printcase: A Stata command for visualizing single observations

Max Weinreb, Austin, TX

Jenny Trinitapoli, Chicago, IL, jennytrini@uchicago.edu

**Abstract.** In this report, we introduce the printcasecommand for outputting 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 all fields associated in table format, with all variables identified by their corresponding labels and all responses their corresponding value labels. We explain how printcase works, give examples of circumstances under which this type of table-based quasi-questionnaire would be useful, and provide Stata code for “printing” single observations.

**Keywords.** survey research, fieldwork, data quality, interviewer training, printcase

# Introduction

printcaseis a Stata command analysts can use to generate a table of variables and values from any .dta file. There are times when examining a single observation can improve comprehension and generate new insights. But in the era of e-tablet 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 quasi-questionnaire generated from responses for a single observation (i.e., 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 particular, 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 go about editing 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; Waal, Pannekoek, and Scholtus 2011). By looking at the complete answers, it can become clear how to recode an outlier. An example 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” 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 only examine data using measures of central tendency. When the underlying dataset has been de-identified, 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 the interaction that informed the answers gathered. Reflecting on more than a decade of fieldwork and data analysis, Bledsoe et al. (1998) remark that their analytical efforts were enhanced when they were able to 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 be able to identify an erroneous response (the “fat fingers” problem) in which the 11 should have been a 1 = “the 8 o’clock news every night,” from a true value of 11, e.g. “Since accident, respondent is bed-ridden & 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 a complement to the quantitative data.

Studying responses from particular individuals is critical for data collection efforts that engage the same respondent multiple times, as 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 (i.e., from synthetic data) – 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, they 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 manifest too deeply in the entire data-collection enterprise.

# 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, which displays: 1) variable name, 2) variable label, and 3) 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 printing longitudinal cases, 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 printcasebuilds include: 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 (i.e., 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 programs allow analysts to suppress rows and control column widths, making printcase output readable and compact for browsing.

### Syntax

The basic syntax of printcase is:

printcase [using *filename*] if id\_variable==value [, options]

where id\_variable is the name of the identifying variable in the dataset, and *value* is the value of id\_variableto print. The id\_variablecan be string or numeric. If id\_variable is unique within the dataset, one case is printed; if it is non-unique, 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 file name by leveraging the using option.

using[filename] species alternate file name and location.

### 2. 2 Options are the following:

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 PDFfile as the output.

font(string) sets the font to be used for the entire document of the output printcase. Any installed font can be specified. If not set, the default is Arial.

noempty suppresses all empty and system missing responses and their variables from the resulting table, if the value label is an empty string (i.e., “”) or a Stata missing value

code (“.”, “.d”, etc.). If not specified, all empty responses will be included. In longitudinal cases, variable rows are only suppressed if variables are empty for all observations.

ignore(“string1” “string2”…) allows users to specify variables to 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 only suppressed 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 help putdocx\_table and help putdocx\_table for the specifics of formatting input.

unit([string]) changes the default title for the response column(s) 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- use

In this example, the researcher first loads the model births recode dataset from the DHS (ICF 2020). 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) doesn’t contain any notes, we add two notes to demonstrate the value of the *addnotes* feature.

**. notes v007: check if this is ethiopian, nepali, or gregorian calendar**

**. notes v130: religious groups are ambiguous in the model dataset**

### Example 1

In Example 1, the analyst calls printcase using the common *if* arguments to specify the id variable and relies on the default filename, saved in the home directory. The options specified here will generate a pdf output format (rather than MS Word document) using Arial font.

**. printcase if id==”1\_13\_3”**

Testing type

ID Variable: id

ID Value: 1\_13\_3

saving doc

successfully created “/Users/id1\_13\_3.docx"

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

### Example 2

In Example 2a, the user likewise produces a MS Word Document, designating a descriptive file name, 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 (i.e., the printed case) is produced by the example do file. Figure 1 shows an extract from the output produced by this example.

**. printcase using "./zzz\_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"

Figure 1. Screenshot of printcase output from Example 2b.

Table

Description automatically generated

Note: Figure 1 shows an extract from example 1, where variable name, labels, and responses are displayed in three columns.

### Example 3

Example 3 shows how analysts 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 to three records in the file. The same logic also applies to longitudinal datasets to which a respondent (id) contributes multiple observations over time. Here, the researcher further prepared 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==.**

After prepping the data, this analyst prints 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 (non-unique 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 only suppress the row if the variables are missing across all associated cases.

**. printcase using "./zzz\_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 replaced "/Users/example3.docx"

Figure 2. Screenshot of printcase output from Example 3.

Table

Description automatically generated

Note: Figure 2 shows an extract from Example 3, where responses from three cases corresponding to a non-unique 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. Oftentimes, 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 over-interpret particular cases. Still, the ability to look closely at a single observation is sometimes valuable. In particular, being able to produce a neat, readable quasi-questionnaire directly from a dataset – without headache – when necessary will enhance the workflow of data collection for many fieldworkers.

# 5 Acknowledgements

We’re 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 References

Bledsoe, C., Fatoumatta Banja, and Allan G. Hill. 1998. “Reproductive Mishaps and Western Contraception: An African Challenge to Fertility Theory.” *Population and Development Review* 24 (1): 15–57.

ICF. “Download Model Datasets.” The DHS Program website. Funded by USAID. [Accessed June 20, 2020]. https://dhsprogram.com/data/Download-Model-Datasets.cfm

Leahey, Erin. 2008. “Methodological Memes and Mores: Toward a Sociology of Social Research.” *Annual Review of Sociology* 34 (1): 33–53. https://doi.org/10.1146/annurev.soc.34.040507.134731.

Leahey, Erin, Barbara Entwisle, and Peter Einaudi. 2003. “Diversity in Everyday Research Practice The Case of Data Editing.” *Sociological Methods & Research* 32 (1): 64–89. https://doi.org/10.1177/0049124103253461.

Pearce, Lisa D. 2002. “Integrating Survey and Ethnographic Methods for Systematic Anomalous Case Analysis.” *Sociological Methodology* 32 (1): 103–32.

Sana, Mariano, and Alexander A. Weinreb. 2008. “Insiders, Outsiders, and the Editing of Inconsistent Survey Data.” *Sociological Methods Research* 36 (4): 515–41. https://doi.org/10.1177/0049124107313857.

Waal, Ton de, Jeroen Pannekoek, and Sander Scholtus. 2011. *Handbook of Statistical Data Editing and Imputation*. John Wiley & Sons.

# 7 About the authors

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

Jenny Trinitapoli is associate professor of sociology at the University of Chicago. Since 2009, she has been PI of the Tsogolo La Thanzi project, spearheading the collection of 12 rounds of data from over 3000 young adults living in Balaka, Malawi. She envisioned and commissioned the printcase command out of necessity, when transitioning from paper-based to e-tablet data collection with an experienced data collection team.