Data Codebook

This codebook follows the structure of the R file 02\_wrangle-data.R in that variables are described in the same order as the corresponding subsections of the R file. However, variables coded but not used in the analysis are not described here. In the last section we discuss missing data imputation.

## Outcome variable - Intimate Partner Violence

We are primarily interested in reports of Intimate Partner Violence at Year 1.

**Question D6**: *Now, think about how (FATHER) behaves towards you. For each statement I read, please tell me how often he behaves this way.*

Answer options: 1 - OFTEN; 2 - SOMETIMES; 3 - NEVER

Statements:

1. He is fair and willing to compromise when you have a disagreement
2. He expresses affections or love for you
3. He insults or criticizes you or your ideas
4. He encourages or helps you to do things that are important to you
5. He tries to keep you from seeing or talking with your friends or family
6. He tries to prevent you from going to work or school
7. He withholds money, makes you ask for money, or takes your money
8. He slaps or kicks you
9. He hits you with his fist or an object that could hurt you
10. He tries to make you have sex or do sexual things you don’t want to do
11. He listens to you when you need someone to talk to
12. He really understands your hurts and joys

**Question E8** is the same as D6, but for those women who are currently romantically involved with a partner other than the father of the child.

**Question D8** are the same as D6, but with all items in the past tense. It is for parents who are no longer together at follow-up, and **specifically asks about the last month of the relationship**.

To construct our IPV outcome variable, we score the items as such:

* Positive items (1, 2, 4, 11, and 12) are coded as Often = 0, Sometimes = 1, and Never = 2.
* Negative items are coded as Often = 2, Sometimes = 1, and Never = 0.

All scores are summed and the total score is divided by 24 to create an IPV index from 0 to 1. For analyses with binary IPV, an index of 0 was taken as 0 and an index greater than 0 was taken as 1. If there are any missing values for the individual items, the total score counts as missing. For the imputed dataset, any missing values are first imputed and only them summarized into an IPV index.

If the woman was still with the father at year 1, the variable captures IPV from the father; if not, we take IPV from their other partner at year 1; if there is no such partner, we take IPV from the father at baseline (question D8). Women who did not report a relationship that would be amenable to IPV score were excluded from the dataset (see sample delineation document).

Three subscales were constructed from these items: - Emotional: items 1,2,3,4,11,12 - Physical: items 8,9,10 - Controlling: items 5,6,7 And the respective indices were calculating by dividing by 6, 3, and 3, respectively.

## Predictors / Covariates

[All measured at baseline, unless otherwise indicated]

### Race

**Question H3**:*Which of these categories best describes your race?*

Answer options: 1 — White, 2 — Black, African-American, 3 — Asian or Pacific Islander, 4 — American Indian, Eskimo, Aleut, 5 — Other, not specified, 101 — Hispanic, -2 — DON’T KNOW

We create the categorical variable m\_race with levels “White” and “Black”, after filtering out those who didn’t identify with either. “White” is the reference category.

### Age

Age of the mother at baseline. From the baseline guide:

*Age is recorded from the Core Baseline survey for mother, father and child and can be retrieved through the constructed variables:* ***cm1age*** *(mother’s age at the interview),* ***cf1age*** *(father’s age at the interview),* ***cm1b\_age*** *and* ***cf1b\_age*** *for the child’s age at the mother and father interview, respectively.* (p. 20)

Age is always a positive integer, and it is treated as a continuous variable.

### Education

Mother’s level of education at baseline.

**Question I1:** *Now I’d like to ask some questions about your education and work experience. What is the highest grade or year of regular school that you have completed?*

Answer options: 1 — No formal schooling, 2 — 8th grade or less, 3 — Some high school (Grades 9, 10, 11, & 12), 4 — High school diploma (Completed 12th grade), 5 — G.E.D., 6 — Some college or 2 year degree, 7 — Technical or trade school, 8 — Bachelor’s degree, 9 — Graduate or professional school

Variable m\_education is a categorical variable, where original answers 1-3 are coded as “Below HS”, and 4-9 as “HS and above”, and all missing values as NA. The reference category is “Below HS”.

### Alcohol use

Mother’s alcohol use during pregnancy.

**Question G2:** *During your pregnancy, about how often did you drink alcoholic beverages?*

Answer options: 1 — NEARLY EVERY DAY, 2 — SEVERAL TIMES A WEEK, 3 — SEVERAL TIMES A MONTH, 4 — LESS THAN ONCE A MONTH, 5 — NEVER

Variable m\_alcohol is a categorical variable with 3 levels: “Never”, “<1 / month”, and “>1 / month” (original values 1-3) with all missing values as NAs and “Never” as the reference level.

### Drug use

Mother’s drug use during pregnancy.

**Question G3:** *During your pregnancy, about how often did you use drugs such as marijuana, crack cocaine, or heroin?*

Answer options: 1 — NEARLY EVERY DAY, 2 — SEVERAL TIMES A WEEK, 3 — SEVERAL TIMES A MONTH, 4 — LESS THAN ONCE A MONTH, 5 — NEVER

Variable m\_drugs is a categorical variable with 3 levels: “Never”, “<1 / month”, and “>1 / month” (original values 1-3) with all missing values as NAs and “Never” as the reference level.

### Employment

We construct the employment variable based on other information of the survey.

**Question I2:** *In what month and year did you last work at a regular job lasting two consecutive weeks or more for which you received a regular paycheck, either full- or part-time?*

Here we either get a month and year (stored in separate variables) or the answer NEVER WORKED FOR TWO CONSECUTIVE WEEKS. If the mother never worked for two consecutive weeks we count her as unemployed. Next we find the difference in months between the date of the interview and the date of last employment with the formula (m\_month + 12\*m\_year) - (m\_employed\_month + 12\*m\_employed\_year). If this difference is greater than 2, then we are sure that the mother has not worked for at least 4 weeks at the date of the interview, and we count her as unemployed. If the difference is 1 or 0, we count her as employed; otherwise there will be a missing value.

Variable m\_employment is a categorical variable with levels “Unemployed” and “Employed”, where “Employed” is the reference level.

### Number of children

Mothers were asked whether they had any other biological children and if so, how many. We created the variable m\_children by pulling information from both items. It is a numeric variable consisting of positive integers (and 0).

### Household income

**Question J3**: Thinking about your income and the income of everyone else who lives with you, what was your total household income before taxes in the past 12 months?

Answer options: 1 — Under $5,000, 2 — $5,000 to $9,999, 3 — $10,000 to $14,999, 4 — $15,000 to $19,999, 5 — $20,000 to $24,999, 6 — $25,000 to $34,999, 7 — $35,000 to $49,999, 8 — $50,000 to $74,999, 9 — Greater than $75,000, -1 — REFUSED, -2 — DON’T KNOW

Variable m\_household\_income is a categorical variable with 3 levels - we collapsed the levels three at a time, and labelled them “Under $5,000”, “$15,000 to $34,999”, and “$35,000 or more”, with “Under $5,000” as the reference level.

### Welfare

Collected at year 1.

**Question H9**: In the past 12 months, have you received income from any of the following programs?

Programs: (1) Welfare or TANF; (2) Food Stamps; (3) Other Assistance such as Unemployment Insurance, or Worker’s Compensation

Answer options: 1 — YES, 2 — NO

If a mother answered YES to any of the three programs, she was coded as “Yes”, if she answered NO to all programs, she was coded as “No”, and otherwise NA.

We created the categorical variable m\_welfare\_last\_year with two categories, with “No” as the reference category.

### Home ownership

**Question F2**: Is the home or apartment where you currently reside…

Answer options: 1 — Owned or being bought by someone in your family?, 2 — Rented?

Variable m\_home is categorical with two levels, “Rented” and “Owned”, with “Owned” as the reference category.

### Religious services

**Question F6**: About how often do you attend religious services?

Answer options: 1 — Once a week or more, 2 — Several times a month, 3 — several times a year, 4 — Hardly ever, 5 — Not at all

Answers were reverse coded so 1 corresponds to lowest level of attendance and 5 to highest. Variable used as continuous in analysis.

### Health

**Question G1**: Now I have some questions about your health. In general, how is your health? Would you say it is…

Answer options: 1 — Excellent, 2 — Very good, 3 — Good, 4 — Fair, 5 — Poor

Answers were reverse coded so 1 corresponds to lowest level of health and 5 to highest. Variable used as continuous in analysis.

### Informal support

**Question E4**: During the next year, if you needed help, could you count on someone in your family to:

Answer options (Yes or No for each): A — Loan you $200?, B — Provide a place to live?, C — Help with babysitting or child care?

We summed the answers of those with complete reponses (with “Yes” = 1, and “No” = 0) and divided by 3 to create an informal support index from 0 to 1. For analyses with binary informal support.

## Data imputation

Missing data were imputed using the missForest::missForest() command. In order for correct imputation to be possible, IPV, informal support, health, and religious services variables were turned into factor – this way missForest would only impute possible values (e.g., 1 or 2 for health but never 1.5). Variables were later turned into numeric for analysis. The IPV and social support indices were only calculated after imputation, and weights were only appended to the dataset after imputation.