

Responses to Open-Ended Survey Questions: What To Do

Paul Pope, Texas Agricultural Extension Service, Texas A&M University System

ABSTRACT

This paper presents simple strategies for handling responses to open-ended survey questions.

INTRODUCTION

What to do with responses to open-ended questions is a common quandary for those involved in survey work. This paper presents some simple strategies for handling responses to open-ended survey questions.

An open-ended survey question allow respondents to provide their own answer to the question. Responses are likely to vary and require categorization in order to obtain a usable summary. Common examples of this question format include OTHER _____, ADDITIONAL COMMENTS, and LISTINGS. The following questions serves as examples.

Q8. WHAT ARE THREE THINGS YOU LIKED MOST ABOUT THE PROGRAM

- A. _____
- B. _____
- C. _____

Q20. ADDITIONAL COMMENTS

ARRANGING RESPONSES TO OPEN-ENDED SURVEY QUESTIONS

At the Texas Agricultural Extension Service, responses to open-ended survey questions are entered in a file (separate from closed-ended responses) with four fields: ID, QUESTION, CATEGORY, and RESPONSE. Each response to an open-ended question is entered as an observation. For example, if a survey instrument has three open-ended questions (say Q3, Q5, and Q99) then the responses of two individuals (with IDs of 234 and 345) would be arranged as follows:

(id)	(question)	(category)	(response)
234	3	1	financial needs
234	5	3	stocking rates
234	99	1	Program was useful but it failed to address management issues
345	3	1	financial concerns
345	5	4	brush and weeds
345	99	1	Need more emphasis on management and ownership concerns

A category code field is included so that a responses can be categorized. An ID field is included so that open-ended responses can be merged with closed-ended responses if need be. Merging would be necessary if frequencies, listings, or other analyses are desired by gender, age category, income level, or other class variable contained in the file with close-ended responses.

If responses were in a ASCII file called OPEN.TXT then the program to list non-missing responses to each question would be as follows:

```
DATA OPEN;
INFILE 'OPEN.TXT' MISOVER LRECL=100;
INPUT ID 1-3 QUESTION 5-6 CATEGORY 8 RESPONSE $ 10-100;
RESPONSE = UPCASE(LEFT(RESPONSE)); /* TO INSURE ALIGNMENT AND CASE CONSISTENCY */;
;

TITLE1 'RESPONSES TO QUESTION 3';
PROC PRINT NOOBS;
  VAR RESPONSE;
  WHERE QUESTION=3 AND RESPONSE NE '';
RUN;

TITLE1 'RESPONSES TO QUESTION 5';
PROC PRINT NOOBS;
  VAR RESPONSE;
  WHERE QUESTION=5 AND RESPONSE NE '';
RUN;

TITLE1 'RESPONSES TO QUESTION 99';
PROC PRINT NOOBS;
  VAR RESPONSE;
  WHERE QUESTION=99 AND RESPONSE NE '';
RUN;
```

Frequency distributions can be obtained using the same strategy:

```
TITLE1 'RESPONSES TO QUESTION 3';
PROC FREQ;
  VAR CATEGORY;
  WHERE QUESTION=3 AND RESPONSE NE '';
RUN;

TITLE1 'RESPONSES TO QUESTION 5';
PROC FREQ;
  VAR CATEGORY;
  WHERE QUESTION=5 AND RESPONSE NE '';
RUN;

TITLE1 'RESPONSES TO QUESTION 99';
PROC FREQ;
  VAR CATEGORY;
  WHERE QUESTION=99 AND RESPONSE NE '';
RUN;
```

The strategy of placing responses to open-ended questions in a file separate from closed-ended questions has at least two advantages. First, open-ended data do not have to be entered initially with the close-ended data, thus expediting the analyses of the close-ended data. Second, it reduces both the number of variables and record length within the closed-ended response file.

WORD FREQUENCY DISTRIBUTION

Knowing the frequency distribution of words within responses to open-ended survey questions can be useful for querying and/or categorization. The program below provides an example of how to produce a frequency distribution for a particular question (Q3 in this case). The maximum number of words for a particular record is set at 70. This parameter can be modified to reflect a shorter or longer record length. The characters listed after WORDS in the COMPRESS functions are considered extraneous and are removed. Of course what constitutes extraneous characters is a function of user discretion.

```
DATA COUNT (KEEP=WORD);
INFILE 'OPEN.TXT' MISSOVER LRECL=100;
INPUT QUESTION 5-6 RESPONSE $ 10-100 a;
IF QUESTION=3 AND COMPRESS(RESPONSE) NE '' THEN
DO;
  LENGTH WORD1-WORD70 $16;
  INPUT (WORD1-WORD70) ($);
  ARRAY TEXT WORD1-WORD70;
  DO OVER TEXT;
    WORDS=TEXT;
    WORD=UPCASE(COMPRESS(WORDS, '.,:;."()?'1234567890*'));
    IF WORD NE '' THEN OUTPUT;
  END;
END;
RUN;

PROC FREQ ORDER=FREQ;
RUN;
```

The following data step can be added to exclude common, unimportant words. Words can be added or deleted from the data step as desired.

```
DATA FINAL;
SET TEXT;
IF WORD='AND' THEN DELETE; ELSE IF WORD='THE' THEN DELETE; ELSE IF
WORD='OR' THEN DELETE; ELSE IF WORD='THERE' THEN DELETE; ELSE IF
WORD='WAS' THEN DELETE; ELSE IF WORD='THIS' THEN DELETE; ELSE IF
WORD='WERE' THEN DELETE; ELSE IF WORD='HE' THEN DELETE; ELSE IF
WORD='IS' THEN DELETE; ELSE IF WORD='SHE' THEN DELETE; ELSE IF
WORD='ARE' THEN DELETE; ELSE IF WORD='I' THEN DELETE; ELSE IF
WORD='BE' THEN DELETE; ELSE IF WORD='AT' THEN DELETE; ELSE IF
WORD='A' THEN DELETE; ELSE IF WORD='AN' THEN DELETE; ELSE IF
WORD='OF' THEN DELETE; ELSE IF WORD='BEEN' THEN DELETE; ELSE IF
WORD='IF' THEN DELETE; ELSE IF WORD='IN' THEN DELETE; ELSE IF
WORD='WITH' THEN DELETE; ELSE IF WORD='AS' THEN DELETE; ELSE IF
WORD='IT' THEN DELETE; ELSE IF WORD='FOR' THEN DELETE; ELSE IF
WORD='DO' THEN DELETE; ELSE IF WORD='DID' THEN DELETE; ELSE IF
WORD='DOES' THEN DELETE; ELSE IF WORD='THEIR' THEN DELETE; ELSE IF
WORD='FROM' THEN DELETE; ELSE IF WORD='NOT' THEN DELETE; ELSE IF
WORD='ON' THEN DELETE; ELSE IF WORD='NO' THEN DELETE; ELSE IF
WORD='SO' THEN DELETE; ELSE IF WORD='THAN' THEN DELETE; ELSE IF
WORD='THAT' THEN DELETE; ELSE IF WORD='THEN' THEN DELETE; ELSE IF
WORD='HAS' THEN DELETE; ELSE IF WORD='HAD' THEN DELETE; ELSE IF
WORD='TO' THEN DELETE; ELSE IF WORD='THEY' THEN DELETE; ELSE IF
WORD='HAVE' THEN DELETE; ELSE IF WORD='WHEN' THEN DELETE;
RUN;
```

This is the result:

WORD FREQUENCY DISTRIBUTION

Q3	Frequency	Percent	Cumulative Frequency	Cumulative Percent
DEVELOPMENT	33	7.0	33	7.0
POLITICAL	31	6.6	64	13.6
DEMOCRATIC	15	3.2	79	16.8
SOCIOECONOMIC	7	1.5	86	18.3
VARIABLES	7	1.5	93	19.7
CAUSAL	5	1.1	98	20.8
COMMUNICATION	5	1.1	103	21.9
ECONOMIC	5	1.1	108	22.9
EDUCATION	5	1.1	113	24.0
EFFECT	5	1.1	118	25.1
RELATIONSHIP	5	1.1	123	26.1
BETWEEN	4	0.8	127	27.0
EQUATION	4	0.8	131	27.8
MODEL	4	0.8	135	28.7
PROCESS	4	0.8	139	29.5
REGRESSORS	4	0.8	143	30.4
VARIABLE	4	0.8	147	31.2
WHICH	4	0.8	151	32.1
ZERO	4	0.8	155	32.9
CORRELATED	3	0.6	158	33.5
DEFINITION	3	0.6	161	34.2
DEMOCRACY	3	0.6	164	34.8
DIFFERENT	3	0.6	167	35.5
EACH	3	0.6	170	36.1
EMPIRICAL	3	0.6	173	36.7
ENDOGENOUS	3	0.6	176	37.4
FACTORS	3	0.6	179	38.0
.				
.				
.				
WII	1	0.2	470	99.8
YEARS	1	0.2	471	100.0

SAS is a registered trademark or trademark of SAS Institute Inc. in the USA and other countries.
® indicates USA registration.

The author encourages you to share your experiences, evaluations, or comments concerning the concepts and applications presented in this paper:

Paul Pope
Texas Agricultural Extension Service
112A Special Services Bldg.
College Station, TX 77843-2468
Phone: (409) 845-3929
Internet: p-pope@tamu.edu