The FIVES Cattell R&D Data

by

Ivan Png

Distinguished Professor
NUS Business School
National University of Singapore

Conditions of Use

The data sets described in this document are made available through the FIVES Project on Firm and Industry Evolution, Entrepreneurship, and Strategy. If you download any of the data sets described in this document, or obtain copies of these data sets that someone else downloaded from FIVES, or obtain modified versions of these FIVES data sets, you agree to abide by the conditions of use set forth here.

The current organizer of the FIVES Project, Constance Helfat (Tuck School of Business at Dartmouth), and the former co-organizer Steven Klepper (deceased, formerly of Carnegie Mellon University), are not responsible for the content of any FIVES Project data sets. The FIVES Project organizers and the FIVES Project data set authors make no representations or warranties of any kind concerning any of the FIVES Project data, including with regard to the absence or presence of errors.

All FIVES Project data sets are protected by copyright. The copyright holders for each data set are indicated on the first page of the documentation for each data set. Permission is granted to reproduce FIVES Project data for non-profit educational and research use only. Requests to reproduce FIVES Project data for other uses should be addressed to: Professor Constance Helfat, Tuck School of Business at Dartmouth, Hanover, NH 03755, constance.helfat@dartmouth.edu.

This document describes the FIVES Cattell R&D Data, which should be referred to by this name in any derivative works. In any written and published work, users of these data should cite this document, the article by the data set contributor that describes the data (Png, Ivan. 2019. U.S. R&D, 1975-1998: A New Data Set, *Strategic Management Journal*), and the FIVE Project: Data Overview (Helfat & Klepper, 2007).

You may create a new data set by modifying a FIVES Project data set, including but not limited to adding or removing data or merging data from different FIVES Project data sets. If you utilize any FIVES Project data to create new data sets, within six years from the date that you obtain the FIVES Project data, you must provide the FIVES Project with a copy of your modified data set by contacting: Constance Helfat, constance.helfat@dartmouth.edu.

Any written documents or statistical estimates that use FIVES Project data, including but not limited to working papers and publications, must be reported to: Constance Helfat, constance.helfat@dartmouth.edu.

FIVES Cattell R&D Data: File List

The FIVES Cattell R&D Data includes digitized Cattell data files, replication programs and data for the *Strategic Management Journal* article by Ivan Png, and the raw data.

A. Data files

1. File: pub-cattell.dta
This is the main Cattell dataset and includes the following variables:

Variable	Description
year	
newid	Parent ID: Created in this project to uniquely identify the organization across years. Note: If company changes name, then it would be assigned another newid.
id	Original parent ID in the Cattell directories: Assigned sequentially in each volume, so, not consistent across years.
cmpy	Name of parent
cmpy1, cmpy2, cmpy3, cmpy4	Name of parent in various alternatives
facy_newid	Facility ID: Created in this project to uniquely identify the facility across years.
facility	Original facility number in the Cattell directories: Assigned sequentially in each volume, so, not consistent across years.
facility_name	Facility name
state	FIPS code for state of facility. FIPS refers to Federal Information Processing Standard Publication 6-4, which provides numeric codes for U.S. states and counties.
zipcode	Zip code of facility from Cattell directories.
level	Level of facility relative to headquarters: 0 = headquarters. Facilities are listed according to hierarchy, so, levels correspond to reporting structure. Missing for about 0.5% of observations.
user	Users served by facility: Years up to 1981: p = R&D for parent organization only, f = Contract R&D for others, c = Consultation for others, t = Testing and analysis for others; Years from 1983: p = R&D for parent organization, g = On contract for government, i = On contract for industry, c = Consultation for others.
prof	Number of professionals in facility. Missing if no information on professionals or doctorates at the facility. If the Cattell directories did not report the number of professionals but did report the number of doctorates, then the number of professionals was stipulated to be the number of doctorates.
doct	Number of professionals with doctorate degree in facility. Missing if no information on professionals or doctorates at the facility.

	If the Cattell directories did not report the number of doctorates but did report the number of professionals, then the number of doctorates was stipulated to be zero.
tech	Number of technicians in facility. Missing if no information on professionals or doctorates at the facility. If the Cattell directories did not report the number of technicians but did report either the number of professionals or the number of doctorates, then the number of technicians was stipulated to be zero.

2. cpstat-match.dta

Match between Cattell and Compustat by name of parent. If required, use id and year to merge (many to one) the Cattell with Compustat data. The Cattell data are organized by company, facility, and year, while the Compustat data are by company and year. Hence the merge is many to one.

Variable	Description
year	
newid	Cattell parent ID
cmpy	Cattell parent name
id	Parent ID in the Cattell directories
gvkey	Compustat company ID
cpstat	Compustat company name
zipbk	Zipcode (Cattell)
zipcp	Zipcode (Compustat)
state	State of headquarters (from Cattell)
score	Reclink score on match of Cattell and Compustat names

3. field-master.dta

List of R&D fields used by Cattell.

Variable	Description
field	Abbreviation of field
field_name	Field (Cattell)
subfield	Abbreviation of sub-field: The abbreviations are not a complete sequence; after checking, some sub-fields were deleted, eg, CHEM-138 was merged into CHEM-135.
subfield_name	Sub-field (Cattell): Specified in 1975 and 1982 and later, but not 1979.
hitech	Indicator of high technology field, roughly based on D. Hecker, "'High-Technology Employment: A Broader View", Monthly Labor Review, 1999.

4. field.dta

R&D fields by facility. If required, use year, Cattell parent company (id), and facility identifier (facility) to merge (one to many) the Cattell data with the field data. Note that Cattell parent and facility identifiers vary by year. Some facilities carry out R&D in multiple fields, and so, the merge is one to many.

Variable	Description
year	
id	Parent ID in the Cattell directories
facility	Facility number in the Cattell directories
field	Abbreviation of field
subfield	Abbreviation of sub-field 1975: This volume lists only subfields (Cattell grouped these into major fields only from 1982 onward). We drop all subfields that are duplicated across fields, e.g., "Acoustics and Acoustic Equipment", then merge the 1975 raw data with the field-master.dta. We then classify the major field of the organization by the most frequent major field among the unique subfields. 1979: This volume lists only major fields, without subfields.

5. state.dta

State abbreviations and FIPS codes.

B. Replication programs and data (*Strategic Management Journal* paper by Png)

1. figtab.do: Stata code to replicate the figures and tables other than maps in main paper.

2. map: Stata dta and shape files to replicate the maps in Figures 3 and 4.

3. patent: Stata dataset of patents for analysis in Table 4.

4. figtab-balance.do: Stata code to replicate test of stability to omitted directories reported in Appendix.

C. Raw data

1. scans.zip: Zip file of scanned directories, including editions not yet digitized. For some years, the file contains only the alphabetical listings and excludes the front matter and R&D fields.

2. OCR. OCR output from scans of Cattell directories.