The Stata Journal (2022)
22, Number 4, pp. 958–968 DOI: 10.1177/1536867X221141022

printcase: A command for visualizing single observations

Max D. Weinreb
University of Texas at Austin
Austin, TX
maxweinreb@utexas.edu

Jenny Trinitapoli
Sociology Department
University of Chicago
Chicago, IL
jennytrini@uchicago.edu

Abstract. In this article, we introduce the `printcase` command, which outputs data from a specific observation into an easy-to-read Microsoft Word or PDF document. `printcase` allows analysts to focus on a single observation within a dataset and view that observation in its entirety. The output displays fields in table format, with all variables identified by their corresponding labels and all responses identified by their corresponding value labels. We explain how `printcase` works, give examples of circumstances under which this type of table-based quasiquestionnaire would be useful, and provide code for printing single observations.

Keywords: dm0109, printcase, survey research, fieldwork, data quality, interviewer training

1 Introduction

`printcase` is a command that analysts can use to generate a table of variables and values from any dataset. Sometimes, examining a single observation can improve comprehension and generate new insights. But in the era of etablet modes of data collection, producing a document that resembles a completed questionnaire is surprisingly difficult. `printcase` addresses this need by providing researchers with an abbreviated quasiquestionnaire generated from responses for a single observation (that is, row) in a data file. When survey items (variables and value labels) are fully labeled, the printed case can proxy a completed questionnaire much like what we used in the days of pencil-and-paper surveys. When leveraged with longitudinal (or otherwise nested) data, `printcase` is additionally useful for analysts who would like to view values over time to understand a trajectory or prepare for a subsequent interview.

We can think of at least four reasons why researchers would want to skim or study responses from a single observation.

First, examining individual questionnaires in their entirety is useful for data cleaning and making judgment calls about unusual values. One of the ways this is done is through examining other responses in the questionnaire to aid in data cleaning to make sure that the answers are inherently consistent. Researchers disagree about whether and how to edit data (Sana and Weinreb 2008), but some argue that leveraging information provided by respondents themselves is superior to even the most sophisticated approaches to imputation (Leahey 2008; Leahey, Entwisle, and Einaudi 2003; de Waal, Pannekoek, and Scholtus 2011). By looking at the complete answers, we can see more clearly how

to recode an outlier. An example that comes from our own data-collection effort in Balaka, Malawi, is that of a woman who said she had never had sex and was "not at all worried about HIV" but reported that she had been tested for HIV six times in the past month. The value was unusual and seemed to be a mistake. However, upon closer examination of the questionnaire, we learned that this woman was part of a peer-to-peer counseling group, in which she would encourage friends to get tested, accompany them to the testing facility, and go through the entire process with them as part of a district-wide effort to increase voluntary testing. This shows us how reading a particular observation vertically can generate insights that are concealed when we examine data only using measures of central tendency. When the underlying dataset has been deidentified, researchers can produce a case for study and contemplation, and that case will also be fully anonymized.

Another reason a researcher would want to study all the answers from one observation is to look at numeric responses in conjunction with the interviewers' notes or other open-ended responses, which may have been collected during interviews or after, to help analysts geographically and temporally separated from the interview understand other aspects of the moment or interaction that informed the answers gathered. Reflecting on more than a decade of fieldwork and data analysis, Bledsoe, Banja, and Hill (1998) remark that their analytical efforts were enhanced when they could juxtapose "open-ended commentary as variables alongside [. . .] quantifiable responses." Sometimes, these open-ended responses provide information that can subsequently be coded up into close-ended responses or may, in fact, require changing a response value. Take, for example, a survey question that asked respondents how many hours of television they watched per day in the past week. The responses range between 0 and 11, with an average of 2.2. The 95th percentile is 5, and analysts are left with questions about how to manage the values of 6 and over. By examining the particular observation, the analyst may identify an erroneous response (the "fat fingers" problem) in which the 11 should have been a 1, for example, "the 8 o'clock news every night", from a true value of 11, for example, "Since accident, respondent is bedridden and watches all day." Given a well-labeled dataset, `printcase` provides similar functionality, in which analysts can leverage interviewer notes, write-in responses, and other qualitative descriptions as complements to the quantitative data.

Studying responses from particular individuals is critical for data-collection efforts that engage the same respondent multiple times because it allows the interviewer to review what the respondent shared previously before reengaging with a respondent. An example of this can be found in Pearce's (2002) approach to systematic anomalous case analysis. In this method, analysts analyze data using traditional regression-based methods, identifying patterns and selecting cases that deviate from the trends for in-depth follow-up research, especially in-depth interviews and ethnography. A careful read of the completed questionnaire is an essential step for preparing to conduct a valuable follow-up interview with the same individual. In the absence of a paper questionnaire to consult, `printcase` can be used to generate a file that sketches the earlier conversation between interviewer and respondent. This would serve as the basis upon which new questions for a follow-up conversation could be generated.

Finally, paper questionnaires are valuable for training interviewers and enumerators. Most studies still use a blank paper version of their questionnaire for interviewer training, emphasizing the scripts that structure transitions between modules and the introductions that cue particular questions and clarify whether responses should be read. Paper versions are easier to browse and skim as a full document, rather than item by item. This is important for teaching skip patterns and familiarizing interviewers with the overarching goals of the particular study they are fielding.

`printcase` cannot replace the designed questionnaire, but it can be used to quickly produce a set of responses—actual or theoretical (that is, from synthetic data)—that interviewers can study as part of interviewer training. In our experience, it is particularly valuable to have interviewers study a completed questionnaire collected while piloting the instrument; this exercise helps prepare interviewers for the kinds of responses they might encounter in the field, and it also helps train them to think about the internal consistency of a narrative during the interview.

Fieldwork supervisors responsible for ensuring data quality may also want to browse printed cases to check the quality of interviewers' work and provide additional support and training where necessary. For example, if one interviewer is entering more "refused to answer" responses than others, he or she may need to introduce a particular topic with more sensitivity or learn how to probe more effectively. By browsing printed cases with a focus on the interviewer's work, field supervisors can catch and remedy interviewer-specific errors before they are manifested too deeply in the entire data-collection enterprise.

2 The printcase command

`printcase` is meant to be used with a dataset that has been adequately labeled with variable names and corresponding variable labels. In cross-sectional datasets, the focal case is called by a unique ID (string or numeric), and the `printcase` output is a table three columns wide that displays 1) the variable name, 2) the variable label, and 3) the response value. The first row contains the column labels, and one row is generated for every variable in the dataset unless exclusions are specified (see below). In longitudinal datasets, additional columns are printed for each observation attached to the focal case. When longitudinal cases are printed, the columns will be printed in the order they appear in the dataset, so the analyst should sort the data before calling `printcase`.

The commands on which `printcase` builds are `putdocx`, `putpdf`, and `levelsof`. The `levelsof` command allows for efficient execution by facilitating quick lookups for matches between variables, given the condition that an ID match (that is, the focal observation) is found. Instead of manually searching the dataset, `levelsof` conveniently returns the result in the `r(levels)` list, which subsequently gets searched. To generate a visually appealing and easy-to-read observation in Word or PDF format, `printcase` draws extensively on `putdocx` and `putpdf` to build the output table and customize the display options. Both commands allow analysts to suppress rows and control column widths, making `printcase` output readable and compact for browsing.

2.1 Syntax

The syntax of `printcase` is

`printcase` [`using` *filename*] `if` *id_variable* `==` *value* [`, pdf font(`*string*`)
`noempty ignore("`*string1*`"` [`"`*string2*`" ...`]`) replace addnotes`
`width(#`[*unit* | `%`] | *matname*`) unit("`*string*`") landscape noheader nofooter`]

`using` *filename* specifies an alternative filename and location.

id_variable is the name of the identifying variable in the dataset and *value* is the value of *id_variable* to print. The *id_variable* can be string or numeric. If *id_variable* is unique within the dataset, one case is printed; if it is nonunique, all rows associated with the target ID are printed in columns in the order they appear in the dataset. If no `using` file is specified, `printcase` saves the file to the home directory and names the file by concatenating the target ID variable name and its value. If a string *id_variable* contains characters that violate standard file-naming conventions, we recommend specifying a filename by leveraging the `using` option.

2.2 Options

`pdf` sets the output of `printcase` to be a PDF file instead of the default Microsoft Word file (`.docx`). Page numbers and footers are not generated in PDF files, whereas they are in Microsoft Word. All other options are unaffected by specifying a PDF file as the output.

`font(`*string*`)` sets the font used for the entire document of the output of `printcase`. Any installed font can be specified. The default is `font(Arial)`.

`noempty` suppresses all empty and system missing responses and their variables from the resulting table if the value label is an empty string (that is, `" "`) or a Stata missing-value code (`"."`, `".d"`, etc.). By default, all empty responses will be included. In longitudinal cases, variable rows are suppressed only if variables are empty for all observations.

`ignore("`*string1*`"` [`"`*string2*`" ...`]`)` allows users to specify that variables be ignored based on the values, for example, missing strings (`" "`), general missing (`"."`), or skipped values (`".s"`) if the dataset distinguishes. (Other datasets use `"99"` or other codes.) The result in the output document will not include those variables or their responses. There is no limit on the number of responses to ignore. In longitudinal cases, variable rows are suppressed only if variables match an ignorable value for all observations.

`replace` overwrites an existing printed case.

`addnotes` includes the first note on any variable in the variable label column. (If the dataset is documented to such a degree, this may help the analyst discern skip

patterns or heed cautions about variables with known problems.) Although multiple notes can be attached to a variable, only the first note is included in the printed case.

`width(`#[*unit* | `%`] | *matname*`)` specifies the width of the columns to be printed with standard Stata syntax. See [RPT] **putdocx table** for the specifics of formatting input.

`unit("`*string*`")` changes the default title for the response columns from `"Response"` to the specified string, such as `"Wave"` or `"Year"`.

`landscape` changes the paper orientation from portrait (the default) to landscape.

`noheader` suppresses the default header in the output document.

`nofooter` suppresses the default footer in a Word document output.

3 Examples of printcase

In this example, we first load the model births recode dataset from the Demographic and Health Survey (DHS) (ICF n.d.). Following the DHS documentation, we generate an identifying string variable by concatenating the three constituent pieces, separated by an underscore.

```
. egen id = concat(v001 v002 v003), punct(_)
```

Because this dataset (like many others) does not contain any notes, we add two notes to demonstrate the value of the `addnotes` option.

```
. notes v007: Check if this is Ethiopian, Nepali, or Gregorian calendar.
. notes v130: Religious groups are ambiguous in the model dataset.
```

3.1 Example 1

In example 1, we call `printcase` using the common `if` arguments to specify the ID variable and rely on the default filename saved in the home directory. The options specified here will generate a PDF output format (rather than Microsoft Word document) using Arial font.

```
. printcase if id == "1_13_3", pdf
ID Variable: id
ID Value: 1_13_3
successfully created "/Users/id1_13_3.pdf"
```

The document produced in example 1 is 28 pages long; it contains many cells that indicate missing data.

3.2 Example 2

In example 2a, we likewise produce a Microsoft Word document, designating a descriptive filename with `using` and suppressing empty variables from the output document using the `noempty` and `ignore()` options. The `addnotes` option embeds any variable notes in the dataset within the second column. Given the length of the full DHS dataset, this shortens the printed case from 28 to 10 pages.

```
. printcase using example2 if id == "1_13_3", noempty ignore(".") addnotes
ID Variable: id
ID Value: 1_13_3
successfully created "/Users/example2.docx"
```

In example 2b, we further improve legibility of the output by specifying a serif font (`Georgia`) and controlling the column width to reduce white space and generate an even more compact case for browsing. The `replace` option allows the user to write over the existing case for this observation. The full table (that is, the printed case) is produced by the example do-file. Figure 1 shows an extract from the output produced by this example.

```
. printcase using example2 if id == "1_13_3", font(Georgia) width(25, 50, 25)
> noempty addnotes ignore(".") replace
ID Variable: id
ID Value: 1_13_3
successfully replaced "/Users/example2.docx"
```

id 1_13_3

User: Redacted

ZZBR62FL.DTA

Date Printed: 6 Jun 2022

Variable Name	Variable Label	Response 1
caseid	case identification	1 13 3
bidx	birth column number	1
v000	country code and phase	ZZ6
v001	cluster number	1
v002	household number	13
v003	respondent's line number	3
v004	ultimate area unit	1
v005	women's individual sample weight (6 decimals)	1057703
v006	month of interview	6
v007	year of interview N: Check if this is Ethiopian, Nepali, or Gregorian calendar.	2015
v008	date of interview (cmc)	1386
v009	respondent's month of birth	4
v010	respondent's year of birth	1990
v011	date of birth (cmc)	1084
v012	respondent's current age	25
v013	age in 5-year groups	25-29
v014	completeness of age information	month and year - information complete
v015	result of individual interview	completed
v016	day of interview	30
v017	cmc start of calendar	1321
v018	row of month of interview	15
v019	length of calendar	66
v019a	number of calendar columns	2
v020	ever-married sample	all woman sample
v021	primary sampling unit	1
v022	sample strata for sampling errors	26
v023	stratification used in sample design	region 2 - rural
v024	region	region 2

1 printcase_1_13_3_ZZBR62FL.DTA

Figure 1. Screenshot of `printcase` output from example 2b.
NOTE: Figure 1 shows an extract from example 2b, where variable name, labels, and responses are displayed in three columns.

3.3 Example 3

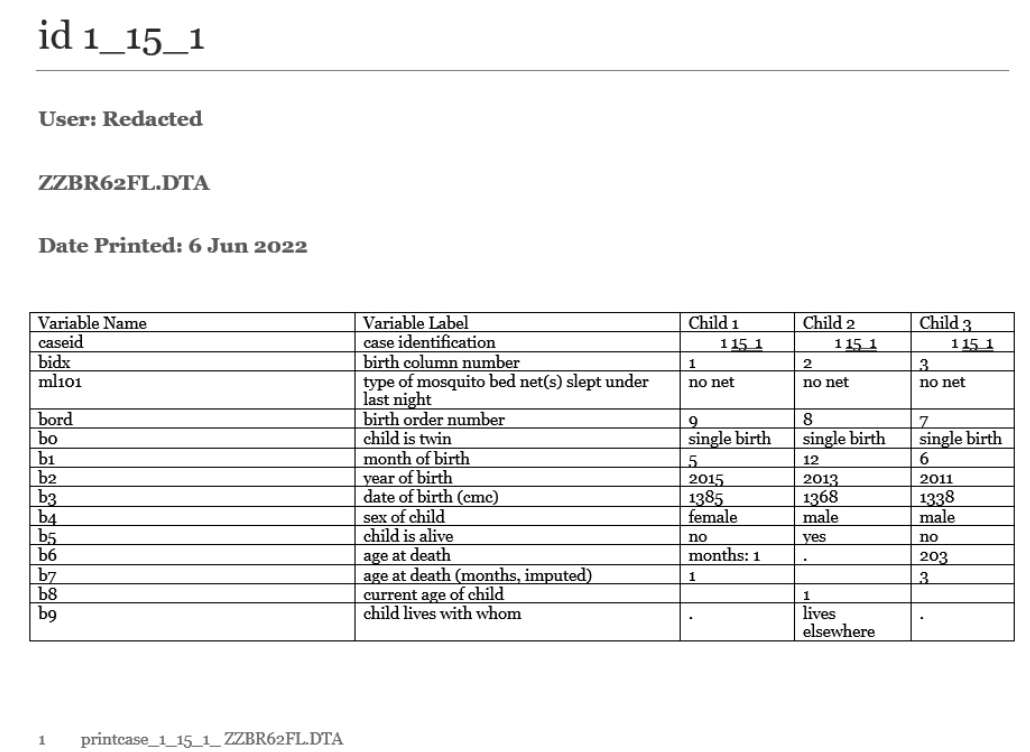
Example 3 shows how we can use `printcase` to view longitudinal or nested data. To illustrate, we use the same dataset, selecting a focal case in which a woman (ID) reports multiple births corresponding with three records in the file. The same logic also applies to longitudinal datasets to which a respondent (ID) contributes multiple observations over time. Here we further prepare the dataset by reducing the number of variables to be examined and deleting rows that indicate children (up to 14) that are not reported in the birth histories (up to 4).

```
. drop aw* v*

. drop if midx == .
(17,698 observations deleted)
```

After prepping the data, we print a case that includes all the corresponding observations in columns and specifies that the units for the columns should be labeled `"Child"`. Because this requires more horizontal space, we use the `landscape` option to rotate the page orientation and the `width()` option to control column width for the three children (observations) reported by this mother (nonunique ID). Figure 2 shows an extract from the output produced by this example, with values for each child appearing in a separate column. The `noempty` and `ignore()` options suppress the row only if the variables are missing across all associated cases.

```
. printcase using example3 if id == "1_15_1", unit("Child") landscape
> width(35, 35, 10, 10, 10) noempty ignore(".") font(Georgia)
ID Variable: id
ID Value: 1_15_1
successfully created "/Users/example3.docx"
```



id 1_15_1

User: Redacted

ZZBR62FL.DTA

Date Printed: 6 Jun 2022

Variable Name	Variable Label	Child 1	Child 2	Child 3
caseid	case identification	1 15 1	1 15 1	1 15 1
bidx	birth column number	1	2	3
ml101	type of mosquito bed net(s) slept under last night	no net	no net	no net
bord	birth order number	9	8	7
b0	child is twin	single birth	single birth	single birth
b1	month of birth	5	12	6
b2	year of birth	2015	2013	2011
b3	date of birth (cmc)	1385	1368	1338
b4	sex of child	female	male	male
b5	child is alive	no	yes	no
b6	age at death	months: 1	.	203
b7	age at death (months, imputed)	1		3
b8	current age of child		1	
b9	child lives with whom	.	lives elsewhere	.

1 printcase_1_15_1_ ZZBR62FL.DTA

Figure 2. Screenshot of `printcase` output from example 3.

NOTE: Figure 2 shows an extract from example 3, where responses from three cases corresponding with a nonunique ID variable are displayed in columns.

4 Conclusions

There are many reasons that reading questionnaires vertically has never caught on as standard practice in survey research. Questionnaires need to be stored carefully and kept confidential. Paper is heavy and difficult to transport. Often, the paper questionnaires have already been destroyed as part of the data security protocol. And the quantitative scholar's goal of making inferences fundamentally rests on our ability to identify statistical regularities—not to overinterpret particular cases. Still, the ability to look closely at a single observation is sometimes valuable. In particular, being able to produce a neat, readable quasiquestionnaire directly from a dataset—without headache—when necessary will enhance the workflow of data collection for many fieldworkers.

5 Acknowledgments

We are grateful to the following people for feedback on the `printcase` command and the text of this article: Kathleen Broussard, Abdallah Chilungo, Ann Moore, Alex Weinreb, and the editors and reviewers from the *Stata Journal*. The Demographic and Health Surveys Program kindly shared DHS model datasets to produce a realistic example.

6 Programs and supplemental materials

To install a snapshot of the corresponding software files as they existed at the time of publication of this article, type

```
. net sj 22-4
. net install dm0109     (to install program files, if available)
. net get dm0109         (to install ancillary files, if available)
```

Alternatively, the package can be downloaded manually from GitHub (https://github.com/theDubW/printcase).

7 References

Bledsoe, C., F. Banja, and A. G. Hill. 1998. Reproductive mishaps and western contraception: An African challenge to fertility theory. *Population and Development Review* 24: 15–57. https://doi.org/10.2307/2808121.

de Waal, T., J. Pannekoek, and S. Scholtus. 2011. *Handbook of Statistical Data Editing and Imputation*. Hoboken, NJ: Wiley.

ICF. n.d. Download model datasets. The DHS Program website. Funded by USAID. Accessed August 31, 2021. https://dhsprogram.com/data/Download-Model-Datasets.cfm.

Leahey, E. 2008. Methodological memes and mores: Toward a sociology of social research. *Annual Review of Sociology* 34: 33–53. https://doi.org/10.1146/annurev.soc.34.040507.134731.

Leahey, E., B. Entwisle, and P. Einaudi. 2003. Diversity in everyday research practice: The case of data editing. *Sociological Methods and Research* 32: 64–89. https://doi.org/10.1177/0049124103253461.

Pearce, L. D. 2002. Integrating survey and ethnographic methods for systematic anomalous case analysis. *Sociological Methodology* 32: 103–132. https://doi.org/10.1111/1467-9531.00113.

Sana, M., and A. A. Weinreb. 2008. Insiders, outsiders, and the editing of inconsistent survey data. *Sociological Methods and Research* 36: 515–541. https://doi.org/10.1177/0049124107313857.

About the authors

Max D. Weinreb is an undergraduate student at the University of Texas at Austin pursuing a BS in computer science with a minor in business. He is the author of the `printcase` command.

Jenny Trinitapoli is an associate professor of sociology at the University of Chicago. Since 2009, she has been principal investigator of the Tsogolo La Thanzi project, spearheading the collection of 12 rounds of data from over 3,000 young adults living in Balaka, Malawi. She envisioned and commissioned the `printcase` command out of necessity, when transitioning from paper-based to etablet data collection with an experienced data-collection team.