Chapter E

Sample Firms and Industry Classification

E.1 Purpose

The empirical analysis developed in this book uses a predetermined sample of 50 U.S. listed firms.

The sample is divided into five broad industry groups:

1. Technology;
2. Healthcare;
3. Industrials;
4. Consumer;
5. Utilities.

Each group contains ten firms, so that

$$N = 5 \times 10 = 50.$$

The principal analysis covers fiscal years 2021–2025. Consequently, the maximum balanced panel contains

$$50 \times 5 = 250$$

firm-year observations.

The purpose of this appendix is to identify the firms included in the empirical analysis, explain the role of the industry classification, and distinguish the predetermined research sample from the smaller effective samples generated by proxy-specific missingness.

E.2 Predetermined Sampling Design

The sample was specified before the principal empirical hypotheses were evaluated.

This is an important feature of the research design.

The sample was not constructed by selecting firms for which the Accounting Bias proxies happened to produce statistically significant relationships.

Nor was it restricted in advance to firms for which every accounting variable was available.

Instead, the empirical design begins with a fixed set

$$\mathcal{S} = \{1, \ldots, 50\},$$

and subsequently allows the effective sample size to vary according to the availability of the variables required by each statistical specification.

Thus,

$$\boxed{\text{Sample Selection} \neq \text{Variable Availability.}}$$

This distinction is particularly important in empirical accounting research, because disclosure and standardized XBRL availability are themselves related to firms' economic characteristics.

E.3 Industry Classification

The five industry groups are intended to represent meaningfully different economic structures through which accounting rules operate.

The classification is deliberately broad.

It is not intended to reproduce a highly granular commercial industry taxonomy. Its purpose is instead to provide economically interpretable cross-sectional variation in exposure to Recognition, Measurement, Timing, and Estimation mechanisms.

E.3.1 Technology

Technology firms provide substantial exposure to internally generated knowledge, research and development, intangible resources, and acquisition activity.

The distinction between internally generated and externally acquired intangibles is particularly relevant for the Recognition dimension of Accounting Bias.

E.3.2 Healthcare

Healthcare firms provide additional exposure to research-intensive activity, intangible resources, business combinations, goodwill, and long development cycles.

The group is therefore particularly informative for Recognition-related mechanisms.

E.3.3 Industrials

Industrial firms generally provide stronger exposure to tangible productive assets, capital expenditure, depreciation, useful-life assumptions, and asset replacement cycles.

They are therefore particularly informative for the Measurement dimension.

E.3.4 Consumer

The Consumer group contains firms with heterogeneous operating structures.

These firms provide variation in tangible assets, brands, acquisitions, working-capital structures, and operating cash-flow patterns.

The group therefore contributes useful cross-sectional variation across several Accounting Bias mechanisms.

E.3.5 Utilities

Utilities are characterized by substantial investment in long-lived tangible assets.

They provide particularly strong exposure to property, plant and equipment, depreciation, asset age, and long-term measurement processes.

Their economic structure makes them especially informative for the Measurement dimension.

E.4 Complete Predetermined Sample

Table E.1 reports the complete sample specified in the empirical data-acquisition program.

The ticker symbols correspond to the identifiers used by the research pipeline.

Because the table contains 50 firms, the `longtable` environment is used so that it can continue across pages without reducing the font to an unnecessarily small size.

Table E.1: Predetermined Firm Sample

Industry Group	Ticker	Company
Technology	AAPL	Apple Inc.
Technology	MSFT	Microsoft Corporation
Technology	GOOGL	Alphabet Inc.
Technology	META	Meta Platforms, Inc.
Technology	ORCL	Oracle Corporation
Technology	CSCO	Cisco Systems, Inc.
Technology	IBM	International Business Machines Corporation
Technology	QCOM	Qualcomm Incorporated
Technology	TXN	Texas Instruments Incorporated
Technology	ADBE	Adobe Inc.
Healthcare	JNJ	Johnson & Johnson
Healthcare	PFE	Pfizer Inc.
Healthcare	MRK	Merck & Co., Inc.
Healthcare	ABBV	AbbVie Inc.
Healthcare	AMGN	Amgen Inc.
Healthcare	GILD	Gilead Sciences, Inc.
Healthcare	BMY	Bristol-Myers Squibb Company
Healthcare	LLY	Eli Lilly and Company
Healthcare	ABT	Abbott Laboratories
Healthcare	TMO	Thermo Fisher Scientific Inc.
Industrials	CAT	Caterpillar Inc.
Industrials	DE	Deere & Company
Industrials	HON	Honeywell International Inc.
Industrials	RTX	RTX Corporation
Industrials	EMR	Emerson Electric Co.
Industrials	ETN	Eaton Corporation plc
Industrials	PH	Parker-Hannifin Corporation
Industrials	CMI	Cummins Inc.
Industrials	F	Ford Motor Company
Industrials	GM	General Motors Company

Continued on next page

Table E.1 – continued from previous page

Industry Group	Ticker	Company
Consumer	WMT	Walmart Inc.
Consumer	COST	Costco Wholesale Corporation
Consumer	HD	The Home Depot, Inc.
Consumer	LOW	Lowe's Companies, Inc.
Consumer	TGT	Target Corporation
Consumer	PG	The Procter & Gamble Company
Consumer	KO	The Coca-Cola Company
Consumer	PEP	PepsiCo, Inc.
Consumer	NKE	NIKE, Inc.
Consumer	MCD	McDonald's Corporation
Utilities	NEE	NextEra Energy, Inc.
Utilities	DUK	Duke Energy Corporation
Utilities	SO	The Southern Company
Utilities	D	Dominion Energy, Inc.
Utilities	AEP	American Electric Power Company, Inc.
Utilities	EXC	Exelon Corporation
Utilities	XEL	Xcel Energy Inc.
Utilities	SRE	Sempra
Utilities	ED	Consolidated Edison, Inc.
Utilities	ETR	Entergy Corporation

E.5 Balanced Panel Structure

The predetermined sample is exactly balanced across the five industry groups:

$$N_{\text{Technology}} = N_{\text{Healthcare}} = N_{\text{Industrials}} = N_{\text{Consumer}} = N_{\text{Utilities}} = 10.$$

With five principal analysis years, each industry contributes a maximum of

$$10 \times 5 = 50$$

firm-year observations.

The maximum industry composition of the panel is therefore

$$\begin{aligned} N_{\text{Technology}} &= 50, \\ N_{\text{Healthcare}} &= 50, \\ N_{\text{Industrials}} &= 50, \\ N_{\text{Consumer}} &= 50, \\ N_{\text{Utilities}} &= 50. \end{aligned}$$

Hence,

$$N_{\text{Total}} = 250.$$

This represents the structure of the predetermined panel.

It does not imply that all 250 observations are available for every Accounting Bias proxy.

E.6 Why the Sample Is Not Defined by Data Availability

An alternative approach would be to construct the research sample only from firms for which every Accounting Bias proxy is observable.

Such a procedure would define

$$\begin{aligned} \mathcal{S}^{Complete} = \big\{(i,t) : {} & RDINT_{it},\, GWINT_{it},\, PPEINT_{it}, \\ & ACCDEPINT_{it},\, ACCRUALINT_{it},\, DTINT_{it} \\ & \text{are all observed}\big\}. \end{aligned}$$

This would create a serious interpretive problem.

Accounting-variable availability is not random.

For example, separate R&D disclosure is closely related to business model and industry.

A sample constructed only from complete observations could therefore become systematically concentrated in industries where particular accounting variables are more frequently reported.

The empirical strategy instead preserves the predetermined sample and treats missingness explicitly.

Accordingly,

$$N_j \leq 250$$

for proxy j, and

$$N_m \leq 250$$

for statistical model m.

The effective sample size is reported separately for each empirical test.

E.7 Dimension-Specific Effective Samples

The structural analysis illustrates the effect of proxy-specific availability.

The complete-case sample for the Recognition dimension is approximately

$$N_R = 114.$$

For Measurement,

$$N_M = 125.$$

For Timing,

$$N_T = 210.$$

The Estimation dimension has substantially lower standardized coverage, with approximately

$$N_S = 15$$

complete observations under the relevant exploratory specification.

These differences are not treated as inconvenient observations to be removed.

They are part of the empirical evidence.

In particular, they demonstrate that some theoretically important accounting mechanisms are considerably more difficult to observe consistently through standardized financial-statement data than others.

E.8 The Restricted All-Proxy Sample

The consequences of complete-case restrictions are especially visible in the multivariate profitability model.

That specification requires simultaneous availability of all six core

Accounting Bias proxies:

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

The resulting sample contains

$$N = 103$$

firm-year observations from

$$23$$

firms.

Its industry composition is approximately

$$\begin{aligned} N_{\text{Healthcare}} &= 45, \\ N_{\text{Technology}} &= 31, \\ N_{\text{Industrials}} &= 17, \\ N_{\text{Consumer}} &= 10, \\ N_{\text{Utilities}} &= 0. \end{aligned}$$

Thus,

$$45 + 31 + 17 + 10 = 103.$$

The absence of Utility observations from the all-proxy complete-case sample is particularly informative.

It does not mean that Utilities were excluded from the predetermined research design.

Rather, the simultaneous data requirements of the six-proxy model eliminate their observations from that particular specification.

This provides a concrete example of the distinction

$$\boxed{\text{Predetermined Sample} \neq \text{Complete-Case Sample.}}$$

E.9 Implications for the Multivariate Model

The change in sample composition has important consequences for interpretation.

The all-proxy regression does not estimate precisely the same cross-sectional population represented by the original balanced panel.

Instead,

$$\text{Complete-Case Restriction} \rightarrow \text{Composition Shift}.$$

For this reason, the multivariate ROA model is treated as a robustness test.

The principal profitability analysis estimates the Accounting Bias proxies separately, allowing each proxy to use the largest valid sample available for that specification.

The multivariate model asks a narrower question:

> Among observations for which all six core Accounting Bias proxies are simultaneously observable, which proxies retain conditional associations with reported profitability?

This is informative, but it should not replace the larger-sample individual models.

E.10 Industry as Economic Structure

Industry plays two related roles in the empirical framework.

First, industry is itself informative about exposure to accounting mechanisms.

The empirical results show statistically significant industry heterogeneity for all six core proxies:

$$\boxed{6/6.}$$

This indicates that the accounting structures measured by the proxies are not distributed randomly across firms.

They are systematically associated with differences in business model and economic activity.

Second, industry indicators are included as controls in the pooled regression models.

The generic specification is

$$\begin{aligned} Y_{it} = \alpha + \beta X_{it} + \boldsymbol{\theta}' \mathbf{C}_{it} \\ + \boldsymbol{\gamma}' Industry_i + \boldsymbol{\delta}' Year_t + \varepsilon_{it}. \end{aligned}$$

The industry coefficients are not interpreted as causal effects.

They absorb broad cross-sectional differences between the five economic groups.

E.11 Industry and Accounting Bias Exposure

The industry design provides a direct empirical connection between economic structure and accounting representation.

Conceptually,

$$\begin{aligned} Industry_i &\rightarrow BusinessModel_i \\ &\rightarrow MechanismExposure_{it} \\ &\rightarrow AccountingRepresentation_{it}. \end{aligned}$$

For example, research-intensive firms are more exposed to recognition rules governing internally generated knowledge.

Capital-intensive firms are more exposed to depreciation, useful-life assumptions, and historical-cost measurement.

Acquisition-intensive firms are more exposed to goodwill recognition and impairment.

The significant industry results are therefore consistent with a central prediction of the Accounting Bias Framework:

Accounting Bias exposure is partly endogenous to economic structure.

The same accounting system can produce systematically different representational effects when applied to firms with different business models.

E.12 Representativeness and External Validity

The sample is structured rather than statistically representative of the entire population of U.S. listed corporations.

No claim is made that the 50 firms constitute a random probability sample.

The empirical design is therefore not intended to estimate population quantities such as

$$P\left(X^{AB} > x\right)$$

for all U.S. corporations.

Instead, the sample is designed to provide economically meaningful variation with which to investigate whether Accounting Bias exposures exhibit

- systematic industry differences;
- persistence through time;
- structural versus dynamic variation;
- relationships within theoretical accounting dimensions;
- conditional associations with reported profitability.

The empirical results should consequently be interpreted within this scope.

Broader population inference would require a substantially larger and more systematically sampled universe of firms.

E.13 Survivorship and Observation Requirements

The empirical panel also depends on firms having sufficiently usable SEC/XBRL information during the relevant period.

The research design does not attempt to estimate the consequences of firm failure, delisting, merger, or other forms of sample attrition outside the constructed panel.

Similarly, a firm-year observation may remain in the general panel while being unavailable for a particular proxy.

Thus, three levels should be distinguished:

$$\begin{aligned}\text{Predetermined Firm} &\rightarrow \text{Firm-Year Observation} \\ &\rightarrow \text{Model-Eligible Observation.}\end{aligned}$$

A firm can belong to the predetermined sample even when one of its firm-year observations lacks a particular accounting fact.

A firm-year can belong to the panel even when it cannot enter a specific regression.

This hierarchical distinction prevents model-specific missingness from being confused with sample exclusion.

E.14 Reproducibility of the Sample

The predetermined ticker list is encoded directly in the data-acquisition program.

Conceptually, the program stores the sample as

```
MAIN_SAMPLE = {
 "technology": [...],
 "healthcare": [...],
 "industrials": [...],
 "consumer": [...],
 "utilities": [...]
}
```

This makes the sampling decision explicit and machine-readable.

A replication of the empirical study can therefore begin from the same 50 tickers rather than requiring the researcher to reconstruct the sample from the published regression tables.

The mapping

$$Ticker_i \to CIK_i$$

is subsequently resolved against SEC identifiers during the data-acquisition stage.

E.15 Summary

The final sample design can be summarized as

$$\begin{aligned} \text{Industries} &= 5, \\ \text{Firms per industry} &= 10, \\ \text{Total firms} &= 50, \\ \text{Analysis years} &= 5, \\ \text{Maximum firm-years} &= 250. \end{aligned}$$

The sample is predetermined and balanced by industry at the design stage.

The effective sample of an individual empirical model can nevertheless be smaller because standardized accounting information is not uniformly available.

The methodological principle is therefore:

Preserve the research sample; report the missingness; do not manufacture completeness.

This distinction is essential to the interpretation of the empirical results reported in the book.