Chapter 34

Data and Sample Construction

Chapter 32 translated the theoretical concept of Accounting Bias into a set of observable, mechanism-based indicators. The next task is to construct a dataset in which those indicators can be measured consistently across firms and through time.

This chapter describes that process.

The empirical analysis is based on publicly available financial-statement data reported to the U.S. Securities and Exchange Commission (SEC). Rather than relying on a commercial financial database, the study retrieves standardized accounting facts directly from SEC EDGAR using the Company Facts infrastructure and XBRL tags contained in corporate filings.

The final research design is a balanced sampling framework consisting of 50 predetermined U.S. listed firms distributed across five broad industry groups. The principal analysis period covers fiscal years 2021 through 2025, producing a maximum theoretical panel of

$$50 \times 5 = 250$$

firm-year observations.

Not every Accounting Bias proxy is available for every firm-year. The effective sample size therefore varies across variables and empirical specifications. Missingness is treated as an empirical feature of the data rather than mechanically replaced by zeros.

The resulting dataset provides the foundation for the structural and hypothesis-testing analyses developed in the following chapters.

34.1 Data Source

The primary source is the SEC EDGAR Company Facts database.

Public companies filing financial statements with the SEC report accounting information using eXtensible Business Reporting Language (XBRL). XBRL associates reported financial-statement quantities with standardized accounting concepts, allowing individual accounting facts to be retrieved programmatically. Prior research nevertheless documents that XBRL filings can contain tagging and data-quality problems, so machine readability should not be treated as equivalent to perfect comparability (Debreceny et al., 2010).

Conceptually, the data pipeline can be represented as

$$\begin{aligned} \text{SEC filings} &\to \text{XBRL facts} \\ &\to \text{firm-year accounting variables} \\ &\to \text{Accounting Bias proxies.} \end{aligned}$$

The advantage of this approach is that the empirical analysis remains closely connected to the accounting information actually reported by firms.

The procedure does not begin with preconstructed financial ratios supplied by a commercial data vendor. Instead, the relevant accounting facts are downloaded and subsequently transformed into the variables required by the Accounting Bias framework.

This distinction is important because the empirical analysis depends on accounting concepts whose construction and availability differ substantially across firms.

34.2 SEC Company Facts

For each reporting entity, the SEC Company Facts data provide observations associated with accounting concepts such as total assets, goodwill, property, plant and equipment, accumulated depreciation, research and development expenditure, net income, operating cash flows, deferred taxes, and related quantities.

A generic accounting fact can be represented as

$$F_{i,j,t},$$

where

$$i = \text{firm},$$

$$j = \text{accounting concept},$$

and

$$t = \text{reporting period.}$$

In practice, a single accounting concept may appear several times within the raw data because filings can contain quarterly observations, annual observations, amended filings, comparative prior-year information, and facts associated with different reporting frames.

The raw SEC data therefore cannot simply be treated as a conventional rectangular dataset.

They first require identification of the economically appropriate annual observation for each firm, accounting concept, and fiscal year.

34.3 Why Direct Extraction Is Non-Trivial

XBRL creates a standardized reporting language, but it does not eliminate differences in corporate reporting practice.

Two firms reporting economically similar quantities may use different XBRL concepts. Even within the same firm, the tag used for a particular accounting quantity may change through time.

Consequently, empirical extraction requires more than requesting one tag for every variable.

For a target accounting quantity Y, the extraction procedure can be represented as

$$Y_{it} = \text{Select}\left(X_{it}^{(1)}, X_{it}^{(2)}, \ldots, X_{it}^{(m)}\right),$$

where

$$X^{(1)}, X^{(2)}, \ldots, X^{(m)}$$

are candidate XBRL concepts capable of representing the required economic quantity.

The objective is to identify a consistent accounting concept while avoiding the incorrect combination of facts that refer to economically different quantities.

This problem is particularly important for specialized variables such as research and development expenditure, accumulated depreciation, intangible assets, allowances, provisions, and deferred-tax positions.

34.4 A Predetermined Firm Sample

The empirical analysis uses a predetermined sample of 50 firms.

The firms were selected before the hypothesis-testing stage and organized into five broad industry groups, with ten firms assigned to each group:

$$5 \text{ industries} \times 10 \text{ firms} = 50 \text{ firms}.$$

This sampling structure serves several purposes.

First, it prevents the empirical sample from being selected after observing the results.

Second, it deliberately introduces variation in firms' underlying economic structures. The theory of Accounting Bias predicts that exposure to accounting mechanisms should depend partly on the economic activities undertaken by firms.

A technology-intensive firm, for example, may be highly exposed to recognition issues surrounding internally generated knowledge resources. A capital-intensive firm may be more exposed to historical-cost measurement, depreciation, and asset-age effects. Acquisition-intensive firms may carry substantial goodwill. Other business models may generate greater importance for accruals, provisions, allowances, or deferred taxes.

Industry variation is therefore not merely a statistical nuisance. It is itself relevant to the theoretical framework.

If Accounting Bias is partly generated by the interaction between accounting rules and economic structure, then we should expect

$$\begin{aligned}\text{economic structure} &\rightarrow \text{accounting exposure}\\ &\rightarrow \text{different empirical bias profiles.}\end{aligned}$$

The industry dimension of the sample makes this proposition empirically testable.

34.5 Time Period

The principal analysis covers fiscal years

$$2021,\ 2022,\ 2023,\ 2024,\ 2025.$$

Thus, the intended analysis panel contains five observations per firm:

$$T = 5.$$

With 50 firms, the maximum number of firm-year observations is therefore

$$N_{\max} = 50 \times 5 = 250.$$

The final constructed panel indeed contains 250 firm-year rows.

However, the raw extraction process also uses information from 2020 when required to construct variables that depend on lagged balance-sheet values or changes between periods.

This distinction gives two relevant time ranges:

$$\text{support period} = 2020\text{–}2025,$$

and

$$\text{analysis period} = 2021\text{–}2025.$$

The support year is not itself part of the main empirical sample. It provides beginning-of-period information necessary for variables such as asset growth and other measures requiring prior-year accounting quantities.

34.6 Firm Identification

Public companies are identified in SEC data through the Central Index Key (CIK).

Because the predetermined sample was initially specified using stock-market tickers, the first stage of the data pipeline maps each ticker into the corresponding SEC CIK.

Formally,

$$Ticker_i \to CIK_i.$$

The CIK then becomes the stable identifier used to retrieve the firm's Company Facts data.

This separation is useful because ticker symbols can change, while the SEC's entity identifier provides a more stable link to the filing system.

The mapping stage was explicitly validated before the main extraction proceeded. Unresolved tickers were treated as errors requiring resolution rather than silently dropped from the sample.

This preserves the predetermined nature of the research design.

34.7 Fiscal-Year Alignment

One of the most important data-construction problems concerns fiscal-year alignment.

SEC Company Facts contain observations associated with filing dates, period-end dates, fiscal years, fiscal periods, and reporting frames. These concepts are related but not identical.

The empirical objective is to construct economically meaningful firm-year observations:

$$(i, t),$$

where t represents the firm's fiscal year rather than simply the calendar year in which a filing happened to be submitted.

For each accounting fact, the extraction procedure therefore considers information such as:

- fiscal year;
- fiscal period;
- period start date;
- period end date;
- filing date;
- form type;
- reporting frame.

Annual financial-statement observations are preferred for the construction of the main panel.

This matters because an annual filing submitted in calendar year $t+1$ may report a fiscal year ending in year t. Assigning the observation according to filing year would create systematic temporal misclassification.

The empirical dataset therefore uses an *economic fiscal year* variable designed to align accounting quantities with the period to which they economically belong.

34.8 Stock and Flow Variables

Accounting variables differ in their temporal structure.

Balance-sheet quantities are stocks measured at a point in time:

$$Stock_{it} = X_i(t_{\text{end}}).$$

Examples include:

- total assets;
- goodwill;
- property, plant and equipment;
- accumulated depreciation;
- deferred-tax balances.

Income-statement and cash-flow quantities are flows measured over an interval:

$$Flow_{it} = \int_{t_0}^{t_1} x_i(s)\, ds.$$

Examples include:

- research and development expenditure;
- net income;
- operating cash flow;
- depreciation expense.

The distinction matters when selecting observations from SEC data.

For stock variables, the relevant fact should correspond to the appropriate fiscal-year-end date.

For flow variables, the observation should correspond to the full annual duration.

Failure to distinguish stocks from flows could, for example, combine a year-end balance-sheet value with a quarterly income-statement amount, generating economically meaningless ratios.

The extraction pipeline therefore treats the temporal structure of accounting variables explicitly.

34.9 Construction of the Base Accounting Panel

After extraction and fiscal-year alignment, the selected accounting facts are combined into a firm-year panel.

Conceptually, the base dataset takes the form

$$\mathcal{D} = \{i, t, X_{1,it}, X_{2,it}, \ldots, X_{K,it}\},$$

where the X_k are accounting quantities required for subsequent proxy construction.

The panel contains one row for each firm-year combination in the predetermined sample.

The intended structure is therefore

$Firm$	$Year$	$Assets$	$Goodwill$	PPE	$\cdots$
1	2021	·	·	·	$\cdots$
1	2022	·	·	·	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
50	2025	·	·	·	$\cdots$

This base panel is distinct from the final Accounting Bias proxy panel.

The first contains extracted accounting facts.

The second transforms those facts into economically interpretable measures of exposure to the mechanisms developed in Chapter 33.

34.10 Scaling Variables

Firm size creates an immediate comparability problem.

Suppose two firms report goodwill of

$$GW_A = 5$$

and

$$GW_B = 50.$$

The second firm does not necessarily have ten times the relative goodwill exposure of the first. It may simply be much larger.

The principal Accounting Bias variables are therefore expressed as intensities.

For an accounting quantity X,

$$XINT_{it} = \frac{X_{it}}{D_{it}},$$

where D_{it} is an economically appropriate scaling variable.

Total assets provide the principal denominator:

$$TA_{it}.$$

This produces ratios such as

$$RDINT_{it} = \frac{R\&D_{it}}{TA_{it}},$$

$$GWINT_{it} = \frac{Goodwill_{it}}{TA_{it}},$$

and

$$PPEINT_{it} = \frac{PPE_{it}}{TA_{it}}.$$

Scaling improves cross-sectional comparability and gives the variables a natural interpretation as the relative importance of particular accounting mechanisms within the firm's reported financial structure.

34.11 Average Assets

Some variables describe activity occurring throughout a fiscal year rather than a balance existing only at year-end.

For such quantities, average assets provide a more natural denominator:

$$\overline{TA}_{it} = \frac{TA_{i,t-1} + TA_{it}}{2}.$$

The use of average assets reduces the mechanical dependence of annual flow ratios on the arbitrary choice of either beginning-of-year or end-of-year balance-sheet size.

This is particularly useful for performance and accrual-based measures.

For example, return on assets is constructed conceptually as

$$ROA_{it} = \frac{NI_{it}}{\overline{TA}_{it}}.$$

Likewise, the principal accrual-intensity measure uses the divergence between net income and operating cash flow relative to the firm's average asset base:

$$ACCRUALINT_{it} = \frac{|NI_{it} - CFO_{it}|}{\overline{TA}_{it}}.$$

The absolute value is used because the variable is intended to capture the *magnitude* of the accounting–cash-flow divergence rather than its direction.

34.12 Core Accounting Bias Proxies

The final empirical panel identifies six variables as the core Accounting Bias proxies:

$$RDINT, \quad GWINT, \quad PPEINT,$$
$$ACCDEPINT, \quad ACCRUALINT, \quad DTINT.$$

These variables cover three of the principal theoretical dimensions with sufficient empirical availability for systematic analysis.

34.12.1 Research and Development Intensity

Recognition exposure associated with internally generated knowledge investment is represented by

$$RDINT_{it} = \frac{R\&D_{it}}{TA_{it}}.$$

The variable identifies firms for which research and development activity is economically important relative to the reported accounting asset base.

34.12.2 Goodwill Intensity

Acquisition-related recognition exposure is represented by

$$GWINT_{it} = \frac{Goodwill_{it}}{TA_{it}}.$$

Goodwill is particularly useful because its accounting existence depends on acquisition activity. It therefore captures a form of recognized intangible value fundamentally different from internally generated knowledge expenditure.

34.12.3 Property, Plant and Equipment Intensity

Exposure to long-lived tangible-asset measurement is represented by

$$PPEINT_{it} = \frac{PPE_{it}}{TA_{it}}.$$

The variable identifies firms for which historical cost, depreciation, impairment, and related measurement rules are especially important to the accounting representation.

34.12.4 Accumulated Depreciation Intensity

A second measurement indicator captures the depreciation state of the firm's tangible asset base.

The empirical variable is denoted

$$ACCDEPINT_{it}.$$

It relates accumulated depreciation to the relevant property, plant and equipment base and therefore provides information about the extent to which recognized tangible assets have already been allocated through accounting depreciation.

Conceptually,

$$ACCDEPINT \uparrow$$

indicates a larger accumulated accounting allocation relative to the underlying depreciable asset base.

This variable complements $PPEINT$. Whereas $PPEINT$ captures the importance of tangible assets within the firm, $ACCDEPINT$ captures characteristics of their accounting measurement through time.

34.12.5 Accrual Intensity

Timing exposure is represented by

$$ACCRUALINT_{it} = \frac{|NI_{it} - CFO_{it}|}{\overline{TA}_{it}}.$$

This measure captures the magnitude of the difference between accrual-based earnings and operating cash-flow realization.

A larger value implies that reported earnings depend more heavily on accounting timing and allocation mechanisms relative to the firm's asset base.

34.12.6 Deferred-Tax Intensity

A second timing-related indicator is deferred-tax intensity:

$$DTINT_{it} = \frac{|DT_{it}|}{TA_{it}},$$

where DT_{it} represents the relevant net deferred-tax position constructed from the available tax-related accounting facts.

Deferred taxes arise because accounting recognition and tax recognition do not always occur in the same periods or at the same measured amounts.

The variable therefore captures a different form of temporal divergence from operating accruals.

This distinction later proves empirically important: two variables can belong to the same broad theoretical dimension without necessarily behaving as manifestations of one homogeneous latent factor.

34.13 Extended Indicators

The data-construction process also produces additional indicators that are theoretically informative but less uniformly available across the sample.

These include

$$INTANINT,$$

representing identifiable intangible-asset intensity, and

$$PPEAGE,$$

which provides information about the effective accounting age of the tangible asset base where the necessary components are available.

These variables are retained for descriptive and structural analysis where appropriate, but they are not given the same role as the six core proxies in the main hypothesis tests.

The distinction between core and extended indicators is driven by empirical availability rather than by an assumption that the extended mechanisms are theoretically unimportant.

34.14 Estimation Indicators

The theoretical framework also motivates variables related to estimation-intensive accounting quantities, including allowance and provision measures.

The empirical construction process therefore considers indicators such as

$$ALLOWINT$$

and

$$PROVINT.$$

These variables are particularly relevant to the Estimation dimension because allowances and provisions depend directly on expectations regarding uncertain future outcomes.

However, standardized XBRL availability for these concepts is substantially more limited across the predetermined sample.

The Estimation dimension consequently has a much smaller complete-case sample than Recognition, Measurement, and Timing.

This limitation has an important methodological consequence.

Rather than forcing a poorly measured Estimation factor into the main statistical analysis, the study treats the available estimation indicators as exploratory.

This preserves the distinction between

$$\text{theoretical relevance}$$

and

$$\text{empirical observability}.$$

A theoretically important mechanism does not become empirically reliable merely because the research design requires a variable for it.

34.15 Control Variables

The empirical models also require variables describing broader firm characteristics that may be correlated with both Accounting Bias exposure and financial outcomes.

Three principal controls are used:

$$SIZE_{it},$$

$$EQUITY\_RATIO_{it},$$

and

$$ASSET\_GROWTH_{it}.$$

Firm size captures systematic differences associated with organizational scale.

Conceptually,

$$SIZE_{it} = \ln(TA_{it}),$$

or the corresponding transformation implemented in the constructed panel.

Capital structure is represented through the equity ratio:

$$EQUITY\_RATIO_{it} = \frac{Equity_{it}}{TA_{it}}.$$

Growth is represented by the annual change in total assets:

$$ASSET\_GROWTH_{it} = \frac{TA_{it} - TA_{i,t-1}}{TA_{i,t-1}}.$$

These controls do not eliminate every source of firm heterogeneity. Their purpose is to reduce the risk that estimated relationships between Accounting Bias proxies and other variables merely reproduce obvious differences in firm scale, capitalization, or growth.

Industry and fiscal-year effects are introduced separately in the empirical models described in the following chapter.

34.16 Return on Assets

The principal accounting outcome used in the later association tests is return on assets.

It is defined as

$$ROA_{it} = \frac{NI_{it}}{TA_{it}}.$$

ROA provides a standardized measure of reported accounting profitability relative to the resources employed by the firm.

Its role in the empirical analysis requires careful interpretation.

The objective is not to estimate a causal equation of the form

$$Accounting\ Bias \to ROA$$

without qualification.

Accounting structure, economic performance, business model, acquisition history, financing, investment, and reporting outcomes are jointly determined in complex ways.

The ROA regressions should therefore be interpreted as tests of conditional association:

$$\text{Accounting Bias exposure} \leftrightarrow \text{reported profitability.}$$

They ask whether firms with different accounting exposure profiles also exhibit systematically different reported profitability after controlling for observable firm characteristics, industry, and time.

34.17 Missing Values

A central issue in the dataset is missingness.

An accounting variable can be absent from the SEC data for several reasons.

The firm may not engage materially in the underlying economic activity. The relevant accounting quantity may not be separately disclosed. The company may use a different XBRL concept. The required components may not be recoverable consistently. Or the economic concept may simply not apply to the firm's business model.

These possibilities are not equivalent.

Most importantly,

$$Missing \neq 0.$$

A missing R&D observation, for example, cannot automatically be interpreted as proof that the firm conducted no research and development.

Likewise, the absence of a standardized provision tag does not establish that the firm's economically relevant provisions were zero.

The construction process therefore preserves missing observations as missing unless the accounting meaning of zero can be established from the underlying facts.

This decision reduces sample sizes for some proxies but avoids creating artificial data.

34.18 Missingness as Economic Information

Missingness is not purely a technical inconvenience.

Because accounting disclosures depend partly on business models, the availability of particular variables may itself be systematically related to industry.

Research and development provides an obvious example. R&D information is much more relevant for some industries than for others.

Consequently, requiring complete observations for every proxy can change the economic composition of the sample:

$$\text{complete-case restriction} \rightarrow \text{industry selection.}$$

This issue becomes especially important when several proxies are included simultaneously.

The empirical analysis therefore reports the number of observations used in each specification rather than treating $N = 250$ as the sample size for every model.

The full panel contains 250 firm-year observations, but each empirical test uses the subset for which its required variables are observed.

34.19 Coverage Across the Empirical Dimensions

The final dataset exhibits substantial differences in coverage across the four conceptual dimensions.

Recognition indicators have sufficient observations for systematic analysis, although research and development is naturally concentrated in particular types of firms.

Measurement indicators also provide a substantial sample.

Timing variables have the broadest coverage among the main dimensions.

Estimation indicators are considerably more restricted.

In the final structural analysis, complete observations across the principal variables of each dimension amount to approximately

$$N_R = 114$$

for Recognition,

$$N_M = 125$$

for Measurement,
and

$$N_T = 210$$

for Timing.

By contrast, the Estimation dimension provides only approximately

$$N_S = 15$$

complete observations under the relevant combined specification.

This difference determines the empirical strategy adopted later.

Recognition, Measurement, and Timing can support systematic structural analysis.

Estimation remains conceptually part of the Accounting Bias framework but is treated empirically as exploratory because the available standardized data do not support equally strong inference.

34.20 No Mechanical Imputation

The empirical pipeline deliberately avoids filling missing accounting facts through mechanical imputation.

For example, the procedure does not replace an unobserved accounting variable with

$$\overline{X}_{industry},$$

$$\overline{X}_{firm},$$

or zero merely to increase the sample size.

Such procedures would create observations that were not actually reported and could artificially strengthen correlations among the Accounting Bias indicators.

Instead, empirical models operate on the observations for which the required variables can be constructed from reported information.

This approach creates unequal sample sizes but preserves a clearer relationship between the empirical measures and the underlying financial statements.

34.21 Data Validation

Automated financial-data extraction requires explicit validation.

Several diagnostic stages were therefore incorporated into the construction pipeline.

First, ticker-to-CIK resolution was checked before the main sample was accepted.

Second, accounting concepts were inspected for availability across firms and years.

Third, variables with unexpectedly low coverage were investigated rather than automatically accepted.

Fourth, selected firms and accounting concepts were examined individually when anomalous extraction patterns appeared.

Fifth, the constructed proxy distributions were evaluated for implausible values and unexpected missingness.

The general validation logic can be represented as

$$\text{download} \rightarrow \text{construct} \rightarrow \text{diagnose} \rightarrow \text{revise} \rightarrow \text{validate}.$$

This iterative procedure was particularly important because XBRL standardization does not guarantee complete economic comparability across issuers.

34.22 Separation of Data Construction and Statistical Analysis

The empirical workflow deliberately separates data engineering from statistical testing.

The principal stages are:

$$\text{Stage 1: SEC extraction,}$$

$$\text{Stage 2: firm-year panel construction,}$$

$$\text{Stage 3: Accounting Bias proxy construction,}$$

$$\text{Stage 4: structural diagnostics,}$$

Stage 5: hypothesis testing,

and

Stage 6: publication output.

This separation is methodologically useful.

The raw-data extraction rules are not modified according to whether a subsequent regression becomes statistically significant.

Likewise, the predetermined sample is resolved and constructed before the main hypothesis-testing stage.

The procedure therefore reduces the possibility that data-construction choices become implicitly optimized to obtain particular empirical results.

34.23 Reproducibility

The entire empirical dataset is generated programmatically.

The workflow begins with the predetermined firm sample and retrieves the required SEC data through a sequence of Python scripts.

The scripts perform tasks including:

- ticker and CIK resolution;
- SEC Company Facts retrieval;
- XBRL concept selection;
- fiscal-year alignment;
- annual observation selection;
- firm-year panel construction;
- proxy construction;
- missingness diagnostics;
- structural analysis;
- regression analysis;
- generation of publication-ready tables and figures.

The resulting empirical process is therefore reproducible in principle from the original public filings.

This is particularly valuable for a study of Accounting Bias because the transformation from reported accounting facts into empirical indicators is itself part of the methodological argument.

The research design does not hide this transformation behind a proprietary database.

34.24 The Final Empirical Panel

The completed dataset contains

$$50$$

firms observed across

$$5$$

analysis years, producing

$$250$$

firm-year rows.

The observations are organized around three identifiers:

$$Ticker_i,$$

$$Industry_i,$$

and

$$FiscalYear_t.$$

Each row then contains the available accounting facts, constructed Accounting Bias indicators, control variables, and accounting outcomes.

The conceptual structure is

$$\mathcal{P}_{it} = \left[ID_i, Industry_i, Year_t, \mathbf{X}_{it}^{AB}, \mathbf{C}_{it}, Y_{it}\right],$$

where

$$\mathbf{X}_{it}^{AB}$$

denotes the Accounting Bias proxy vector,

$$\mathbf{C}_{it}$$

the control variables, and

$$Y_{it}$$

the accounting outcome variables.

The core empirical proxy vector is

$$\mathbf{X}_{it}^{AB} = \begin{bmatrix} RDINT & GWINT & PPEINT \\ ACCDEPINT & ACCRUALINT & DTINT \end{bmatrix}_{it}.$$

The controls are

$$\mathbf{C}_{it} = [SIZE, EQUITY\_RATIO, ASSET\_GROWTH]_{it},$$

while the principal profitability outcome is

$$Y_{it} = ROA_{it}.$$

This panel becomes the common empirical foundation for the analyses that follow.

34.25 Limits of the Dataset

The use of SEC XBRL data offers transparency and reproducibility, but it also imposes limitations.

First, standardized tags are not perfectly uniform across firms.

Second, disclosure availability varies by business model and industry.

Third, some theoretically attractive Accounting Bias indicators cannot be constructed with sufficient coverage.

Fourth, the sample contains 50 predetermined firms rather than the entire population of U.S. listed companies.

Fifth, the five-year period is sufficient to distinguish persistent from time-varying patterns to some extent, but it is not a long historical panel.

Sixth, accounting quantities remain accounting representations. The dataset does not independently observe the true economic value of each asset, liability, revenue, or expense.

The empirical analysis therefore does not claim to reconstruct the unobservable vector

$$E_{it}.$$

Instead, it investigates the observable structure predicted by the theory:

$$\begin{aligned}\text{business model} &\rightarrow \text{accounting mechanism exposure}\\ &\rightarrow \text{observable accounting patterns.}\end{aligned}$$

These limitations define the scope of inference and will be considered explicitly when interpreting the results.

34.26 Conclusion

This chapter has described the construction of the dataset used to investigate Accounting Bias empirically.

The study uses publicly available SEC EDGAR Company Facts and XBRL data rather than relying on preconstructed commercial financial ratios. A predetermined sample of 50 firms is distributed across five industry groups and observed over fiscal years 2021–2025.

The resulting panel contains a maximum of

$$N = 250$$

firm-year observations.

Raw accounting facts are retrieved, aligned to economic fiscal years, classified according to their stock or flow nature, and transformed into mechanism-based Accounting Bias indicators.

The six core proxies are

Core proxies: $RDINT$, $GWINT$, $PPEINT$, $ACCDEPINT$, $ACCRUALINT$, and $DTINT$.

and represent Recognition, Measurement, and Timing mechanisms.

Additional indicators extend the analysis toward identifiable intangibles, asset age, allowances, and provisions, but uneven XBRL availability means that these variables play a more limited role in the main hypothesis tests.

Three principles govern the construction of the empirical panel:

$$\boxed{Missing \neq 0}$$