SELECTED POPULAR QUESTIONS (REPEATEDLY ASKED QUESTIONS):

MISSING VALUE QUESTIONS:

In Compustat, many widely used items (e.g., R&D/XRD, CAPEX/CAPX, PPE/PPEGT, advertising, receivables, special items, deferred taxes, intangibles) are often blank/missing for some firm-years—how should researchers interpret these blanks: as non-disclosure/not reported/not applicable, or as zero/immaterial? Are there specific items where S&P recommends treating missing as zero, and is there a reliable way (flags/metadata) to distinguish a true reported zero from a missing field? Finally, what is S&P's recommended best-practice approach for handling these missing values in panel research (vs. zero-filling or other imputation)?

- Siyuan Fang @ Georgia Institute of Technology
  - Many firms have missing XRD values in Compustat.
    Are these missing values intended to reflect non-disclosure rather than zero R&D, and how should they be handled in research?
- Chaehwan Lim @Michigan State University
  - What is appropriate ways to impute missing values in panel data (replacing NaN with zero, mean, etc) ?
- Shibao Liu @ University of Illinois Chicago
  - How to deal with missing values for variables such as R&D(xrd), PPE(ppegt), CAPEX(capx), etc.?
- LukasLensch @ Georg-August-University Göttingen
  - I am especially interested in the following question: When a field is blank, is it really zero? Interpreting missing values for R&D, special items, deferred taxes, intangibles, and other frequently misused items.
- Dmitri Byzalov @ Temple U
  - Missing values for R&D, advertising, receivables, etc -- for which items can we treat them as immaterial/zero?

GENERAL ANSWERS

Compustat Guidance for Interpreting Blank or Missing Compustat Items

1. By default, do not interpret missing values as zero.

   A blank Compustat field typically indicates that a reliable numeric value was not available for that firm-year item. It does not generally signify zero or immateriality.

2. Most blank fields reflect the absence of a collectible or reliable value, rather than an economic absence.

   Missing values often occur because the item is not disclosed, not collected for a specific context or industry, not separately reported, or not reliably extractable from statements. For example, R&D (XRD) is frequently unavailable for banks and utilities in Compustat, even though these firms may still engage in innovation. If Compustat determines that a value is truly zero, or if there is an explicit rule to record zero, it will typically store a numeric 0 instead of leaving the field blank.

3. Zero values may be explicitly recorded for certain items, but this is not a universal rule.

   In some cases, Compustat records a zero by rule, such as when an item is deemed insignificant for certain fields. However, this policy does not apply universally and should not be assumed for items like XRD, CAPX, PPEGT, or advertising.

4. Use data codes or metadata when available. Codes 4 and 8 are particularly informative.

   For many items, Compustat provides an accompanying data code (such as *_DC) that indicates why the numeric field is absent. Two particularly useful values are:

   - DC = 4 (“combined figure”): the value is rolled into another line item; do not treat as zero.
   - DC = 8 (“insignificant figure”): the item is immaterial. Depending on research design, treating it as zero may be appropriate, but this approach should be disclosed and tested for robustness.
   - In practice, *_DC fields are often sparsely populated and account for only a small portion of missing values. For example, in a WRDS sample covering Compustat data from approximately 2000 to 2025, most missing observations are not assigned an informative code, even when *_DC exists. Only about 1.2% of missing XRD observations are coded as “insignificant” (DC=8). For XAD, about 1.2% are coded as “combined” (DC=4) and about 1.0% as “insignificant” (DC=8).

WRDS Recommended Best Practices for Handling Missing Data in Panel Research

- Do not automatically fill all missing values with zero.
- When possible, classify missing values by type, such as numeric zero, DC=4, DC=8, or unclassified missing.
- If you choose to treat certain missing values as zero, do so explicitly and transparently. Include robustness checks, such as comparing results with and without zero-filling, or adding a missing-indicator variable to distinguish between non-disclosure and true zero values.

CONSOLIDATION QUESTIONS

In Compustat, why can the same firm appear in one product but not another (U.S./North America vs Global), and why do variable availability and even the “same” variable definitions differ across these formats? Relatedly, how should researchers interpret multiple entries for the same firm in comp.funda (e.g., multiple statements/standards, consolidation levels), and what is the recommended best practice for handling parent–subsidiary structures and headquarters vs *subsidiaries—especially for cross-listed firms/ADRs and for U.S. vs Canada vs Global coverage (consolidated vs standalone reporting)?*

- Rong Zeng @ University of Manitoba
  - Some US firms are excluded in one but included in another (Global vs. US). Why?
    The variable lists in US and Global differ to some extent. How about the definitions of the same variables?
    How to deal with Headquarters and subsidiaries in Compustat?
- Steven Ho@University of Nevada
  - why do some firms have more than one entry in comp.funda, is it because the accounting standard used is different 2. how does compustat handle companies traded as ADRs?
- Raj Mohan@ Indian Institute of Management Indore
  - My question is related to the third point in the topic to be covered -
    How “standard” is standardized?
    What researchers should know about U.S. vs. Global formats?
    US firms consolidated statement, while rest of the words its standalone statement?

GENERAL ANSWERS

Why does the same firm appear in one product but not the other (North America vs Global)?

The two products are designed to cover different populations.

- Compustat North America covers the U.S. and Canada, focusing on major exchanges and firms meeting specific criteria. It also includes a limited number of international companies through ADRs.
- Compustat Global provides fundamental data for companies outside North America, using separate internal and third-party data sources.

ADRs are a primary reason for overlap or gaps between the two products. North America coverage explicitly includes international companies listed as Level II or III ADRs.

As a result, a single economic firm might appear:

- in North America as an ADR representation, and/or
- in Global as the native/local filer.

Why does variable availability, including the same variable, differ between U.S. and Global formats?

Even when mnemonics appear identical, availability and meaning may differ due to variations in data organization and delivery rules:

- The products use different data format families and restatement handling rules.

Quarterly restatements are handled differently in each product:

- North America: if a restatement exists, the restated quarterly income statement overwrites the prior values (originals are available via Snapshot).
- International/Global: Compustat retains both originally reported quarterly data (HIST_STD) and restated quarterly data (RST_STD).
- SUMM_STD is a condensed North America format that includes fewer items than STD.

This difference alone can affect variable availability across products. There are also differences in industry format coverage.

- Compustat supports multiple industry formats. INDL serves as the universal baseline, while FS is annual-only, limited to certain index constituents, and supplements rather than replaces INDL.
- Selecting an inappropriate industry format may result in missing data items.

Standardization aims for comparability, not identical accounting. Standardization is intended to support comparability across firms and time, aligning with U.S. GAAP and IFRS. However, it may require judgment, and some metrics are intentionally non-GAAP.

In summary, standardization ensures comparability but does not guarantee that all country-specific or accounting nuances are perfectly aligned.

Why are there multiple entries for the same firm in funda, and how should they be interpreted?

This is a key point regarding data mechanics:

In Xpressfeed-style Compustat, a firm-year may have multiple records because records are keyed by both firm/date and format dimensions. To avoid double-counting, it is often necessary to combine the key filtering fields (GVKEY, DATADATE, POPSRC, INDFMT, DATAFMT, CONSOL, and sometimes FYR). The key dimensions are defined as follows:

- POPSRC: Indicates domestic versus international, and distinguishes ADR from native presentations.
- DATAFMT: Specifies the data format, such as STD, HIST_STD, or RST_STD.
- INDFMT: Identifies the industry format, such as INDL, FS, or others.
- CONSOL: Indicates the consolidation level, such as consolidated or non-consolidated.

Therefore, multiple entries usually reflect different format or consolidation versions, not multiple accounting standards.

Recommended best practices for consolidation, parent/subsidiary relationships, and ADRs in research

Best Practice A: Select a single research keyset and filter your data accordingly.

- A common WRDS research convention is to restrict analysis to one standardized consolidated record, for example:
- North America default: indfmt='INDL', datafmt='STD', popsrc='D', consol='C'
- This approach is often referenced when addressing duplicate records.
- Global default (common starting point): indfmt='INDL', datafmt='HIST_STD' (or RST_STD for restated quarterlies), popsrc='I', consol='C'
- If you are combining North America and Global data, use caution. The structures differ and should not be merged without careful harmonization.

Best Practice B**:** Do not mix ADR-based and native-based representations unless it is intentional.

If a firm is cross-listed or has an ADR, it may appear in both representations:

- North America ADR representation, and Global native filer representation.

Use POPSRC and your sample design to select one representation and avoid double-counting.

Best Practice C**:** Parent versus subsidiary, or headquarters versus subsidiaries

- For most firm-level panel research, the default is to use consolidated financials (consol='C') at the main company level.
- Subsidiary-level standalone statements may appear as non-consolidated (consol='N') or as distinct entities, depending on coverage. Include these only if your research specifically focuses on subsidiaries.