

Data-Processing Strategies
The 1977 Baseline Interviews of the Earthquake Watch in California, 1977-1979
(SPSS Dataset of EWCA77BASE.SAV)

This document describes how we identified and extracted 1,450 respondents who represented the baseline sample of Los Angeles County households for the study conducted by Turner and colleagues in 1977. The original data were recorded on IBM cards.

We started with an SPSS data file called EWCA77BASE.SAV, which is archived at <http://www.sscnet.ucla.edu/issr/da/earthquake/erthqkstudies2.datasets.htm>. The total sample of the dataset did not represent a probability sample because the researchers deliberately oversampled certain populations. We had to extract 299 oversampled cases so that we were left with the probability sample of 1,450 cases. The various components of the samples are described in Turner.samplinginfo.pdf.

Unfortunately, EWCA77BASE.SAV does not contain the raw ID numbers in spite of the fact that the original ID numbers must be used to differentiate the County probability sample from the oversampled populations, which include residents from San Fernando Valley where the 1971 San Fernando earthquake occurred, residents who lived in areas vulnerable to inundation, and Black residents (Turner, Nigg & Paz, 1986).

Furthermore, the EWCA77BASE.SAV does not provide clear information regarding which resident aged 18 and older was selected when there were multiple adults in a household. Consequently, despite the fact that the dataset includes socio-demographic characteristics (i.e. gender, age, marital status and employment status) for all adult residents in an interviewed household, the characteristics of the actual respondent, except for gender¹, are left uncertain.

We investigated the roster information in EWCA77BASE.SAV, card image files which documented the sampling procedure and respondents' contact information, and the raw data. We succeeded in extracting the 1,450 respondents who completed the baseline survey, excluding the over-sampled populations. Simultaneously, we identified each respondent's age, marital status, and employment status. In addition, from the study files and the raw data, zipcodes of interviewed households were also identified. The zipcode data enabled Akiko Sato, the doctoral student who performed this extraction, to conduct geographical analyses and modeling². Table 1 summarizes the data and files used to extract the 1,450 cases.

¹ EWCA77BASE.SAV provides a variable of a respondent's gender (i.e., v114 in the SPSS dataset) aside from the variables of socio-demographic characteristics of adults residents in a household

² The zipcodes are original data. Therefore, the data are not adjusted based on the changes in zipcode boundaries over time.

Table 1. Descriptions of data files

	Name	Type	Key features
Raw data	DTA491.file0031.EQDATA	Card image	<ul style="list-style-type: none"> ○ 18 punch cards with 80-columns are used for each respondent. ○ Each digit or intentionally blank space is a “column.” ○ Both raw and log ID numbers are contained. ○ We used raw ID numbers 1-3000 to correct the over-sampling used in the original study³.
Original files from the study	Turner.samplinginfo	PDF file	<ul style="list-style-type: none"> ○ This file provides a list of files from the study and descriptions of the file contents are provided.
	DTA900.file2711.EWCA77BASE	PDF file	<ul style="list-style-type: none"> ○ This file shows the study questionnaire with variable names in the SPSS dataset.
	DTA494.file0171.HHSAMPS	Text file	<ul style="list-style-type: none"> ○ This file reports respondents’ raw ID numbers and contact information. ○ A code of 90 in the first line indicates a completed interview. ○ Three lines are assigned per completed respondent. Two lines are assigned per respondent who did not complete the interview. ○ The last four-digit number in the first line for each respondent is their raw ID number. ○ Columns 15-17 illustrate a geographic code. The first line of each geographic code provides the area name. A zipcode is on the last five columns of the line.
Dataset	EWCA77BASE	SPSS. SAV file	<ul style="list-style-type: none"> ○ There are 649 variables and 1,749 cases. ○ Variable ‘v2’ provides log ID numbers. ○ Raw ID numbers are not provided.

The new EWCA77BASE dataset, named “EWCA77BASE_v2base” contains the identified information. The following describes the procedures used to create the new dataset.

³ According to the file “Turner.samplinginfo”, 0000 < raw ID < 3000 are “basic county representatives”; 3000 < raw ID < 4000 are the “San Fernando Valley supplement”; 4000 < raw ID < 5000 are the “other inundation supplement”; and 5000 < raw ID are the “Black supplement”.

1) Zip Code variable

- Save DTA494.file0171.HHSAMPS as a Microsoft Excel spreadsheet
- Delete all fields, except for completion codes (Columns 70-71), raw ID numbers⁴ and zipcodes (Columns 75-79)
- Delete all rows of respondents who did not complete the interview (i.e. those who have a *non-90* code in the first line), and the second and third rows of each respondent who completed the interview (i.e. those who have a code of 90 in the first line)
- Create a new column called “zipcode” and add zipcode information:
rawID_zip_DTA494.file0171.HHSAMPS.xls
- Convert DTA491.file0031.EQDATA into a Microsoft Excel spreadsheet, and delete all columns except for raw ID numbers (Columns 1-4) and log numbers (Columns 29-32 in the first card of each respondent):
raw_log_ID_DTA491.file0031.EQDATA.xls
- Merge the two Microsoft Excel spreadsheets based on raw ID numbers:
MERGED_raw_log_ID_zip.xls
- Merge the combined Microsoft Excel spreadsheets into the dataset based on log ID numbers (v2 in the SPSS data and log_ID in the merged Excel spreadsheet):
EWCA77BASE_v2.sav
- Extract “basic county representatives” by selecting respondents who have a raw ID number lower than 3000: **EWCA77BASE_v2base**⁵

2) Selected residents and their age, marital status, and employment status

- A selected respondent’s line number on a household roster is shown on Columns 63-64 in the second punch card (see DTA491.file0031.EQDATA). The variable v48 in the SPSS data also shows line numbers of selected residents (also see 4th page (Page 1) in DTA900.file2711.EWCA77BASE). For example, for the household with an original ID number of 4 (i.e. a Log ID number of 747), the second person on the roster of adult residents in the household, called “P2” in the dataset, was selected.
 - The numeric part in variable labels in the SPSS data shows a card number and columns where the information were originally recorded in punch cards. The card-number and column information are also indicated in the questionnaire: a number following “deck” indicates a card number, whereas numbers in or over a response space per question are column numbers.

⁴ Raw ID numbers are Column 76-79.

⁵ Raw ID numbers range from 4 to 2565. Log ID numbers range from 1 to 1749.

- Table 2 shows where the information on a respondent's age, marital status, and employment status are provided.

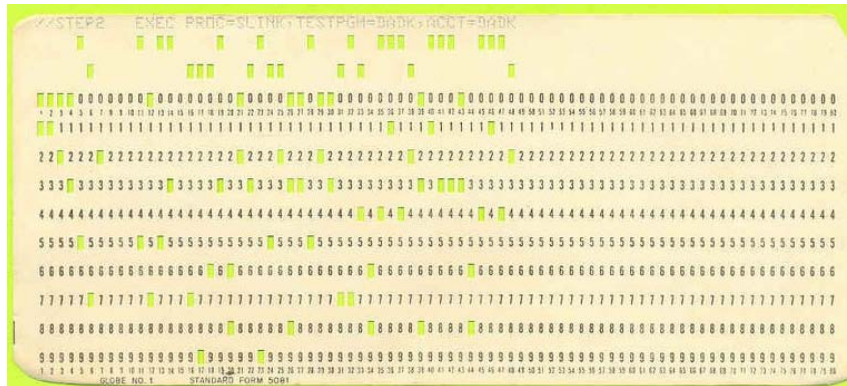
Table 2. Card numbers and columns in the raw data, and the dataset variables that list respondent's age, marital status, and employment status. The Baseline Interviews of the Earthquake Watch in California.

		Raw Data DTA491.file0031.EQDATA		Dataset EWCA77BASE
		Card	Column	Variable Name
Age	P1	2	14-15	V13
	P2	2	16-17	V14
	P3	2	18-19	V15
	P4	2	20-21	V16
	P5	2	22-23	V17
	P6	2	24-25	V18
	P7	2	26-27	V19
Marital Status	P1	2	49	V34
	P2	2	50	V35
	P3	2	51	V36
	P4	2	52	V37
	P5	2	53	V38
	P6	2	54	V39
	P7	2	55	V40
Employment Status	P1	2	56	V41
	P2	2	57	V42
	P3	2	58	V43
	P4	2	59	V44
	P5	2	60	V45
	P6	2	61	V46
	P7	2	62	V47

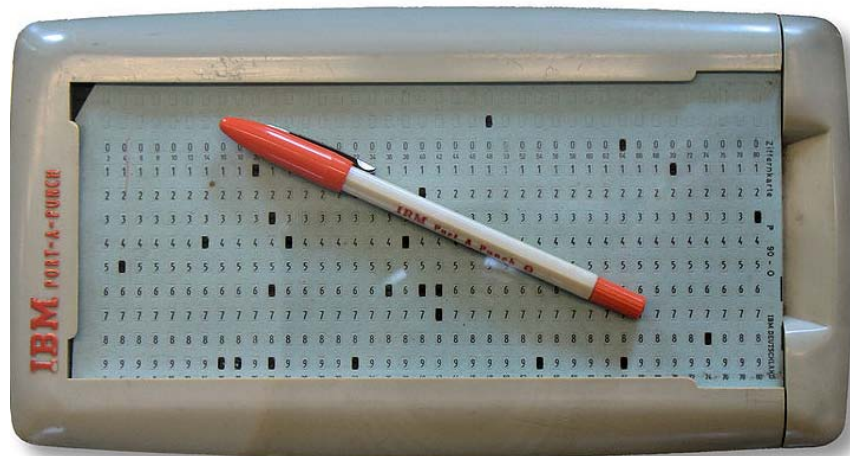
- Of 1,450 respondents residing in Los Angeles County, 955 are listed on the first row on a household roster (i.e. P1). There are 416 'P2's, 65 'P3's, 11 'P4's, two 'P5's and one 'P6'. There is no 'P7'.
- New variables for selected residents' age (R_age), marital status (R_mar), and employment status (R_emp) are created by extracting the information from v13-v47.

Manual Punch Cards

Manual punch cards were used to input, process, and store digital data in the 20th Century.



Each digit was labeled as a separate column, including blank spaces. Letters and special characters were signified with multiple punches. For example, using the Extended Binary Coded Decimal Interchange Code, “a” was represented by punching 1, 2, and 9 in a single column, whereas “A” was represented by punching 1, 9, and 3. There were several codes that evolved. IBM was the most dominant company using these cards.



Early digital computers were able to translate these cards, but the data remained stored on these cards until computer technology improved during the 1970s. It wasn't until the 1980s that these became obsolete. Using older data sources may require interpreting data originally stored on these cards.

References

Turner, R., Nigg, J. M., & Paz, D. H. (1986). *Waiting for disaster: Earthquake watch in California*. Berkley, CA: University of California Press.