Package: CATAcode (via r-universe)

June 4, 2026

Title Explore and Code Responses to Check-All-that-Apply Survey Items

Version 1.0.0

Description Analyzing responses to check-all-that-apply survey items often requires data transformations and subjective decisions for combining categories. 'CATAcode' contains tools for exploring response patterns, facilitating data transformations, applying a set of decision rules for coding responses, and summarizing response frequencies.

License GPL (>= 3)

URL https://github.com/knickodem/CATAcode, https://knickodem.github.io/CATAcode/

BugReports https://github.com/knickodem/CATAcode/issues

Depends R (>= 3.6)

Imports rlang, dplyr (>= 1.1.0), tidyr, ggplot2

Suggests knitr, rmarkdown, testthat (>= 3.0.0)

Encoding UTF-8

LazyData true

Roxygen list(markdown = TRUE)

RoxygenNote 7.2.3

Config/testthat/edition 3

VignetteBuilder knitr

Config/pak/sysreqs libicu-dev

Repository https://knickodem.r-universe.dev

Date/Publication 2025-10-02 23:16:09 UTC

RemoteUrl https://github.com/knickodem/catacode

RemoteRef HEAD

RemoteSha c1f7ae854ec936b3e264b5b4b45f2f4dc130151f

Contents

`cata_code` *Code check-all-that-apply responses into a single variable*

Description

In a cross-sectional or longitudinal context, select a set of decision rules to combine responses to multiple categories from a check-all-that-apply survey question into a single variable.

Usage

```
cata_code(
  data,
  id,
  categ,
  resp,
  approach,
  endorse = 1,
  time = NULL,
  priority = NULL,
  new.name = "Variable",
  multi.name = "Multiple",
  sep = "-"
)
```

Arguments

`data`	A data frame with one row for each `id` (by `time`, if specified) by category combination. If `data` are currently in "wide" format where each response category is its own column, use `cata_prep()` first to transform `data`into the proper format. See *Examples*.
`id`	The column in `data` to uniquely identify each participant.
`categ`	Unquoted column in `data` indicating the check-all-that apply category labels.
`resp`	Unquoted column in `data` indicating the check-all-that apply responses.
`approach`	One of "all", "counts", "multiple", "priority", or "mode". See *Details*.
`endorse`	The value in `resp` indicating endorsement of the category in `categ`. This must be the same for all categories. Common values are 1 (default), "yes", TRUE, or 2 (for SPSS data).

time	The column in data for the time variable; used to reshape longitudinal data with multiple observations for each id.
priority	Character vector of one or more categories in the categ column indicating the order to prioritize response categories when approach is "priority" or "mode".
new.name	Character; column name for the created variable.
multi.name	Character; value given to participants with multiple category endorsements when approach %in% c("multiple", "priority", "mode").
sep	Character; separator to use between values when approach = "all".

Details

For all approach options, participants with missing data for all categories in categ are removed and not present in the output.

There are two options for approach that provide summary information rather than a single code for each id.

"all" returns a data frame with new.name variable comprised of all categories endorsed by separated by sep. The time argument is ignored when approach = "all". Rather, if data includes a column for time, then output includes a row for each id at each time point. This approach is a useful exploratory first step for identifying all of the response patterns present in the data.

"counts" is only relevant for longitudinal data and returns a data frame with the number of times an id endorsed a category. Only categories with >= 1 endorsement are included for a particular id. As with "all", the time argument is ignored and instead assumes data is in longer format with a row for each id by time combination. If not, the column of counts will be 1 for all rows.

The three remaining options for approach produce a single code for each id. The output is a data frame with one row for each id. The choice of approach is only relevant for participants who selected more than one category whereas participants who only selected one category will be given that code in the output regardless of which approach is chosen.

"multiple" If participant endorsed multiple categories within or across time, code as multi.name.

"priority" Same as "multiple" unless participant endorsed category in priority argument at any point. If so, then code in order specified in priority.

"mode" Participant is coded as the category with the mode (i.e., most common) endorsement across all time points. Ties are coded as as the value given in multi.name. If the priority argument is specified, these categories are prioritized first, followed by the mode response. The "mode" approach is only relevant if time is specified. When time = NULL it operates as "priority" (when specified) or "multiple".

Value

data.frame

Examples

```
# prepare data
data(sources_race)
sources_long <- cata_prep(data = sources_race, id = ID, cols = Black:White, time = Wave)
```

```
# Identify all unique response patterns
all <- cata_code(sources_long, id = ID, categ = Category, resp = Response,
approach = "all", time = Wave, new.name = "Race_Ethnicity")
unique(all$Race_Ethnicity)


# Coding endorsement of multiple categories as "Multiple
multiple <- cata_code(sources_long, id = ID, categ = Category, resp = Response,
approach = "multiple", time = Wave, new.name = "Race_Ethnicity")

# Prioritizing "Native_American" and "Pacific_Islander" endorsements
# If participant endorsed both, they are coded as "Native_American" because it is listed first
# in the priority argument.
priority <- cata_code(sources_long, id = ID, categ = Category, resp = Response,
approach = "priority", time = Wave, new.name = "Race_Ethnicity",
priority = c("Native_American", "Pacific_Islander"))

# Code as category with the most endorsements. In the case of ties, code as "Multiple"
mode <- cata_code(sources_long, id = ID, categ = Category, resp = Response,
approach = "mode", time = Wave, new.name = "Race_Ethnicity")

# Compare frequencies across coding schemes
table(multiple$Race_Ethnicity)
table(priority$Race_Ethnicity)
table(mode$Race_Ethnicity)
```

---

cata_prep *Prepare data for* cata_code()

Description

A helper function to transform data into a longer format in preparation for use in cata_code().

Usage

```
cata_prep(
  data,
  id,
  cols,
  time = NULL,
  names_to = "Category",
  values_to = "Response",
  ...
)
```

Arguments

data	A data frame where rows are participants or participant by time combinations if time is specified.
id	The column in data to uniquely identify each participant.
cols	<tidy-select> The columns in data indicating the check-all-that-apply categories to combine. Endorsement of the category should be indicated by the same value (e.g., 1, "Yes") across all columns included here. Columns are typically dichotomous variables with the two values indicating endorsement or not, but this is not a requirement.
time	The column in data for the time variable; used to reshape longitudinal data with multiple observations for each id.
names_to	Character. The name for the new column of category labels (i.e., names of the cols columns), which is passed to pivot_longer().
values_to	Character. The name for the new column of responses (i.e., cell values in the cols columns), which is passed to pivot_longer().
...	Optional additional arguments passed to pivot_longer().

Value

An object of the same type as data with one row for each id (by time, if specified) by response category combination.

Examples

```
data(sources_race)
cata_prep(data = sources_race, id = ID, cols = Black:White, time = Wave)
```

CATAcode — *CATAcode: Explore and Code Responses to Check-All-that-Apply Survey Items*

Description

Analyzing responses to check-all-that-apply survey items often requires data transformations and subjective decisions for combining categories. CATAcode contains tools for exploring response patterns, facilitating data transformations, applying a set of decision rules for coding responses, and summarizing response frequencies.

Author(s)

Maintainer: Kyle Nickodem <kyle.nickodem@gmail.com> (ORCID)

Authors:

- Gabriel J. Merrin <gjmerrin@syr.edu>

See Also

Useful links:

- https://github.com/knickodem/CATAcode
- https://knickodem.github.io/CATAcode/
- Report bugs at https://github.com/knickodem/CATAcode/issues

---

`sources_race` *Sources of Strength Race/Ethnicity Data*

Description

Responses to the check-all-that-apply race/ethnicity question at four time points from a randomized controlled trial of the Sources of Strength program.

Usage

```
sources_race
```

Format

A data frame with 16,922 rows and 9 columns:

ID Subject identification number

Wave Data collection time point

Black, Native_American, Asian, Hispanic, Multiracial, Pacific_Islander, White Indicator variables for check-all-that-apply responses where 1 = endorsement

Source

doi:10.15139/S3/EZ8ILC

Index