

# A Comparison of Results from Two Sampling Approaches in the South African National HIV Prevalence, Incidence and Behavior Survey, 2012

**Khangelani Zuma<sup>1,2\*</sup>, Olive Shisana<sup>3</sup>, Dorina Onoya<sup>4</sup>, Thomas Rehle<sup>5</sup>, Leickness Simbayi<sup>6,7</sup>, Sean Jooste<sup>6</sup>, Nompumelelo Zungu<sup>1,8</sup>**

<sup>1</sup>Social Aspects of Public Health Research Programme, Human Sciences Research Council, Pretoria, South Africa

<sup>2</sup>School of Public Health, University of the Witwatersrand, Johannesburg, South Africa

<sup>3</sup>University of Cape Town, Cape Town, South Africa

<sup>4</sup>Health Economics and Epidemiology Research Office, Department of Internal Medicine, School of Clinical Medicine, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

<sup>5</sup>Centre for Infectious Disease Epidemiology and Research, School of Public Health and Family Medicine, University of Cape Town, Cape Town, South Africa

<sup>6</sup>Social Aspects of Public Health Research Programme, Human Sciences Research Council, Cape Town, South Africa

<sup>7</sup>Department of Psychiatry and mental Health, University of Cape Town, Cape Town, South Africa

<sup>8</sup>Department of Psychology, University of Pretoria, Pretoria, South Africa

Email: \*kzuma@hsrc.ac.za

---

**How to cite this paper:** Zuma, K., Shisana, O., Onoya, D., Rehle, T., Simbayi, L., Jooste, S. and Zungu, N. (2020) A Comparison of Results from Two Sampling Approaches in the South African National HIV Prevalence, Incidence and Behavior Survey, 2012. *Open Journal of Statistics*, **10**, 87-96.

<https://doi.org/10.4236/ojs.2020.101007>

**Received:** November 12, 2019

**Accepted:** February 22, 2020

**Published:** February 25, 2020

Copyright © 2020 by author(s) and Scientific Research Publishing Inc. This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

## Abstract

**Background:** South Africa implements variations of second generation surveillance surveys to monitor human immunodeficiency virus (HIV) epidemic.

**Objective:** This paper compares HIV estimates from two design variations: *take all* approach and *sub-sampling* approach to ascertain if any changes in HIV epidemic are due to methodological changes or inherent evolution of the epidemic. **Methods:** A multi-stage stratified cluster sample of 1000 census enumerator areas was implemented with 15 households systematically sampled within each census enumerator area. In each household, every member was invited to participate (*take all* approach). To compare to the previous survey designs, a *sub-sampling* approach of at most four people from each household was implemented by randomly sampling one person from each age group: <2 years, 2 - 14 years, 15 to 24 years and 25 years and above. **Results:** HIV estimates were comparable with no systematic pattern. Prevalence estimates were slightly higher 12.2% [11.4% - 13.1%] in the *take all* compared to 11.6% [10.6% - 12.6%] in the *sub-sampling* approach. Estimates from sub-sampling approach were more variable. The design effects in the take all

approach were also slightly higher than those obtained in the sub-sampling. The overall synthetic measure of homogeneity for both methods was  $\rho = 0.10$ . *Conclusion:* In conclusion, as the household size increases the number of people living with HIV in each household increases thus increasing intraclass correlation. Similarity of resulting HIV estimates is re-assuring. However, the take all approach is more preferable than sub-sampling approach as it allows for detailed analyses of HIV data such as estimating discordance between sexual partners and parent-child pair.

## Keywords

Second Generation Surveillance, Human Immunodeficiency Virus (HIV), Sub-Sampling, Response Rate

## 1. Introduction

Reliable data level on human immunodeficiency virus (HIV) and associated determinants at population is crucial in understanding the dynamics of HIV. Countries with generalized HIV epidemics obtain estimates from surveillance systems such as antenatal care surveillance surveys. These surveillance systems have their limitations [1] [2]. Population based second generation surveillance surveys have been used by a number of countries repeatedly to monitor the epidemic of HIV and are now considered a gold standard [2] [3]. These surveys address some of the weaknesses that are encountered in antenatal care surveillance surveys [2] [4] and thus greatly enhance surveillance systems [5]. Second generation surveillance surveys combine behavioural data, socio-economic data and biomedical data which provide together greater explanatory power to assess the HIV epidemic in a country [6].

In South Africa, several small-scale focused HIV surveys have been conducted [7] [8] [9]. In the past 10 years Human Sciences Research Council (HSRC) has conducted a series of population based second generation surveillance surveys [10] [11] [12] [13]. The sampling design of these surveys varies from country to country with varying challenges [14] [15].

The design of the South African second generation surveys has had some variations from one wave of the survey to another over the years. In 2002 and 2005 one person aged two years and above was randomly selected in each age group 2 to 14 years, 15 to 24 and 25 years and above in each sampled household [10] [11]. In 2008, persons younger than two years were further included in the survey [12]. This resulted in a sample of at most four people in each household from four distinct age groups, that is younger than 2 years, 2 to 14 years, 15 to 24 years and 25 years and above. In 2012, all household members were eligible to participate in the survey [13].

In the analyses of surveys where inference is drawn on trends over time, it is important to ascertain whether any differences in the results are true differences

or are an artifact of variations in methodological design. For example, inclusion or exclusion of high risk groups in some setting can lead to biased estimates [5]. The objective of this paper is to compare the HIV prevalence estimates when all persons in the sampled households were invited to participate [13] to those results obtained when one person is randomly sampled in each age group (younger than two years, 2 to 14 years, 15 to 24 years and 25 years and above) as has been carried out in the previous survey of 2008 using the population-based survey data of South African 2012 national HIV prevalence survey reported in Shisana, Rehle, Simbayi, *et al.* 2014 [13].

## 2. Methodology

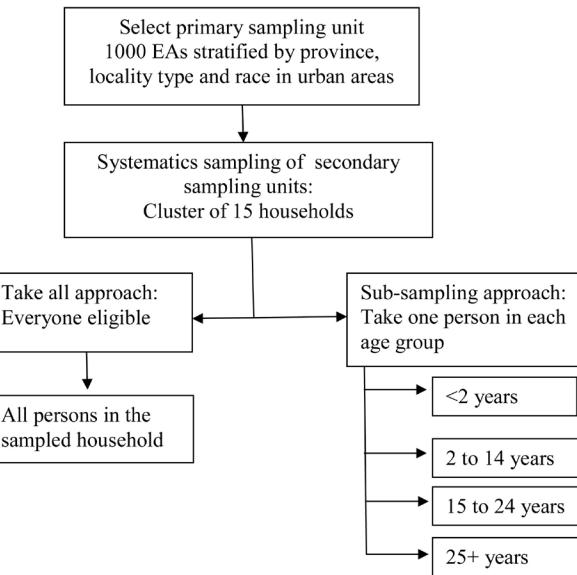
### 2.1. Design of Sampling Frame

Complex survey designs involve a combination of a number of design components including stratification, multistage sampling and selection with unequal probabilities or weighting. The design of the South African national HIV household surveys is complex and based on a multi-stage stratified cluster sample design. A random sample of 1000 census enumerator areas from a national database of 86,000 enumeration areas (EA) used during the 2001 census [16] served as the Master Sample of the primary sampling units. The master sample was explicitly stratified by province and locality type of the EAs. Locality types were urban formal, urban informal, rural formal (including commercial farms) and rural informal (tribal authority areas). In the formal urban areas, race was also used as an additional stratification variable [13]. In each sampled EA, a cluster of 15 households was randomly sampled to form the secondary sampling unit. In each sampled household all persons residing at the household including visitors who spent a night before were invited to participate and referred to as “take all approach”.

**Figure 1** presents the two sampling approaches used in the comparative analysis.

The “take all” approach implies that the designs of the previous HSRC surveys can be deduced from the database. Using the captured data, individuals within each household were grouped by age group as presented in **Figure 1**. In each age group at each household, a sampling scheme was implemented using *proc survey select* in Statistical Analysis System (SAS) version 9.3 (SAS Institute) to randomly select one person in each age group to be considered in further analysis (“sub-sampling”). The sub-sampled data (three-stage sampling at EA, dwelling and sub-sampled persons) mimics the previous HSRC survey designs for all practical purposes. In this way, if the sampled person was a refusal, the results were recorded as such in the sub-sampled data.

A consequence of implementing complex survey design is that sampling errors of the survey estimates cannot be computed using standard formulae found in standard statistical texts since they are based on independently and identically distributed random variables. Complex methods of estimating variances for



**Figure 1.** Schematic representation of the sample design.

complex sample designs are used [17] [18] which are often larger than those obtained from standard formulae. A design effect defined as the ratio of the complex variance estimate and variance obtained from standard formulae is computed to shed useful light on the precision of survey estimates between the two designs [19]. In this design, a household is considered a cluster and a synthetic measure of homogeneity within clusters ( $\rho$ ) is computed to measure the level of homogeneity within household for each determinant [20].

## 2.2. Weighting and Benchmarking of the Sample

Owing to the multi-stage stratified sampling design of the survey, some individuals have a greater or lesser probability of being selected than others. To correct for potential bias due to unequal sampling probabilities, sample weights were introduced at the EA, household, and individual levels and also to adjust for non-response. The final sampling weight was thus equal to the final EA weight multiplied by the final VP sampling weight adjusted for individual non-response in the take all approach. In the sub-sampling the final sampling weight was thus equal to the final EA weight multiplied by the final VP sampling weight multiplied by the sampling weight of each person in the household in each age group adjusted for individual non-response. Thus, the sampling weights corrected for unequal number of household members within each age group.

The final individual weights were benchmarked to 2012 mid-year population estimates by age, race, sex, and province [21]. This process produced a final sample representative of the population in South Africa for sex, age, race, and province.

## 2.3. Data Capturing, Management and Analysis

Survey data from questionnaires were double entered and verified by the Data

Capturing Unit (DCU) at the HSRC using Census Survey Processing (CS Pro) software Version 5.0 (U.S. Census Bureau). A database was designed with range restrictions to ensure that data captured were not out of range. Exploratory analysis was conducted in SPSS version 17. Final data analysis was conducted in STATA version 12 (Stata Corporation, College Station, TX) taking into account the complex design aspects of the survey via a suite of *svy* commands. The *svy* commands were used to obtain the estimates of HIV prevalence or proportions of responses or absolute estimated totals and confidence intervals (95% CI).

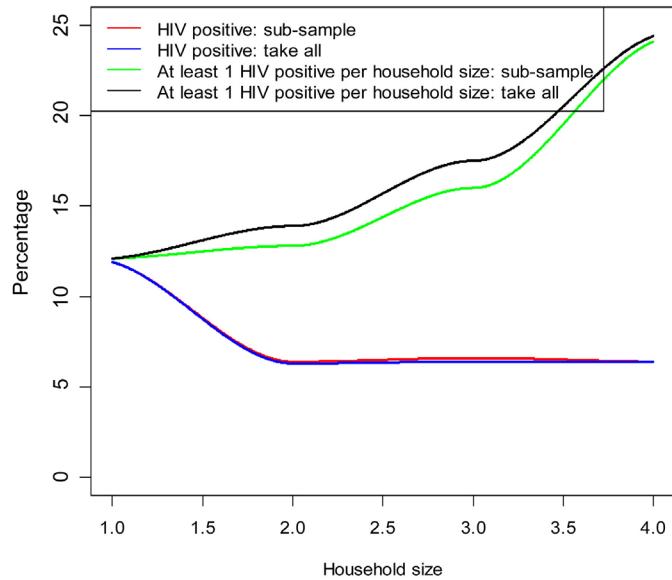
### 3. Results

In total 15,000 households were sampled from 1000 EAs. The households that were found to have been destroyed were not considered as non-response. Of these sampled households 13,083 (87.2%) were valid and occupied households whilst 1266 were invalid. Of the valid households, 11,079 were interviewed resulting in the 2012 survey household response rate of 84.7% [13]. Proportions of non-response at household level varied from 1347 (10.3%) who refused to take part in the survey, 657 (5.0%) with valid households but empty after required visits. Households in urban formal areas had the lowest response rate, 80.3%, and households in other geographical areas had response rates above 91%. All provinces had a visiting point response rate of 80% and above, except Gauteng (78.4%) and Western Cape (78.9%). **Table 1** presents individual response rate for both specimen and questionnaire.

The results in **Table 1** show that there are no considerable differences in the questionnaire and HIV testing response rates between the *take all* and *sub-sampling* approaches. The differences in all other determinants such as age, race, sex and geotype are less than two percent.

In the take all approach, the household size ranged from 1 to 18 people whilst in the sub-sample ranged from 1 to 4 people since at most four people could be sampled from each household. The crude percentage of people infected with HIV per household size was practically similar between the take all approach and sub-sample approach (**Figure 2**). The percentages of households that had at least one person infected with HIV between the two approaches are consistently and systematically different. As the household size increases there is consistently an increasing likelihood that at least one person is infected with HIV. This was more pronounced in the take all approach than sub-sampling approach. This indicates an increased likelihood of similar HIV positive status among individuals from the same household in the take all approach compared to subsampling approach.

**Table 2** presents key comparison results between the two sampling approaches when assessing the validity of the HIV results. The HIV estimates between the two methods are very comparable with no consistent pattern in any direction. These results are in agreement with consistent similar proportions of HIV positives between the two methods in **Figure 2**. However, the estimates based on the



**Figure 2.** Percentages of HIV positives and households with at least one person infected with HIV per household size.

**Table 1.** Response rate for take all and sub-sampled approaches.

	Take all		Sub-sampled	
	Total	Percentage	Total	Percentage
<b>Questionnaire interview response</b>				
Questionnaire completed	38,431	89.5	20,738	90.4
Refused to be interviewed	2295	5.3	1109	4.8
Absent from household	2284	5.2	1089	4.8
<b>HIV testing response</b>				
Specimen collected	28,997	67.5	15,497	67.6
Interviewed but not tested	9434	22.0	5241	22.9
Refused both interview and testing	2295	5.3	1109	4.8
Absent from household	2224	5.2	1089	4.7
<b>HIV testing response by sex of the respondent</b>				
Male	20,089	64.2	10,428	64.0
Female	22,547	70.3	11,674	69.0
<b>HIV testing response by race of the respondent</b>				
African	25,402	73.3	13,546	72.3
White	4033	43.0	2135	42.5
Coloured	8079	69.9	4123	69.1
Indian	4859	54.0	2221	51.5
<b>HIV testing response by age group in years</b>				
0 to <2	1714	51.6	1603	53.8
2 to 14	10,191	70.2	5547	68.4
15 to 24	8221	71.6	5191	71.1
25+	22,526	65.8	10,546	67.6
<b>HIV testing response by geotype</b>				
Urban formal	24,844	59.7	13,296	60.3
Urban informal	4582	72.7	2611	72.1
Rural informal	9752	80.0	4994	79.2
Rural formal	3771	80.8	2034	80.8

**Table 2.** HIV prevalence among participants age 0 years and older, socio-demographic characteristics, coefficient of variation, and the design effect.

<b>Variables</b>	Take all sample design						Sub-sampled design				
	Tested	HIV + (%) [95% CI]	SE	CV	DEFT	Tested	HIV + (%) [95% CI]	SE	CV	DEFT	
Total	28,997	12.2 [11.4 - 13.1]	0.4	0.03	2.0	14,908	11.6 [10.6 - 12.6]	0.5	0.04	2.0	
<b>Age group (Years)</b>											
0 - 14	7154	2.4 [1.9 - 2.9]	0.3	0.12	1.5	4193	2.3 [1.6 - 3.3]	0.4	0.17	1.8	
15 - 24	5890	7.1 [6.2 - 8.1]	0.5	0.07	1.7	3656	7.9 [6.5 - 9.6]	0.8	0.10	1.5	
25 - 49	8830	25.2 [23.2 - 27.3]	1	0.04	2.5	4232	23.3 [20.9 - 25.9]	1.3	0.06	2.2	
50+	5986	7.6 [6.5 - 8.8]	0.6	0.08	2.0	2824	6.9 [5.6 - 8.6]	0.8	0.12	1.4	
25+	14,816	19.9 [18.3 - 21.6]	0.8	0.04	2.7	7056	18.4 [16.5 - 20.4]	1.0	0.05	2.0	
15 - 49	14,720	18.8 [17.5 - 20.3]	0.7	0.04	2.3	7888	17.9 [16.3 - 19.6]	0.9	0.05	2.0	
<b>Sex</b>											
Male	12,896	9.9 [8.9 - 11.0]	0.5	0.05	2.5	6674	10.0 [8.6 - 11.6]	0.8	0.08	2.2	
Female	15,794	14.4 [13.3 - 15.6]	0.6	0.04	2.0	8060	13.7 [12.3 - 15.3]	0.8	0.06	1.9	
<b>Race</b>											
Black African	18,629	15.0 [14.0 - 15.9]	0.5	0.03	1.7	9798	14.5 [13.3 - 15.8]	0.6	0.04	2.0	
white	1733	0.3 [0.1 - 0.8]	0.2	0.67	1.0	908	0.4 [0.2 - 1.2]	0.2	0.50	1.2	
Coloured	5625	3.1 [2.2 - 4.2]	0.5	0.16	2.5	2850	3.7 [2.3 - 5.7]	0.8	0.22	1.6	
Indian / Asian	2626	0.8 [0.5 - 1.4]	0.2	0.25	1.0	1144	1.2 [0.5 - 2.5]	0.5	0.42	0.8	
<b>Locality type</b>											
Urban formal	14,821	10.1 [8.8 - 11.7]	0.7	0.07	3.5	7543	9.1 [7.7 - 10.8]	0.8	0.09	2.3	
Urban informal	3329	19.9 [17.4 - 22.7]	1.4	0.07	2.3	1859	19.8 [15.5 - 25.0]	2.4	0.12	2.1	
Rural informal	7801	13.4 [12.2 - 14.7]	0.6	0.04	1.5	3913	13.4 [11.9 - 15.1]	0.8	0.06	1.8	
Rural formal	3046	10.4 [7.4 - 14.4]	1.8	0.17	3.0	1593	8.4 [5.9 - 11.9]	1.5	0.18	1.5	
<b>Province</b>											
Western Cape	3372	5.0 [3.5 - 7.1]	0.9	0.18	2.3	1710	4.8 [3.3 - 7.0]	0.9	0.19	1.8	
Eastern Cape	4102	11.6 [10.0 - 13.5]	0.9	0.08	2.3	2167	11.7 [9.7 - 14.0]	1.1	0.09	1.5	
Northern Cape	2381	7.4 [4.4 - 12.0]	1.9	0.26	3.8	1228	5.4 [3.7 - 7.9]	1.1	0.20	0.8	
Free State	2098	14.0 [11.0 - 17.7]	1.7	0.12	2.4	1211	15.0 [9.6 - 22.8]	3.3	0.22	2.6	
KwaZulu-Natal	7052	16.9 [15.2 - 18.6]	0.9	0.05	2.3	3205	16.3 [14.1 - 18.7]	1.2	0.07	1.7	
North West	2038	13.3 [11.5 - 15.3]	1	0.08	1.4	1087	10.7 [8.3 - 13.6]	1.3	0.12	1.4	
Gauteng	3007	12.4 [10.2 - 15.1]	1.2	0.10	3.1	1622	11.5 [9.1 - 14.4]	1.3	0.11	2.5	
Mpumalanga	2274	14.1 [11.7 - 16.9]	1.3	0.09	1.9	1181	11.8 [9.1 - 15.1]	1.5	0.13	1.6	
Limpopo	2673	9.26.9 - 12.2]	1.3	0.14	2.6	1497	10.0 [7.2 - 13.7]	1.6	0.16	2.1	

sub-sampling approach are more variable than those from the take all approach. The design effects in the take all approach are also slightly higher than those obtained in the sub-sampling. The design effects vary proportionally with the syn-

thetic measure of homogeneity ( $\rho$ ) indicating a higher intraclass correlation within households with respect to HIV in the take all approach compared to sub-sampling approach. The overall synthetic measure of homogeneity for both methods is  $\rho = 0.10$ .

The HIV prevalence estimates is slightly higher 12.2% in the take all approach compared to 11.6% in the subsampling approach. All other estimates for age, sex, race, geotype and province are in the same direction and order of magnitude (**Table 2**).

#### 4. Conclusions

The results of the paper from the take all approach used in the 2012 survey [13] and the sub-sampling design implemented in the previous surveys are compared. The calculated response rate was similar for both methods. The findings show that the estimates of HIV are comparable for all key determinants. However, the estimates from sub-sampling are more variable than those from the take all approach. This could be a function of cluster (household) sample size and intraclass correlation within each cluster which leads to practically less effective sample size due to high correlation. In the generalised epidemic settings like South Africa, the risk of HIV infection is likely to be clustered within households [22] due to both heterosexual transmission among sexual partners within households and vertical transmission to their children. The overall estimate of intraclass correlation and design effect are similar for both methods. However, for various determinants, the estimates for intraclass correlation and design effects are moderately higher for the take all approach than sub-sampling.

The comparison of the two methods is subject to some limitations. The sub-sampling arm of the study conditions on the household roster for the take all alternative. In actual implementation of sub-sampling method, there could be non-coverage and differential non-response. Thus the simulated experiment might not replicate exactly the outcome that would be obtained under two real survey conditions.

In conclusion, the two approaches yield similar results for all practical purposes. However, even though with high intraclass correlation resulting in lesser effective sample size, the take all approach is more preferable than sub-sampling approach. The take all approach allows for further analyses of data such as estimating discordance between sexual partners and parent-child pair.

#### Acknowledgements

The data used in this article comes from a study supported by the President's Emergency Plan for AIDS Relief (PEPFAR) through the Centers for Disease Control and Prevention (CDC) under the terms of 3U2GGH000570 and the South African National AIDS Council. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of CDC or SANAC.

## Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

## References

- [1] Stover, J., Ghys, P.D. and Walker, N. (2004) Testing the Accuracy of Demographic Estimates in Countries with Generalized Epidemics. *Aids*, **18**, S67-S73. <https://doi.org/10.1097/00002030-200406002-00008>
- [2] Garcia-Calleja, J.M., Gouws, E. and Ghys, P.D. (2006) National Population Based HIV Prevalence Surveys in Sub-Saharan Africa: Results and Implications for HIV and AIDS Estimates. *Sexually Transmitted Infections*, **82**, iii64-iii70. <https://doi.org/10.1136/sti.2006.019901>
- [3] Boerma, T.J., Ghys, P.D. and Walker, N. (2003) Estimates of HIV-1 Prevalence from National Population-Based Surveys as a New Gold Standard. *The Lancet*, **362**, 1929-1931. [https://doi.org/10.1016/S0140-6736\(03\)14967-7](https://doi.org/10.1016/S0140-6736(03)14967-7)
- [4] WHO (2002) Guidelines for Second-Generation HIV Surveillance. Geneva.
- [5] Mishra, V., Vaessen, M., Boerma, T.J., Arnold, F., Way, A., Barrere, B., et al. (2006) HIV Testing in National Population-Based Surveys: Experience from the Demographic and Health Surveys. *Bulletin of the World Health Organization*, **84**, 537-545. <https://doi.org/10.2471/BLT.05.029520>
- [6] Rehle, T., Lazzari, S., Dallabetta, G. and Asamoah-Odei, E. (2004) Second Generation HIV Surveillance: Better Data for Decision Making. *Bulletin of the World Health Organization*, **82**, 121-127.
- [7] Auvert, B., Bolland, R., Campbell, C., et al. (2001) HIV Infection among Youth in a South African Mining Town Is Associated with Herpes Simplex Virus-2 Seropositivity and Sexual Behaviour. *Aids*, **15**, 5885-5989. <https://doi.org/10.1097/00002030-200105040-00009>
- [8] Colvin, M., Abdool Karim, S.S., Connolly, C., Hoosen, A.A. and Ntuli, N. (1998) HIV Infection and Asymptomatic Sexually Transmitted Infections in a Rural South African Community. *International Journal of STD & AIDS*, **9**, 548-550. <https://doi.org/10.1258/0956462981922683>
- [9] MacPhail, C., Williams, B. and Campbell, C. (2002) Relative Risk of HIV Infection among Young Men and Women in a South African Township. *International Journal of STD & AIDS*, **13**, 331-342. <https://doi.org/10.1258/0956462021925162>
- [10] Shisana, O. and Simbayi, L. (2002) Nelson Mandela/HSRC Survey of HIV/AIDS: South African National HIV Prevalence. Behavioural Risks and Mass Media Household Survey 2002. Human Sciences Research Council (HSRC) Press, Cape Town.
- [11] Shisana, O., Rehle, T., Simbayi, L.C., Parker, W., Zuma, K. and Bhana, A. (2005) South African National HIV Prevalence, Incidence, Behaviour and Communication Survey. HSRC Press, Cape Town.
- [12] Shisana, O., Rehle, T., Simbayi, L.C., Zuma, K., Jooste, S., et al. (2009) South African National HIV Prevalence. Incidence and Communication Survey 2008: A Turning Tide among Teenagers? HSRC Press, Cape Town.
- [13] Shisana, O., Rehle, T., Simbayi, L.C., Zuma, K., Jooste, S., et al. (2014) South African National HIV Prevalence, Incidence and Behaviour Survey, 2012. HSRC Press, Cape Town.

- [14] Vaessen, M., Thiam, M. and Thanh, L. (2004) The Demographic and Health Surveys. In: *Household Sample Surveys in Developing Countries: Design, Implementation and Analysis*, United Nations, Department of Economic and Social Affairs, New York, 495-522.
- [15] ORC Macro. (2005) Anemia and HIV Testing Filed Manual: Demographic and Health Surveys. ORC Macro, Calverton.
- [16] Statistics South Africa (2001) South Africa Census 2001. Pretoria.
- [17] Rust, K.F. (1985) Variance Estimation for Complex Estimators in Sample Surveys. *Journal of Official Statistics*, **1**, 381-397.
- [18] Rust, K.F. and Rao, J.N.K. (1996) Variance Estimation for Complex Surveys Using Replication Techniques. *Statistical Methods in Medical Research*, **5**, 283-310.  
<https://doi.org/10.1177/096228029600500305>
- [19] Kish, L. (1965) Survey Sampling. Wiley, New York.
- [20] Karlton, G. (1977) Practical Methods for Estimating Survey Sampling Errors. *Bulleting of the International Statistical Institute*, **47**, 495-514.
- [21] Statistics South Africa (2013) South African Population Mid-Year Estimates. Pretoria.
- [22] Zuma, K., Lurie, M. and Jorgensen, M. (2006) Analysis of Interval-Censored Data from Circular Migrant and Non-Migrant Sexual Partnerships Using the EM Algorithm. *Statistics in Medicine*, **26**, 309-319. <https://doi.org/10.1002/sim.2539>