Documentation of HAALSI Wave 3 Sample Weights

Report prepared by

David Kapaon, Data Manager for HAALSI

Harvard Center for Population and Development Studies
T.H. Chan School of Public Health
Cambridge, MA

April 2023

Table of Contents

Introduction and Overview

We constructed sample weights to adjust for non-response due to mortality within the cohort, as well as non-mortality-related attrition from Wave 3 of HAALSI. We also created weights to adjust for the non-trivial amount of data lost from the anthropometric and point of care measurements due to respondent non-consent. All weights are weighted back to the baseline wave. In constructing the weights, HAALSI team members met to discuss the different options and approaches. The team agreed on inverse probability weights, similar to those constructed for Wave 2. While we did not manually multiply any variables together to create combined weights for Wave 3, the Wave 3 weights can be multiplied together to depending on which missingness needs to be accounted for in a given analysis.

The final agreed upon models and weight creation are described below, in turn. We present the models used to create the inverse probably weights in the Appendix, along with each model's respective C-statistic (concordance statistic, or C-index). The C-statistic is equal to the area under a ROC curve, and is a measure of goodness of fit for binary outcomes in a logistic regression model.

We also tested for some key interactions in the models (Age x Sex; Age x Grip Strength; Age x Cognition; Sex x Grip Strength; Sex x Cognition; Cognition x Grip Strength), and included those with significant coefficents.

Mortality Weight

Variable name: w3_ipw_mortality

We created an inverse probability weight to account for mortality between waves 1 and 3 using the inverse of the predicted value of survival. This weight is intended to adjust for nonresponse due to respondent death in models that rely on wave 3 data. The full logistic regression model that was used to predict survival is shown in Table A1; the weight variable is the inverse of this resulting probability.

The following is a list of the variables included in the model, with a description and brief justification for inclusion of each. When mean/mode imputation is noted to have been used, this indicates that the mean/mode value for respondents with non-missing data was imputed for respondents who has missing data on that item. Note that we considered truncating the mortality weight at the 99th percentile, but we did not do so because there were no extreme outlying individual weights.

Sex. This variable is a binary measure to indicate respondent's sex (1=Male, 2=Female). This was included to account for the gender differences in mortality risk.

Age at wave 1 (continuous). This variable is a measure of continuous age (precise to the year) of the respondent as of wave 1. This measure was included because older people were less likely to have survived to wave 3.

Born in South Africa. This binary variable indicates whether or not the respondent was born inside of South Africa, to account for in-migration of HAALSI respondents to Agincourt (largely from Mozambique). It is coded as 1 if the respondent was born in South Africa, and 0 if they were not born in South Africa. Mode imputation was used to impute missing information for 5 respondents. Respondents born outside South Africa were expected to be disadvantaged relative to those born inside South Africa, leading to differing mortality risks.

Years of education at wave 1. This variable indicates total years of education reported by the respondent. Missing data in HAALSI for 17 respondents was filled in using their information from the 2017 Agincourt census. Education was included because we expected more educated people to face lower mortality risk.

Literacy at wave 1. This variable indicates whether the respondent could both read and write (coded as 1), or could not do either (coded as 0). Mode imputation was used to impute missing information for 3 respondents. Similar to education, literacy was included due to higher risk of mortality among illiterate respondents.

Marital status at wave 1. The model included a series of dummy variables to indicate respondents' marital status at wave 1: never married, currently married or living with a partner, separated/deserted, divorced, or widowed. Mode imputation was used to impute missing information for 4 respondents. Marital status was included in the model to account for the disadvantage that non-married groups face in survival, relative to being married.

Employment status at wave 1. The model included a dummy variable to indicate respondents' primary employment status at wave 1: employed (full or part time), or not working/homemaker. Mode imputation was used to impute missing information for 14 respondents. Employment status was included because of the expectation that employment status is associated with risk of health outcomes and mortality.

Household consumption quintiles at wave 1. The model included a series of dummy variables to indicate quintiles of total household consumption per capita. Household consumption was calculated by aggregating data on regular and nonregular expenditures of different food products, goods, and services, as well as home-based food production. Consumption captures what was consumed by the household, whether it is produced or purchased, and tends to be an appropriate measure of living standards in developing countries (Deaton, 2000). Consumption was also chosen, rather than current household income, because it represents the living standard of a household and accounts for inter–temporal cash transfers. It can be regarded as a measure of long run or permanent income for the household if it smooths consumption over short run income shocks (Deaton, 2000). In addition, the household income data in HAALSI has missing values in many cases for the labor income of household members who are not the financial respondent, while the consumption data is more complete. The level of consumption was scaled according to the number of people living in the household (Hentschel and Lanjouw 1996). Households were then categorized into quintiles of household consumption. This measure was included in the model to create the mortality weight because it tends to be predictive of health in this setting (Riumallo-Herl et al., 2019).

Total cognitive score at wave 1. The model included a measure of cognition that is coded to mirror the Health and Retirement Study (HRS) total cognitive score as closely as possible. This measure ranges from 0 to a possible total of 26 points. Cognitive tests that were summed to create this measure included (1) orientation (correct reporting of the day, month, year, and current president; up to 4 points), immediate word recall of 10 words (up to 10 points), delayed recall of the same 10 words (up to 10 points), counting correctly from 1 to 20 (up to 1 point), and correct response to a numerical patterning item (up to 1 point). Participants who could not count to 20 were automatically assigned zero points on the numerical patterning test. Proxy respondents received a code of 0 because the need for a proxy interview implies impairment of the respondent. Mean imputation was used to impute missing information for 44 respondents. Cognitive function is an indicator of dementia-related disease and was included due to its association with mortality risk.

CESD-8 depression scale at wave 1. This variable indicates the 8-item CES-D depression scale, ranging from 0 to 8 possible depressive symptoms. Mean imputation was used to impute missing information for 14 respondents. The CES-D is an indicator of mental health, and was included due to its association with mortality risk.

Grip strength at wave 1. This variable indicates the respondent's grip strength, coded as the highest value (or maximum grip strength) of up to four possible measurements taken (if they did not report having surgery in the last 3 months, arthritis, or pain on each respective arm, hand, or wrist, then two measurements were taken per hand). Mean imputation was used to impute missing information for 208 respondents. Grip strength is an important measure of physical ability, and was included due to its association with mortality risk.

Average walk time at wave 1. This variable indicates the time that it took respondent to walk 2.5 meters. Respondents were asked to walk 2.5 meters twice, and this variable is the average of those two measurements. If the either of the two measurements were missing, the non-missing measure was used. Mean imputation was used to impute missing information for 221 respondents. Walk time is an important measure of physical health and was included due to its association with mortality risk.

HIV positive at wave 1. The model included a variable to indicate whether the respondent was HIV positive, as confirmed through dried blood spots. Respondents with indeterminate blood tests or missing information due to non-consent to blood at wave 1 (n=499) were coded as zero. A second measure was included to indicate missing on HIV, with those missing cases coded as 1 and non-missing coded as 0. These measures were included due to the importance of HIV, in this setting, in predicting mortality risk.

HIV viral load at wave 1. This variable indicates respondent's HIV viral load. Categories of viral load included (1) 0, (2) <100, (3) 100-400, (4) 400-1000, (5) 1000-10,000, or (6) >10,000 copies/mL. Mean imputation was used to impute missing information for 485 respondents. HIV viral load was included because of its importance in predicting mortality risk in this setting – those with a greater value on viral load are at great risk of mortality due to HIV.

Proxy interview at wave 1. This variable indicates whether or not the interview was completed by a proxy respondent. Respondents who required a proxy were more likely to be in poor health, and therefore less likely to survive to wave 3.

Attrition Weight

Variable name: w3_ipw_attrit

We also created an inverse probability weight to account for attrition due to reasons other than mortality between waves 1 and 3 (i.e., refusal or incomplete interviews, or inability to be found for contact). This weight is intended to be used to adjust for this nonresponse in models that rely on wave 3 data. The full logistic regression model that was used to predict non-attrition is shown in Table A2; the weight variable is the inverse of this resulting probability.

The following is a list of the variables included in the model, with a description and brief justification for inclusion of each. When mean/mode imputation is noted to have been used, this indicates that the mean/mode value of for respondents without missing data was imputed for that item. Many of the same measures were used in this model as in the model used to create the mortality weight. Note that we considered truncating the attrition weight at the 99th percentile, but we did not do so because there were no extreme outlying individual weights.

Sex. This variable is a binary measure to indicate respondent's sex (1=Male, 2=Female). This is included to account for possible gender differences in attrition risk.

Age at wave 1 (categorical). This variable is a series of dummy measures to indicate decade of respondent's age as of wave 1: 40-49, 50-59, 60-69, 70-79, or 80 plus. Age is included due to its expected association with willingness to participate in a second wave of data collection.

Born in South Africa. This binary variable indicates whether or not the respondent was born inside of South Africa, to account for in-migration of HAALSI respondents to Agincourt (largely from Mozambique). It is coded as 1 if the respondent was born in South Africa, and 0 if they were not born in South Africa. Mode imputation was used to impute missing information for 3 respondents. Respondents born outside South Africa were expected to be face greater attrition risk.

Years of education at wave 1. This variable indicates total years of education reported by the respondent. Missing data in HAALSI for 12 respondents was filled in using their information from the 2017 Agincourt census. Education was included due to its expected association with attrition risk.

Literacy at wave 1. This variable indicates whether the respondent could both read and write (coded as 1), or could not do either (coded as 0). Mode imputation was used to impute missing information for 2 respondents. Similar to education, literacy was included due to the different expected risk of attrition by literacy status.

Marital status at wave 1. The model included a series of dummy variables to indicate respondents' marital status at wave 1: never married, currently married or living with a partner, separated/deserted, divorced, or widowed. Mode imputation was used to impute missing information for 3 respondents. Marital status was included because of expected differences in attrition risk across these groups.

Employment status at wave 1. The model included a dummy variable to indicate respondents' primary employment status at wave 1: employed (full or part time), or not working/homemaker. Mode imputation was used to impute missing information for 12 respondents. Employment status is included because of its possible association with attrition risk.

Household consumption quintiles at wave 1. The model included a series of dummy variables to indicate quintiles of total household consumption per capita. Household consumption was calculated by aggregating data on regular and nonregular expenditures of different food products, goods, and services, as well as home-based food production. Consumption captures what was consumed by the household, whether it is produced or purchased, and tends to be an appropriate measure of living standards in developing countries (Deaton, 2000). Consumption was also chosen, rather than current household income, because it represents the living standard of a household and accounts for inter–temporal cash transfers. It can be regarded as a measure of long run or permanent income for the household if it smooths consumption over short run income shocks (Deaton, 2000). In addition, the household income data in HAALSI has missing values in many cases for the labor income of household members who are not the financial respondent, while the consumption data is more complete. The level of consumption was scaled according to the number of people living in the household (Hentschel and Lanjouw 1996). Households were then categorized into quintiles of household consumption. This measure was included in the model to create the attrition weight because of its expected association with attrition risk (Riumallo-Herl et al., 2019).

Total cognitive score at wave 1. The model included a measure of cognition that is coded to mirror the Health and Retirement Study (HRS) total cognitive score as closely as possible. This measure ranges from 0 to a possible total of 26 points. Cognitive tests that were summed to create this measure included (1) orientation (correct reporting of the day, month, year, and current president; up to 4 points), immediate word recall of 10 words (up to 10 points), delayed recall of the same 10 words (up to 10 points), counting correctly from 1 to 20 (up to 1 point), and correct response to a numerical patterning item (up to 1 point). Participants who could not count to 20 were automatically assigned zero points on the numerical patterning test. Proxy respondents received a code of 0 because the need for a proxy interview implies impairment of the respondent. Mean imputation was used to impute missing information for 38 respondents. Cognitive function is included due to the expectation that respondents with greater impairment may be more likely to refuse to participate.

Proxy interview at wave 1. This variable indicates whether or not the interview was completed by a proxy respondent. Respondents who required a proxy at wave 1 are less likely to participate in a second wave of data, as those respondents were more likely to have had a physical or mental impairment that increases their chance of refusing to participate.

Migration status at wave 3. The model includes a measure to indicate whether the respondent was a migrant at wave 3. This measure includes both internal migrants (i.e., individuals who moved out of Agincourt but still remained in the province of Mpumalanga) and external migrants (i.e., individuals who had moved outside of Mpumalanga, either to another part of South Africa, or to a different country). This measure was included because migrants are more likely to attrite from follow-up surveys, as they are more difficult to find/contact.

Participation in AWI-Gen. The model includes a measure capturing whether the respondent participated in the AWI-Gen study, which went into the field for a follow-up wave just prior to HAALSI entering the field for wave 3. This measure was included because we expected that respondents who had recently participated in AWI-Gen would feel greater burden and be more likely to refuse to participate in HAALSI wave 3.

Month of first contact at wave 3. This variable indicates the month of first contact made with either the respondent’s household (in most cases) or the individual respondent, in the fieldworker’s attempt to administer the interview. Some months were combined due to the small number of first contacts made during individual months (e.g., July and August). This measure was included because respondents’ availability and willingness to participate in the survey may be dependent on the time of year that the fieldworker first made the request. Fieldworkers made several contact attempts, if prior attempts did not result in a refusal to participate. We use information from the first attempt because it is assumed to have had the most important impact on the final interview outcome.

Time of day of first contact at wave 3. This variable indicates the time of day of first contact made with either the respondent’s household (in most cases) or the individual respondent, in the fieldworker’s attempt to administer the interview. The measure indicates whether the first contact was made in the morning or the afternoon. This measure is included because respondents' availability and willingness to participate in the survey may be dependent on the time of day that the fieldworker first made the request. As above, fieldworkers made several contact attempts, and we use information from the first attempt because it is assumed to have had the most important impact on the final interview outcome.

Missing and Non-Consents to Measurements

We also created two inverse probability weights to account for missing data due to respondent non-consent to anthropometric (height and weight) or point of care blood measurements (glucose, hemoglobin, and cholesterol) during the wave 3 survey interview. The next two sections of this document focus on those weights, with a separate section dedicated to each.

At the start of the wave 3 survey, respondents were asked whether or not they would consent to several measurements administered by fieldworkers later in the interview. Respondents could choose to either consent to all anthropometric (i.e., height and weight) and point of care measurements, or none of them. This differed from the wave 2 survey, where each participants was presented with a list and respondents could either choose to consent, or not, to each individual measurement. In wave 3, since respondents were not explicitly asked to consent to have their height and weight measured, here we created a weight for missingness on these anthropemetric measures. For point of care tests we created a weight to for non-consent to these measurements, which included: drops of blood for glucose, drops of blood for hemoglobin, drops of blood for cholesterol, and drops of blood for HIV.

The following sections will go into more detail about each of these two weights.

Anthropometric Weight
Variable name: w3_ipw_anthropometric

We created an inverse probability weight in order for researchers to be able to adjust for missingness on two anthropometric measurements in wave 3. This weight is based on respondents missing values of height or weight, often due to not agreeing to have these measurements taken. It is intended to be used in models that rely on wave 3 data of these anthropometric measurements, since the models would otherwise be biased due to the large amount of missingness. The full logistic regression model that was used to predict consent to anthropometric measurements is shown in Table A5; the weight variable is the inverse of this resulting probability.

As stated above, the weight is based on whether or not respondents had missing values on at least one of the following measurements: height and weight. Although blood pressure was not included in the coding of this weight for wave three, due to high correlations between blood pressure and height/weight, this weight could also be used to account for missingness on that measure as well.

The following is a list of the variables included in the model, with a description and brief justification for inclusion of each. When mean/mode imputation is noted to have been used, this indicates that the mean/mode value for respondents with non-missing data was imputed for respondents who has missing data on that item. Note that we considered truncating the anthropometric weight at the 99th percentile, but we did not do so because there were no extreme outlying individual weights.

Sex. This variable is a binary measure to indicate respondent's sex (1=Male, 2=Female). This was included to account for the gender differences in likelihood of agreeing to anthropometric measurements.

Age at wave 1 (categorical). This variable is a series of dummy measures to indicate decade of respondent's age as of wave 1: 40-49, 50-59, 60-69, 70-79, or 80 plus. Age is included due to its expected association with willingness to have anthropometric measurements taken.

Years of education at wave 1. This variable indicates total years of education reported by the respondent. Missing data in HAALSI for 10 respondents was filled in using their information from the 2017 Agincourt census. Education was included in the model due to its expected association with likelihood of missing anthropometric measurements.

Literacy at wave 1. The model includes a measure to indicate respondents' literacy, coded as 1 if they could both read and write and 0 if they could not do either or could only read OR write. Mode imputation was used to impute missing information for 2 respondents. Literacy was included because the odds of agreeing to anthropometric measurements was expected to differ by literacy status.

Marital status at wave 1. The model included a series of dummy variables to indicate respondents' marital status at wave 1: never married, currently married or living with a partner, separated/deserted, divorced, or widowed. Mode imputation was used to impute missing information for 3 respondents. Marital status was included in the model because of expected differenced in likelihood of missing anthropometric measurements across these groups.

Employment status at wave 1. The model included a dummy variable to indicate respondents' primary employment status at wave 1: employed (full or part time), or not working/homemaker. Mode imputation was used to impute missing information for 11 respondents. Employment status is included because of its possible association with likelihood of missing anthropometric measurements.

Household consumption quintiles at wave 1. The model included a series of dummy variables to indicate quintiles of total household consumption per capita. Household consumption was calculated by aggregating data on regular and nonregular expenditures of different food products, goods, and services, as well as home-based food production. Consumption captures what was consumed by the household, whether it is produced or purchased, and tends to be an appropriate measure of living standards in developing countries (Deaton, 2000). Consumption was also chosen, rather than current household income, because it represents the living standard of a household and accounts for inter–temporal cash transfers. It can be regarded as a measure of long run or permanent income for the household if it smooths consumption over short run income shocks (Deaton, 2000). In addition, the household income data in HAALSI has missing values in many cases for the labor income of household members who are not the financial respondent, while the consumption data is more complete. The level of consumption was scaled according to the number of people living in the household (Hentschel and Lanjouw 1996). Households were then categorized into quintiles of household consumption. This measure was included in the model to create the anthropometric weight because people with different standards of living may have different likelihood of agreeing to anthropometric measurements (Riumallo-Herl et al., 2019).

Total cognitive score at wave 1. The model included a measure of cognition that is coded to mirror the Health and Retirement Study (HRS) total cognitive score as closely as possible. This measure ranges from 0 to a possible total of 26 points. Cognitive tests that were summed to create this measure included (1) orientation (correct reporting of the day, month, year, and current president; up to 4 points), immediate word recall of 10 words (up to 10 points), delayed recall of the same 10 words (up to 10 points), counting correctly from 1 to 20 (up to 1 point), and correct response to a numerical patterning item (up to 1 point). Participants who could not count to 20 were automatically assigned zero points on the numerical patterning test. Proxy respondents received a code of 0 because the need for a proxy interview implies impairment of the respondent. Mean imputation was used to impute missing information for 36 respondents. We included this measure due to expected differences in willingness to have anthropometric measurements taken based on cognitive status.

Obese at wave 1. This variable indicates whether the respondent is obese, as defined by BMI classifications: respondents with a BMI of 30 or higher received a code of 1 on this measure, and 0 otherwise. Mode imputation was used to impute missing information for 196 respondents. We included this measure because of the expectation that respondents who were obese at wave 1 may have been less willing having their weight measured at wave 3.

Hypertension at wave 1. This variable to indicate hypertension is based on a combination of self-reports and lab measurements. A respondent was considered hypertensive if their systolic blood pressure was greater than or equal to 150 mmHg, or their diastolic blood pressure was greater than or equal to 90 mmHg, or they reported using anti-hypertensive medication at the time of the wave 1 interview. Mode imputation was used to impute missing information for 62 respondents. This measure was included to account for the expectation that respondents with elevated blood pressure at wave 1 may have been less willing to having their blood pressure measured at wave 3.

Proxy interview at wave 1. This variable indicates whether or not the interview was completed by a proxy respondent. This measure was included due to expected an expected association between proxy interview status and missing anthropometric measurements.

Participation in AWI-Gen. The model includes a measure capturing whether the respondent participated in the AWI-Gen study, which went into the field for a follow-up wave just prior to HAALSI entering the field for wave 3. This measure was included because we expected that respondents who had recently participated in AWI-Gen would feel greater burden and be less willing to have their anthropometric measures taken.

Month of first contact at wave 3. This variable indicates the month of first contact made with either the respondent's household (in most cases) or the individual respondent, in the fieldworker's attempt to administer the interview. Some months were combined due to the small number of first contacts made during individual months (e.g., July and August). This measure was included because respondents' willingness have their anthropometric measurements taken may be dependent on the time of year that the fieldworker first made the request.

Time of day of first contact at wave 3. This variable indicates the time of day of first contact made with either the respondent's household (in most cases) or the individual respondent, in the fieldworker's attempt to administer the interview. The measure indicates whether the first contact was made in the morning or the afternoon. This measure is included because respondents' willingness to have their anthropometric measurements taken may be dependent on the time of day that the fieldworker first made the request.

Interviewer at wave 3. This series of dummy variables indicates the ID of the fieldworker who conducted the interview. Due to a small number of interviews conducted by some fieldworkers, some fieldworker IDs (IDs 30 + 50) were combined into a single category. We treat ID 29 as the reference category. This measure was included because respondent likeliness to consent to point of case measurements may have depended upon the fieldworker/interviewer with whom they were interacting.

Point of Care Weight
Variable name: w3_ipw_point_of_care

We created an inverse probability weight in order for researchers to be able to adjust for missingness due to respondent non-consent to several point of care measurements in wave 3, similar to the anthropometric weight described above. This weight is based on respondent non-consent to the point of care tests (glucose, hemoglobin, and cholesterol). Similarly, Although dried blood spots were not included in the coding of this weight for wave three, due to high correlations between dried blood spots and the other point of care tests, this weight could also be used to account for missingness on that measure as well. It is intended to be used in models that rely on wave 3 data of these measurements, since the models would otherwise be biased due to the large amount of missingness. The full logistic regression model that was used to predict consent to point of care measurements is shown in Table A6; the weight variable is the inverse of this resulting probability.

The following is a list of the variables included in the model, with a description of each variable as well as a brief justification for inclusion of each. When mean/mode imputation is noted to have been used in creating a variable, this indicates that the mean/mode value for respondents or respondents with non-missing data was imputed for respondents who has missing data on that item. Note that we considered truncating the biomarker weight at the 99th percentile, but we did not do so because there were no extreme outlying individual weights.

Sex. This variable is a binary measure to indicate respondent's sex (1=Male, 2=Female). This is included to account for likely gender differences in consent to point of care measurements.

Age at Wave 1 (categorical). This variable is a series of dummy measures to indicate decade of respondent's age as of wave 1: 40-49, 50-59, 60-69, 70-79, or 80 plus. Age is included due to its expected association with willingness to consent to point of care measurements.

Years of education at wave 1. This variable indicates total years of education reported by the respondent. Missing data in HAALSI for 10 respondents was filled in using their information from the 2017 Agincourt census. Education was included in the model due to its expected association with likelihood of consenting to point of care measurements.

Literacy at wave 1. The model includes a measure to indicate respondents' literacy, coded as 1 if they could both read and write and 0 if they could not do either or could only read OR write. Mode imputation was used to impute missing information for 2 respondents. Literacy was included because the odds of not consenting to point of care measurements was expected to differ by literacy status.

Marital status at wave 1. The model included a series of dummy variables to indicate respondents' marital status at wave 1: never married, currently married or living with a partner, separated/deserted, divorced, or widowed. Mode imputation was used to impute missing information for 3 respondents. Marital status was included in the model because of expected differenced in likelihood of consenting to point of care measurements across these groups.

Employment status at wave 1. The model included a dummy variable to indicate respondents' primary employment status at wave 1: employed (full or part time), or not working/homemaker. Mode imputation was used to impute missing information for 11 respondents. Employment status is included because of its possible association with likelihood of missing point of care measurements.

Household consumption quintiles at wave 1. The model included a series of dummy variables to indicate quintiles of total household consumption per capita. Household consumption was calculated by aggregating data on regular and nonregular expenditures of different food products, goods, and services, as well as home-based food production. Consumption captures what was consumed by the household, whether it is produced or purchased, and tends to be an appropriate measure of living standards in developing countries (Deaton, 2000). Consumption was also chosen, rather than current household income, because it represents the living standard of a household and accounts for inter–temporal cash transfers. It can be regarded as a measure of long run or permanent income for the household if it smooths consumption over short run income shocks (Deaton, 2000). In addition, the household income data in HAALSI has missing values in many cases for the labor income of household members who are not the financial respondent, while the consumption data is more complete. The level of consumption was scaled according to the number of people living in the household (Hentschel and Lanjouw 1996). Households were then categorized into quintiles of household consumption. This measure was included in the model to create the point of care weight because people with different standards of living may have different likelihood of consenting to point of care measurements (Riumallo-Herl et al., 2019).

Total cognitive score at wave 1. The model included a measure of cognition that is coded to mirror the Health and Retirement Study (HRS) total cognitive score as closely as possible. This measure ranges from 0 to a possible total of 26 points. Cognitive tests that were summed to create this measure included (1) orientation (correct reporting of the day, month, year, and current president; up to 4 points), immediate word recall of 10 words (up to 10 points), delayed recall of the same 10 words (up to 10 points), counting correctly from 1 to 20 (up to 1 point), and correct response to a numerical patterning item (up to 1 point). Participants who could not count to 20 were automatically assigned zero points on the numerical patterning test. Proxy respondents received a code of 0 because the need for a proxy interview implies impairment of the respondent. Mean imputation was used to impute missing information for 36 respondents. We included this measure due to expected differences in willingness to consent to point of care measurements by cognitive status.

Anemic at wave 1. This variable indicates respondent's anemia status, and is based on official South African measurements (Shisana et al., 2013). Male respondents were considered to be anemic if they had hemoglobin levels less than 12.9 g/dl, and females less than 11.9 g/dl. Mode imputation was used to impute missing information for 426 respondents. It was included because of the expectation that respondents who were anemic at wave 1 may have been less likely to consent to having their anemia status measured at wave 3.

Diabetic at wave 1. This binary variable indicates respondent's diabetes status based on a combination of self-reports and glucose from blood. A respondent is considered to be diabetic if they self reported diabetes treatment (cm007_males, cm007_females); or had a glucose measurement that met the threshold (≥ 7 mmol/l [126 mg/dL] in fasting group [defined as > 8 hours], and ≥11.1 mmol/l [200 mg/dL] in nonfasting ["random or casual"] group). Individuals with missing fasting information were considered to be not fasting. Mode imputation was used to impute missing information for 308 respondents. This measure was included to capture the likelihood that respondents' decision to consent to have their diabetes status measures depended on their status at wave 1.

Proxy interview at wave 1. This variable indicates whether or not the interview was completed by a proxy respondent. This measure was included due to expected an expected association between proxy interview status and consent to point of care measurements.

Participation in AWI-Gen. The model includes a measure capturing whether the respondent participated in the AWI-Gen study, which went into the field for a follow-up wave just prior to HAALSI entering the field for wave 3. This measure was included because we expected that respondents who had recently participated in AWI-Gen would feel greater burden and be less likely to consent to have their point of care measurements taken in HAALSI wave 3.

Month of first contact at wave 3. This variable indicates the month of first contact made with either the respondent's household (in most cases) or the individual respondent, in the fieldworker's attempt to administer the interview. Some months were combined due to the small number of first contacts made during individual months (e.g., July and August). This measure was included because respondents' willingness to consent to point of care measurements may be dependent on the time of year that the fieldworker first made the request.

Time of day of first contact at wave 3. This variable indicates the time of day of first contact made with either the respondent's household (in most cases) or the individual respondent, in the fieldworker's attempt to administer the interview. The measure indicates whether the first contact was made in the morning or the afternoon. This measure is included because respondents' willingness consent to point of care measurements may be dependent on the time of day that the fieldworker first made the request.

Interviewer at wave 3. This series of dummy variables indicates the ID of the fieldworker who conducted the interview. Due to a small number of interviews conducted by some fieldworkers, some fieldworker IDs (IDs 30 + 50) were combined into a single category. We treat ID 29 as the reference category. This measure was included because respondent likeliness to consent to point of case measurements may have depended upon the fieldworker/interviewer with whom they were interacting.

Appendix

Table A1. Logistic regression predicting survival at Wave 3 from Wave 1 characteristics

Observations = 5,059
Pseudo R2 = 0.1680
C-statistic: 0.7731

	Odds Ratio	*SE*	*P value*	*C.I.*
Sex (1=Male, 2=Female)	0.401	0.037	0.000	[0.334,0.481]
Age at wave 1 (continuous)	0.942	0.004	0.000	[0.934,0.949]
Born in South Africa	0.801	0.071	0.012	[0.673,0.952]
Years of education at wave 1	0.985	0.014	0.269	[0.958,1.012]
Literacy at wave 1	1.115	0.118	0.300	[0.907,1.371]
Marital status at wave 1				
Married or living with partner (reference category)	-	-	-	-
Never married	0.616	0.111	0.007	[0.434,0.876]
Separated or deserted	0.884	0.134	0.414	[0.657,1.189]
Divorced	0.584	0.102	0.002	[0.415,0.822]
Widowed	0.866	0.088	0.158	[0.709,1.058]
Employed at wave 1	1.201	0.167	0.187	[0.915,1.576]
Household consumption quintile at wave 1				
Least consumption (reference category)	-	-	-	-
Less consumption	0.978	0.116	0.850	[0.774,1.235]
Middle	0.915	0.109	0.454	[0.724,1.156]
More consumption	0.878	0.105	0.279	[0.694,1.111]
Most consumption	0.991	0.126	0.944	[0.772,1.272]
Total cognitive score at wave 1	1.026	0.010	0.009	[1.006,1.045]
CESD-8 depression scale at wave 1	0.909	0.021	0.000	[0.869,0.950]
Grip strength at wave 1	1.023	0.005	0.000	[1.014,1.032]
Average walk time at wave 1	0.976	0.012	0.047	[0.952,1.000]
HIV positive at wave 1	0.139	0.127	0.03	[0.023,0.829]
Missing on HIV	0.845	0.108	0.188	[0.658,1.085]
HIV viral load at wave 1				
0	-	-	-	-
<100	6.032	5.481	0.048	[1.016,35.801]
100-400	5.943	5.710	0.064	[0.904,39.072]
400-1000	7.910	7.731	0.034	[1.165,53.722]
1000-10,000	5.421	5.088	0.072	[0.861,34.117]
>10,000	2.297	2.130	0.370	[0.373,14.138]
Proxy interview at wave 1	0.231	0.051	0.000	[0.149,0.356]
_cons	241.982	86.842	0.000	[119.758,488.945]

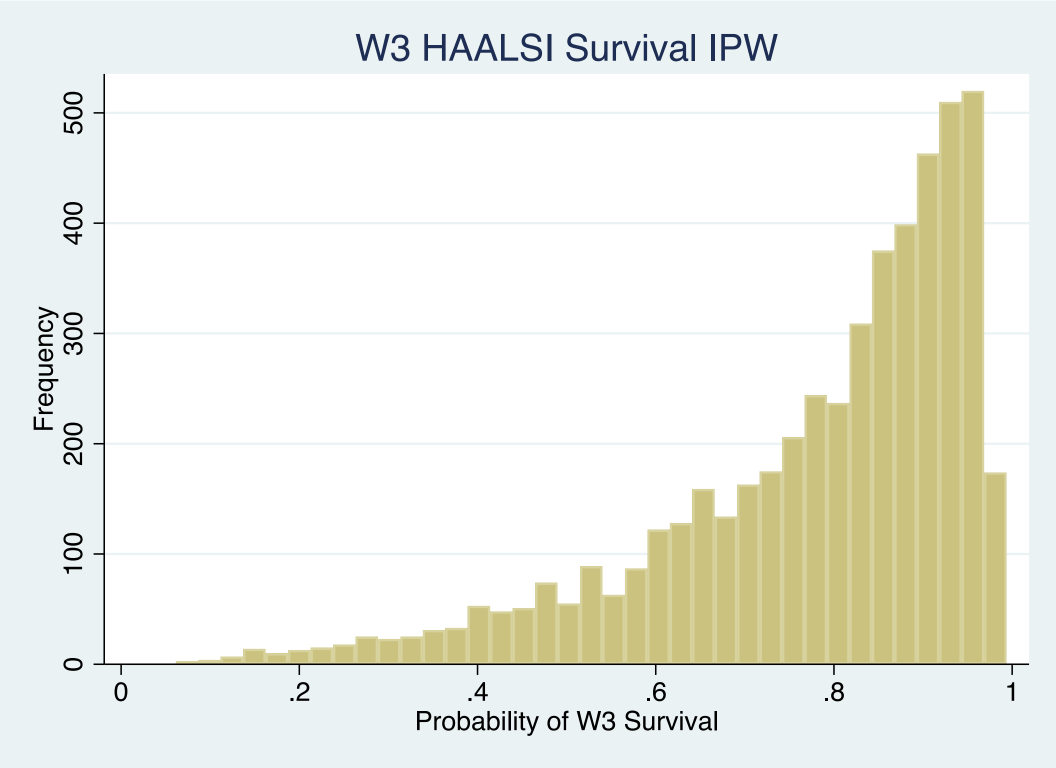

Figure A1. Histogram of the probability of survival, from logistic regression in Table A1.

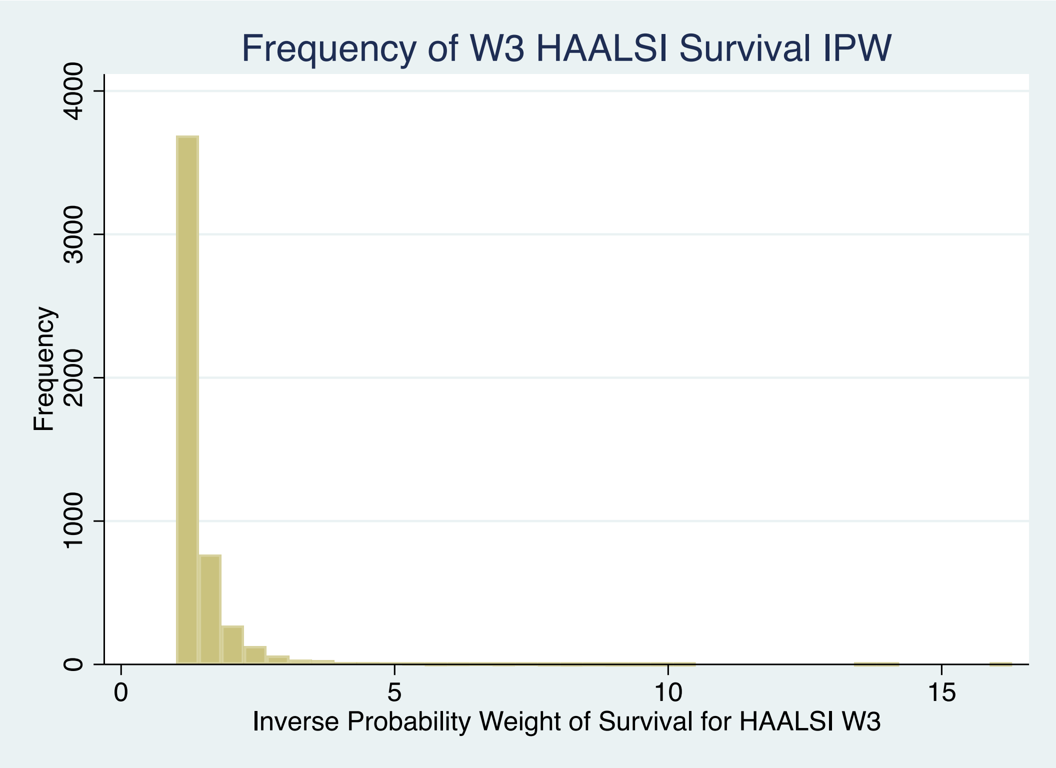

Figure A2. Histogram of mortality weight

Table A2. Logistic regression predicting non-attrition to refusal or not being found at Wave 3

Observations = 3,938
Pseudo R2 = 0.2810
C-statistic: 0.8801

	Odds Ratio	*SE*	*P value*	*C.I.*
Sex (1=Male, 2=Female)	2.474	1.270	0.077	[0.905,6.765]
Total cognitive score at wave 1	1.013	0.027	0.615	[0.962,1.067]
Male X Total cognitive score at wave 1	0.923	0.031	0.019	[0.864,0.987]
Age at wave 1 (categorical)				
40-49 (reference category)	-	-	-	-
50-59	1.180	0.236	0.408	[0.797,1.747]
60-69	1.785	0.446	0.020	[1.094,2.912]
70-79	1.664	0.497	0.088	[0.927,2.988]
80+	1.716	0.700	0.185	[0.772,3.816]
Born in South Africa	1.578	0.305	0.018	[1.080,2.306]
Years of education at wave 1	0.940	0.022	0.008	[0.898,0.984]
Literacy at wave 1	1.535	0.333	0.048	[1.004,2.348]
Marital status at wave 1				
Married or living with partner (reference category)	-	-	-	-
Never married	0.855	0.266	0.616	[0.465,1.575]
Separated or deserted	0.979	0.258	0.937	[0.585,1.641]
Divorced	0.512	0.164	0.037	[0.273,0.961]
Widowed	1.140	0.245	0.543	[0.748,1.738]
Employed status at wave 1	0.643	0.122	0.002	[0.443,0.933]
Household consumption quintile at wave 1				
Least consumption (reference category)	-	-	-	-
Less consumption	0.696	0.166	0.127	[0.436,1.109]
Middle	0.760	0.184	0.258	[0.473,1.222]
More consumption	1.301	0.358	0.339	[0.759,2.229]
Most consumption	0.671	0.164	0.102	[0.416,1.082]
Proxy interview at wave 1	3.218	3.449	0.275	[0.394,26.289]
Migration status at wave 3	0.276	0.070	0.000	[0.168,0.455]
Participation in AWIGEN	0.980	0.155	0.901	[0.719,1.338]
Month of first contact at wave 3				
February + March 2022 (reference category)	-	-	-	-
July + August 2021	0.100	0.041	0.000	[0.045,0.222]
September 2021	22.136	23.783	0.004	[2.695,181.815]
October 2021	5.087	3.567	0.020	[1.287,20.106]
November 2021	4.905	3.968	0.049	[1.005,23.943]
December 2021	0.704	0.351	0.481	[0.265,1.870]
Time of day of first contact at wave 3				
Morning (reference category)	-	-	-	-
Afternoon/Evening	2.256	0.581	0.002	[1.361,3.738]
_cons	36.553	22.085	0.000	[11.185,119.458]

Note: conditional on being alive at wave 3

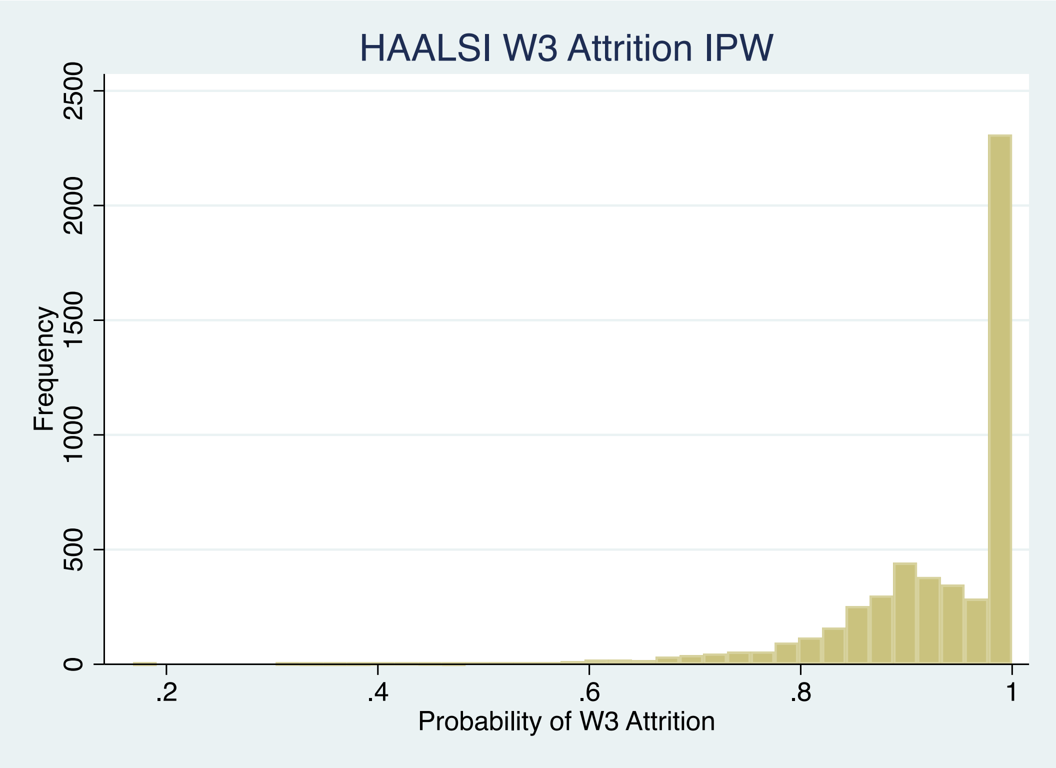

Figure A3. Histogram of the probability of non-attrition, from logistic regression model in Table A2.

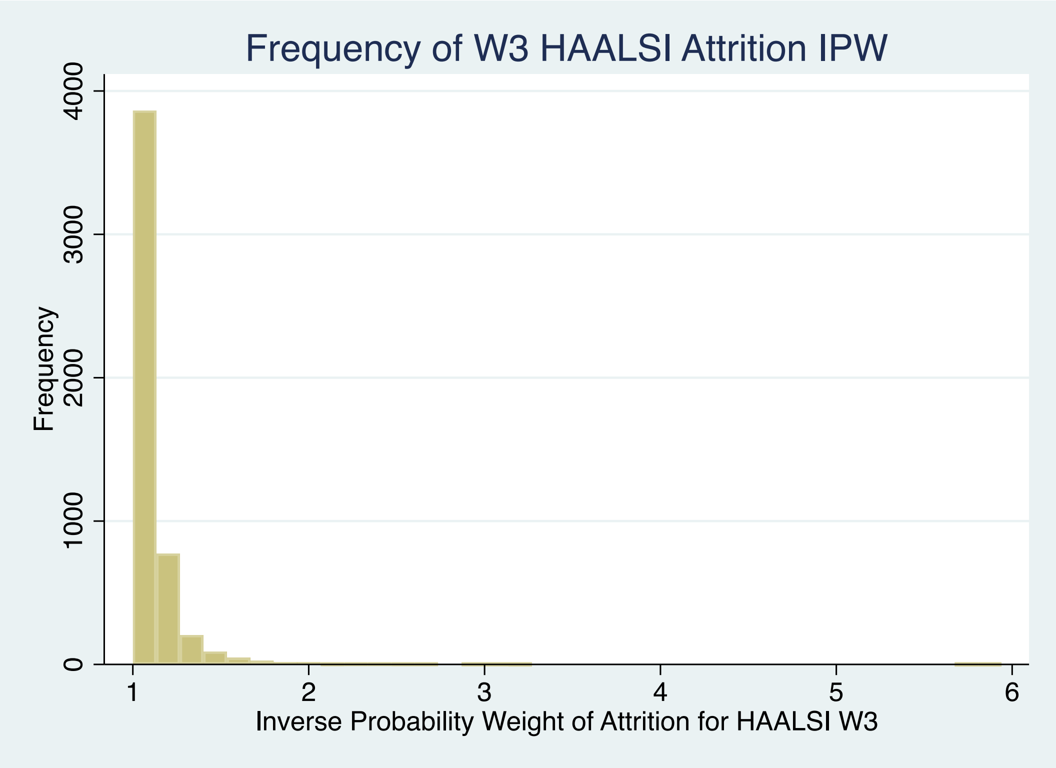

Figure A4. Histogram of attrition weight

Table A5. Logistic regression predicting non-missing on anthropometric measurements

Observations = 3,707
Pseudo R2 = 0.1475
C-statistic = 0.7438

	Odds Ratio	***SE***	***P value***	***C.I.***
Sex (1=Male, 2=Female)	0.477	0.089	0.000	[0.330,0.688]
Age at wave 1 (categorical)				
40-49 (reference category)	-	-	-	-
50-59	0.790	0.136	0.171	[0.564,1.107]
60-69	0.846	0.162	0.383	[0.582,1.231]
70-79	0.627	0.134	0.029	[0.412,0.953]
80+	0.384	0.095	0.000	[0.237,0.624]
Male X Age at wave 1 (categorical)				
Male X 40-49 (reference category)	-	-	-	-
Male X 50-59	1.414	0.336	0.145	[0.887,2.253]
Male X 60-69	1.787	0.446	0.020	[1.096,2.914]
Male X 70-79	2.412	0.674	0.002	[1.395,4.170]
Male X 80+	4.088	1.496	0.000	[1.995,8.377]
Years of education at wave 1	0.983	0.013	0.198	[0.957,1.009]
Literacy at wave 1	1.071	0.122	0.543	[0.858,1.338]
Marital status at wave 1				
Married or living with partner (reference category)	-	-	-	-
Never married	0.592	0.104	0.003	[0.420,0.834]
Separated or deserted	1.003	0.156	0.986	[0.739,1.360]
Divorced	0.983	0.206	0.933	[0.651,1.483]
Widowed	0.921	0.103	0.460	[0.739,1.146]
Employed at wave 1	1.277	0.155	0.044	[1.007,1.619]
Household consumption quintile at wave 1				
Least consumption (reference category)	-	-	-	-
Less consumption	0.996	0.127	0.975	[0.776,1.279]
Middle	1.024	0.133	0.852	[0.795,1.321]
More consumption	0.968	0.126	0.805	[0.750,1.250]
Most consumption	1.149	0.157	0.309	[0.879,1.503]
Total cognitive score at wave 1	1.023	0.011	0.027	[1.003,1.045]
Obese at wave 1	1.144	0.112	0.171	[0.944,1.387]
Hypertension at wave 1	0.822	0.073	0.027	[0.691,0.978]
Proxy interview at wave 1	0.499	0.184	0.060	[0.242,1.029]
Participation in AWIGEN	1.436	0.125	0.000	[1.211,1.703]
Month of first contact at wave 3				
February + March 2022 (reference category)	-	-	-	-
July + August 2021	18.184	4.314	0.000	[11.421,28.950]
September 2021	15.404	3.644	0.000	[9.689,24.489]
October 2021	11.799	2.863	0.000	[7.333,18.986]
November 2021	7.265	1.808	0.000	[4.460,11.833]
December 2021	4.054	1.049	0.000	[2.441,6.732]
Time of day of first contact at wave 3				
Morning (reference category)	-	-	-	-
Afternoon/Evening	0.675	0.070	0.000	[0.550,0.828]
Interviewer at wave 3				
Interviewer 1	0.137	0.041	0.000	[0.076,0.246]
Interviewer 2	0.361	0.113	0.001	[0.196,0.665]
Interviewer 3	0.311	0.096	0.000	[0.170,0.569]
Interviewer 4	1.006	0.355	0.987	[0.503,2.009]
Interviewer 5	0.126	0.038	0.000	[0.070,0.227]
Interviewer 6	0.342	0.107	0.001	[0.185,0.633]
Interviewer 7	0.489	0.164	0.033	[0.254,0.942]
Interviewer 8	0.811	0.274	0.536	[0.418,1.574]
Interviewer 9	0.807	0.274	0.527	[0.415,1.569]
Interviewer 10	0.615	0.203	0.141	[0.322,1.174]

Interviewer 11	0.407	0.123	0.003	[0.224,0.738]
Interviewer 12	0.500	0.158	0.028	[0.270,0.927]
Interviewer 13	2.126	0.850	0.059	[0.970,4.656]
Interviewer 14	0.595	0.218	0.156	[0.290,1.219]
Interviewer 15	0.593	0.182	0.089	[0.325,1.083]
Interviewer 16	1.143	0.423	0.717	[0.554,2.359]
Interviewer 17	0.267	0.087	0.000	[0.140,0.507]
Interviewer 18	0.435	0.135	0.007	[0.237,0.801]
Interviewer 19	0.312	0.094	0.000	[0.173,0.564]
Interviewer 20	0.641	0.208	0.170	[0.339,1.210]
Interviewer 21	0.412	0.127	0.004	[0.225,0.753]
Interviewer 22	0.432	0.132	0.006	[0.238,0.784]

Table A5 continued…

Interviewer 23	0.839	0.280	0.599	[0.437,1.613]
Interviewer 24	0.922	0.312	0.809	[0.475,1.788]
Interviewer 25	0.291	0.088	0.000	[0.161,0.525]
Interviewer 26	0.556	0.177	0.065	[0.298,1.038]
Interviewer 27	0.376	0.125	0.003	[0.196,0.723]
Interviewer 28	0.764	0.241	0.393	[0.411,1.418]
Interviewer 29 (reference category)	-	-	-	-
Interviewer 30 + 50	0.354	0.106	0.001	[0.197,0.636]
_cons	0.556	0.207	0.114	[0.268,1.151]

Note: conditional on participating in wave 3

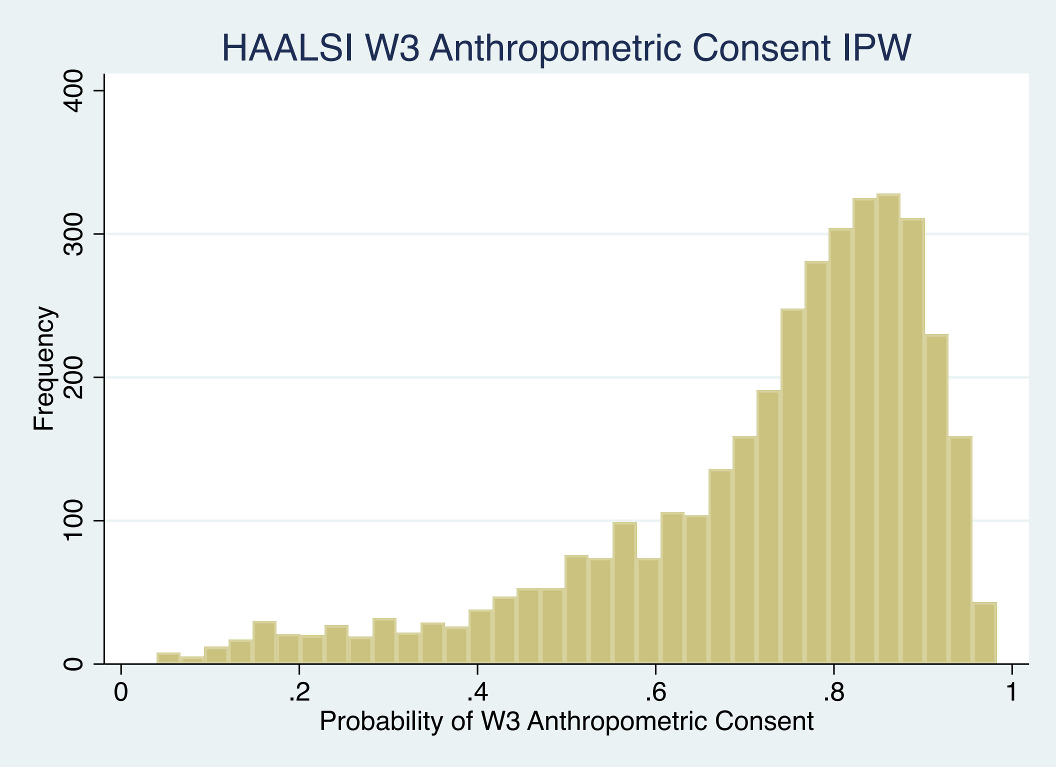

Figure A5. Histogram of the probability of consent to anthropometric measurements, from logistic regression model in Table A3.

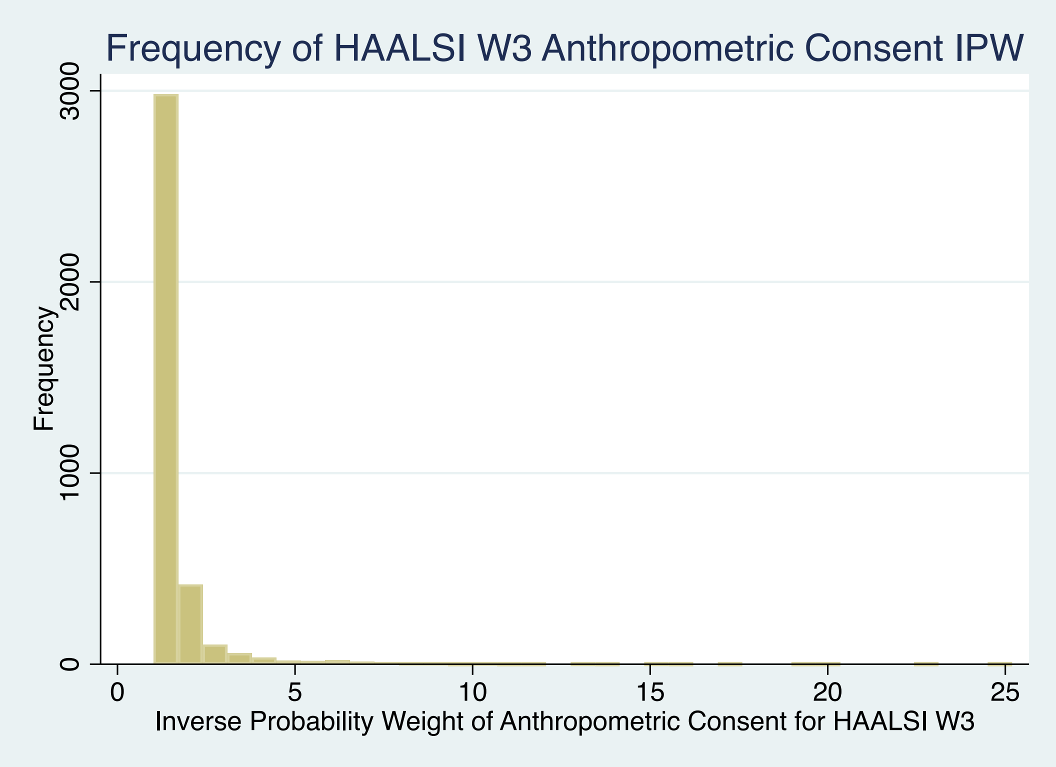

Figure A6. Histogram of anthropometric weight

Table A6. Logistic regression predicting consent to point of care measurements

Observations = 3,707
Pseudo R2 = 0.1182
C-statistic = 0.7130

	Odds Ratio	***SE***	***P value***	***C.I.***
Sex (1=Male, 2=Female)	0.770	0.258	0.435	[0.399,1.485]
Age at wave 1 (categorical)				
40-49 (reference category)	-	-	-	-
50-59	0.798	0.132	0.173	[0.576,1.104]
60-69	0.835	0.155	0.332	[0.581,1.202]
70-79	0.735	0.157	0.149	[0.483,1.116]
80+	0.511	0.127	0.007	[0.314,0.832]
Male X Age at wave 1 (categorical)				
Male X 40-49 (reference category)	-	-	-	-
Male X 50-59	1.344	0.312	0.204	[0.852,2.120]
Male X 60-69	1.529	0.378	0.085	[0.942,2.481]
Male X 70-79	1.718	0.491	0.058	[0.981,3.008]
Male X 80+	3.666	1.432	0.001	[1.705,7.881]
Total cognitive score at wave 1	1.033	0.014	0.015	[1.006,1.061]
Male X Total cognitive score at wave 1	0.975	0.018	0.174	[0.941,1.011]
Years of education at wave 1	0.979	0.013	0.121	[0.954,1.006]
Literacy at wave 1	1.101	0.122	0.389	[0.885,1.368]
Marital status at wave 1				
Married or living with partner (reference category)	-	-	-	-
Never married	0.509	0.086	0.000	[0.365,0.709]
Separated or deserted	0.883	0.132	0.406	[0.658,1.184]
Divorced	0.833	0.166	0.359	[0.563,1.232]
Widowed	1.007	0.111	0.949	[0.811,1.250]
Employed at wave 1	1.147	0.134	0.240	[0.913,1.441]
Household consumption quintile at wave 1				
Least consumption (reference category)	-	-	-	-
Less consumption	1.017	0.126	0.889	[0.798,1.297]
Middle	1.067	0.135	0.607	[0.833,1.367]
More consumption	1.000	0.127	0.999	[0.780,1.282]
Most consumption	1.297	0.175	0.053	[0.996,1.690]
Anemic at wave 1	1.220	0.104	0.020	[1.032,1.443]
Diabetic at wave 1	1.315	0.202	0.075	[0.973,1.777]
Proxy interview at wave 1	0.476	0.168	0.036	[0.238,0.952]
Participation in AWIGEN	1.406	0.119	0.000	[1.190,1.661]
Month of first contact at wave 3				
February + March 2022 (reference category)	-	-	-	-
July + August 2021	11.119	2.561	0.000	[7.080,17.462]
September 2021	9.434	2.169	0.000	[6.012,14.805]
October 2021	7.776	1.837	0.000	[4.893,12.356]
November 2021	5.803	1.419	0.000	[3.594,9.370]
December 2021	3.563	0.905	0.000	[2.166,5.862]
Time of day of first contact at wave 3				
Morning (reference category)	-	-	-	-
Afternoon/Evening	0.751	0.078	0.006	[0.613,0.920]
Interviewer at wave 3				
Interviewer 1	0.223	0.064	0.000	[0.126,0.393]
Interviewer 2	0.406	0.121	0.002	[0.226,0.728]
Interviewer 3	0.586	0.182	0.086	[0.319,1.078]
Interviewer 4	1.506	0.545	0.257	[0.742,3.060]
Interviewer 5	0.250	0.073	0.000	[0.141,0.442]
Interviewer 6	0.429	0.131	0.005	[0.236,0.779]
Interviewer 7	0.702	0.232	0.284	[0.367,1.341]
Interviewer 8	0.951	0.311	0.878	[0.501,1.806]
Interviewer 9	0.939	0.306	0.847	[0.495,1.780]
Interviewer 10	0.667	0.209	0.196	[0.362,1.231]

Interviewer 11	0.399	0.115	0.001	[0.227,0.702]
Interviewer 12	0.680	0.210	0.211	[0.372,1.244]
Interviewer 13	0.923	0.294	0.801	[0.494,1.724]
Interviewer 14	0.293	0.093	0.000	[0.157,0.548]
Interviewer 15	0.553	0.161	0.042	[0.312,0.979]
Interviewer 16	0.893	0.296	0.733	[0.467,1.709]
Interviewer 17	0.388	0.125	0.003	[0.207,0.729]
Interviewer 18	0.593	0.180	0.086	[0.327,1.076]
Interviewer 19	0.444	0.130	0.006	[0.250,0.790]
Interviewer 20	0.857	0.273	0.627	[0.459,1.599]
Interviewer 21	0.521	0.155	0.028	[0.291,0.933]
Interviewer 22	0.608	0.182	0.096	[0.339,1.093]

Table A5 continued…

Interviewer 23	0.764	0.237	0.385	[0.416,1.402]
Interviewer 24	1.328	0.452	0.404	[0.682,2.586]
Interviewer 25	0.521	0.156	0.029	[0.290,0.936]
Interviewer 26	0.569	0.172	0.062	[0.315,1.028]
Interviewer 27	0.618	0.207	0.151	[0.320,1.193]
Interviewer 28	1.246	0.399	0.493	[0.665,2.335]
Interviewer 29 (reference category)	-	-	-	-
Interviewer 30 + 50	0.226	0.064	0.000	[0.129,0.394]
_cons	0.457	0.175	0.041	[0.216,0.969]

Note: conditional on participating in wave 3

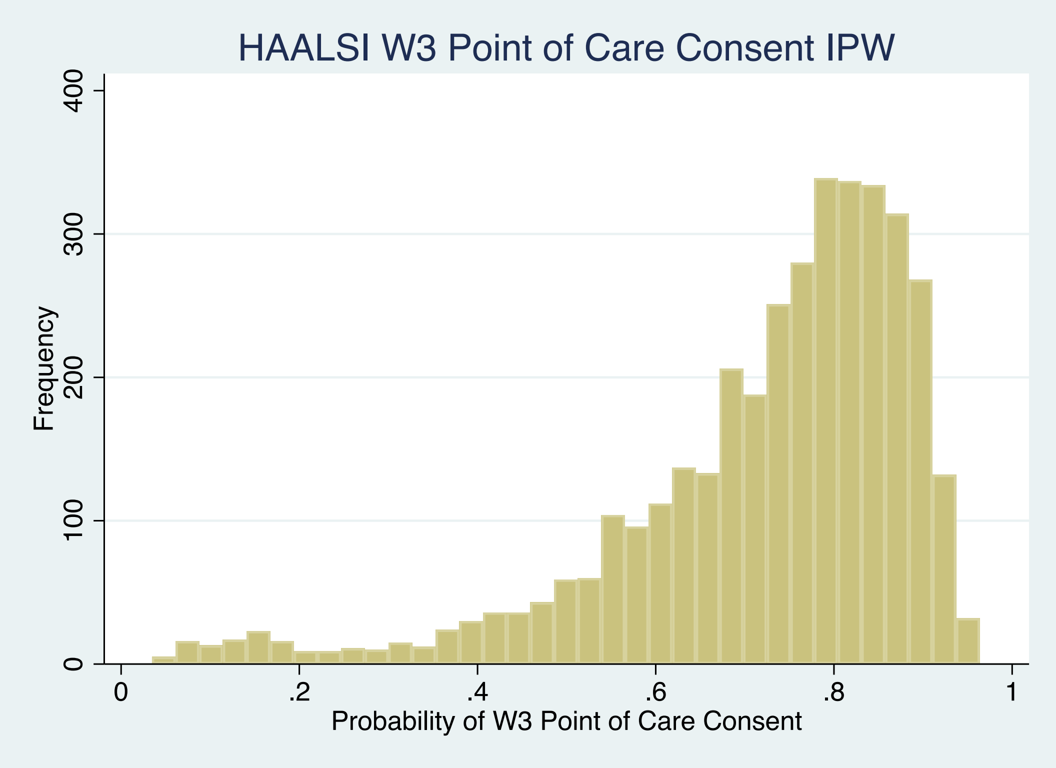

Figure A7. Histogram of the probability of consent to point of care measurements, from logistic regression model in Table A4.

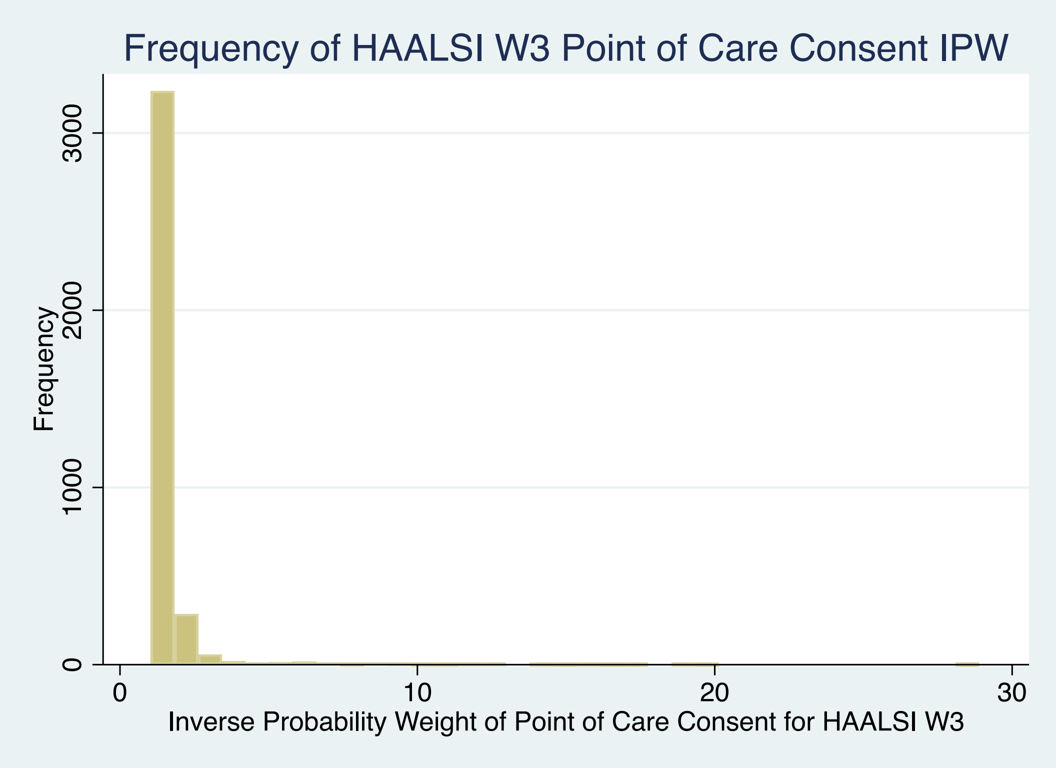

Figure A8. Histogram of point of care weight

References

Deaton, A. (2000). Consumption. In M. Grosh & P. Glewwe (Eds.), *Designing Household Survey Questionnaires for Developing Countries: Lessons from Ten Years of LSMS Experience* (Vol. 1, pp. 91–134). The World Bank.

Grosh, E. M., & Glewwe, P. (n.d.). *Designing Household Survey Questionnaires for Developing Countries: Lessons from Ten Years of LSMS Experience*. 78.

Hentschel, J., & Lanjouw, P. (1996). *Constructing an indicator of consumption for the analysis of poverty: Principles and illustrations with reference to Ecuador*. The World Bank. https://doi.org/10.1596/0-8213-3584-7

Riumallo-Herl, C., Canning, D., & Kabudula, C. (2019). Health inequalities in the South African elderly: The importance of the measure of social-economic status. *The Journal of the Economics of Ageing*, *14*, 100191. https://doi.org/10.1016/j.jeoa.2019.01.005

Shisana, O., Labadarios, D., Rehle, T., Simbayi, L., Zuma, K., Dhansay, A., Reddy, P., Parker, W., Hoosain, E., Naidoo, P., Hongoro, C., Mchiza, Z., Steyn, N., Dwane, N., Makoae, M., Maluleke, T., Ramlagan, S., Zungu, N., Evans, M., … SANHANES-1 Team. (2013). *The South African National Health and Nutrition Examination Survey (SANHANES-1)*. Human Sciences Research Council.