Appendix A: Creation of the socioeconomic status indicator

We wanted to create an indicator of the participants' socioeconomic status (SES), since this is an attribute that shapes most health outcomes in South Africa. Most of the participants in the study came from impoverished communities by global standards, so we decided that the SES category to which a participant belonged to would be relative to other participants in the study, rather than everyone else in the world.

The questionnaire that was administered to AGYW included a number of categorical variables related to SES. We chose 13 variables to derive our SES indicator, several of which are commonly used in other surveys to create similar indices:[1] 1. AGYW was away from home for more than one month in past 12 months (internal migration has been shown to cause and be caused by poverty[2]); 2. Has piped water in household; 3. Has flushing toilet in household; 4. Household has working electricity; 5. Household has a car; 6. Household has a computer; 7. Household has the internet; 8. Household has a refrigerator; 9. Household has a stove; 10. AGYW or member of her household went a day/night without eating in the past month; 11. AGYW has own money; 12. AGYW saves money; and 13. AGYW owes money.

We used cluster analysis with the K-modes algorithm[3] to divide participants into two SES categories based upon their responses to the 13 variables listed above. Cluster analysis is an exploratory and unsupervised machine learning technique that allows analysts to divide data into meaningful groups based upon shared features. We considered using principle component analysis (PCA) to create this indicator, but because our variables were all categorical, this ruled out that approach. Multiple correspondence analysis (MCA) was also considered. However, some variables (e.g. asset ownership variables) did not have sufficient variance among the population and there was quite a bit of clustering of responses. When this happens, MCA no longer becomes a viable method to categorise people into groups. We therefore decided to use a cluster analysis (with the K-modes algorithm, due to the categorical variables) since this approach seemed to work best given the data we had.

Finally, it is important to note that this algorithm allows the analyst to specify how many groups you want to classify participants into. We decided on two groups because diagnostic analysis showed that as you increase the number of clusters, the within-cluster differences stop having large rates of decrease after two clusters. In order to be parsimonious and improve interpretation we therefore opted to use only two clusters: 'relatively high' and 'relatively low' SES.

Stata 15.1 (StataCorp, USA) and R version 3.5.0 were used to perform the analyses.[4] Specifically, the package 'klaR' was used for the cluster analysis.[5]

References

1. Rutstein SO, Johnson K, Macro ORC, Measure/Dhs, United States Agency for International Development. The DHS wealth index. 2004.
2. Crush J, Frayne B. Surviving on the Move: Migration, Poverty and Development in Southern Africa. 2010.

3. Huang Z. A fast clustering algorithm to cluster very large categorical data sets in data mining. In: Lu H, Matoda H, Luu H, eds. KDD: Techniques and Applications. Singapore: World Scientific, 1997: 21-34.
4. R Core Team. A language and environment for statistical computing. Vienna, Austria, 2018. https://www.R-project.org/
5. Weihs C, Ligges U, Luebke K, Raabe N. klaR analyzing German business cycles. In: Baier D, Decker R, Schmidt-Thieme L, eds. Data Analysis and Decision Support. Berlin: Springer-Verlag, 2005:334-343.