

Sample Construction and SEC/XBRL Extraction

B.1 Overview

The empirical dataset is constructed from publicly available information reported to the U.S. Securities and Exchange Commission.

The principal source is SEC Company Facts, which organizes financial-statement information reported through XBRL.

The complete transformation can be represented as

SEC Filing $\rightarrow$ XBRL Facts
 $\rightarrow$ Firm-Year Accounting Data
 $\rightarrow$ Accounting Bias Proxies.

B.2 Predetermined Sample

The analysis uses

50

predetermined U.S. listed companies.

The firms are distributed across five broad industry groups, with ten firms in each group:

$$5 \times 10 = 50.$$

The sample is fixed before the principal hypothesis tests are evaluated.

This prevents firms from being selected or removed merely because their inclusion changes statistical significance.

B.3 Analysis Period

The principal empirical period is

2021–2025.

With 50 firms and five fiscal years, the maximum panel contains

$$50 \times 5 = 250$$

firm-year rows.

A support year,

2020,

is additionally downloaded when required for lagged calculations.

Thus,

support period = 2020–2025,

while

analysis period = 2021–2025.

B.4 Firm Identification

The original sample is specified using stock-market tickers.

SEC retrieval is performed using Central Index Keys.

The first identification step is therefore

$$Ticker_i \rightarrow CIK_i.$$

Unresolved sample firms are treated as data-construction errors rather than silently dropped.

This preserves the predetermined sampling design.

B.5 Raw Company Facts

For each firm, Company Facts observations contain information including

- XBRL taxonomy;
- accounting concept;

- unit;
- value;
- period start;
- period end;
- fiscal year;
- fiscal period;
- filing form;
- filing date;
- accession information.

A single accounting concept may appear more than once. Reasons include

- annual and quarterly reporting;
- comparative prior-year amounts;
- amended filings;
- multiple reporting contexts;
- taxonomy differences.

The raw data therefore require explicit selection rules before a firm-year panel can be constructed.

B.6 Stock and Flow Variables

The extraction process distinguishes between stock and flow variables.

A stock variable is measured at a point in time:

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

Examples include

- total assets;
- goodwill;
- PPE;
- accumulated depreciation;

- deferred-tax balances.

A flow variable is measured over an interval:

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

Examples include

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

Annual flow observations must represent the full fiscal-year duration.

Stock observations must correspond to the relevant fiscal-year-end date.

B.7 Fiscal-Year Alignment

SEC filing dates do not necessarily coincide with the fiscal year represented by the accounting data.

An annual report submitted in calendar year $t + 1$ may contain the financial statements for fiscal year t .

The empirical panel therefore aligns observations to an economic fiscal year rather than simply using filing year.

The selection process considers information such as

- fiscal year;
- fiscal period;
- period start;
- period end;
- filing form;
- filing date.

B.8 XBRL Concept Resolution

Companies do not always report economically similar information under identical taxonomy elements.

For a target variable Y , the extraction procedure therefore searches a predefined hierarchy of candidate concepts:

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

The procedure is:

1. search the preferred XBRL concept;
2. search predefined fallback concepts;
3. apply the appropriate annual-period filters;
4. select the appropriate observation;
5. preserve missingness if no sufficiently comparable observation exists.

Fallback definitions are determined as part of data construction rather than adapted after statistical results are observed.

B.9 Missing Observations

The pipeline applies the rule

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

An absent standardized XBRL observation does not necessarily imply that the economic quantity is zero.

Missingness can result from

- different taxonomy choices;
- aggregation with another line item;
- lack of separate disclosure;
- economic non-applicability;
- insufficiently comparable available concepts.

The empirical dataset therefore avoids mechanical replacement of missing values with

0,

firm means, or industry means.

B.10 Missingness and Sample Composition

Missingness can itself be economically structured.

Research and development disclosure, for example, is much more common among some industries than others.

Consequently,

complete-case restriction $\rightarrow$ sample composition change.

This is especially important in the all-core-proxy specification.

The full panel contains

250

firm-year rows, but the sample size of each statistical model depends on the availability of the variables required by that model.

B.11 Coverage of the Empirical Dimensions

The dimension-specific complete-case samples used in the structural analysis are approximately

$$N_R = 114$$

for Recognition,

$$N_M = 125$$

for Measurement,
and

$$N_T = 210$$

for Timing.

The Estimation dimension provides only approximately

$$N_S = 15$$

complete observations under the relevant exploratory specification.

The empirical role of each dimension therefore reflects both its theoretical importance and its observed data coverage.

B.12 Validation

Automated extraction is followed by diagnostic validation.

The pipeline checks, where appropriate,

- ticker-to-CIK resolution;
- duplicate firm-year observations;
- unexpected missingness;
- invalid denominators;
- accounting ratios with implausible values;
- concept coverage across firms and years;
- year alignment;
- selected firm-specific anomalies.

The general construction logic is

download $\rightarrow$ construct $\rightarrow$ diagnose $\rightarrow$ revise $\rightarrow$ validate.

B.13 Final Panel Structure

The final panel can be represented as

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

where

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

the control vector is

$$\mathbf{C}_{it} = [SIZE, EQUITY_RATIO, ASSET_GROWTH]_{it},$$

and the principal outcome is

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

The complete panel contains

	50
firms,	
	5
analysis years, and	
	250
firm-year rows.	