

# GE Public Data

## Data description

### a) Admin

- (1) *GE\_admin\_village\_randomization* – contains the variables used for randomization and the final treatment assignment at the village level. Can be merged to the caregiver and girl level datasets using the *village* variable.
- (2) *GE\_admin\_attendance\_payments* – contains information on attendance for each session of the GE program and on the payments made to the girls and caregivers.

### b) Baseline

- (1) *GE\_baseline\_caregiver\_survey* – this is the reshaped baseline caregiver dataset. The caregiver has been reshaped from wide to long because some caregivers corresponded to more than one girl respondent. At baseline 1,136 caregivers were surveyed, 9 caregivers cannot be matched to a girl. The variable *cgidnumber* uniquely identifies caregivers.
- (2) *GE\_baseline\_girl\_survey* – this is the final clean baseline girl individual dataset. At baseline 1,216 girls were surveyed, of which 57 could not be matched to a caregiver. The variable *girldidnumber* uniquely identifies girls.

### c) Endline

- (1) *GE\_endline\_caregiver\_survey* – this is the reshaped endline caregiver dataset. The caregiver has been reshaped from wide to long because some caregivers corresponded to more than one girl respondent. At endline 1,082 caregivers were surveyed, 30 caregivers cannot be matched to a girl. The variable *cgidnumber* uniquely identifies caregivers.
- (2) *GE\_endline\_girl\_survey* – this is the final clean endline girl individual dataset. At baseline 1,179 girls were surveyed, of which 98 could not be matched to a caregiver. The variable *girldidnumber* uniquely identifies girls.

## Notes

### Main variables

- The unique identifier in the girl individual surveys is the variable girlidnumber.
- The treatment identifier is the variable treatment or treatment bl when looking at baseline data- 1 corresponds to control, 2 to GE and 3 to GE+
- The unique identifier in the caregiver survey is again the girl id number: girlidnumber, the caregiver id: cgidnumber is not unique in the reshaped caregiver dataset as one respondent may have been the caregiver for more than one girl

### Questionnaires and data

- **Revisit data**
  - In the original endline survey the sexual and physical violence questions had not been translated into local languages so a revisit was conducted to ask the violence and the CRIES8 question again. During the revisit surveys we used audio recordings in local languages to administer the surveys. The girl respondents listened to the audio recordings on the PDAs and enumerators had translated responses at hand.
- **Randomization**
  - The randomization is block-stratified by village size and school enrollment, which are proxied by the number of girls surveyed at baseline in each village and the enrollment rate amongst the surveyed girls during the schoolyear 2014-15, respectively. These variables required to reproduce random assignment are included in the *GE\_admin\_village\_randomization* dataset. The variable that is used for blocking is called enrollment and is the combination of village size and school enrollment.