

Validation of Data for the Mexico Panel Study 2006

Original data entry

Information from the completed questionnaires (i.e., the paper copies) was entered into a spreadsheet by a team of professional data entry personnel employed full-time by *Reforma* newspaper. Electronic data were then checked for errors.

Data were then validated by the polling team at *Reforma* and by the research team at the Massachusetts Institute of Technology (MIT)

Validation of data by *Reforma*

In the first wave, information for each of the polling points was checked against the original sample. The following variables were validated: precinct (*sección*), state (*estado*), county (*municipalidad*), electoral district (*distrito*), town or village (*localidad*), urbanization (urban, rural, or mixed, according to Federal Electoral Institute categories) and zip code (*código postal*). Results of validation are as follows:

Validated variable	Number of files with errors or imprecision	Observations
Town	4	Corrected in database
District	15 (all at one polling point)	Corrected in database
Zip Code	4	Corrected in database

In second wave of the panel, socio-demographic data were checked against those obtained in the first round. The following variables were verified: gender (*sexo*), age (*edad*), date of birth (day / *día*, month / *mes*, and year / *año*), and education (*educación*). Where inconsistencies were found, electronic data were checked against paper copies of the questionnaires to determine whether the error was the product of interviewer miscoding or data entry. All data entry errors were corrected; discrepancies between the paper copies were not. Results were as follows:

Validated variable	Percentage of files with mistakes or imprecision	Action taken
Gender	0.17%	Checked against data in the paper copy of the questionnaire and corrected
Age*	15%	Corrected as long as difference was based on data entry error
Education*	12%	Corrected when difference was based on data entry error
Birth date	17%	Corrected when difference was based on data entry error

*Cases were investigated whenever the difference in the responses between Wave 1 and Wave 2 was +/- one unit or more.

In the third wave, data from all odd-numbered files – that is, 50% of the sample – were entered into a separate database, and the new data were checked against the original electronic data. (The total amount of data re-entered consisted of 204 variables from 801 files.) Paper copies of the questionnaires were inspected for various sources of inconsistency: data entry errors, instances where the boxes in which interviewers recorded the interviewees’ responses contained values that fell outside of the pre-set range of responses on the questionnaire; cases in which interviewers failed to follow the “skip to” instructions in the questionnaires for certain items (for instance, if interviewers inadvertently asked respondents who said they did not watch any television news program which television news program they watched), and so forth. The results of this exercise are summarized in Table 1 below:

Table 1: Transcription, coding, and data entry errors

Number of re-entered questionnaires	801
Number of questions	91
Number of items in these questions	204
Number of items re-entered	163,404
Number of errors detected (breakdown below)	266
Data entry inconsistencies / errors	128
Number in the coding box recorded by interviewer was illegible	64
Coding errors on paper copy of questionnaire by interviewer	37
Interviewer circled one response on the questionnaire but recorded a different number in the box	30
Instructions to skip certain questions were not followed properly	7
AVERAGE ERROR:	0.16%

Statistics are based on: Average error = (Number of detected errors (266) *100) / Number of captured variables (163,404). Number of items re-entered = Number of items in questions (204) * Number of re-captured questionnaires (801).

All errors found were corrected in the final database.

Independent validation of data by MIT

After the final database was compiled by *Reforma*, the research team at MIT performed checks for other potential sources of error. The first was a check for “frame-shift” errors: that is, instances in which a specific entry may have been inadvertently omitted or in which data from a particular response was inadvertently split between two cells, thus causing all data in the individual file after the error to be incorrect. No such frame-shift errors were found. Second, the electronic database was then scanned for cases where stray data were entered into cells that should presumably have been left blank. Two such cases were found in the data from the cross-sectional poll that accompanied the second wave of the panel; these entries were deleted. Third, the entire database was checked for cases where values for individual entries fell outside the coding range used in the final electronic version (which includes responses volunteered by some respondents that were not originally listed in the questionnaire). No such errors were found. The fourth check examined “internal control” variables, such as gender of interviewer (i.e., the person conducting the interviews) and postal code of the respondent. A handful of files were found to have coding errors for the gender of the interviewer, and in the third wave a few clusters of files had small differences in postal codes that did not match the previous two waves due to data entry errors. All of these discrepancies were checked against the original questionnaires and corrected.

Ex-post examination of potential errors in re-contacting

A separate set of checks was performed to discover whether any remaining inconsistencies in the data could be attributed to interviewers’ inadvertently interviewing the wrong individual in the second or third panel waves. To this end, all files for the second and third waves were analyzed for discrepancies in demographic characteristics: gender (coded by the interviewer), skin color of the respondent (as coded by the interviewer), economic status of the respondent’s dwelling (as coded by the interviewer), the accent of the respondent (coded by the interviewer), age (reported by the respondent), and education level (reported by the respondent). In the third wave, birth dates were not recorded, even though they were asked by the interviewer; rather interviewers were instructed to end the interview if the day and month of birth provided by the respondent did not match that of the birth date recorded for respondents in earlier rounds. Interviewer coding of the economic status of the respondent’s dwelling was not included in the second round and age was not asked in the third round. Consequently, these variables could not be used for matching.

No inconsistencies were detected for respondent gender across panel waves. As noted above, however, a number of inconsistencies were detected for age and education, as well as for skin color, economic status of the respondent’s dwelling, and accent. In the case of age, approximately half of the respondents in the first wave of the panel reported being one year older in subsequent waves. As the first wave was conducted five months before the second wave and nine months before the third wave, these inconsistencies were disregarded. Also disregarded were cases in which levels of variables subjectively coded by the interviewer varied from one wave to the next by only one unit – for instance, where a person coded as “light brown” in one wave was coded as “dark brown” in the next or where the economic level of the respondent’s dwelling

was coded as “C” (lower-middle class) rather than “D” (lower class). Finally, as education levels were ranges that allowed for some ambiguity, and misreporting on education is not unheard of, discrepancies of only one unit on that scale were also disregarded. The remaining discrepancies are summarized in Table 2, below.

Table 2: Inconsistencies in demographic variables

	Wave 1- Wave 2	Wave 1- Wave 3	Wave 2- Wave 3
Age discrepancies (> 1 year older, > 0 years younger)	223	--	--
Total N	1768	--	--
Percent discrepant	12.6%	--	--
Age discrepancies (> 2 years older, > 1 year younger)	74	--	--
Total N	1768	--	--
Percent discrepant	4.2%	--	--
Skin color discrepancies (> 1 unit off)	20	32	20
Total N	1763	1575	1362
Percent discrepant	1.1%	2.0%	1.5%
Education discrepancies (> 1 unit off)	100	86	63
Total N	1762	1584	1369
Percent discrepant	5.7%	5.4%	4.6%
Economic level discrepancies (> 1 unit off)	--	44	--
Total N	--	1568	--
Percent discrepant	--	2.8%	--
Accent of respondent discrepancies (> 1 unit off)	31	40	22
Total N	1593	1770	1375
Percent discrepant	1.9%	2.3%	1.6%
<i>At least one demographic discrepancy</i>	<i>327</i>	<i>179</i>	<i>101</i>
<i>Percent with at least one demographic discrepancy</i>	<i>18.8%</i>	<i>11.5%</i>	<i>7.5%</i>
<i>At least two demographic discrepancies</i>	<i>26</i>	<i>5</i>	<i>1</i>
<i>Percent with at least two demographic discrepancies</i>	<i>1.5%</i>	<i>0.3%</i>	<i>0.1%</i>
<i>Three of more demographic discrepancies</i>	<i>0</i>	<i>0</i>	<i>0</i>
<i>Percent with at least three discrepancies</i>	<i>0%</i>	<i>0%</i>	<i>0%</i>
<i>N</i>	<i>1740</i>	<i>1561</i>	<i>1249</i>

These inconsistencies do not by themselves indicate that the wrong respondent was interviewed, as interviewers may have coded the same responses differently to certain questions, and errors could have been made in completing the forms or in data entry

that were not detected in subsequent checks. Respondents may also have misreported their age or education level, and codings of items like economic status of the dwelling are subject to broad interpretation. Finally, in a few cases, respondents' education levels, economic status, and even skin tone may have actually changed during the course of the panel. (Respondents who were interviewed in the late spring and summer were less slightly likely to be coded as white than they were when interviewed in the fall – the “Acapulco effect”.)

For these reasons, data were checked to see if the respondents had multiple discrepancies across demographic variables. (In this case, as noted above, these discrepancies were counted only if the difference was greater than one unit.) These results are also shown in Table 2; they indicate that errors did not tend to cluster on certain respondents. Only a very small number of respondents had two inconsistencies, and none had more than two.

As a third check, inconsistencies were cross-tabulated with cases in which the respondent was interviewed by a different interviewer across panel waves. These results are shown in Table 3 below.

Table 3: Inconsistencies by interviewer

		Percent without inconsistencies	Percent with one inconsistency	Percent with two inconsistencies
Wave 1 to Wave 2				
	Same interviewer	81%	18%	1%
	Different interviewer	79%	19%	2%
Wave 1 to Wave 3				
	Same interviewer	92%	8%	0%
	Different interviewer	87%	13%	0%
Wave 2 to Wave 3				
	Same interviewer	93%	7%	0%
	Different interviewer	91%	9%	0%

As the data suggest, there is some evidence of greater inconsistency where interviews were conducted by a different interviewer across panel waves. For instance, 13% respondents who were re-interviewed in Wave 3 by different interviewers showed at least one discrepancy, against only 8% of those who were re-contacted by the original interviewer. Likewise, all five subjects who had two discrepancies in Wave 1-Wave 3 were interviewed by a new interviewer in Wave 3 (though these five constituted only 0.4% of the total number interviewed by different people in the second wave). In general, however, inconsistency rates are not dramatically greater for respondents who were interviewed by different interviewers.

In the next step, cases of inconsistency were cross-tabulated with interviewer code, to see if errors clustered in certain interviewers. No systematic pattern of clustering was found, though five interviewers had somewhat higher levels of inconsistencies.

Subjects interviewed by Interviewer 53, who was only employed in the first wave, had the highest levels of inconsistencies: 4 cases of two inconsistencies and 11 cases of one inconsistency out of 32 interviews between Wave 1 and Wave 2, as well as 1 double inconsistency and 6 inconsistencies out of 22 interviews between Wave 1 and Wave 3. However, all of the interviews conducted by Interviewer 53 in Wave 1 were supervised. Respondents who were interviewed by Interviewer 518 in Wave 2 also showed elevated levels of inconsistency; these inconsistencies were substantially more common and pronounced in the unsupervised interviews, all of which took place in rural Chiapas. However, these apparent discrepancies were driven primarily by different codings of accent (“indigenous” versus “typical”) in a heavily indigenous region where local accents are quite distinct from other areas of the country. Respondents in this area were also of very low education levels – typically incomplete primary or no formal schooling – which would tend to increase misreporting. Respondents interviewed in Wave 2 by Interviewers 65 and 401 (who were not employed in the first wave) also showed higher levels of discrepancies, though in these cases there was no difference in the number of discrepancies between interviews that were supervised and those that were not. Subjects interviewed by Interviewer 110, who was employed only in Wave 3, showed somewhat higher levels of inconsistencies, but all of the cases that showed discrepancies were supervised after the fact.

As a corollary check, individual files were screened for two types of problems. The first type involved files where at least one demographic variable showed a pronounced discrepancy (for example, from complete primary to complete secondary school) and at least one other demographic variable changed by at least one unit. The second type involved files with only small discrepancies (e.g., one unit) across the board. The few cases falling into these groups were Respondents 966, 1070, 1458, 1472 and 1789 in Wave 2, as well as Respondents 27, 1177, and 1721 in Wave 3. Of these cases, interviews with Respondent 1721 in Wave 3 and Respondent 1789 in Wave 2 were supervised in Wave 1, but not in the other waves. All of the other cases were independently supervised either during or after the interview in question. Interviews with Respondent 966, whose file showed the most discrepancies, were supervised during the interview for all three panel waves.

All told, no compelling evidence of mismatch was found. Consequently, no cases were deleted from the dataset because of suspected mismatch.

Point errors detected through demographic matching

Checking for inconsistencies in demographic data revealed two instances that seem likely to have resulted from misreporting by the respondent, coding error on the part of the interviewer, or some combination of the two. In the most extreme inconsistency for age, Respondent 1619 was coded as 62 years old in Wave 1 and 82 in Wave 2. Interviews with this Respondent, who spoke only a Mayan language, were conducted via interpreter; thus, it is possible that the answer to the age question was partly lost in translation. Because age in Wave 1 agreed with the respondent’s reported birth date, that datum is more likely to be accurate than the response for Wave 2. In the final dataset, the age was changed to 62 years for both waves.

The most extreme inconsistency in education occurred with Respondent 1087, a middle-aged housewife from Nuevo León state who is coded as having a primary school education (“1”) in Wave 1 and a university education (“9”) in Wave 3. Other demographic variables show no inconsistencies, and variables in the first wave that would normally correlate with education – such as living standards, as measured by an index of household items – were much more consonant with a university-level education than a primary-level background. In the final dataset, education is coded as a “9” (indicating university education) for all three waves.

The second largest inconsistency in education concerns Respondent 702. A housewife from Puebla state in her early 20’s, she is coded as having completed her secondary / technical education in Wave 1 and subsequently as having begun university – a jump of three units on the educational scale. However, some people with a technical degree can go directly to university in Mexico without attending high school (*preparatoria*), and the respondent in question is of appropriate age to do so. This inconsistency was thus not corrected.

Other apparent inconsistencies in the database were likewise not changed.

Impact of corrections on replication

Revisions to the dataset discussed in this report were made after members of the Mexico 2006 Panel Study submitted book chapters or articles using an earlier version of the dataset. Because the changes noted here are minor, the results should be substantively the same as they would be using the final dataset. However, precise coefficients might differ slightly upon replication.