 2.3). The study therefore examines associations with the index and not with fraud itself.

The purpose is to develop and apply a panel-based screening procedure for first-digit conformity when each observation rests on few figures, using twelve Indonesian banks from 2020 to 2025. Four research questions (RQs) guide the study. RQ1 asks to what extent the pooled and bank-level first-digit distributions depart from the Benford distribution. RQ2 asks how the index differs between banks and between years, and what share of its variation lies between banks. RQ3 asks to what extent proxies for pressure, capability, arrogance and collusion are associated, within banks over time, with the index. RQ4 asks how stable the RQ3 associations are under year effects, alternative inference, alternative measures, restricted samples and a benchmark that reproduces the arithmetic dependence between regressors and index.

RQ3 is addressed through one hypothesis and three exploratory questions. H1 states that, within banks, higher leverage (PRESS, total liabilities divided by total assets), a proxy for pressure, is

associated with higher BENFORD_MAD. The exploratory questions ask whether years with a documented board change (E1, proxy CAP), a disclosed political or public-office background among directors (E2, ARO) and disclosed government-project or cooperation activity (E3, COL) are associated with BENFORD_MAD, in either direction. Only H1 is posed as directional and primary, because PRESS varies in every bank and year whereas the binary proxies change within only a few banks (Section 3.3), so that a directional test of them would rest on very few transitions; E1 to E3 are therefore reported descriptively without a verdict.

1.3 Contribution and scope

The contribution is principally methodological: a rotation benchmark for judging first-digit conformity and digit-based associations in a small panel, five decision rules (R1 to R5; Section 3.7), and a minimum detectable effect (MDE) reported beside every coefficient. A simulation reports the benchmark's false-positive rate and power against controlled digit manipulations (Section 3.9). The Indonesian panel serves to illustrate the procedure and not to test the fraud hexagon: the extracted amounts and the binary proxies are not validated (Sections 3.2 and 3.3), and the empirical results show how the benchmark behaves on real panel data. The benchmark is a safeguard against false alarms and against arithmetic dependence, not a detector of manipulation. It indicates whether an observed deviation or digit-based coefficient exceeds what the scale of the figures and their accounting relations would produce; with five amounts per bank-year, its power to reveal altered leading digits is limited (at most 0.07 for a single bank and 0.26 for the pooled panel), and the MDE makes this limit visible to the reader.

Calibrating first-digit statistics to the data at hand is not new. The conformity bands of Nigrini (2012a) are fixed values, and Barney and Schulzke (2016) showed that they generate false alarms unless the number of figures tested is taken into account. Morrow (2014) proposed test values that

depend on sample size, and Lesperance et al. (2016) showed how sample size governs what goodness-of-fit statistics and simultaneous confidence intervals can reject. Cho and Gaines (2007) introduced a distance-based screening statistic for large transaction files, and Cerioli et al. (2019) calibrate by simulation, building the null distribution from idealised non-cheating traders with the same number of transactions and products. These procedures treat the units tested as exchangeable draws, either from the Benford distribution or from a pool of market data. To our knowledge, none addresses a panel in which a few amounts per unit are linked by an accounting identity and persist from year to year, and none calibrates regression coefficients of a digit statistic on regressors built from the same figures.

The present procedure adds three things to that literature. First, the reference distribution is conditional on each bank's own observed amounts: it randomises only the overall scale, which leaves each bank's ratios, balance relations and growth path unchanged, and it rests on the scale invariance of the significant-digit law (Hill 1995; Pinkham 1961). Second, the same device is applied to the coefficients of the fixed-effects models, so that an association that arithmetic alone could produce between a ratio such as leverage and the digit statistic is separated from a behavioural one. Third, the bank-level benchmark is computed exactly, by enumerating at most 240 breakpoints on the unit interval, and its false-positive rate and power are reported (Section 3.9).

The study is relevant to researchers who combine digit tests with accounting ratios, to auditors and supervisors who need to know what a first-digit screen on a few aggregate amounts can discriminate, and to scholars of the fraud hexagon, for whom it shows how far hexagon proxies can be examined in annual-report data from an emerging banking system.

Three delimitations bound the study. The twelve banks were selected purposively on the availability of complete annual reports, and only five amounts per bank-year are tested. Four of the six hexagon components are examined through proxies, with opportunity only as a control and rationalisation not examinable. The analysis conditions on, and does not test, the assumptions that hexagon conditions make manipulation more likely and that manipulation disturbs leading digits. A deviation from the Benford pattern is a signal for further examination and not evidence of fraud. The paper is organised as follows. Section 2 reviews the literature and develops the hypotheses, Section 3 describes the data, index, models, benchmark and decision rules, Section 4 reports the results, Section 5 discusses them and their limitations, and Section 6 concludes.

2. Theoretical review and hypothesis development

This section reviews Benford's Law (Section 2.1) and the fraud hexagon in Indonesian banking (Section 2.2) and then develops the hypotheses that connect them (Section 2.3).

2.1 Benford's Law and numerical anomalies in financial reporting

Benford's Law states that, in many naturally occurring collections of numbers, the leading digit d appears with probability $\log_{10}(1 + 1/d)$, so that the digit 1 occurs far more often than the digit 9 (Benford 1938). Digit tests compare the observed distribution of leading digits with this pattern; the primary tests and the MAD-based conformity criteria are set out by Nigrini (2012a, b). In accounting, a discrepancy can serve as an initial signal that certain data warrant scrutiny (Durtschi et al. 2004), and the logic has been extended to financial risk (Cerqueti et al. 2024) and banking (Grammatikos and Papanikolaou 2021).

A deviation is nevertheless not, by itself, evidence of manipulation (Aybars and Ataunal 2016). Conformity depends on the nature of the data, which should span several orders of magnitude (Durtschi et al. 2004), and on accounting conventions such as conditional conservatism (Nguyen

and Duong 2022). It also depends on the number of figures tested: with five amounts the MAD cannot fall below 0.0669 (Section 3.4), so departure from the conventional bands is expected even for unremarkable data (Barney and Schulzke 2016).

Data-mining and machine-learning methods (Aftabi et al. 2023; Huang et al. 2022; Liu 2023; Papík and Papíková 2022; Zhou et al. 2024) may offer greater predictive power than digit tests, but an observation they classify as unusual need not reflect manipulation when entities are observed repeatedly and carry persistent characteristics (Del Gaudio et al. 2022). Digit-based screening and behavioural theory therefore remain complementary: the screen indicates where to look, and the fraud hexagon offers a candidate explanation of why manipulation might occur (Vousinas 2019).

2.2 The fraud hexagon and fraud risk in Indonesian banking

The fraud hexagon extends the fraud triangle of pressure, opportunity and rationalisation (Skousen et al. 2009) with capability from the fraud diamond (Wolfe and Hermanson 2004), arrogance from the fraud pentagon (Situngkir and Triyanto 2020) and collusion (Vousinas 2019). It describes conditions that create incentives, opportunities and capacity for fraud, not numerical patterns.

Indonesian applications cover state-owned enterprises (Achmad et al. 2022; Sudrajat et al. 2023), government institutions (Sukmadilaga et al. 2022), listed firms (Sagala and Siagian 2021; Wilantari and Ariyanto 2023) and banking (Achmad et al. 2023; Najmuddin and Widihardimas 2026; Sare and Kurniawati 2025). Results in banking are mixed: Achmad et al. (2023) report that only external pressure and arrogance are related to fraudulent financial reporting, and Najmuddin and Widihardimas (2026) describe the evidence as inconclusive. None of these studies uses first-digit deviations from reported amounts as the outcome.

2.3 Hypothesis development

The link between the two approaches is a proposition about conditions within a bank. If the hexagon dimensions describe conditions under which manipulation becomes more likely, and if manipulation is accompanied by adjustments that disturb the leading-digit pattern (Durtschi et al. 2004; Nigrini 2012a), then a bank in years of stronger fraud-risk conditions should, other things equal, show a larger departure from the Benford pattern. This within-bank implication is the one the data can examine most credibly, because it compares each bank with itself and removes stable differences in ownership, size and business model; associations across banks are confounded with those features and are only described. The proposition rests on the premise that manipulation disturbs leading digits in a way the index can register, and all tests are conditional on it (Section 5.1).

Pressure arises when management faces incentives, often linked to financial stress or high leverage, to present more favourable results (Skousen et al. 2009). It is proxied by total liabilities divided by total assets, which reflects the construct only partially and, being built from two tested amounts, is evaluated against the benchmark of Section 3.6. H1 is the primary, directional test.

H1: Higher leverage (PRESS), used as a proxy for financial pressure, is associated with higher BENFORD_MAD within banks.

Capability concerns the traits and positions needed to commit and conceal fraud, and board changes may bring in or empower individuals able to override controls (Umar et al. 2020; Wolfe and Hermanson 2004). It is proxied by a documented change in the Board of Directors, as in Achmad et al. (2023) and Sagala and Siagian (2021). This and the two proxies below change within few banks, so they are posed as exploratory questions.

E1 (exploratory): To what extent are years in which a board change is documented (CAP), a proxy for capability, associated with BENFORD_MAD within banks, in either direction?

Arrogance reflects an attitude of superiority that may lead management to disregard controls (Situngkir and Triyanto 2020). Because attitude cannot be observed, Achmad et al. (2023) use the frequency of the chief executive's photographs in the annual report; here it is approximated by a disclosed political or public-office background among directors, a status indicator whose link to arrogance has not been validated. Because Sagala and Siagian (2021) use political connections as a proxy for collusion, E2 and E3 overlap conceptually.

E2 (exploratory): To what extent are years in which directors disclose a political or public-office background (ARO), a proxy for arrogance, associated with BENFORD_MAD within banks, in either direction?

Collusion refers to cooperation that can circumvent controls designed to rely on independent checks (Achmad et al. 2022; Khamainy et al. 2022). It is proxied by disclosure of government-project or government-cooperation activity, which indicates exposure to relationships in which checks may be weaker but does not measure collusion. Government-related activity is part of the business model of the government-owned banks (all but one of the twelve), so the within-bank design identifies the proxy only from changes in disclosure, which may reflect altered reporting practice as much as altered exposure.

E3 (exploratory): To what extent are years in which government-project or cooperation activity is disclosed (COL), a proxy for collusion exposure, associated with BENFORD_MAD within banks, in either direction?

Opportunity enters only as a control, because its available proxy, operating income divided by total assets, measures operating return and not weakness in internal control. The rationalisation

proxy, a change of external audit firm, takes the same value in all 72 bank-years and cannot be identified (Section 3.3). Both omissions reflect the data and not a judgement about these dimensions.

3. Research method

3.1 Research design

The study is quantitative and observational, and all findings are interpreted as associations conditional on the data. RQ1 compares conformity with a benchmark for the same number of figures, at the pooled level (360 amounts) and the bank level (30 amounts per bank). RQ2 and RQ3 use the bank-year panel of 72 observations (twelve banks, six years), in which BENFORD_MAD serves as a relative index, and RQ4 re-estimates the panel models under alternative year effects, inference methods, measures and samples. Because conventional bands and significance levels ignore the arithmetic dependence among the tested amounts, inference for RQ1, RQ3 and RQ4 rests on the rotation benchmark of Section 3.6. The analysis plan, including the decision rules of Section 3.7, was written with the data in hand and was not preregistered; it is a transparent protocol and not a confirmatory test, and Rule R4 was revised once after a first run.

3.2 Population, sample and data

Twelve Indonesian banks were selected purposively on one criterion, that a complete set of annual reports for 2020 to 2025 could be obtained, giving a balanced panel of 72 bank-years (Table 1). Because the extracted amounts are not fully reconciled with the source statements (see below), the banks are reported under neutral labels (Banks A to L), and because a first-digit deviation or a position above a benchmark percentile is not evidence of misreporting and naming a listed bank in that connection could be read as an allegation. The key is deposited in a restricted file in the

OSF project and is available to verify the results. Ownership and listing categories are summarised in Table 1.

Table 1 Listing status and ownership of the twelve banks, by category, 2020–2025

Banks (n)	Listing status	Ownership category
6	2 listed, 4 unlisted	Regional development banks (BPD), regional-government-owned
5	5 listed	State-owned or state-controlled banks
1	1 listed	Private national bank

Note: Listing status refers to the Indonesia Stock Exchange; the four unlisted banks were unlisted in all years.

Five amounts were taken from each bank-year as reported in its annual report or annual financial statements: total assets (TA), total liabilities (TL), total equity (TE), operating income (OI) and net income (NI). This gives 30 amounts per bank and 360 in total, the only amounts to which the primary Benford tests are applied. The leading digit of an amount is the first non-zero digit of its absolute value, so the two negative operating income and three negative net income figures are evaluated by magnitude. All amounts are in Indonesian rupiah (IDR) millions: 48 bank-years are reported in millions, 18 in billions (three banks) and six in thousands (one bank), and the latter two groups were converted by powers of 1,000, which does not alter leading digits. Where a later report restated a prior-year figure, the restated amount was used. For Bank A, the consistently disclosed operating-revenue line serves as operating income, which makes that bank's operating-return control not strictly comparable with the others. The extracted figures were checked before analysis for internal consistency only: derived variables and leading digits were recomputed independently, and the balance identity was examined for every bank-year. They were not matched to the source statements by a second person, and 33 bank-years with a balance-identity gap above 1% remain unreconciled (Section 4.1).

Corrections to the extracted figures are recorded in a log deposited in the Open Science Framework (OSF). They concern the units and operating income (2020 to 2025) of Bank F, the operating and net income of Bank C for 2021 to 2023 (misaligned by one year) and the net income of Bank A

for 2020 to 2022. One 2020 liability figure of Bank J was normalised from IDR billions and has not been checked against the source report, although its leading digit is unaffected. Bank F, the only unit classified as a departure in Section 4.2, is among the banks whose figures were corrected. Two features of the source documents limit inference. First, four pairs of adjacent bank-years (two at Bank K and two at Bank G) draw on a shared source report, so those eight observations are not fully independent. Second, the extracted amounts did not always satisfy the balance identity. Defining the gap as the absolute value of TA minus the sum of TL and TE, divided by TA, the gap is below 0.1% in 39 of the 72 bank-years and above 1% in the other 33, of which 24 exceed 5%. A gap may reflect items presented outside liabilities and equity, such as the temporary syirkah funds reported by a bank with sharia operations, or an extraction error such as a misidentified component line or wrong year; we could not confirm either explanation against the source statements. At one bank the gap exceeds 60% of assets in every year, a systematic difference in presentation or a missing component line and not random extraction error (Section 4.1). Because the same extracted amounts feed both the digit index and the leverage ratio, an extraction error would alter the two together, in a direction we cannot determine. We therefore make no claim about the reporting behaviour of any bank, and the empirical results show only how the benchmark behaves on real panel data. To show how far the gaps matter, Section 4.5 re-estimates the main model on the 39 reconciled bank-years alone, and the Appendix lists the 33 bank-years with a gap above 1%. Every bank-year with a gap above 1% should be re-examined against its source statement and documented off-balance items treated consistently; this has not been done, so the estimates describe the panel as extracted and are provisional.

3.3 Variable measurement

The dependent variable, BENFORD_MAD (written MAD_{it} in the models), is the mean absolute deviation of Equation 1 computed for the five amounts of bank *i* in year *t*. PRESS, used for H1, is TL divided by TA; because items outside liabilities and equity can make it non-comparable across banks, one minus TE divided by TA, which does not use the liabilities line, is the alternative in the robustness tests. OPP, the control for opportunity, is OI divided by TA, with NI divided by TA as its alternative. Both ratios are unchanged when a bank's amounts are multiplied by a common factor, a property used in Section 3.6. A pressure proxy that uses none of the five tested amounts, such as the capital adequacy, non-performing loan or loan-to-deposit ratio, would share no inputs with the index and give the cleanest test of H1, but such ratios were not entered for any bank-year, so this test is not estimated.

The three binary variables are defined as follows. CAP equals one when a change in the Board of Directors is documented in the year, ARO equals one when board biographies disclose a political or public-office background, and COL equals one when the report discloses government-project or government-cooperation activity. They were coded by an automated procedure applied to annual-report text, carrying the nearest available year forward or backward where a report was scanned (Table S4 of Online Resource 1 reproduces the coding rules as documented and their status). The affected bank-years were not flagged individually, so all 72 observations are labelled preliminary. The keyword list behind ARO and the rule behind COL were not retained, so the codes cannot be regenerated from the reports; they are treated as given data and are replicable as deposited. CAP, ARO and COL equal one in 57, 45 and 39 of the 72 bank-years, with four, five and nine within-bank transitions (Section 4.1).

Because a binary regressor in a fixed-effects model is identified only from banks in which it changes, and carry-forward can create or remove those changes, the findings for E1 to E3 are provisional. A sixth variable, RAT, equals one when the external audit firm differs from the previous year; it equals zero in all 72 bank-years, which may reflect the coding procedure and not the banks' audit histories, so it is excluded and rationalisation is not tested.

3.4 Benford statistic and reference values

Benford's Law gives the probability that the first significant digit is d as $\log_{10}(1 + 1/d)$, for d = 1, …, 9. For any set of n amounts the mean absolute deviation is

$$MAD = \frac{1}{9}\sum_{d=1}^{9}\left|p(d) - \log_{10}\left(1 + \frac{1}{d}\right)\right| \qquad (1)$$

where p(d) is the observed proportion of amounts with first digit d. The conventional conformity bands are 0.000 to 0.006 (close conformity), 0.006 to 0.012 (acceptable), 0.012 to 0.015 (marginal) and above 0.015 (nonconformity) (Nigrini 2012a). The bands serve only as a descriptive reference. The conventional bands cannot be applied directly to a bank-year, because the MAD of a handful of figures is bounded away from zero. With five amounts, each first-digit proportion is a multiple of 0.2, so the MAD cannot fall below 0.0669, a value reached when two leading digits are 1 and one each is 2, 3 and 4. Under independent Benford draws, the MAD of five amounts has an exact distribution, obtained by enumerating all 1,287 possible digit-count patterns with their multinomial probabilities: its mean is 0.1098, its standard deviation 0.0221, its median 0.1069 and its 5th to 95th percentile range 0.0765 to 0.1491, so every draw lies above 0.015. These values serve for scale only, because the amounts of a bank are linked and persistent; decisions rest on the benchmark of Section 3.6. Chi-square statistics are not used for decisions, because they assume independent observations and, at the bank level, the smallest expected cell count is 1.37; they and

the second-digit and first-two-digit tests appear only in Table S1 of Online Resource 1. For RQ2 to RQ4 the bank-year MAD is a continuous index and not a classification.

3.5 Empirical models and inference

A pooled model serves as the reference for the panel specifications:

$$MAD_{it} = \gamma_0 + \beta_1 PRESS_{it} + \beta_2 CAP_{it} + \beta_3 ARO_{it} + \beta_4 COL_{it} + \theta OPP_{it} + \varepsilon_{it} \qquad (2)$$

The main specification for H1 and E1 to E3 (RQ3) replaces the intercept with bank fixed effects α_i, which absorb all stable differences between banks:

$$MAD_{it} = \alpha_i + \beta_1 PRESS_{it} + \beta_2 CAP_{it} + \beta_3 ARO_{it} + \beta_4 COL_{it} + \theta OPP_{it} + \varepsilon_{it} \qquad (3)$$

Year effects λ_t, which capture shocks common to all banks, are added for RQ2 and RQ4:

$$MAD_{it} = \alpha_i + \lambda_t + \beta_1 PRESS_{it} + \beta_2 CAP_{it} + \beta_3 ARO_{it} + \beta_4 COL_{it} + \theta OPP_{it} + \varepsilon_{it} \qquad (4)$$

To show how the between-bank association differs from the within-bank one, a correlated-random-effects specification is estimated (Wooldridge 2019):

$$MAD_{it} = \gamma_0 + \lambda_t + \sum_{k=1}^{5} \beta_k \, x_{k,it} + \sum_{k=1}^{5} \delta_k \, \bar{x}_{k,i} + \varepsilon_{it} \qquad (5)$$

where x_k are the five regressors and the bar terms are their bank means. The β_k reproduce the within-bank estimates of Equation 4, and δ_k measures how far the between-bank association differs from them. Under bank fixed effects, CAP, ARO and COL are identified only from four, five and nine transitions in three, four and five banks, respectively (Section 4.1), whereas PRESS varies in all twelve banks.

Models are estimated by least squares with bank and year indicators and HC3 heteroskedasticity-robust standard errors (Long and Ervin 2000). HC3 does not adjust for serial correlation within banks, and clustering by bank is unreliable with only twelve clusters (Cameron and Miller 2015). A wild cluster bootstrap-t with bank clusters, the null imposed, Webb six-point weights and 9,999

replications (Cameron et al. 2008) is therefore reported as secondary evidence, although it too can reject too often or too rarely when a regressor changes in few clusters (MacKinnon and Webb 2018). For each coefficient the MDE is approximately 2.8 times the standard deviation of its benchmark distribution, the factor corresponding to a two-sided test at the 5% level with 80% power. Because that spread reflects the variability generated by randomising scale and not the sampling error of the estimator under an alternative, an MDE based on the HC3 standard error is reported beside it (Section 4.4). The benchmark p-values of the four coefficients of interest are adjusted jointly by the Holm procedure (Bretz et al. 2016), with unadjusted values alongside.

3.6 The rotation benchmark

Two features of the data make a benchmark necessary. The five amounts of a bank-year are tied together by the balance identity and a common scale, so a ratio such as leverage is built from figures whose leading digits are tested and the two can be related by arithmetic alone. In addition, a bank's amounts change slowly, so the position of each amount within its decade, and hence its leading digit, is persistent. The median year-to-year change in the position of total assets within its decade was 0.041, against 0.25 for independent draws, so a null that redraws amounts each year would overstate the variability of the index. The benchmark preserves both features by randomising only the overall scale of each bank. For bank *i* a shift u_i is drawn uniformly from [0, 1), and every amount of that bank in every year is multiplied by 10 raised to u_i:

$$\tilde{x}_{it} = 10^{u_i} \times x_{it} \qquad (6)$$

Intuitively, rescaling every figure of a bank by the same factor changes which digit each figure begins with but not how the figures relate to one another, so the benchmark shows what the index would look like if only the arbitrary scale of the balance sheet determined the leading digits. The rescaling preserves the observed ratios among the five amounts, the balance relations and the

growth path, while the logarithmic mantissa of every amount becomes uniform, so each shifted amount follows the Benford distribution in expectation (Hill 1995). The benchmark models no particular form of manipulation.

The regression benchmark is computed as follows. In each of 5,000 replications, twelve shifts are drawn, the first digits of all 360 shifted amounts are recomputed, BENFORD_MAD is recomputed for every bank-year, Equation 3 is estimated with the observed regressors, and the five coefficients are stored. The regressors are not shifted, because PRESS and OPP are ratios that a common factor leaves unchanged and the binary variables do not depend on scale. The distribution of each coefficient shows what the arithmetic and growth structure could produce without a behavioural relation.

For RQ1, the bank-level benchmark multiplies all 30 amounts of a bank by one random factor and is computed exactly, and the pooled benchmark gives each bank its own factor. The first digit of an amount x under shift u changes only when the fractional part of $\log_{10}|x| + u$ crosses $\log_{10} d$ for $d = 2, \ldots, 9$, so the MAD is constant between at most 240 breakpoints on [0, 1). The probability of each attainable MAD value is the total length of the intervals on which it occurs. The 95th percentile is the smallest value whose cumulative probability reaches 0.95, the percentile of the observed MAD is the benchmark probability of a value at or below it, and the benchmark probability P(≥ MAD) is the probability of a value at or above it. The pooled distribution depends on twelve shifts and is simulated with 1,000,000 draws.

3.7 Decision rules

Each decision rule links one research question to an explicit verdict, so that the reading of a result does not depend on judgement made after the numbers are seen.

Rule R1 (RQ1). A pooled or bank-level MAD is a detectable departure from the Benford distribution when it lies above the 95th percentile of the same statistic under the rotation benchmark. The conventional bands are shown for reference and do not enter the decision.

Rule R2 (RQ2). Persistent differences between banks, or common year effects, are present when the F-test comparing Equation 2 with Equation 3 (bank effects) or Equation 3 with Equation 4 (year effects) has a p-value below 0.05. Because a bank's amounts change slowly, a significant bank effect is expected and is not by itself evidence of differences in reporting behaviour, so the raw share of the total sum of squares lying between banks, the intraclass correlation ((F − 1)/(F + 5) from the one-way analysis-of-variance F-ratio, with six years per bank) and the within-bank standard deviation of the index, compared with its benchmark distribution, are also reported.

Rule R3 (H1). H1 is supported when the Equation 3 coefficient lies above the median of its benchmark distribution and its Holm-adjusted benchmark p-value is below 0.05, and it is contrary to the prediction when the coefficient lies below that median with an adjusted p-value below 0.05. Otherwise it is labelled "not supported", and additionally "inconclusive" when its 95% interval, the observed coefficient minus the benchmark median plus or minus 1.96 times the benchmark standard deviation, contains zero; "inconclusive" means that the sample could not distinguish the hypothesised association from none beyond arithmetic and chance, and is not evidence of absence. For E1 to E3 no verdict is assigned; each is reported with its coefficient, benchmark p-value, interval and MDE, and its dependence on a single bank is judged from the leave-one-bank-out results.

Rule R4 (RQ4). Stability is assessed for all four coefficients. An estimate is stable to year effects when the Equation 4 coefficient keeps its sign and lies within the 95% HC3 interval of the Equation 3 coefficient. The sign is also examined in each leave-one-bank-out sample, in each restricted

sample (excluding bank-years with gaps above 5% and above 1%, and dropping the earlier or the later member of each shared-source pair) and under the alternative measures of PRESS and OPP; a coefficient with the same sign in every specification is stable in sign, and otherwise sign-sensitive. A coefficient supported under Rule R3 is additionally robust if its benchmark p-value is below 0.10 in every specification. This restriction was adopted after a first run, in which the original requirement, applied to unsupported coefficients, was failed mechanically and conveyed nothing about stability.

Rule R5. For each coefficient in Equation 3 the benchmark p-value is twice the smaller of the share of simulated coefficients at or below the observed value and the share at or above it, capped at one. A coefficient is corroborated when its sign agrees with the sign of its difference from the benchmark median and both the HC3 and the wild cluster bootstrap p-values are below 0.10; otherwise, it is sensitive to the method of inference.

3.8 Robustness and diagnostics

Stability (RQ4) is assessed under the specifications listed in Rule R4 and by estimating the correlated-random-effects model (Equation 5), with the benchmark recomputed for each leave-one-bank-out sample. Diagnostics comprise variance inflation factors and correlations on raw and bank-demeaned regressors, a first-difference serial-correlation test (Drukker 2003), Cook's distance and DFBETAS, and Shapiro–Wilk and Breusch–Pagan tests of the residuals.

3.9 Performance of the rotation benchmark

A calibrated reference is useful only if it rarely flags units without a departure and does flag units with one, so a simulation examines the false-positive rate and the power of Rule R1 at the bank level (30 amounts) and the pooled level (360 amounts). The false-positive rate is estimated for three processes without departure: the observed amounts of a randomly chosen bank multiplied by

a common random factor (A1, the rotation null; at the pooled level each of the twelve banks receives its own factor), independent Benford leading digits (A2), and a synthetic panel with fixed ratios among five amounts, a growth path and a random scale, which reproduces the persistence of Section 3.6 (A3). In each replication Rule R1 is applied as in Section 4.2, and the proportion of flagged units is compared with the nominal 0.05.

Power is examined by replacing the leading digit of a share π of the amounts in the rescaled data, with π equal to 0.10, 0.20 and 0.30, under two mechanisms: shifting each selected digit to the next higher digit (M1; the digit 9 is unchanged), or replacing it with a digit drawn uniformly from 1 to 9 (M2). The benchmark is recomputed from the manipulated amounts, as an analyst would do. Each cell uses 5,000 bank-level and 500 pooled replications, each pooled replication drawing its benchmark 1,000 times. The mechanisms are stylised and do not represent actual misreporting.

3.10 Replication and ethics

The analysis used Python 3.13.16, NumPy 2.1.3 (Harris et al. 2020), pandas 2.2.3 (McKinney 2010), SciPy 1.16.3 (Virtanen et al. 2020) and statsmodels 0.15.0 (Seabold and Perktold 2010), with seed 12345 throughout. Replication requires the 360 amounts as corrected (Section 3.2), the binary codes (Section 3.3) and the computations of Sections 3.4 to 3.9 in order. The exact benchmarks, digit-based statistics and least-squares estimates reproduce exactly, whereas quantities based on random draws (5,000 regression draws, 1,000,000 pooled draws, 9,999 bootstrap replications and the simulations) replicate within Monte Carlo error.

The study uses only information that Indonesian banks publish in their annual reports under the disclosure requirements of the Financial Services Authority of Indonesia (Otoritas Jasa Keuangan, OJK) and the Indonesia Stock Exchange (IDX), and the authors collected no information from individuals. The ARO variable draws on publicly disclosed biographies of identifiable directors

but is aggregated to a bank-year indicator; no director is named or described individually, and the variable is not used to evaluate any individual. Claude 3.5 Sonnet (Anthropic) was used to debug the analysis code and check Python syntax. In the final revision, Claude was used to draft and edit passages of the abstract, Sections 1.3, 3.2, 4.4, 4.5, 5.2 and 6 and the Appendix, and to cross-check reported values against the analysis notebook. The authors reviewed, tested and edited all output and agree that the text reflects their own work (see also Statements and Declarations).

4. Results

Results are reported in the order of RQ1 to RQ4, followed by the performance check of the benchmark, after Section 4.1 states the status of the data on which every result is conditional. Conformity and association are judged against the rotation benchmark, with conventional bands and significance levels secondary. All values come from one complete run of the analysis on the panel of Section 3.2.

4.1 Data status and sample characteristics

The analysis uses the balanced panel of twelve banks, 72 bank-years and 360 non-zero amounts, with all five regressors available for every observation. Every derived variable and all 360 leading digits matched an independent recomputation from the extracted amounts, which checks the calculation and not the extraction against the source reports. The balance identity qualifies the data (Section 3.2). At Bank G the gaps lie between 63.4% and 67.2% of assets in all six years, a size and stability that point to a systematic difference in how items are presented and not to random extraction error. Smaller gaps, from 1.7% to 12.8% of assets, occur at seven other banks. None of the 33 bank-years has been matched to its source statement, and the results are therefore treated as provisional.

CAP changes in three banks (Banks A, B and H) through four transitions, ARO in four banks (Banks A, B, E and I) through five transitions, and COL in five banks (Banks C, D, E, H and J) through nine transitions, whereas PRESS and OPP vary within all twelve banks. RAT is zero in all 72 bank-years, so rationalisation is not tested. The non-amount pressure proxies and the line-item statements for the extended index were not entered, so the cleanest test of H1 and the extended index could not be run.

4.2 First-digit conformity of reported amounts (RQ1)

Fig. 1 compares the pooled distribution of the 360 leading digits with the Benford expectation. The digit 1 leads 134 amounts (37.2%) against an expected 30.1%, and the digit 4 only 19 (5.3%, against 9.7%). The pooled MAD is 0.0255, in the nonconformity range of the conventional bands, but against the rotation benchmark it lies at the 67th percentile of a distribution with a median of 0.02145 and a 95th percentile of 0.03836, so Rule R1 classifies it as within the benchmark range. Table 2 reports the same comparison for the pooled amounts and for each bank.

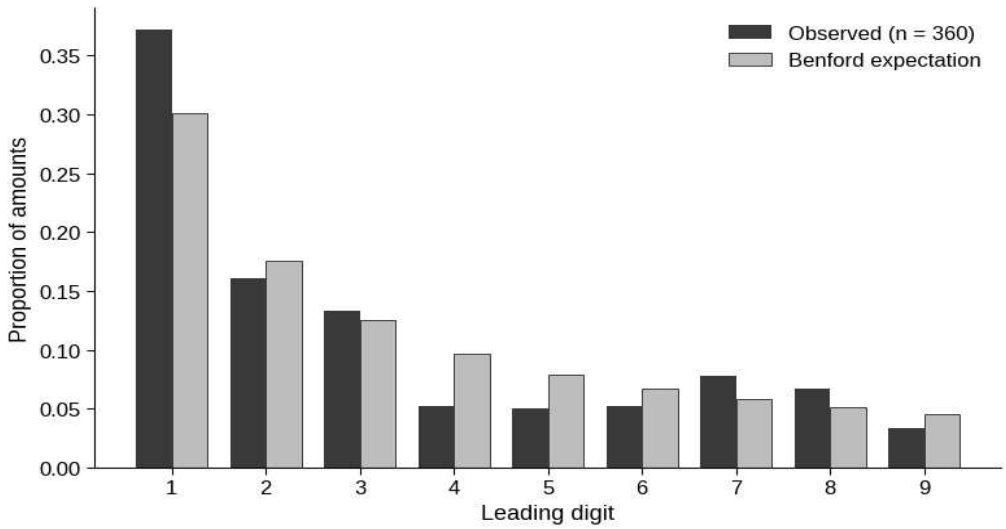

Fig. 1 Pooled first-digit distribution of the 360 reported amounts compared with the Benford expectation. Observed proportions are based on five amounts (total assets, total liabilities, total equity, operating income and net income) for each of 72 bank-years

Table 2 First-digit mean absolute deviation (MAD) against the rotation benchmark, pooled and by bank

Unit	n	MAD	Bench. Median	Bench. 95th pct.	Percentile of MAD	P (bench. ≥ MAD)	Rule R1
Pooled	360	0.02553	0.02145	0.03836	67.4	0.326	Within range
Bank A	30	0.10879	0.10138	0.11275	89.4	0.207	Within range
Bank B	30	0.04795	0.04884	0.06824	45.0	0.554	Within range
Bank C	30	0.11620	0.12570	0.14051	27.8	0.785	Within range
Bank D	30	0.07984	0.08102	0.09397	41.8	0.609	Within range
Bank E	30	0.12107	0.11088	0.12174	93.4	0.095	Within range
Bank F	30	0.10144	0.08935	0.10138	98.6	0.024	Departure
Bank G	30	0.05355	0.05535	0.06773	44.8	0.574	Within range
Bank H	30	0.03940	0.04009	0.05880	49.5	0.508	Within range
Bank I	30	0.04585	0.05349	0.07057	16.8	0.834	Within range
Bank J	30	0.04750	0.05156	0.06359	29.6	0.715	Within range
Bank K	30	0.10347	0.12570	0.13655	12.1	0.928	Within range
Bank L	30	0.06690	0.07916	0.09239	14.7	0.900	Within range

Note: n is the number of amounts. Banks are identified by neutral labels because the extracted amounts are not fully reconciled (Section 3.2). Bank-level benchmark values are exact and the pooled benchmark uses 1,000,000 draws; five decimals are shown because bank-level values can lie as little as 0.00002 apart.

At the bank level the MAD ranges from 0.0394 at Bank H to 0.1211 at Bank E, and every value lies in the nonconformity range. Bank F's MAD of 0.10144 lies above the 95th percentile of its benchmark (0.10138), and the benchmark probability of a MAD at least this large is 0.0244, so Rule R1 classifies it as a detectable departure. Because the figures of this bank were among those corrected during extraction (Section 3.2), the classification is a prompt to verify its source figures and not a finding about its reporting. Bank F is the only one of 13 units above its benchmark's 95th percentile, against 0.65 expected by chance; because the 13 units are not independent and no adjustment for multiple comparisons has been made, this figure is approximate. In response to RQ1, the first-digit distributions of the pooled amounts and of every bank depart from the Benford expectation by conventional criteria, but against the rotation benchmark the pooled amounts and eleven of the twelve banks lie within the range the benchmark produces.

4.3 Variation of the bank-year index (RQ2)

The index averages 0.1146 across the 72 bank-years, against 0.1098 expected for five independent Benford draws, and its standard deviation of 0.0260 is 1.18 times the 0.0221 expected under independent draws. The variation lies between banks: bank means range from 0.0865 (Bank B)

and 0.0887 (Bank J) to 0.1456 (Bank K) and 0.1465 (Bank E), whereas year means move only between 0.1056 (2022) and 0.1199 (2021).

The F-test for bank effects (Equation 2 against Equation 3) gives $F(11, 55) = 4.78$ with $p < 0.001$, whereas the test for year effects (Equation 3 against Equation 4) gives $F(5, 50) = 0.97$ with $p = 0.446$. The raw share of the total sum of squares lying between banks is 57.3% (0.0275 of 0.0481), and the intraclass correlation is 0.51. The within-bank standard deviation of the index, 0.0185, lies at the 62nd percentile of its benchmark distribution (median 0.0179; 5th to 95th percentile 0.0150 to 0.0209), close to what the benchmark produces. Bank C has the same index (0.1162) in all six years because its five amounts always begin with 1 or 2, with at least two 1s and one 2. In response to RQ2, the index varies mainly between banks, common year effects are not detected, and within-bank variation is close to what a change of scale alone produces.

4.4 Hexagon proxies and the index within banks (RQ3)

Table 3 reports the bank fixed-effects estimates (Equation 3), their benchmark and the secondary inference of Rule R5, in units of the index, whose within-bank standard deviation is 0.0185.

Table 3 Within-bank associations between the hexagon proxies and the index (Equation 3; 72 bank-years; 55 residual degrees of freedom)

Variable	Coeff.	Bench. median	Bench. p	Holm p	MDE	MDE (HC3)	HC3 p	Bootstrap p
PRESS (H1)	0.1023	0.0173	0.313	0.940	0.238	0.316	0.368	0.113
CAP (E1)	−0.0130	−0.0083	0.612	0.940	0.026	0.027	0.177	0.091
ARO (E2)	−0.0133	−0.0026	0.201	0.805	0.024	0.023	0.114	0.084
COL (E3)	−0.0078	−0.0017	0.415	0.940	0.021	0.028	0.438	0.751
OPP (control)	0.0761	−0.3792	0.139	–	0.839	1.882	0.910	–

Note: Bench. p is twice the smaller tail share of 5,000 rotation draws. Holm p is adjusted across the four coefficients of interest for description; only H1 receives a verdict. MDE is approximately 2.8 times the standard deviation of the benchmark distribution (bench.) or the HC3 standard error. Bootstrap p is the wild cluster bootstrap-t p-value (12 clusters, 9,999 replications, Webb weights). Fit: overall $R^2 = 0.614$ (adjusted 0.501); within $R^2 = 0.096$ (adjusted 0.014); residual standard error = 0.0184; whole-model $F(16, 55) = 5.46$, $p < 0.0001$; joint $F(5, 55) = 1.17$, $p = 0.336$.

H1 (primary). The coefficient on PRESS is positive, as hypothesised (0.102), and its benchmark median is 0.017, so about 17% of it is reproduced when the scale of each bank's amounts is randomised. The difference from the median, 0.085, is about one benchmark standard deviation, and the observed coefficient lies at the 84th percentile of its benchmark distribution (Fig. 2), giving

a benchmark p-value of 0.313 and a Holm-adjusted value of 0.940. The 95% interval for the difference, −0.081 to 0.252, contains zero, so under Rule R3 H1 is not supported and is inconclusive. The MDE, 0.238, is 2.3 times the coefficient. HC3 (p = 0.368) and the bootstrap (p = 0.113) do not alter this reading, and under Rule R5 the result is sensitive to the method of inference.

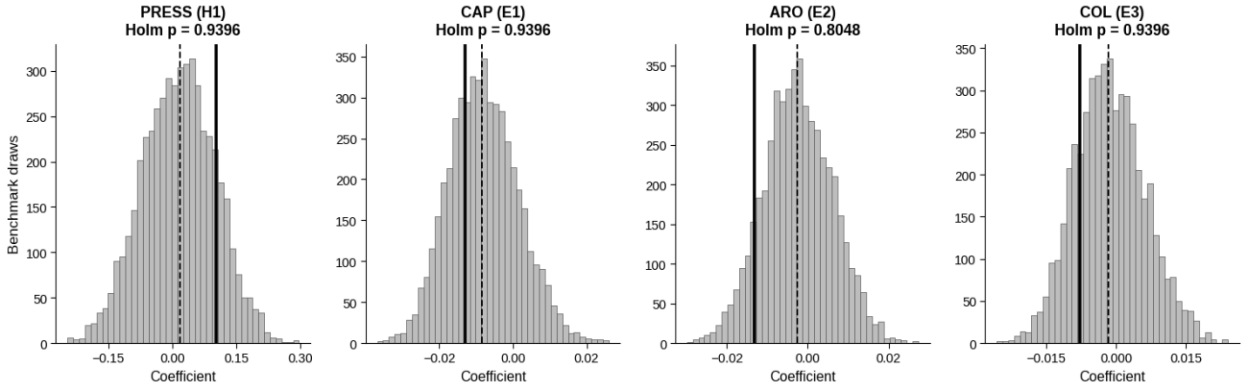

Fig. 2 Observed Equation 3 coefficients compared with their rotation-benchmark distributions for (a) PRESS (H1), (b) CAP (E1), (c) ARO (E2) and (d) COL (E3). Histograms show the coefficient obtained when only a bank-level shift of scale generates the index (5,000 draws); the solid line is the observed coefficient and the dashed line the benchmark median. The observed coefficients lie at the 84th (PRESS), 31st (CAP), 10th (ARO) and 21st (COL) percentiles

E1, E2 and E3 (exploratory). The coefficients are all negative, whereas the hexagon rationale of Section 2.3 would suggest positive signs. Their benchmark p-values are 0.612, 0.201 and 0.415 (Holm-adjusted 0.940, 0.805 and 0.940), and they receive no verdict. The 95% intervals for the difference from the benchmark median (−0.023 to 0.014 for CAP, −0.027 to 0.006 for ARO and −0.021 to 0.009 for COL) all contain zero, and the MDEs (0.026, 0.024 and 0.021) are 2.0, 1.8 and 2.7 times the absolute coefficients. On the HC3 standard error the MDEs are 1.7 to 3.6 times the absolute coefficients of all four proxies, so the conclusion that the sample could not detect an association of the observed size does not depend on the standard error used. The CAP and ARO coefficients are below 0.10 under the bootstrap (p = 0.091 and 0.084) but not under HC3 (p = 0.177 and 0.114), so no coefficient is corroborated. The sign of the COL coefficient depends on Bank E,

whose exclusion turns it positive, whereas the CAP and ARO coefficients keep their sign in every leave-one-bank-out sample.

Bank effects account for most of the fit, whereas the regressors explain little within banks. Residual diagnostics raise no concern: the Shapiro–Wilk and Breusch–Pagan tests do not reject normality or homoskedasticity (p = 0.412 and 0.353), the largest variance inflation factor is 1.43, the largest absolute correlation 0.37, and the first-difference test does not detect serial correlation (p = 0.212). In response to RQ3, the test cannot discriminate: the leverage coefficient is not distinguishable from what arithmetic and chance produce, so Rule R3 returns not supported and inconclusive for H1, and the coefficients for E1 to E3 are likewise not distinguishable from what arithmetic and chance produce.

4.5 Stability of the estimates (RQ4)

The non-amount-proxy and extended-index checks were not implemented (Section 4.1). Because H1 is not supported and E1 to E3 receive no verdict, Rule R4 characterises each coefficient only as stable in sign or sign-sensitive; stability to the method of inference is read from Table 3. Table 4 and Fig. 3 summarise the other specifications.

Table 4 Stability of the coefficients across specifications

Specification	PRESS (H1)	CAP (E1)	ARO (E2)	COL (E3)
Equation 3	0.1023	−0.0130	−0.0133	−0.0078
Equation 4	0.1299	−0.0121	−0.0112	−0.0063
Alternative measure (benchmark p)	−0.0579 (0.758)	−0.0125 (0.564)	−0.0107 (0.283)	−0.0068 (0.706)
Leave-one-bank-out range	0.0634 to 0.1465	−0.0178 to −0.0101	−0.0181 to −0.0065	−0.0142 to 0.0045
Gaps ≤ 5% (n = 48)	0.1611	−0.0109	−0.0002	0.0042
Gaps ≤ 1% (n = 39)	0.0850	−0.0189	−0.0081	0.0086
Later year of each pair dropped (n = 68)	0.1372	−0.0135	−0.0141	−0.0078
Earlier year of each pair dropped (n = 68)	0.0458	−0.0121	−0.0119	−0.0077
Stability	Sign-sensitive	Stable in sign	Stable in sign	Sign-sensitive

Note: Equation 4 adds year effects. In the alternative-measure row, PRESS is replaced by one minus equity over assets and OPP by net income over assets. Leave-one-bank-out results are the minimum and maximum over 12 reduced samples. No bank-year has a gap between 0.1% and 1%, so the sample with gaps ≤ 1% is the sample of bank-years that reconcile within 0.1% of assets.

Adding year effects does not change the conclusions: in all four cases the Equation 4 coefficient keeps its sign and lies within the 95% HC3 interval of the Equation 3 coefficient (for PRESS,

−0.124 to 0.328). The between-bank component of the correlated-random-effects specification differs in sign from the within-bank coefficient for PRESS (−0.035), CAP (0.018) and ARO (0.008) and agrees with it for COL (−0.029). The PRESS coefficient changes sign when leverage is replaced by one minus equity over total assets (−0.058 against 0.102) but is positive in every other specification, and the COL coefficient turns positive when Bank E is excluded and in both gap-restricted samples, so both are sign-sensitive. The CAP and ARO coefficients keep their negative sign in every specification, although the ARO coefficient shrinks from −0.013 to −0.0002 and −0.008 when bank-years with larger gaps are excluded. In response to RQ4, CAP and ARO are stable in sign, PRESS and COL are sign-sensitive, and because no coefficient is supported, none is described as robust.

Because no bank-year has a gap between 0.1% and 1%, the sample with gaps ≤ 1% in Table 4 is also the sample of 39 bank-years that reconcile within 0.1% of assets. They come from eight banks, because Banks C, E, G and I have no reconciled year. Re-estimating Equation 3 on this sample, with the benchmark recomputed on it, gives coefficients of 0.085 for PRESS, −0.019 for CAP, −0.008 for ARO and 0.009 for COL, with benchmark p-values of 0.97, 0.57, 0.78 and 0.99 (Holm-adjusted 1.00 for all four). No coefficient differs from its benchmark, and the MDEs rise from 0.238 to 1.009 for PRESS, from 0.026 to 0.068 for CAP, from 0.024 to 0.087 for ARO and from 0.021 to 0.024 for COL, so the reconciled sample is less informative than the full panel. The conclusions are therefore unchanged in substance, and only the sign of the COL coefficient differs. The Appendix (Table A1) lists the bank-years excluded.

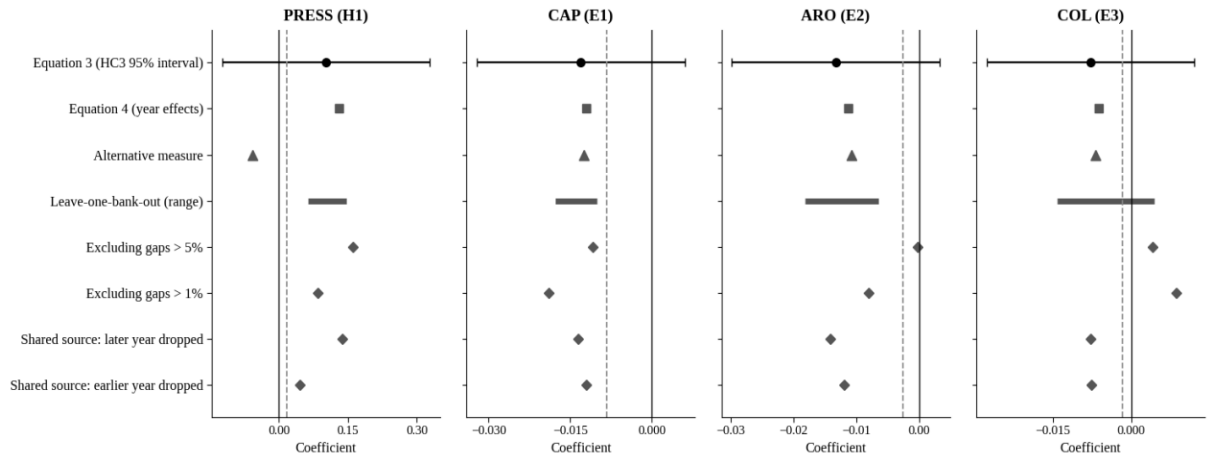

Fig. 3 Coefficient estimates across specifications for (a) PRESS (H1), (b) CAP (E1), (c) ARO (E2) and (d) COL (E3). The dot and horizontal line show the Equation 3 estimate and its 95% HC3 interval, the bar the leave-one-bank-out range, the square Equation 4, the triangle the alternative measure and the diamonds the restricted samples. The dashed vertical line is the median of the Equation 3 benchmark and the solid line marks zero

4.6 Performance of the rotation benchmark

Online Resource 1 reports the false-positive rate of Rule R1 (Table S2) and its detection rate against controlled digit departures (Table S3). Without departures, Rule R1 flagged between 3.7% and 3.9% of bank-level replications and between 5.6% and 7.2% of pooled replications, against the nominal 0.05; the pooled rates lie at most about 2.3 Monte Carlo standard errors above it (the standard error is about 0.01 with 500 replications), so a modest excess cannot be excluded. Against the controlled departures, detection ranged from 0.040 to 0.068 at the bank level and from 0.054 to 0.260 pooled, and these rates should be read against the false-positive rate of the same data process (0.037 and 0.056). At the bank level the rate stayed close to that benchmark in every scenario except the replacement of 30% of leading digits by uniform digits (0.068), so Rule R1 had little ability to detect altered amounts in a single bank. At the pooled level it rose with the share of altered amounts, to 0.194 under the next-higher-digit mechanism and 0.260 under uniform replacement at a share of 0.30, and stayed near the false-positive rate at a share of 0.10.

5. Discussion

5.1 What the Benford screen does and does not reveal

The divergence between the conventional and the benchmarked reading of conformity is the main methodological result. Fixed bands classify every unit as nonconforming, whereas the benchmark places the pooled amounts and eleven of the twelve banks within the range their own structure produces. Benchmark medians span 0.040 to 0.126, so raw MAD values cannot be compared across banks as measures of conformity (Barney and Schulzke 2016; Durtschi et al. 2004; Nguyen and Duong 2022).

The one flagged unit among 13 is close to the 0.65 expected by chance, and its figures were among those corrected during extraction (Section 3.2), so Bank F is a prompt to check the source figures and not evidence about its reporting. Equally, the absence of a flag shows nothing about manipulation, because bank-level detection was at most 0.07 even when 30% of leading digits were altered.

Most of the variance of the index lies between banks, common year effects are not detected, and within-bank variation resembles what a change of scale produces. The index is therefore best treated as a relative measure of numerical structure that may reflect the size and composition of a balance sheet more than reporting behaviour, although no size measure is available to test this. Because no observed manipulation is linked to any bank-year, any result for or against H1 or an exploratory question bears on fraud only to the extent that the premise of Section 2.3, that manipulation disturbs leading digits, holds.

5.2 Arithmetic dependence and the pressure hypothesis

The result for H1 bears on the difficulty identified in Section 1, that a ratio computed from the tested amounts can be associated with the digit statistic by construction. The benchmark

reproduces about 17% of the leverage coefficient, and the test cannot discriminate between an arithmetic and a behavioural association, mainly because the benchmark distribution is as wide as the coefficient is large. The dependence is clearer still for the opportunity control, whose benchmark median (−0.379) is about five times its observed coefficient (0.076) and has the opposite sign.

The power of the design limits every result. The MDEs are 1.7 to 3.6 times the estimates (Section 4.4), so a true association of the size observed could not have been distinguished from none, and every null result is uninformative. The leverage coefficient also changes sign when leverage is replaced by one minus equity over total assets, and because an extraction error would affect the digit statistic and the leverage ratio together, its direction should not be interpreted until a pressure proxy that shares no inputs with the index is available.

5.3 The exploratory proxies and the fraud hexagon

The three exploratory proxies have negative coefficients, whereas the hexagon rationale would suggest positive ones, but none is distinguishable from zero, and their MDEs (0.021 to 0.026) are at least as large as the within-bank standard deviation of the index (0.0185). The evidence is also fragile: each proxy changes in few banks, the collusion coefficient depends on Bank E, and the arrogance coefficient shrinks almost to zero when bank-years with gaps above 5% are excluded.

Only four of the six components could be examined, and the index is not a fraud outcome, so the absence of a detectable association does not weaken the hexagon framework. The null results are compatible with the mixed evidence for Indonesian banks (Achmad et al. 2023; Najmuddin and Widihardimas 2026) and with a systematic review of 622 articles that finds no consensus on standardised measures of financial statement fraud (Farooq et al. 2026). The proposition that banks

show larger departures in years of stronger fraud-risk conditions therefore remains an untested conjecture.

5.4 Methodological contribution and implications

The contribution is a screening procedure whose parts can be separated from the application: the rotation benchmark replaces fixed bands with a distribution reflecting the number of figures tested and their dependence, rules R1 to R5 attach a verdict to each result, and the MDE shows what the sample could not detect. Because detection was at most 0.07 for a single bank and 0.26 for the pooled panel, while the false-positive rate stayed near nominal at the bank level and at or modestly above it at the pooled level, the benchmark is a calibrated reference for conformity and has not been shown to detect bank-level manipulation.

For digit-based research, the study suggests judging conformity against such a benchmark, benchmarking or replacing regressors built from tested amounts, and reporting the MDE. For auditors and supervisors, a first-digit test on a handful of aggregate amounts per bank-year has little capacity to discriminate; it is useful only if applied to more line items, read against a bank's own history, and used to direct examination and not to classify banks.

5.5 Limitations and directions for future research

The conclusions are conditional on the data and the design. The amounts have not been reconciled with the source statements, and an extraction error would alter the index and the leverage ratio together, in a direction that cannot be determined. Restricting the panel to the 39 bank-years that reconcile did not change the conclusions but lowered the power further (Section 4.5). The automated coding of the binary proxies is unassessed and cannot be repeated from the reports, and carry-forward can create or remove the transitions that identify the coefficients. The estimates are, moreover, associations and not causal effects, because bank fixed effects remove stable

characteristics but not time-varying factors such as size, product mix, accounting policy or mergers. The panel is small (twelve purposively selected banks, six years, 55 residual degrees of freedom), spans the pandemic and regulatory change, and rests on five amounts per bank-year, and with no external outcome such as an enforcement action or restatement the study offers no validated detection model.

Future work should reconcile the amounts, re-code the binary proxies under a retained, documented rule with a hand-checked validation sample, repeat the pressure test with a capital adequacy, non-performing loan or loan-to-deposit ratio, build the extended index, enlarge the panel and add an external outcome such as regulatory sanctions (Del Gaudio et al. 2022). Pre-specifying the analysis would address the lack of preregistration (Brodeur et al. 2024).

6. Conclusion

This study developed a panel-based rotation benchmark for judging first-digit conformity when each observation rests on a few reported amounts, and applied it to twelve Indonesian banks with proxies for four fraud hexagon components. The benchmark preserves each bank's ratios and growth path while randomising its scale, so it separates a deviation that the scale and arithmetic of the figures produce from one they do not.

Applied to the panel, fixed conformity bands classify the pooled amounts and every bank as nonconforming, whereas against the benchmark the pooled amounts and eleven of the twelve banks fall within the range their own figures produce. The index varies mainly between banks, and the proxies are indistinguishable from zero, with minimum detectable effects well above the estimates. These results therefore neither show nor rule out associations between the proxies and reporting behaviour.

The contribution is methodological. Researchers using digit statistics on ratios built from the tested amounts should benchmark their associations against arithmetic dependence and report what the sample could have detected. Supervisors should use a first-digit screen on a few aggregate amounts to direct attention and not to classify banks. The empirical findings are exploratory, depend on the data status described in Section 4.1, and are not evidence of manipulation by any bank. Future work should reconcile the amounts, verify the coding of the binary proxies and test pressure with a proxy that shares no inputs with the index.

Supplementary information. Online Resource 1 contains supplementary tables S1 to S4. S1: indicative chi-square, second-digit and first-two-digit statistics. S2: false-positive rates of Rule R1. S3: detection rates of Rule R1 against controlled leading-digit departures. S4: coding rules and status of the binary variables.

Statements and Declarations

Funding. This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests. The authors have no relevant financial or non-financial interests to disclose.

Ethics approval. The study uses only information published in bank annual reports. No human participants or animals were involved and no individual-level data were collected, so ethics approval and informed consent were not required.

Data availability. The amounts analysed come from the twelve banks' publicly available annual reports, obtainable from each bank's investor-relations page and the Indonesia Stock Exchange website. The derived dataset, correction log, codebook and binary codes are deposited in the OSF project at https://doi.org/10.17605/OSF.IO/574S6 under MIT License. The 33 bank-years with a balance-identity gap above 1% of assets are listed under neutral labels in the Appendix. The key

linking the bank labels (Banks A to L) to bank names is held in the same project as a restricted file. Access is granted on request to the authors solely to verify the reported results, including after publication.

Code availability. The Python code that reproduces all exact statistics and, within Monte Carlo error, all random-draw results from the deposited data (random seed 12345) is deposited in the same OSF project, https://doi.org/10.17605/OSF.IO/574S6, under the MIT License. The binary codes CAP, ARO and COL enter as deposited data, because the rule that generated them was not retained (Section 3.3).

Use of AI tools. During the preparation of this manuscript, the authors used Claude 3.5 Sonnet (Anthropic) for debugging the analysis code and checking Python syntax and, in the final revision, Claude to draft and edit parts of the abstract, introduction, data, results, discussion and conclusion text and the Appendix, and to cross-check reported values against the analysis notebook. The authors reviewed, tested and edited all output, verified that all reported values reproduce from the code, and take full responsibility for the content of this publication.

Author contributions. Dien Noviany Rahmatika: Conceptualisation, Methodology, Formal analysis, Writing – original draft. Dewi Amaliah Nafiati: Investigation, Data curation, Writing – review and editing. Foster Mesreh: Software, Visualisation, Writing – review and editing. All authors approved the final manuscript and agree to be accountable for all aspects of the work.

Appendix

Table A1 lists the 33 bank-years whose balance-identity gap, the absolute value of total assets minus the sum of total liabilities and total equity divided by total assets, exceeds 1% (24 of them exceed 5%). The other 39 bank-years have a gap of at most 0.1%, and none lies between 0.1% and 1%. Banks C, E, G and I have no reconciled year. Gaps are given in year order.

Table A1 Bank-years with a balance-identity gap above 1% of total assets

Bank	Years (n)	Gap, % of total assets
Bank A	2022–2025 (4)	1.82, 2.28, 12.77, 10.34
Bank B	2020 (1)	1.66
Bank C	2020–2025 (6)	4.45, 4.62, 4.62, 4.92, 4.61, 5.21
Bank D	2020 (1)	3.62
Bank E	2020–2025 (6)	5.49, 6.12, 6.18, 6.18, 6.23, 6.04
Bank G	2020–2025 (6)	63.36, 67.24, 64.95, 64.84, 63.95, 65.61
Bank I	2020–2025 (6)	5.90, 10.23, 10.44, 10.41, 9.85, 10.25
Bank K	2023–2025 (3)	8.65, 6.90, 7.45

Note: Banks are identified by the neutral labels of the main text. Gaps are percentages of total assets, in the order of the years shown; the gap at Bank G, between 63.36% and 67.24% of assets in every year, is described in Section 4.1.

References

Achmad T, Ghozali I, Pamungkas ID (2022) Hexagon Fraud: Detection of Fraudulent Financial Reporting in State-Owned Enterprises Indonesia. Economies 10(1):13. https://doi.org/10.3390/economies10010013

Achmad T, Ghozali I, Helmina MRA, Hapsari DI, Pamungkas ID (2023) Detecting Fraudulent Financial Reporting Using the Fraud Hexagon Model: Evidence from the Banking Sector in Indonesia. Economies 11(1):5. https://doi.org/10.3390/economies11010005

Aftabi SZ, Ahmadi A, Farzi S (2023) Fraud detection in financial statements using data mining and GAN models. Expert Systems with Applications 227:120144. https://doi.org/10.1016/j.eswa.2023.120144

Aybars A, Ataunal L (2016) An application of Benford's law to fundamental accounting figures reported by Borsa Istanbul (BIST) companies. Journal of Economics, Finance and Accounting 3(3):234–243. https://doi.org/10.17261/Pressacademia.2016321991

Barney BJ, Schulzke KS (2016) Moderating "cry wolf" events with excess MAD in Benford's law research and practice. Journal of Forensic Accounting Research 1(1):A66–A90. https://doi.org/10.2308/jfar-51622

Benford F (1938) The Law of Anomalous Numbers. Proceedings of the American Philosophical Society 78(4):551–572.

Bretz F, Hothorn T, Westfall P (2016) Multiple comparisons using R. CRC Press, Boca Raton

Brodeur A, Cook NM, Hartley JS, Heyes A (2024) Do preregistration and preanalysis plans reduce p-hacking and publication bias? Evidence from 15,992 test statistics and suggestions for improvement. Journal of Political Economy Microeconomics 2(3):527–561. https://doi.org/10.1086/730455

Cameron AC, Miller DL (2015) A practitioner's guide to cluster-robust inference. Journal of Human Resources 50(2):317–372. https://doi.org/10.3368/jhr.50.2.317

Cameron AC, Gelbach JB, Miller DL (2008) Bootstrap-based improvements for inference with clustered errors. Review of Economics and Statistics 90(3):414–427. https://doi.org/10.1162/rest.90.3.414

Cerioli A, Barabesi L, Cerasa A, Menegatti M, Perrotta D (2019) Newcomb–Benford law and the detection of frauds in international trade. Proceedings of the National Academy of Sciences 116(1):106–115. https://doi.org/10.1073/pnas.1806617115

Cerqueti R, Maggi M, Riccioni J (2024) Statistical methods for decision support systems in finance: How Benford's law predicts financial risk. Annals of Operations Research 342:1445–1469. https://doi.org/10.1007/s10479-022-04742-z

Cho WKT, Gaines BJ (2007) Breaking the (Benford) law: Statistical fraud detection in campaign finance. The American Statistician 61(3):218–223. https://doi.org/10.1198/000313007X223496

Del Gaudio BL, Salerno D, Sampagnaro G, Verdoliva V (2022) Misconduct risk in banking services: Does a propensity to be sanctioned exist? International Review of Financial Analysis 81:102081. https://doi.org/10.1016/j.irfa.2022.102081

Drukker DM (2003) Testing for serial correlation in linear panel-data models. Stata Journal 3(2):168–177.

Durtschi C, Hillison W, Pacini C (2004) The effective use of Benford's law to assist in detecting fraud in accounting data. Journal of Forensic Accounting 5(1):17–34.

Farooq U, Nasir A, Khan KI (2026) Towards a 4P framework of financial statement fraud: A systematic literature review. International Journal of Management Reviews 28(2):e70005. https://doi.org/10.1111/ijmr.70005

Grammatikos T, Papanikolaou NI (2021) Applying Benford's law to detect accounting data manipulation in the banking industry. Journal of Financial Services Research 59(1):115–142. https://doi.org/10.1007/s10693-020-00334-9

Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D et al (2020) Array programming with NumPy. Nature 585:357–362. https://doi.org/10.1038/s41586-020-2649-2

Hill TP (1995) A statistical derivation of the significant-digit law. Statistical Science 10(4):354–363. https://doi.org/10.1214/ss/1177009869

Huang L, Abrahams A, Ractham P (2022) Enhanced financial fraud detection using cost-sensitive cascade forest with missing value imputation. Intelligent Systems in Accounting, Finance and Management 29(3):133–155. https://doi.org/10.1002/isaf.1517

Khamainy AH, Amalia MM, Cakranegara PA, Indrawati A (2022) Financial Statement Fraud: The Predictive Relevance of Fraud Hexagon Theory. JASF: Journal of Accounting and Strategic Finance 5(1):110–133. https://doi.org/10.33005/jasf.v5i1.249

Lesperance M, Reed WJ, Stephens MA, Tsao C, Wilton B (2016) Assessing conformance with Benford's law: Goodness-of-fit tests and simultaneous confidence intervals. PLOS ONE 11(3):e0151235. https://doi.org/10.1371/journal.pone.0151235

Liu R (2023) Tracking disclosure change trajectories for financial fraud detection. Production and Operations Management 32(2):584–602. https://doi.org/10.1111/poms.13888

Long JS, Ervin LH (2000) Using heteroscedasticity consistent standard errors in the linear regression model. The American Statistician 54:217–224. https://doi.org/10.1080/00031305.2000.10474549

MacKinnon JG, Webb MD (2018) The wild bootstrap for few (treated) clusters. The Econometrics Journal 21(2):114–135. https://doi.org/10.1111/ectj.12107

McKinney W (2010) Data structures for statistical computing in Python. In: van der Walt S, Millman J (eds) Proceedings of the 9th Python in Science Conference, pp 56–61. https://doi.org/10.25080/Majora-92bf1922-00a

Morrow J (2014) Benford's law, families of distributions and a test basis. CEP Discussion Paper No. 1291. Centre for Economic Performance, London School of Economics and Political Science, London

Najmuddin AB, Widihardimas TJ (2026) Detection Fraudulent Financial Reporting Using Hexagon Fraud Theory in the Banking Sector. Jurnal Akuntansi 15(1). https://doi.org/10.46806/ja.v15i1.1697

Nguyen TT, Duong CM (2022) Is conditional conservatism a source of deviations of financial statements from Benford's Law? Journal of Applied Accounting Research 23(5):1024–1046. https://doi.org/10.1108/JAAR-02-2021-0037

Nigrini MJ (2012a) Assessing conformity to Benford's Law. In: Benford's Law: applications for forensic accounting, auditing, and fraud detection. Wiley, Hoboken. https://doi.org/10.1002/9781119203094.ch7

Nigrini MJ (2012b) Primary Benford's Law tests. In: Benford's Law: applications for forensic accounting, auditing, and fraud detection. Wiley, Hoboken. https://doi.org/10.1002/9781119203094.ch4

Papík M, Papíková L (2022) Detecting accounting fraud in companies reporting under US GAAP through data mining. International Journal of Accounting Information Systems 45:100559. https://doi.org/10.1016/j.accinf.2022.100559

Pinkham RS (1961) On the distribution of first significant digits. Annals of Mathematical Statistics 32(4):1223–1230

Sagala SG, Siagian V (2021) Pengaruh Fraud Hexagon Model terhadap Fraudulent Laporan Keuangan pada Perusahaan Sub Sektor Makanan dan Minuman yang Terdaftar di BEI Tahun 2016-2019 [The effect of the Fraud Hexagon model on financial statement fraud in food and beverage sub-sector companies listed on the IDX, 2016-2019]. Jurnal Akuntansi 13(2):245–259. https://doi.org/10.28932/jam.v13i2.3956

Sare MTS, Kurniawati AD (2025) Assessing the Risk of Fraudulent Financial Reporting in Indonesian Banking Companies Through the Lens of Hexagon Fraud Analysis. KINERJA 29(2). https://doi.org/10.24002/kinerja.v29i2.12289

Seabold S, Perktold J (2010) Statsmodels: econometric and statistical modeling with Python. In: van der Walt S, Millman J (eds) Proceedings of the 9th Python in Science Conference, pp 92–96. https://doi.org/10.25080/Majora-92bf1922-011

Shahana T, Lavanya V, Bhat AR (2023) State of the art in financial statement fraud detection: A systematic review. Technological Forecasting and Social Change 192:122527. https://doi.org/10.1016/j.techfore.2023.122527

Situngkir NC, Triyanto DN (2020) Detecting Fraudulent Financial Reporting Using Fraud Score Model and Fraud Pentagon Theory: Empirical Study of Companies Listed on the LQ 45 Index. The Indonesian Journal of Accounting Research 23(3). https://doi.org/10.33312/ijar.486

Skousen CJ, Smith KR, Wright CJ (2009) Detecting and Predicting Financial Statement Fraud: The Effectiveness of the Fraud Triangle and SAS No. 99. Corporate Governance and Firm Performance 13(2). https://doi.org/10.1108/S1569-3732(2009)0000013005

Sudrajat S, Suryadnyana NA, Supriadi T (2023) Fraud Hexagon: Detection of Fraud of Financial Report in State-owned Enterprises in Indonesia. Jurnal Tata Kelola Dan Akuntabilitas Keuangan Negara 9(1):87–102. https://doi.org/10.28986/jtaken.v9i1.1358

Sukmadilaga C, Winarningsih S, Handayani T, Herianti E, Ghani EK (2022) Fraudulent Financial Reporting in Ministerial and Governmental Institutions in Indonesia: An Analysis Using Hexagon Theory. Economies 10(4):86. https://doi.org/10.3390/economies10040086

Umar H, Partahi D, Purba RB (2020) Fraud Diamond Analysis in Detecting Fraudulent Financial Reports. International Journal of Scientific and Technology Research 9(3):6638–6646.

Virtanen P, Gommers R, Oliphant TE et al (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods 17:261–272. https://doi.org/10.1038/s41592-019-0686-2

Vousinas GL (2019) Advancing theory of fraud: The S.C.O.R.E. model. Journal of Financial Crime 26(1):372–381.

Wilantari NM, Ariyanto D (2023) Determinan Fraud Hexagon Theory dan Indikasi Financial Statement Fraud [Determinants of Fraud Hexagon theory and indications of financial statement fraud]. E-Jurnal Akuntansi 33(1):87–103. https://doi.org/10.24843/EJA.2023.v33.i01.p07

Wolfe DT, Hermanson DR (2004) The fraud diamond: Considering the four elements of fraud. The CPA Journal 74(12):38–42.

Wooldridge JM (2019) Correlated random effects models with unbalanced panels. Journal of Econometrics 211:137–150. https://doi.org/10.1016/j.jeconom.2018.12.010

Zhou Y, Xiao Z, Gao R, Wang C (2024) Using data-driven methods to detect financial statement fraud in the real scenario. International Journal of Accounting Information Systems 54:100693. https://doi.org/10.1016/j.accinf.2024.100693

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ESM1.pdf