# Guide the OHS-GHS CEW Series

Amy Thornton Martin Wittenberg

February 2022

## 1 Introduction

The OHS-GHS CEW series consists of a set of survey weights constructed by Amy Thornton and Martin Wittenberg for the OHSs 1995-1999 and GHSs 2002-2011. The weights were constructed in order to address known weaknesses with the weights in the StatsSA public releases of the same data sets. These weaknesses are described in detail in Thornton & Wittenberg (2021) and Thornton (2021), along with the calibration procedure for our new weights. The new weights stop in 2011 because this was the latest year for which we were able to obtain design weights for the GHS from StatsSA, which we use in our calibration procedure. We hope to update the time period eventually. The data set we currently provide consists of:

1. **hhid:** original StatsSA unique household identifiers for the OHS and GHS
2. **year:** a year variable ranging from 1995-2011
3. **cewgt:** the newly constructed cross-entropy weight

Researchers can use these variables to merge the weight into a wave or series of waves of the OHS or GHS between 1995 and 2011. The CEW weight is integrated, meaning it is the same for every member of the household and the OHS-GHS CEW series has therefore been released as a household-level data set. Researchers can merge 1:1 in Stata into household files of the OHS or GHS using the year and hhid variable. If one is using a person file of the OHS or GHS, then researchers can merge m:1 in Stata on the same variables. The OHS hhid variables have been converted to strings in the OHS-GHS CEW series, although the OHS mainly uses numeric hhid variables. Researchers will either need to convert the OHS hhid variables to strings, or the OHS-GHS CEW hhid variables to numeric variables (for the OHS years) in order to merge. The merges should match completely in most years. However in the few cases where merges are not complete, this happens because these observations partly or totally lacked the demographic information we needed to run our procedure (e.g. sex, race, province, age, household headship status, original StatsSA calibrated weight). The OHS and GHS surveys can be downloaded from DataFirst's website.

## 2 Problems with the existing StatsSA weights

Thornton & Wittenberg (2021) and Thornton (2021) describe issues with the existing StatsSA weights in detail. Very briefly, the three main issues are:

1. **Representativeness:** StatsSA calibrate two weights for both the OHS and GHS series, a person weight released in the person file and a household weight released in the household file. These weights are calibrated on completely different information and are never equal to each other. The person weight is calibrated on person information (age, sex, race, province), and the household weight is calibrated on similar information for the household head only. This procedure is at odds with how the surveys are collected. The sampling design implies the data should be released with a single integrated weight: the household weight should equal the person weight which should be equal within households. By breaking with this, StatsSA's two-weight system creates conceptual problems because researchers have to choose between one weight that yields a representative household universe, and one that yields a representative person universe, but not both at the same time even though the weights apply to the same sample. This system also creates empirical problems because researchers can extract multiple estimates of the same statistic (e.g. two household counts for the same year). It is also unclear which weight applies to statistics that combine person- and household-information (e.g. per capita household income).

2. **Household counts:** A series of the total number of households in the country per year cannot be reliably extracted from a stacked series of the OHS and GHS for different reasons. The OHS household weights were not benchmarked to a total household count, so the series of counts jumps around unrealistically in this period. The GHS household weights were benchmarked to household counts using the 2001 and 2011 censuses and the 2007 Community Survey. In this case, inconsistent treatment of the worker hostel sub-population plus distortionary effects of the 2007 Community Survey has resulted in household count estimates that are too low, at a rate of about 5% per year by our calculations.

3. **Small households:** With each new Master Sample in the OHS and GHS series, coverage of small households worsened (after healthy levels of coverage at the beginning of the GHS). And, the existing StatsSA weights make no compensation for this undersampling. The household weights cannot be relied on to provide an accurate trend in the numbers of single-, two-, and three-person households, and by 2011 underestimate the number of single-person households by 25% compared to the census. Since households are the unit that are sampled, if certain types of households are systematically missed this can bias other estimates. Missing small households is also a problem as the size of South African households shrinks over time and small households become more important as a topic in social research agendas.

# 3 Advantages of the CEW

The CEW series of weights was therefore created specifically to address these three issues. The CEW series combines the information StatsSA use in two separate calibration procedures into one procedure to create a single integrated weight. This solves many of the issues created by the two-weight system because the CEW is representative of both households and the people who comprise them at the same time. The CEW is calibrated using both person and household auxiliary information meaning both person and household counts are benchmarked. We have also improved the quality of the series of household count benchmarks that went into the calibration. As a result, household counts using the CEW are slightly higher than using the GHS household weight, and are more in line with the census. Additionally, the CEW is benchmarked on counts of one-, two-, and three-person households enabling researchers to extract reliable trends for small households. Advantages of the CEW weights are:

1. A consistent series of person and household counts benchmarked on both person and household auxiliary information

2. The benchmarks for the household counts include worker hostels in all census years, and omit the 2007 Community Survey which is clearly out of step with the census trends

3. A benchmarked series of total household counts for the OHS era for the first time

4. Benchmarked series of counts for one-, two-, and three-person households[1]

5. An internally consistent weight to use for statistics that combine person- and household-level information, e.g. household size, per capita household income

## References

Machemedze, T., Kerr, A. & Wittenberg, M. (2007), Recalibrating the OHSs to adjust for sampling changes, DataFirst Technical Paper No. 28, DataFirst, University of Cape Town, South Africa.

Thornton, A. (2021), Household formation in post-apartheid South Africa, 1995-2011: measurement and trends, PhD thesis, School of Economics, University of Cape Town, Cape Town, South Africa.

Thornton, A. & Wittenberg, M. (2021), Reweighting the OHS and GHS to improve data quality: representativeness, household counts, and small households, SALDRU Working Paper No. 283, Southern African Labour and Development Research Unit, University of Cape Town, South Africa.

[1] Machemedze et al. (2007) previously supplied such for the OHS era, this series includes the OHS and the GHS era until 2011.