DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 33, No. 1, 2017, pp. 249–274, http://dx.doi.org/10.1515/JOS-2017-0013

# Cost-Benefit Analysis for a Quinquennial Census: The 2016 Population Census of South Africa

*Bruce D. Spencer*[1], *Julian May*[2], *Steven Kenyon*[3], *and Zachary Seeskin*[4]

The question of whether to carry out a quinquennial Census is faced by national statistical offices in increasingly many countries, including Canada, Nigeria, Ireland, Australia, and South Africa. We describe uses and limitations of cost-benefit analysis in this decision problem in the case of the 2016 Census of South Africa. The government of South Africa needed to decide whether to conduct a 2016 Census or to rely on increasingly inaccurate post-censal estimates accounting for births, deaths, and migration since the previous (2011) Census. The cost-benefit analysis compared predicted costs of the 2016 Census to the benefits of improved allocation of intergovernmental revenue, which was considered by the government to be a critical use of the 2016 Census, although not the only important benefit. Without the 2016 Census, allocations would be based on population estimates. Accuracy of the postcensal estimates was estimated from the performance of past estimates, and the hypothetical expected reduction in errors in allocation due to the 2016 Census was estimated. A loss function was introduced to quantify the improvement in allocation. With this evidence, the government was able to decide not to conduct the 2016 Census, but instead to improve data and capacity for producing post-censal estimates.

*Key words:* Demographic statistics; fiscal allocations; loss function; population estimates; post-censal estimates.

## 1. Introduction

### *1.1. Background on Costs and Benefits of Mid-Decade Censuses*

At all times, but especially in challenging economic times, governments considering investment in an accurate census or other social information face simultaneous decision problems of how much to invest and how much accuracy to seek. In the United States, the constitutional requirement of a census every ten years has been met, at increasing cost, and with varying degrees of accuracy. On the other hand, the US Congress has never provided

[1] Department of Statistics and Institute for Policy Research, Northwestern University, Evanston, IL 60208-4070, U.S.A. Email: bspencer@northwestern.edu
[2] DST-NRF Centre of Excellence in Food Security, University of the Western Cape, Bellville, South Africa. Email: jmay@uwc.ac.za
[3] National Treasury, Pretoria, South Africa. Email: steven.kenyon@treasury.gov.za
[4] NORC at the University of Chicago, Chicago, IL, 60603, U.S.A. Email: Seeskin-Zachary@norc.org
**Acknowledgments:** This research was supported a grant from the US National Science Foundation (SES-1129475). We thank Thamsanqa Mohale of the South African National Treasury and Diego Iturralde and Louis van Tonder of Statistics South Africa, for the financial and demographic statistics inputs that they compiled for this report, and Ben Mphahlele of the Statistics Council for recognizing and convincing others of the need for a cost-benefit analysis. Comments received from the Statistics Council, the Statistician General, Mr. Pali Lehohla, and the Associate Editor are also gratefully acknowledged.

funds for a mid-decade census, despite the legal requirement that a mid-decade census be carried out "in the year 1985 and every ten years thereafter" (Census Act of 1976, PL 94-521; 13 USC §141(d)). Since the 2010/2011 round of census-taking, media reports suggest that the timing and format of national censuses is being debated in several countries including Australia (The Guardian 2015), Canada (The Globe and Mail 2011), Ireland (The Journal 2012), and Nigeria (Nigerian Tribune 2016).

In South Africa, the Statistics Act of 1999 requires "a population census to be taken in the year 2001 and every five years thereafter . . . unless the Minister, on the advice of the Statistician-General . . . determines otherwise." The Act further provides for an independent Statistics Council to advise both the Minister and the Statistician-General on a wide range of matters pertaining to official statistics, with the taking of a population census specifically identified. In accordance with the law, censuses were taken in 2001 and 2011, but not in 2006, under the advice of the Statistics Council. The analysis described in this article was prepared as part of the evaluation of the 2011 Census to help the Statistics Council advise the Statistician-General and the Minister responsible for official statistics on whether a 2016 Census should be carried out by the government statistical agency, Statistics South Africa (Stats SA).

Considering the costs and benefits of government data programs, such as the 2016 Census, is essential for making informed decisions on how much to invest in such data programs. In November 2009, representatives of national statistical agencies and UN agencies met in Dakar to discuss improving the provision of statistics in the context of the United Nations Millennium Development Goals. The Dakar Declaration on the Development of Statistics that followed from this meeting proposed that official statistics are a public good, and that their production and dissemination is a core responsibility of all governments. Considering the costs and benefits of data programs is necessary because the market does not lead to socially optimal investment in public goods (Sims 1984). The cost of a 2016 Census is estimated to be at least ZAR 3 billion, which was the cost of the 2011 Census. (All amounts are in 2011 prices and at the time of the census, the South African Rand was equivalent to USD 0.14.) Note that the value of the census really refers to the added value of the census, compared with the value of alternatives, in particular a large sample survey to provide data on inter-provincial migration since the 2011 Census. The more accurately population change can be measured without a census, the less is the 2016 Census's value, ceteris paribus.

Benefits of data programs arise largely from their use, and understanding the causal pathways by which outputs from the data program affect outcomes is enormously complex. In particular, we would want to predict the outcomes if the 2016 Census were to be conducted and the outcomes if it were not conducted. The benefit of the 2016 Census reflects the difference in the value of the outcomes in the two scenarios, and therefore outcomes that would be the same in both scenarios can be ignored in the analysis. Even so, to consider all actions or outcomes by carrying out or not carrying out the 2016 Census is not feasible. Furthermore, assigning values (e.g., monetary values) to outcomes is challenging with regard to many uses of statistics (Spencer 1982a).

The impossibility of studying all the benefits of a major data program, such as a census, implies that cost-benefit analysis of the program must, necessarily, be incomplete in that

some benefits – perhaps even the majority of the benefits – will be unmeasured. Our analysis is a partial cost-benefit analysis, in that not all benefits are considered. As discussed below, we focus on just a single use of the census data: allocation of national funds to subnational jurisdictions by formulas. There are many other uses of census statistics which may be important. For consideration of other benefits from the South African Census, see May et al. (2013). The Office for National Statistics in the United Kingdom explicitly considered costs and benefits of the 2011 Census after receiving the recommendation of the "House of Commons Treasury Select Committee . . . that: "any future Census should also be justified in cost-benefit terms" (Cope 2015, 2). However, the detailed "business case" that was developed to "make the case" for the 2011 Census is not publicly available, only a high-level summary (Parliament of the United Kingdom 2009) and links thereto discussing some identified uses. The business case analysis for the 2011 Scotland Census is available, and it contains an analysis of shifts in fund allocations to Health Board Areas that would have occurred with a 2001 Census and without a 2001 Census (in which case post-censal estimates would have been used). (General Register Office for Scotland 2006, 27–34). Bakker (2014) analyzes costs and benefits of the New Zealand census. However, in all of these studies, the quantification of benefit of nonallocative uses of census statistics typically is highly uncertain.

The earliest identified cost-benefit analysis of a quinquennial census is that of Redfern (1974), who focused on benefits of more accurate fund allocations arising from a mid-decade census in England and Wales; the analysis did not appear to support carrying out a mid-decade census there (Spencer 1980a, 13–17; Alho and Spencer 2005, 368). Spencer (1980a) conducted a cost-benefit analysis comparing two alternative versions of the 1970 US Census. Seeskin and Spencer (2015) analyze benefits of improved allocations of funds and political representation under alternative accuracy profiles of the 2020 US Census. May and Lehohla (2005) discuss reasons for cost increases in South Africa's 2001 Census, but only describe some of the benefits.

Assigning values to alternative outcomes is a challenge for cost-benefit analysis of data programs. To compare costs and benefits most directly, it is convenient for benefits to be quantified in the same units as costs. However, when such a comparison is not feasible, the issue should not be forced. Instead, summaries can be prepared showing what benefits are attainable at what costs. Savage (1985) and Sims (1984) offer cautionary critiques of misguided attempts to force benefits of data programs to be measured in units comparable to those used by costs.

A partial cost-benefit analysis of a data program should not be narrowly interpreted as a formal set of calculations that will point to the "correct" or "optimal" decision (Savage 1985, 4). Cost-benefit analysis in the narrow sense can be misleading when applied to data programs, as pointed out by the National Research Council (1985).

> Cost-benefit analysis, as we understand and use the term, means describing a program as a set of commodities produced (benefits) and a set of commodities consumed (costs) and aggregating those using prices, market prices when possible, otherwise "shadow prices" that emerge from calculations based on assumptions of optimization, either by individuals or by components of a market economy.

> With information dissemination programs, this analytical framework is not helpful. Technical analysts can determine some of the political and economic decisions to which the information is relevant, and they can look for alternative pathways through which the information might flow, if the program were reduced or eliminated. But these efforts will involve tracing out the operation of incomplete and imperfect markets and of nonmarket information transfer mechanisms; the usual practices of relying on market prices and on the uniqueness of the values of traded goods will not be available. Trying to proceed nonetheless to attach dollar values to the effects of the information will nearly always lead to guesswork and arbitrary assumptions that obscure, rather than clarify, the analysis. (54–55)

We use the term cost-benefit analysis in the broad sense of providing a way of thinking about, and a way of organizing information on, some of the benefits and costs of a data program. There should be no automatic presumption that the measured benefits will outweigh the measured costs, even in a data program that is implemented in full, in the sense that the difference between its actual benefits and its actual costs is greater than for other programs. Failure to demonstrate that measured benefits exceed costs does not mean that the data program is unjustified or should not be carried out. The value of a cost-benefit analysis is a reduction in the uncertainty concerning the benefits and costs, and in an ideal world this would improve decisions concerning statistical programs. However, there is a risk in this approach that decision-makers may conclude that a data program is not worth funding if the partial cost-benefit analysis does not show benefits exceeding costs.

Although additional practical constraints on statistical agencies could, in principle, be incorporated into the cost function, factors other than cost may influence whether a data program is carried out. These potential factors include the capacity of the responsible institution to undertake data collection, competing demands by other data collection programs, and anticipated technology or methodology changes that improve the accuracy of estimating the population. In the case of capacity constraints, the institution may opt to reprioritize its work program, delaying or suspending other data collection activities in order to undertake the activity which it deems a priority. In the case of technology or methodology changes, improvements in the capacity to sample, such as satellite imagery, may permit the institution to opt for a large survey rather than a full census, thereby affecting the cost function of an alternative to a census.

There are major limitations in scope to partial cost-benefit analyses that must be communicated by researchers. If incorrectly interpreted, a partial cost-benefit analysis could do more harm than good. Key assumptions must be presented in a transparent way. Decision-makers within the statistical agency should be aware of all the limitations. In their communication with decision-makers and the general public, the researchers should explain the limitations in an understandable, albeit abbreviated form.

### *1.2. Legal Context for the Census in South Africa*

In South Africa, census-taking has a longstanding and sometimes controversial history dating back to the 18th century. However, most Statistics Acts (1976, 1978, and 1980) and censuses were designed during the apartheid regime, and therefore considered to be too narrow and insufficient to protect and promote the rights of all citizens of South Africa. To

address the limitations of the previous Acts, the current democratic South African Government designed the 1999 Statistics Act (Act No. 6 of 1999). The Act provides for "a Statistician-General as head of Statistics South Africa, who is responsible for the collection, production and dissemination of official and other statistics, including the conducting of a census of the population, and for coordination among producers of statistics; to establish a Statistics Council and provide for its functions; to repeal certain legislation; and to provide for connected matters." The first responsibility of the Statistician-General specified in the Act is to "cause a population census to be taken in the year 2001 and every five years thereafter . . . unless the Minister [of Finance, or other Minister as chosen by the President], on the advice of the Statistician-General . . . determines otherwise."

### *1.3. Uses of Census Data*

The additional information that a 2016 Census would provide about the population would lead to changes of various kinds, including, but not limited to the following.

1. Under South Africa's system of multi-tier government, funds are allocated by the national government to provinces and municipalities on the basis of population and other data. Fund allocations will differ depending on whether a 2016 Census is carried out or not.
2. Additional social information about population sizes (for groups classified by geography, ethnicity, and other criteria) would be provided, along with information about internal migration and migration between South Africa and other countries. Such information is important for understanding, and may or may not lead to identifiable changes in actions or outcomes. May et al. (2013) discuss a survey conducted to yield some limited insight into this.
3. Surveys carried out by Stats SA and by other survey organizations can be designed more efficiently (using updated sampling frames) based on information that the 2016 Census will provide. The survey analysis is also improved by the availability of more accurate population totals for various and diverse subgroups, which can be used to calibrate the survey data.
4. Policy analyses in all spheres of government will change to some degree as a result of having the 2016 Census data available.
5. Social planning and allocation of funding for electricity, water, sanitation, education facilities, and telecommunications can be based on more accurate data about population distribution.
6. Businesses may make different decisions about where to locate, about product design, or about risk assessment.

In addition, a census can have an important ceremonial aspect and be taken as a symbol of government efficiency (or inefficiency, depending on point of view), as observed by Kruskal (1984) and confirmed in the survey of data users as discussed by May et al. (2013, viii).

Uses of census data for formula-based allocation of funds are perceived as important in the context of a multi-tier government system such as the one adopted by South Africa,

and are the focus of the benefit analysis in this article. Subsection 1.4 provides further context.

Uses of census data for policy analysis (item 4 in the list) appear to be important as well. McCaa et al. (2006) discuss the strategic importance of the census in providing demographic, economic, and social data pertaining, at a specified time, to all persons in a country or a well-defined part of the country. They further note that a census helps in undertaking efficient management of economic and social policies or programmes, and one infers that census information is a key element in evidence-based policymaking. Indeed, concern for effects of population change and numbers is reflected by the South African Government's *White Paper on Population Policy*, which emphasizes "the need for reliable and up-to-date information on the population and human development situation in the country to inform policy making and programme design, implementation, monitoring and evaluation" (Ministry for Welfare and Population Development 1998, 16).

Understanding how data affect policy development and analysis is a challenge, and may require careful case studies of policy processes. Although we did not attempt this in full, we considered how changes in population numbers would affect outputs from the kind of microsimulation analyses that would be produced in the policy context, and found moderate impact (May et al. 2013, 32–35). The findings were communicated to the Statistics Council, the Statistician-General and the Minister responsible for Stats SA, but will not be further discussed in this article.

### *1.4. Formula-Based Allocations of Funds*

The South African Constitution considers various aspects of intergovernmental fiscal relations, including the devolution of certain revenue and expenditure assignments to subnational governments. Responsibility for revenue generation is unequally distributed between the national, provincial and local spheres of government. The national government has a wide variety of tax instruments available for raising revenue. In contrast, the provinces have limited options for taxation, and the municipalities largely rely on property taxes and service charges. Although the revenue-generating power of municipal governments was strengthened following the Municipal Property Rates Act (2004), the bulk of national revenue accrues to national government (Yemek 2005, 9). To address this, the Constitution also provides for a nonpartisan Financial and Fiscal Commission (FFC) that advises parliament and subnational governments on a variety of issues concerning intergovernmental fiscal relations, including the allocation of revenue among the three spheres of government, that is, national, provinces, and municipalities. According to Section 214 of the Constitution, one of the two main instruments for transferring revenue from the national sphere to the other two spheres is the "equitable shares" program. The provincial equitable share accounts for around 80% of transfers to provinces and the local government equitable share accounts for over half of the transfers to municipalities (National Treasury 2015).

The provincial shares and local government shares are divided between the provinces and the municipalities according to revenue-sharing formulae that are revised periodically. The Provincial Equitable Share (PES) and Local Government Equitable Share (LGES) formulas are based on the demographic and economic profiles of the subnational jurisdictions, as

revealed by population sizes and other statistics. To align with the mandated responsibilities of these jurisdictions, the PES has included the following components: an education share based on the average size of the school-age population (ages 5 to 17) and the number of learners enrolled in public ordinary schools; a health share based on the use of the public health system and the number of people without medical aid or health insurance; a general component based on population size. The LGES formula depends mainly on population numbers from the latest census, since updated population statistics are not available at municipal level in non-census years. This article is based on the LGES formula that was used prior to 2013, as this was the formula in use at the time the research was conducted. The new formula, introduced in 2013, is still driven mainly by the number of poor households in each municipality (National Treasury 2013, 34–43).

### *1.5. Refining the Set of Choices*

In a cost-benefit or other decision analysis, it is important to specify the alternative choices and underlying assumptions. We assume that a census will be taken in 2021 irrespective of whether or not a census is taken in 2016. Further, we assume that if the 2016 Census is not taken, Stats SA will conduct a large sample survey in 2016 similar to the Community Survey undertaken in 2007, which sampled 300,000 households. This will provide data on inter-provincial migration since the 2011 Census. Uses of population numbers in 2016 will be unaffected by the 2016 Census, since the census results would not yet be available. Users of population numbers for 2022 and beyond will rely on the 2021 Census numbers. Although post-2021 analysis of population dynamics would still be improved by the availability of 2016 Census data, we assess the benefits of the improvement to be relatively small in comparison to other benefits of the 2016 Census. These considerations lead us to focus on benefits arising from uses of population numbers for the five-year period, 2017–2021.

If a 2016 population census is not carried out, province-level population numbers for 2017–2021 will be available from the mid-year population estimates, which are derived by allowing for births, deaths, and net movements into and out of each province since the time of the 2011 Census (Stats SA 2011). The first two are derived from civil registration of vital statistics, but the last item can only be estimated, as internal migration is not recorded and there is a potentially substantial unrecorded international immigration. Thus, in the absence of a 2016 Census, the mid-year estimates for provinces will need to account for 6–10 years of population change since the 2011 Census; the Community Survey will be useful for this. If the 2016 Census is conducted, the population numbers for provinces in 2017–2021 will again be provided by the mid-year population estimates. However, these need only account for 1–5 years of population change since the 2016 Census, and official population numbers below the province level will be 1–5 years out of date instead of 6–10 years. Mid-year estimates are not available below the province level. Thus, municipal population numbers for 2017–2021 will be based either on the 2016 census, if it is conducted, or on the 2011 Census, if no 2016 Census is carried out.

### *1.6. Organization of Article*

As noted, we focus on the benefits of the 2016 Census that arise from improved allocations from the LGES and PES over the period 2017–2021. For this analysis, we treat the PES

allocations as correct if the input data for the allocation calculations were entirely correct. A loss function for measuring the aggregate discrepancy between the calculated allocations, $\hat{\theta}$, and the correct (or "true") allocations, $\theta$, is developed (Section 2). We consider two alternative ways in which $\hat{\theta}$ can be developed, according to the construction of mid-year population estimates for 2017–2021: the "*cen16*" alternative uses the 2016 Census results either as the estimates (LGES) or as the base for mid-year estimates (PES), whereas the "*nocen16*" alternative relies on the 2011 Census for municipal estimates (LGES) and as the base for the mid-year estimates for provinces (PES), supplemented by a 2016 Community Survey. To model the accuracy of the two alternative sets of mid-year population estimates, we assess the past performance of mid-year population estimates by comparing them to the 2011 Census results (Section 3), then we model their accuracy for 2017–2021 (Section 4). The distributions of PES and LGES allocations are then derived under the "*cen16*" and "*nocen16*" alternatives (Section 5), leading to estimates of improvement in allocation as a result of the 2016 Census. Limitations of the analysis are described (Section 7). After discussing census cost (Section 8), we discuss the benefits in light of the costs (Section 9). The article concludes with a brief discussion of the decisions made concerning the 2016 Census and alternatives (Section 10).

## 2. Use of Loss Functions to Measure Improvement in the Allocations of Funds

### *2.1. Loss Functions for Errors in Allocations*

An important identified use of population census data in South Africa is the allocation of funds using a formula with inputs from statistics of various kinds and with an output that specifies the share that each province should receive. As already noted, the formula is called the Provincial Equitable Share (PES). A similar important use is the allocation of funds to municipalities using the Local Government Equitable Share (LGES) formula. The design and weighting of the formulas are agreed by intergovernmental forums that include provincial and municipal representatives. The formulas are also reviewed by an independent constitutional advisory institution, the Fiscal and Financial Commission (FFC). These formulas are used annually by the National Treasury to allocate shares of a total that is not affected by population statistics.

Distortions in the allocations arise from error in the data used to compute the allocations. We will use a loss function, as applied in statistical decision theory, to accomplish two purposes. First, the loss function will reflect rankings over alternative patterns of errors in allocation, with smaller loss corresponding to higher ranking and greater preference (National Research Council 1980, 84ff; Spencer 1980c). The loss functions considered here all take the value zero when there is no error in allocations arising from statistical error. The loss function is thus the negative of a utility function and satisfies the properties of a regret function (Berger 1985, 46ff, 376ff). The scale of the utility function is chosen (at least in theory), so that preferences under uncertainty, including risk aversion, are automatically taken into account when expected utility (or expected loss) is considered. Alternative axioms for preferences under uncertainty lead to focus on minimizing the maximum regret rather than expected regret or loss (Manski 2011). More generally, providing the probability distribution of loss – either the full multivariate distribution or

the marginal distributions for each of the recipients (e.g., local governments) can be informative. Second, we will use the loss function to compare costs of improving data to the benefits in terms of improved allocations (Spencer 1980a, 31–33).

Different perspectives have been taken in the literature on the effects of the distortions in allocations. One perspective addresses inequities that arise because the allocations differ from those that would arise if the legislated formulas were applied to error-free data. A second perspective looks at inefficiencies and reductions in social welfare that are believed to arise when the allocations are based on data with error instead of error-free data. Our analysis will focus on inequities because we believe that measuring changes in social welfare caused by distortions in allocations arising from data error is simply too difficult.

### 2.2. *Loss in Social Welfare from Errors in Allocations*

Analyses of benefits of censuses arising from increased "utility" or social welfare have been conducted recently for England and Wales (Cope 2015) and New Zealand (Bakker 2014). Although the details of the analysis for England and Wales could not be discovered by the authors, Cope mentions differences in utility from overallocations and underallocations and refers to the sum of net differences as "efficiency loss." More details are available for the analysis of the value of the New Zealand (NZ) census and associated population statistics. Bakker (2014, 50–53) considered distortions in allocations with the NZ health funding formula. The analysis assumed that the allocations based on error-free population data maximized the welfare of NZ residents. In particular, let $H_a$ and $\hat{H}_a$ denote the health expenditure allocations to area $a$ with error-free data and actual data, respectively, and let $X_a$ denote other final consumption expenditure to area $a$, with $a = 1, \ldots, A$. The analysis specified that the social welfare $W$ from health formula allocations $\hat{H}_a$ and other final consumption expenditures $X_a$ has the form $W(\hat{\mathbf{H}}, \mathbf{X}) = \sum_a X_a + u_a(\hat{H}_a)$ with $\hat{\mathbf{H}} = (\hat{H}_1, \ldots, \hat{H}_A)$, $\mathbf{X} = (X_1, \ldots, X_A)$, and $u_a(\hat{H}_a) = H_a \log(\hat{H}_a)$. This social welfare specification implies that the optimal distribution of a fixed sum equal to $\sum_a H_a$ occurs when the allocation to area $a$ is indeed equal to $H_a$. The total loss from distortions in health expenditure allocations was taken to be $W(\mathbf{H}, \mathbf{X}) - W(\hat{\mathbf{H}}, \mathbf{X})$. This is non-negative and is equal to $\sum_a u_a(H_a) - u_a(\hat{H}_a)$ or $= \sum_a H_a[\log(H_a) - \log(\hat{H}_a)]$. It is important to note that, other than the assumptions of optimality and decreasing marginal utility from health-funding allocations as reflected by $u_a(\cdot)$, the analysis made no attempt to justify the specifications involving $W$ and $u_a(\cdot)$. Different specifications would lead to different assessments of loss from distortions in allocations due to data error.

The assumption that an allocation formula is optimal should not be made casually. The United States' experience indicates diverse ways that formulas fail to be optimal (Buehler and Holtgrave 2007). The National Research Council (2003) report, *Statistical Issues in Allocating Funds by Formula*, commissioned several papers examining the design, development, structure, and inherent compromises in intergovernmental aid formulas. Downes and Pogue (2002) discuss the "often contradictory aid objectives . . . [and] assess the extent to which, in practice, formulas deviate from the ideal" (National Research Council 2003, 97). Zaslavsky and Schirm (2002) describe formula complexities such as hold-harmless provisions, floors, ceilings, and inconsistent data sources; they describe how their effects can be difficult to predict and can "produce allocations that don't line up

with original intentions" (National Research Council 2003, 97). Similar critiques appear in Spencer (1982b). Melnick (2002) describes the legislative process by which allocation formulas "pass the test for face validity while generating the necessary political support" (National Research Council 2003, 97). Possibly, legislators are motivated to secure the most funding for their constituents, but by including factors representing need, capability, and effort, the formulas appear as if they are addressing program goals. A legislator who participated in the development of a complex formula for General Revenue Sharing, a program that would distribute more than USD 55 billion in the United States between 1972 and 1980, described the process this way: "We finally quit, not because we hit on a rational formula, but because we were exhausted. And finally we got one that almost none of us could understand at the moment. We were told that the statistics were not available to run the [computer] print on it. So we adopted it, and it is here for you today" (quoted in Spencer 1980a, 152). Furthermore, even if the formula could be regarded as optimal when the input data were error-free, the formula allocations may not be optimal, for example, if the allocations also depended on other data series that contained error. For example, Schirm et al. (1999) discuss estimation error for local governments.

In conclusion, analysis of benefits of improved data in terms of increased social welfare arising from more accurate formula-based allocation of funds should be used with caution, unless the formula can be demonstrated to be optimal and the form of the social welfare function can be justified.

### 2.3. *Loss from Inequity in Allocations Due to Data Error*

The very names of the PES and LGES, Provincial Equitable Share and Local Government Equitable Share, indicate the importance of equitable allocations in South Africa. Therefore, we did not attempt a social welfare analysis based on assumptions of formula optimality. Instead, we considered which patterns of distortions of allocations would lead to larger increases in inequity for the local governments and their people.

For the purposes of the analysis, the allocations will be considered to be correct if there is no error in the statistics used as inputs to the allocation formulas. We will index the $n$ units (provinces or municipalities) receiving allocations by $i = 1, \ldots, n$. The correct allocation to recipient unit $i$ will be denoted by $\theta_i$ and the allocation based on statistics will be denoted by $\hat{\theta}_i$. The arrays of allocations are respectively denoted by $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ and $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_n)$. The component loss function for misallocation to unit $i$ is denoted by $\ell_i(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ and the aggregate loss equals the sum of the component losses,

$$\sum_{i=1}^{n} \ell_i(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}). \tag{1}$$

Summing the component losses to the recipients, as in (1), is consistent with a utilitarian view of social welfare measurement (Spencer 1985, 816–817). In addition to considering aggregate loss, it is important to also ensure that the expected component loss $E\ell_i(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ is not excessive for any recipient $i$. This principle could be extended to see that the upper quantiles of the component loss functions are not excessive for any recipient.

To motivate the form of the component loss functions $\ell_i(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ consider the asymmetry of the recipients' views regarding positive and negative errors in allocation. If the error in the allocation, $\hat{\theta}_i - \theta_i$, is negative (an underpayment), the recipient unit suffers a shortfall

equal to that amount. A simple measure of loss in this case is $a(\theta_i - \hat{\theta}_i)$ with $a > 0$. If the error in the allocation, $\hat{\theta}_i - \theta_i$, is positive (an overpayment), the recipient is receiving a positive benefit. In this case, a simple measure of loss is $-b(\hat{\theta}_i - \theta_i)$ with $b > 0$. A simple component loss function for recipient unit $i$ that takes this perspective into account is

$$\ell_i(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = a(\theta_i - \hat{\theta}_i)^+ - b(\hat{\theta}_i - \theta_i)^+, \tag{2}$$

where $(x)^+ = \max\{x, 0\}$. Perceiving an underpayment to be somewhat more consequential than an overpayment of the same magnitude, we have $a > b \geq 0$, but the ratio $b/a$ will not be too small. For the PES and LGES, the fact that the total amount allocated is fixed implies that the sum of the overallocations must equal the sum of the underallocations, and hence

$$\sum_{i=1}^{n} a(\theta_i - \hat{\theta}_i)^+ - b(\hat{\theta}_i - \theta_i)^+ = c\sum_{i=1}^{n} |\hat{\theta}_i - \theta_i|, \tag{3}$$

with $c = (a - b)/2$. The non-negativity of $b$ implies $c \leq a/2$. The value of $c$ is considered further in Section 9.

The loss function (3) refers to one year's allocation at a time. To account for multiple years of allocation, we sum the loss functions for the individual years from 2017 through 2021 to obtain the aggregate loss function

$$\ell(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = c \sum_{y=2017}^{2021} \sum_{i=1}^{n} |\hat{\theta}_{iy} - \theta_{iy}|. \tag{4}$$

In effect, this treats the years independently and does not allow for cancellation of a recipient unit's underpayment one year by equivalent overpayment the following year. However, the factor $c$ does account for offsetting of underpayments and overpayments to different units in the same year. The benefit of reducing errors in allocations is measured by the reduction in the expected value of the aggregate loss when $\hat{\boldsymbol{\theta}}$ is developed with the availability of the 2016 Census data, versus when 2016 Census data are not available.

### *2.4. Additional Rationale for the Loss Function*

In applying statistical decision theory, the optimality criterion should lead to the desired choices. The loss function (3) satisfies the criterion of Fisher-consistency, in that minimization of loss occurs precisely when the allocations are correct, that is, when $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$ (Spencer 1980a, 36). If Fisher-consistency is violated, then minimization of expected loss would lead to statistical inaccuracy being optimal, which is contrary to the principles of statistical agencies. A generalization of (2) is given by

$$\ell_i(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = w_i\big[aH\big((\theta_i - \hat{\theta}_i)^+\big) - bH\big((\hat{\theta}_i - \theta_i)^+\big)\big], \tag{5}$$

with $w_i > 0$, $H(0) = 0$, and $H$ strictly increasing on $[0, \infty)$. The criterion of Fisher-consistency imposes strong restrictions on the weights $w_i$ and the shape of $H$ in (5). Requiring that (1) remain Fisher-consistent for an arbitrary number $n$ of recipients and any

size errors in allocation leads to the conditions that

$$\frac{\max w_i}{\min w_i} < \frac{a}{b} \tag{6}$$

and

$$A \leq \frac{H(x)}{x} \leq B \tag{7}$$

for $x \geq 1$ and for positive constants $A$, $B$ not depending on $n$ (Spencer 1980a, 41–46). Condition (6) implies that the weights cannot be inversely proportional to $\theta_i$, for example, because the values of $\theta_i$ vary widely. Provided that condition (6) holds, it is possible that the weights might be inversely proportional to per capita income in the areas. On the other hand, distribution of income (or wealth) within the recipient units (provinces or municipalities) could also be important, motivating alternative weights. Condition (7) rules out choice of a nonlinear power function for $H$. Thus, choosing more complicated component loss functions either leads to violation of Fisher-consistency or to component loss functions similar to (2).

## 3. Accuracy of Mid-Year Population Estimates for Provinces, 2002–2011

### *3.1. Overview and Motivation*

The performance of mid-year estimates based on the 2001 Census and accounting for ten years of change can be assessed by comparing with the 2011 Census results. The error structure observed for the 2001–2011 period will be extrapolated to the 2011–2021 period (Section 4). As in other evaluations of population estimates to account for post-censal change, we find that the estimates under-predict growth or decline in shares of the population (Subsection 3.2). To estimate the variances of mid-year estimates that account for ten years of change, we analyze deviations in the average errors for provinces in which the relative share of the population was growing or shrinking. To model the variances for time spans less than ten years, we consider two models of year-to-year correlation between estimates of yearly population change, independence or correlation equal to 1.

Thus, in the absence of a 2016 Census, mid-year estimates will need to account for 6–10 years of population change since the 2011 Census. Evidence of their accuracy is derived from the analysis of accuracy of the mid-year population estimates produced using the 2001 Census as a base, as discussed in Subsection 3.2.

### *3.2. Biases of Estimates of Population*

Denote the (mid-year) estimate of a province's population size $t$ years after the census by $\hat{P}_t$, and denote the actual population size by $P_t$. Thus, $P_0$ denotes the population size at the time of the last census. Numerical values for $P_0$ and $P_{10}$ are taken from prior census results (Stats SA 2012a, Table 2.1 for 2001 population and Table 2.9 for 2011 population), with undercount adjustments for both censuses (Stats SA 2012b, section 5). All censuses in South Africa (at least since the 1996) have had undercount adjustments based on data from

a post-enumeration survey. Following a matching procedure to identify persons who should have been enumerated (and those that should not have been), the adjustments are predicted using Chi-square Automatic Interaction Detection, CHAID, technique using race, geographic category, sex, and age. These are applied to produce adjustment classes. Summing up the adjusted population across adjustment classes produces a separate ratio estimate of a total, from which the national adjusted population could be calculated. At the municipal level, the effect of adjustment will vary according to the share of different adjustment classes present in that municipality; see Stats SA (2012b, Section 5).

Figure 1 plots estimated percent change based on the mid-year population estimates for 2011, $(\hat{P}_t - P_0)/P_0$, against the observed percent change based on the 2011 Census adjusted for undercount, $(P_t - P_0)/P_0$. Note that small changes are overestimated and larger changes are underestimated.

Let $\varepsilon_t$ denote the relative error in the estimate of population change $t$ years since the last census,

$$\varepsilon_t = \frac{(\hat{P}_t - P_0) - (P_t - P_0)}{P_t - P_0} = \frac{\hat{P}_t - P_t}{P_t - P_0}.$$

Figure 2 plots $\varepsilon_t$ versus the observed relative change in population. The relative errors are positive for relative changes below 15% and are negative for changes above 15%. Calculations based on the 2011 mid-year population estimates ($\hat{P}_t$), adjusted census counts

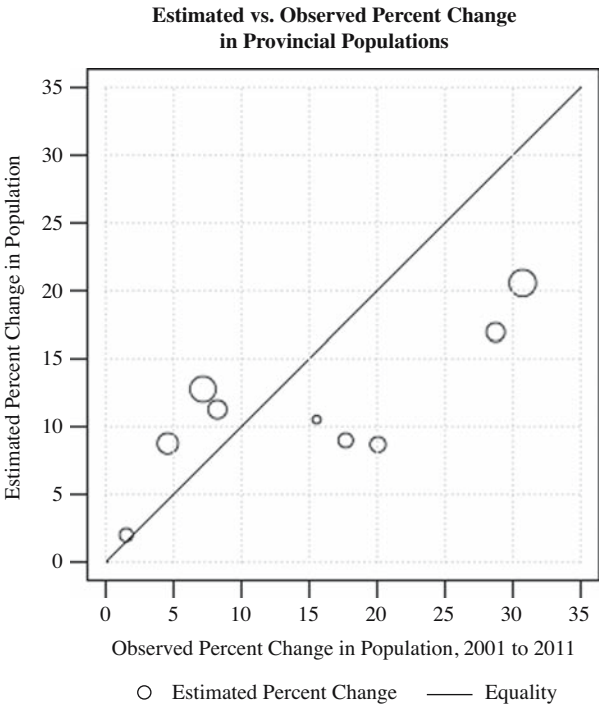


*Fig. 1. Estimated versus observed change in province population size ten years after 2001 Census. Area of circle is proportional to average of 2001 and 2011 population sizes.*

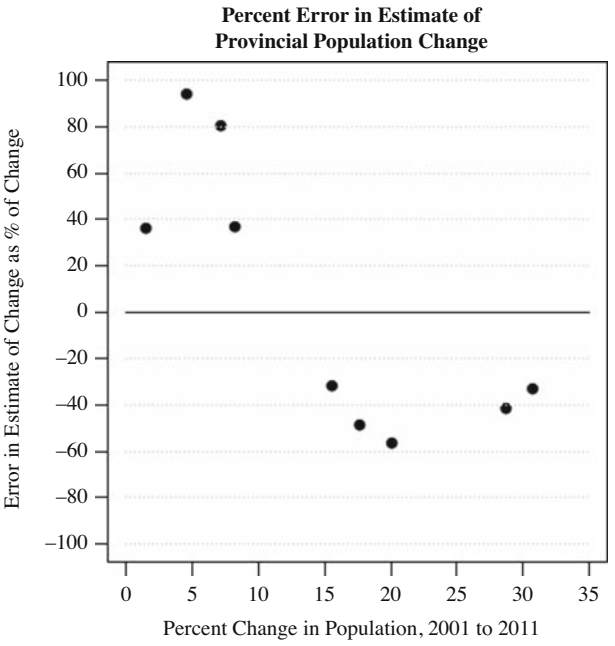


*Fig. 2. Relative error of estimate of change versus observed percent change in province population size ten years after 2001 Census.*

for 2011 ($P_t$), and the 2001 Census counts ($P_0$) for provinces show that the average value of $\varepsilon_{10}$ is $+0.42$ for the five provinces that grew by more than 12% in size between 2001 and 2011 and is $-0.62$ for the four provinces that grew by less than 12% over that period. Given the small number of observations and the similarity in magnitudes, we decided to specify the same magnitude for provinces growing faster and slower than average, leading to the model that the expected value of $\varepsilon_{10}$ is

$$E(\varepsilon_{10}) \approx 0.52 \times \text{sgn}\,(dP_0 - P_{10}), \tag{8}$$

with $d = 1.12$ and $\text{sgn}\,(x) = 1$ if $x > 0$, $\text{sgn}\,(x) = -1$ if $x < 0$, and $\text{sgn}\,(x) = 0$ if $x = 0$.

Specifying the mean of $\varepsilon_t$ for intermediate times $1 \le t < 10$ requires some assumptions, since we have direct information only about $\varepsilon_{10}$. Denote the incremental error in the estimate of annual change by $\delta_t = (\hat{P}_t - \hat{P}_{t-1}) - (P_t - P_{t-1})$, with $\hat{P}_0 = P_0$ by assumption. It follows that

$$\hat{P}_t - P_t = (P_t - P_0)\varepsilon_t = \sum_{s=1}^{t} \delta_s.$$

Given the short time span, it is reasonable to use the simple approximation that the expected incremental error for a province is the same for each of the ten years, that is,

$$E(\delta_t) = (P_{10} - P_0)E(\varepsilon_{10})/10, \quad 1 \le t \le 10. \tag{9}$$

### *3.3. Variances of Estimates of Population*

The average squared deviation regarding the mean for the relative errors observed for 2011 was 0.01067, leading to the model that the variance of $\varepsilon_{10}$ is

$$V(\varepsilon_{10}) = 0.01067.$$

As in the case of the mean, specifying the variance of $\varepsilon_t$ for $1 \leq t < 10$ requires assumptions, since we have direct information only about $\varepsilon_{10}$. A simple model for the variances of the incremental errors is that $V(\delta_s)$ does not change with $s$. If the incremental errors in a province are independent over time, then $V\left(\sum_{s=1}^{t} \delta_s\right)$ grows linearly with $t$, and hence $V(\delta_s) = 0.001067(P_{10} - P_0)^2$. On the other hand, if the incremental errors in a province are perfectly correlated over time, then $V\left(\sum_{s=1}^{t} \delta_s\right)$ is quadratic in $t$, and $V(\delta_s) = 0.0001067(P_{10} - P_0)^2$. We are assuming that the incremental errors in different provinces are mutually independent. To summarize, we have two alternative models for the variances of sums of incremental errors within provinces, the independent increments model

$$V\left(\sum_{s=1}^{t} \delta_s\right) = 0.001067t\,(P_{10} - P_0)^2, \tag{10}$$

and the dependent increments model,

$$V\left(\sum_{s=1}^{t} \delta_s\right) = 0.0001067t^2\,(P_{10} - P_0)^2. \tag{11}$$

### *3.4. Accuracy of Estimates of School-age Population of Provinces*

The observed errors in mid-year estimates of ten-year change in the school-age population (i.e., persons aged 5–17) from 2001 to 2011 were all positive. The magnitudes were proportional to the error in the estimated ten-year change in the total province population, with different constants of proportionality for overestimates and underestimates of total population change. Estimates of those proportionality constants are 0.80 and $-0.26$, respectively. This means that the prediction of error in the school-age population estimate is 0.80 times the predicted error in the estimate of the total province population if the predicted error is positive. The prediction of error in the school-age population estimate is $-0.26$ times the predicted error in the estimate of total province population if the predicted error is negative. In both cases, the predicted error in the estimate of school-age population is positive.

## 4. Distributions of Mid-Year Population Estimates, 2017–2021

### *4.1. Overview*

If no 2016 Census is carried out, population estimates for 2017–2021 must account for 6–10 years of change since the 2011 Census. If the 2016 Census is carried out, mid-year estimates for 2017–2021 will need to account for only 1–5 years of population

change since the 2016 Census. Error distributions for the two sets of estimates are based on the analysis of Section 3. To specify the distributions of the estimates for 2017–2021, we add the errors to the specified true values of the population. The true values of the future population are developed in Subsection 4.2. Then, the distributions of estimates without a 2016 Census (Subsection 4.3) and with a 2016 Census (Subsection 4.4) are developed.

### 4.2. *Specifications of True Population of Provinces, 2017—2021*

To specify true values of total population and school-age population (ages 5–17) in provinces, we utilized projections of future population prepared in 2003 by the Actuarial Society of South Africa (ASSA). These projections are referred to as the ASSA2003 population projections, and they are prepared using the 2001 Census as a base (after adjustment for undercount). We assume that the true population sizes are unaffected by whether or not a 2016 Census is carried out. This is a nontrivial assumption since more accurate population data may lead to better provision of services, which can in turn influence fertility and mortality rates, as well as migration flows, as migrants seek access to better resourced areas that can provide better services. For example, in the case of the former, HIV/AIDS, low birth weight, and diarrheal diseases accounted for more than 60% of under age five deaths in South Africa at the time of the 2001 Census (Bradshaw et al. 2003). A range of primary health and basic service interventions has been found to have a direct impact on these causes (Bhutta et al. 2013). Many of these would be affected by inequalities arising from inaccurate population data and inadequate resource allocation to the authority responsible for their implementation (Say and Raine 2007). This includes vitamin A supplementation, the provision of Antiretroviral Therapy, the availability of healthcare workers, and the provision of adequate sanitation and protected water.

We use the ASSA projections in two alternative ways to specify true future population values. One specification is simply the total population as projected by the ASSA, and the other specification multiplies ASSA forecasts by the ratio of the undercount-adjusted 2011 Census figure for the province to the ASSA forecast for the 2011 population of the province. The latter "calibrated" population thus coincides with the undercount-adjusted census number for 2011. For school-age population (ages 5–17), one specification was derived from the ASSA2003 projections for five-year age groups, with population numbers disaggregated by single age based on the Sprague multiplier software on the Stats SA website. As with total population, a second specification was developed by ratio-adjusting (calibrating) the school-age population forecasts to agree with undercount-adjusted 2011 Census school-age population numbers. The two alternative sets of true values are denoted by the indicator $k$ taking values 1 (uncalibrated) and 2 (calibrated).

### 4.3. *Specifications of True Population of Municipalities, 2017–2021*

The true values of total population for municipalities as used in the LGES can be taken to be the values for 2016, because no updating for post-censal population change is used in the LGES. Lacking ASSA projections of 2016 values for municipalities, we carried out a simple modeling of future values by extrapolating the 2001–2011 trends in the statistical

inputs to the formula to 2016. This was subject to the constraint that the change from 2011 to 2016 could not exceed 50% of the 2011 total population size of the municipality.

### *4.4. Distribution in the Absence of a 2016 Census*

For province estimates in the no-2016-census scenario, the variances of sums of incremental errors in mid-year population estimates are given by (10) or alternatively by (11). Using the independent increments assumption, we model the ten values $\delta_1, \ldots, \delta_{10}$ as independently normally distributed with means given by (8) and (9) with $d = 1$ and variances given by (10). Expression (10) can be evaluated because the modeling described in Subsection 4.2 specifies $P_0$ and $P_{10}$. Alternatively, using the dependent increments assumption, we model $\delta_1$ as normally distributed, with means given by (8) and (9) and variance given by (11), and $\delta_{10} = \cdots = \delta_1$. The two alternative independence assumptions are denoted by the indicator $l$ taking values 1 (independence) and 2 (perfect dependence).

The population estimate for province $i$, in year $y$, for dependence model $l$, corresponding to true value specification $k$ (indicating uncalibrated or calibrated forecast), is denoted by $\hat{P}_{iykl}^{nocen16}$ when no 2016 Census is conducted and by $\hat{P}_{iykl}^{cen16}$ when a 2016 Census is conducted.

For municipalities, the error in the population estimate for municipality $m$ in year $y, 2017 \leq y \leq 2021$, is equal to $P_{m2011} - P_{m2016}$ in the no-census scenario, since errors in census numbers are ignored.

### *4.5. Accuracy in the Presence of a 2016 Census*

In a scenario with a 2016 Census, mid-year estimates for provinces for $2011 + t$, $6 \leq t \leq 10$ are based on the 2016 Census, and therefore account for only $t - 5$ years of population change. This is in contrast to the estimates in the no-2016-census scenario, which must account for the full $t$ years of population change. Therefore, in a 2016 Census scenario, for each province the joint distribution of $\delta_t$, $6 \leq t \leq 10$ is equal to the joint distribution of the corresponding values of $\delta_{t-5}$ in the no-2016 census scenario.

For municipalities, the error in the population estimate for municipality $m$ in year $y$, $2017 \leq y \leq 2021$, is identically zero under the 2016 Census scenario, since errors in census numbers are ignored.

## 5. Distributions of PES and LGES Allocations

### *5.1. Hypothetical True Values*

In the analysis, the true values of the allocations are allowed to change over time as the true population changes (Subsection 4.2). LGES allocations depend only on the population numbers for municipalities according to the latest census. PES allocations depend not only on population statistics, but on other statistics as well. To fully model the joint distribution of the various statistics and their underlying true values would have involved substantial additional work and would have added to the complexity of the analysis. Instead, our analysis conditions on (i.e., takes as fixed) the values of the nonpopulation statistics which served as inputs to the 2011 PES allocations.

The true allocation for province $i$, in year $y$, for true value specification $k$ (indicating uncalibrated or calibrated forecast) is denoted by $\theta_{iyk}$.

### 5.2. Specifying and Simulating Errors in PES Allocations

The errors in PES allocations are functions of population numbers only, because any nonpopulation statistics are held fixed. The joint distributions of the true and estimated allocations are determined by the joint distributions of the true and estimated populations. Recall that for the population of a province in a given year, in both the 2016 Census and no-census scenario, there are four alternative specifications, depending on whether or not the forecasts specifying the true values were calibrated and whether the estimates of year-to-year change are independent or perfectly dependent over time. For each of the eight specifications, we randomly generated four independent replications, which we denote by $r = 1, \ldots, 4$. These yielded four replications of population estimates $\hat{P}^{cen16}_{iyklr}$ and $\hat{P}^{nocen16}_{iyklr}$, respectively, in the case when a 2016 Census is and is not taken. (To increase the precision of estimated reduction in expected loss due to the 2016 Census, we set not only the distributions, but the realizations of $\delta_t$, $6 \leq t \leq 10$, equal to realizations of $\delta_{t-5}$ in the no-2016 census scenario.) Each replication of population estimates leads to a replication of the allocation, $\hat{\theta}^{cen16}_{iyklr}$ and $\hat{\theta}^{nocen16}_{iyklr}$, respectively. The corresponding errors in allocation are $\hat{\theta}^{cen16}_{iyklr} - \theta_{iyk}$ and $\hat{\theta}^{nocen16}_{iyklr} - \theta_{iyk}$.

### 5.3. Specifying and Simulating Errors in LGES Allocations

As with the PES, the errors in LGES allocations are functions of population numbers only, because any nonpopulation statistics are held fixed. The joint distributions of the true and estimated allocations are determined by the joint distributions of the true and estimated populations. Recall that there are only two possible alternative estimates for the population of a municipality $m$ in year $y$, $P_{m2016}$ and $P_{m2011}$, corresponding to the 2016 Census scenario and the no-census scenario. The corresponding allocations to municipality $m$ in year $y$ are denoted by $\hat{\theta}^{cen16}_{my}$ and $\hat{\theta}^{nocen16}_{my}$.

## 6. Estimating Improvement in the Allocations As a Result of the 2016 Census

### 6.1. Estimating Reduction in Expected Loss from Errors in PES Allocations

The reduction in expected loss from errors in PES allocations when the 2016 Census is conducted is $E[\ell(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{nocen16}) - \ell(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{cen16})]$, where the loss function is specified by (4). To estimate this reduction in expected loss, we use the scaling constant $c$ times

$$\frac{1}{16}\sum_{y=2017}^{2021}\sum_{i=1}^{9}\sum_{k=1}^{2}\sum_{l=1}^{2}\sum_{r=1}^{4}\left|\hat{\theta}^{nocen16}_{pyklr} - \theta_{pykl}\right| - \frac{1}{16}\sum_{y=2017}^{2021}\sum_{i=1}^{9}\sum_{k=1}^{2}\sum_{l=1}^{2}\sum_{r=1}^{4}\left|\hat{\theta}^{cen16}_{pyklr} - \theta_{pykl}\right|. \quad (12)$$

Expression (12) shows the model averaging approach used to manage the different options for calculating the true population (calibrated or not) and the two variance options.

For practical considerations arising from tight decision deadlines, instead of computing the allocations for each year from 2017 to 2021, we computed the allocations just for 2021

for both $\hat{\theta}$ and $\theta$, and we used those values for each year. This likely led to a modest overstatement of the reduction in expected loss due to the 2016 Census, since the accuracy of the mid-year population estimates is at its lowest in 2021. The calculated value of (12) is ZAR 4.8 billion.

One technical point is worth noting. By ignoring error in any nonpopulation statistics in the allocation formulas, we are, in effect, approximating $E[\ell(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{nocen16}) - \ell(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{cen16})]$ by $E[\ell(\boldsymbol{\theta}', \hat{\boldsymbol{\theta}}^{nocen16}) - \ell(\boldsymbol{\theta}', \hat{\boldsymbol{\theta}}^{cen16})]$, where $E[\,\cdot\,]$ denotes expectation and $\boldsymbol{\theta}'$ denotes the array of allocations when the population statistics have no error, but the other statistics are observed with possible error. Research in progress suggests that the approximation either overstates or only modestly understates the reduction in expected loss.

### *6.2. Estimating Reduction in Expected Loss from Errors in LGES Allocations*

Recall that the LGES allocations for 2017–2021 will be based on the 2011 Census, if the 2016 Census is not conducted, and on the 2016 Census if it is conducted. As was the case for the PES, we approximate $E[\ell(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{nocen16}) - \ell(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{cen16})]$ by $E[\ell(\boldsymbol{\theta}', \hat{\boldsymbol{\theta}}^{nocen16}) - \ell(\boldsymbol{\theta}', \hat{\boldsymbol{\theta}}^{cen16})]$, where $E[\,\cdot\,]$ denotes expectation and $\boldsymbol{\theta}'$ denotes the array of allocations to municipalities when the population statistics have no error, but the other statistics are observed with possible error. By construction, $\boldsymbol{\theta}' = \hat{\boldsymbol{\theta}}^{cen16}$ and so we estimate the reduction in expected loss by scaling constant $c$ times

$$5\sum_{m=1}^{278}\left|\hat{\theta}_{m2016}^{nocen} - \hat{\theta}_{m2016}^{cen}\right|, \tag{13}$$

where $m$ indexes the 278 municipalities and the allocations are calculated for 2016. The calculated value of (13) is ZAR 32.1 billion. As the LGES is assumed to allocate only 1/15 as much money as the PES program over the five-year period, ZAR 38.9 billion for the LGES versus ZAR 600 billion for the PES, it is surprising that (13) is more than six times as large as (12). The explanation is much larger differences in LGES allocations in the presence or absence of the 2016 Census. Even though mid-year estimates do not estimate population change accurately, PES allocations are based on the total population level. Mid-year estimates predict population levels much more accurately than population change, whereas in the LGES, the municipal estimates of population levels are not updated at all in the absence of a 2016 Census.

## 7. Limitations

Several limitations to the analysis of reduction in PES and LGES misallocations arising from the 2016 Census may be noted.

1. The specifications for true values for province populations, which depend on the ASSA projections, are inaccurate to an unknown degree.
2. The true values of population for the LGES allocation are taken to be for 2016 rather than the true population sizes for 2017–2021.
3. We are ignoring the effects of errors in nonpopulation statistics that are used to calculate PES and LGES allocations. This may well increase the estimated magnitude of improvement in allocations in conducting the 2016 Census.

4. Distribution of error in mid-year population estimates 2017–2021 could be different than in last decade, due either to changes in patterns of population growth or decline or to differences in quality of data used to estimate births, deaths, and net migration among provinces.
5. Errors in 2011 Census numbers used in the analysis can cause errors in estimates of error in mid-year population estimates for 2011 (Spencer 1980b). Our analysis ignores possible error in the 2011 Census numbers.
6. Hold-harmless provisions in the allocation formulas were not taken into account.

The effects of the limitations noted in points 1, 2, and 5 might be slightly reduced by use of a prior distribution to specify uncertainty about true values, as part of a full Bayesian decision theoretic analysis. However, it is unlikely that this will greatly change the estimates of expected loss.

## 8. 2016 Census Cost

The costs of census-taking include the investment cost, the amount spent on the collection, capture, cleaning, and data assurance (quality control). Other costs that are often forgotten include data curatorship, which refers to looking after, updating and maintaining the data and the ongoing assistance provided to the users of this data. Finally, dissemination and publicity also carry costs. Nonetheless, for a standard cost-benefit analysis, estimating the direct costs of a Census is relatively straight-forward, to the extent that reliable and up-to-date expenditure data are available from the appropriate government departments. Some indirect costs, such as calculating cost of the time taken by respondents to complete a census questionnaire, may be more complex, but can be estimated using an appropriate shadow wage rate. This has not been undertaken for this study.

In the absence of a 2016 Census, it is assumed that some variation of the 2007 Community Survey would be conducted, and it is assumed that the cost of the mid-year estimates program is essentially unchanged regardless of whether the census or the Community Survey is taken in 2016. The net additional cost of the 2016 Census (over and above the 2016 Community Survey) was predicted to be on the order of ZAR 3 billion.

## 9. Measuring Benefit from Improvement in Allocations

The measures of reduction in absolute values of misallocations, such as (12) and (13) should not be interpreted directly as measures of benefit. In monetary terms, the sum of the overallocations equals the sum of the underallocations, or equivalently, one area's loss is another area's gain. As discussed in Subsection 2.3, the benefit arises from reduction of inequity of the allocations. The translation from (12) and (13) to benefit, or reduction of expected loss, is achieved through the scaling constant $c$ in the loss function (4). The scaling constant $c$ should reflect the sensitivity of society or the decision-makers to misallocations. Logically, the value of $c$ should not be as large as 1, as in the cautionary example of *Jarndyce v. Jarndyce* (Dickens 1985). If overallocations are viewed as beneficial or benign for the local governments that receive them, then $c \leq 0.5$, as noted in Subsection 2.1. Ultimately, however, the magnitude of $c$ depends on the decision-maker's preferences regarding tradeoffs for equitable allocations versus spending money to

achieve the equitable allocations. If it is just worth spending ZAR ten million to reduce the sum of absolute misallocations by ZAR one billion, then $c = 0.01$. If it is just worth spending ZAR 100 million to reduce the sum of absolute misallocations by ZAR one billion, then $c = 0.10$, and if it is just worth spending ZAR 500 million to reduce the sum of absolute misallocations by ZAR one billion, then $c = 0.50$.

We believe that the specification of $c$ is inherently subjective and should be openly addressed. People's values are not objectively determined, and the choice of the scaling constant $c$ involves a question of values – how much is it worth spending to achieve more equitable allocations. Our analysis has drawn on technical analyses to compute the expected loss as parameterized by the scaling constant $c$. However, the specific choice for $c$ reflects the willingness of the decision-makers to use tax dollars to reduce inequity in fund allocations. Having a single, easily interpretable parameter for social values conveys the additional advantage of providing transparency to the analysis.

Spencer (1980a) suggested that, $c = 0.01$ in the 1970s context of General Revenue Sharing in the United States The rationale, as discussed in Subsection 2.3 above, was that if $x = \hat{\theta}_i - \theta_i$ and $x < 0$, then local government $i$ incurs a deficit of $|x|$, and if $x > 0$ then it incurs a surplus of $|x|$. If a deficit of $|x|$ is incurred, local government is assumed to borrow an amount equal to the shortfall, to be repaid in the next fiscal period. If the interest rate for the period was $a - 1$, the monetary loss to the local government would be $a|x|$. We neglected long-term effects because they are hard to trace and because the local government cannot make adjustments for the deficit before the end of the current period, but it can make adjustments after the period. Conversely, if $x > 0$, so that a surplus is produced, the local government invests $|x|$ for the period at interest rate $b - 1$. The local government's monetary loss incurred is $-b|x|$, a negative loss (i.e., a gain). From (3), $c = (a - b)/2$, and so we may interpret $c$ as half of the difference between the local governments' interest rates for investing versus borrowing for the period. Spencer (1980a) took the period to be one year and the difference between interest rates to be 0.02, leading to a specification that $c = 0.01$. In this scenario, the choice of $c$ would reflect economic conditions and the length of the period that the local government would need to adjust for the shortfall.

Table 1 shows the expected improvement in allocation when a 2016 Census is conducted, for various values of $c$.

To illustrate, if $c = 0.06$ is a reflection of the preference tradeoff between PES and LGES equity on the one hand and expenditure on the other, the benefit, in terms or more equitable allocation of funds, is ZAR 2.2 billion. If a 2016 Census will cost an additional ZAR three billion (beyond the cost of a 2016 Community Survey), then the improvement in allocation of funds justifies about three quarters of the census cost. Other uses of the data would need to justify the remaining ZAR 800 million of the census cost. If $c > 0.08$, then the benefit of improvement in allocation of funds equals or exceeds the census cost, in which case the analysis would provide strong support for a 2016 Census.

## 10. Discussion

As mentioned at the outset, the decision to fund a Census in 2016 is not only dependent upon the costs and anticipated financial benefits involved, and the South African

*Table 1. Effects of 2016 Census on Improvement of Allocation of Funds to Provinces and Municipalities, 2017–2021, with Alternative Levels of Scaling.*

| Expected reduction from misallocations when a 2016 Census is carried out (ZAR millions) | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Scaling constant, c* | | | | | | | | | | |
| | 1.00 | 0.50 | 0.30 | 0.20 | 0.15 | 0.10 | 0.08 | 0.06 | 0.04 | 0.02 | 0.01 |
| PES | 4,800 | 2,400 | 1,440 | 960 | 720 | 480 | 384 | 288 | 192 | 96 | 48 |
| LGES | 32,121 | 16,061 | 9,636 | 6,424 | 4,818 | 3,212 | 2,570 | 1,927 | 1,285 | 642 | 321 |
| Total | 36,921 | 18,461 | 11,076 | 7,384 | 5,538 | 3,692 | 2,954 | 2,215 | 1,477 | 738 | 369 |

Government made the decision not to undertake a census in 2016 (Stats SA 2014a, 21). Instead, the Government decided to improve its data collection program. An enlarged Community Survey, with a sample size increased from 300,000 households to one million households, is being undertaken in 2016 and is projected to cover all enumerator areas in the country (Parliament of the Republic of South Africa 2014, 3424). Furthermore, the agency has focused on improving civil registration of vital statistics to be able to better estimate the mid-year population (Stats SA 2014a, 59). Further considerations include a long-term strategy to introduce a continuous population survey that will collect population and other social statistics on an ongoing basis. The methodology described above permitted this decision to be evidence-based, up to the subjective specification of the parameter $c$, and to confront the possible effects of error (Stats SA, 2014b). Indeed, the impact of prior error resulting from the ten-year gap between 2001 and 2011 has been taken into account in South Africa's most recent government budget. The Annex to the Budget notes that by not properly accounting for migration, the division of revenue between provinces has become inequitable, with receiving provinces such as Gauteng and the Western Cape being allocated less resources than would have been provided with accurate data. However, as the National Treasury (2015:17–18) acknowledges, provinces which have been receiving more resources need time to adjust to revised allocations, and a total ZAR 4.2 billion has had to be added to the PES over the three years from 2013 to 2015 to cushion the impact of the census data. The results of this partial cost-benefit analysis of South African census-taking contributed to greater awareness of the role played by official statistics in the allocation of resources, greater awareness of the wider costs of error, and of assessing the 'value for money' of official statistics. The decision to triple the size of the Community Survey in 2016 and the introduction of methodological improvements by Stats SA to improve cost effectiveness are examples of ongoing reflection concerning official statistics in South Africa and elsewhere (Stats SA 2016). The cost-benefit approach used in this article is applicable to other data programs as well, such as improvements in sample surveys and vital registration statistics, provided uses of the statistics are sufficiently understood.

## 11. References

Alho, J.M. and B.D. Spencer. 2005. *Statistical Demography and Forecasting*. New York: Springer.

Bakker, C. 2014. *Valuing the Census*. Wellington: Statistics New Zealand. Available at: www.stats.govt.nz.

Berger, J.O. 1985. *Statistical Decision Theory and Bayesian Analysis. 2nd ed.* New York: Springer.

Bhutta, Z.A., J.K. Das, A. Rizvi, M.F. Gaff, N. Walker, S. Horton, P. Webb, A. Lartey, and R.E. Black. 2013. "Evidence-based Interventions for Improvement of Maternal and Child Nutrition: What Can Be Done and at What Cost?" *Lancet* 382: 452–477.

Bradshaw, D., D. Bourne, and N. Nannan. 2003. "What Are the Leading Causes of Death among South African Children?" MRC Policy Brief, No 3., Medical Research Council, Bellville.

Buehler, J.W. and D.R. Holtgrave. 2007. "Challenges in Defining an Optimal Approach to Formula-Based Allocations of Public Health Funds in the United States." *BMC Public Health* 7: 44. Doi: http://dx.doi.org/10.1186/1471-2458-7-44.

Cope, I. 2015. The Value of Census Statistics in England and Wales. Note by the Office for National Statistics, United Kingdom. United Nations Economic and Social Council, 4 September 2015. Report ECE/CES/GE.41/2015/16. Available at: https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.41/2015/mtg1/CES_GE.41_2015_16_-_UK.pdf (accessed 25 March 2016).

Dickens, C. 1985. *Bleak House.* 1852–53. New York: Penguin.

Downes, T.A. and T.E. Pogue. 2002. "How Best to Hand Out Money: Issues in the Design and Structure of Intergovernmental Aid Formulas." *Journal of Official Statistics* 18: 329–352.

General Register Office for Scotland. 2006. 2011 Census Business Case. Prepared by John Aldridge, Consultant, July 2006. Available at: https://www.whatdotheyknow.com/request/8345/response/20302/attach/3/business%20case.pdf (accessed 8 March 2016).

Kruskal, W.H. 1984. "The Census as a National Ceremony." In *Federal Statistics and National Needs* prepared for the Subcommittee on Energy, Nuclear Proliferation and Government Processes, an arm of the Committee on Government Affairs of the United State Senate, by the Congressional Research Service of the Library of Congress, 177–180. Washington DC: U.S. Government Printing Office.

Manski, C.F. 2011. "Actualist Rationality." *Theory and Decision* 71: 195–210.

May, J., M. Dimbabo, J. Tamri, G. Wright, Z. Seeskin, and B.D. Spencer. 2013. *Cost Benefit Analysis of South Africa's Population Census, Final Report*, 21 May, 2013. Bellville, South Africa: Institute for Social Development, University of the Western Cape.

May, J. and P. Lehohla. 2005. "Counting the Costs of a 21st Century Census: South Africa's Census 2001." *Development Southern Africa* 22: 215–232.

McCaa, R., A. Esteve, S. Ruggles, and M. Sobek. 2006. "Using Integrated Census Microdata for Evidence-Based Policy Making: The IPUMS-International Global Initiative." *The African Statistical Journal* 2: 83–100.

Melnick, D. 2002. "The Legislative Process and the use of Indicators in Formula Allocations." *Journal of Official Statistics* 18: 353–370.

Ministry for Welfare and Population Development. 1998. *White Paper on Population Policy.* 7 September 1998. Pretoria: *Government Gazette*.

National Research Council. 1980. *Estimating Population and Income of Small Areas*. Panel on Small-Area Estimates of Population and Income, Committee on National Statistics. Washington DC: The National Academies Press.

National Research Council. 1985. *Natural Gas Data Needs in a Changing Regulatory Environment*. Panel on Statistics on Natural Gas, Committee on National Statistics. Washington DC: The National Academies Press.

National Research Council. 2003. *Statistical Issues in Allocating Funds by Formula.* Panel on Formula Allocations, edited by T.A. Louis, T.B. Jabine, and M.A. Gerstein. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

National Treasury. 2013. Annexure W1 to the Budget Review: Explanatory Memorandum to the Division of Revenue, National Treasury, Pretoria. Available at: http://www.treasury.gov.za/documents/national%20budget/2013/review/Annexure%20W1.pdf (accessed 1 August 2016).

National Treasury. 2015. Website Annexure to the 2015 Budget Review: Explanatory Memorandum to the Division of Revenue, National Treasury, Pretoria. Available at: http://www.treasury.gov.za/documents/national%20budget/2015/review/Annexure%20W1.pdf (accessed 1 March 2015).

Nigerian Tribune. 2016. "Towards a Credible Census." Available at: http://tribuneonlineng.com/towards-credible-census/ (accessed 16 January 2017).

Parliament of the Republic of South Africa. 2014. Announcements, Tablings and Committee Reports No. 95–2014 [First Session, Fifth Parliament, 19 November 2014], 3402, Cape Town. Available at: http://www.parliament.gov.za/live/commonrepository/Processed/20141124/593735_1.pdf (accessed 2 March 2015).

Parliament of the United Kingdom. 2009. Draft Census (England and Wales) Order 2009 etc – Merits of Statutory Instruments Committee. Available at: http://www.publications.parliament.uk/pa/ld200809/ldselect/ldmerit/176/17606.htm (accessed 8 March 2016).

Redfern, P. 1974. "The Different Roles of Population Censuses and Interview Surveys, Particularly in the U.K. Context." *International Statistics Review* 42: 131–146.

Savage, I.R. 1985. "Hard-Soft Problems." *Journal of the American Statistical Association* 80: 1–7.

Say, L. and R. Raine. 2007. "A Systematic Review of Inequalities in the Use of Maternal Health Care in Developing Countries: Examining the Scale of the Problem and the Importance of Context." *Bulletin of the World Health Organization* 85: 812–817.

Schirm, A.L., A.M. Zaslavsky, and J.L. Czajka. 1999. "Large Numbers of Estimates for Small Areas." *Proceedings of the 1999 FCSM Research Conference*. Available at: https://fcsm.sites.usa.gov/files/2014/05/IV-A_Schirm_FCSM1999.pdf (accessed 7 March 2016).

Seeskin, Z.H. and B.D. Spencer. 2015. "Effects of Census Accuracy on Apportionment of Congress and Allocations of Federal Funds." Institute for Policy Research Working Paper WP- 15-05. Evanston, IL: Northwestern University. Available at: http://www.ipr.northwestern.edu/publications/papers/2015/ipr-wp-15-05.html (accessed 8 March 2016).

Sims, C.A. 1984. "Can We Measure the Benefits of Data Programs?" In *Proceedings of the Social Statistics Section: American Statistical Association*, 60–67. Washington, D.C.: American Statistical Association.

Spencer, B.D. 1980a. *Benefit-Cost Analysis of Data Used to Allocate Funds*. New York: Springer.

Spencer, B.D. 1980b. "Effects of Biases in Census Estimates on Evaluation of Postcensal Estimates." In National Research Council, 1980, *Estimating Population and Income of Small Areas*. Panel on Small-Area Estimates of Population and Income: 232–6. Committee on National Statistics, Assembly of Behavioral and Social Sciences. Washington, D.C.: The National Academy Press.

Spencer, B.D. 1980c. "Implications of Equity and Accuracy for Undercount Adjustment: A Decision- Theoretic Approach." In U.S. Bureau of the Census, *Conference on Census*

*Undercount: Proceedings of the 1980 Conference*: 204–216. Washington, D.C. U.S. Department of Commerce.

Spencer, B.D. 1982a. "Feasibility of Benefit-Cost Analysis of Data Programs." *Evaluation Review* 6: 649–672.

Spencer, B.D. 1982b. "Technical Issues in Allocation Formula Design." *Public Administration Review* 4: 524–529.

Spencer, B.D. 1985. "Statistical Aspects of Equitable Apportionment." *Journal of the American Statistical Association* 80: 815–822.

Stats SA. 2011. *Mid-Year Population Estimates, 2011*. Report P0302. Pretoria: Statistics South Africa.

Stats SA. 2012a. *Census 2011, Census in Brief*. Report 03-01-41. Pretoria: Statistics South Africa.

Stats SA. 2012b. CENSUS 2011: Post Enumeration Survey, Report 03-01-46. Pretoria: Statistics South Africa.

Stats SA. 2014a. Annual Report 2013/2014 (Book 1). Pretoria: Statistics South Africa. Available at: http://www.gov.za/sites/www.gov.za/files/STATSAnnual_Report_2013-2014.pdf (accessed 2 March 2015).

Stats SA. 2014b. "Population Household Surveys in the Case of South Africa." Presentation at the Workshop on Strengthening the Collection and Use of International Migration Data for Development, 18–21 November 2014, Addis Ababa, Ethiopia. Available at: http://www.un.org/en/development/desa/population/migration/events/other/workshop/docs/Session%20VI%20South%20Africa.pdf (accessed 3 March 2015).

Stats SA. 2016. Community Survey 2016: Technical Report 03-01-01, Statistics South Africa, Pretoria. Available at: http://cs2016.statssa.gov.za/wp-content/uploads/2016/06/CS-2016-Technical-report_Web.pdf (accessed 1 August 2016).

The Globe and Mail. 2011. "Traditional Census the Only Option for 2016, Statistics Canada Says." Available at: http://www.theglobeandmail.com/news/politics/traditional-census-the-only-option-for-2016-statistics-canada-says/article556243/ (accessed 4 March 2015).

The Guardian. 2015. "Census in Doubt as 10-year Data Collection Is Considered." Available at: http://www.theguardian.com/world/2015/feb/19/census-in-2016-in-doubt-as-10-year-data-collection-considered (accessed 4 March 2015).

The Journal. 2012. "2016 Census May Be Delayed in Government Spending Review." Available at: http://www.thejournal.ie/2016-census-may-be-delayed-in-government-spending-review-455663-May2012/ (accessed 4 March 2015).

Yemek, E. 2005. *Understanding Fiscal Decentralisation in South Africa*. IDASA Budget Information Service Occasional Paper. Cape Town: Institute for Democratic Alternatives in South Africa. Available at: http://www.gsdrc.org/docs/open/CC107.pdf (accessed 1 March 2015).

Zaslavsky, A.M. and A.L. Schirm. 2002. "Interactions between Survey Estimates and Federal Funding Formulas." *Journal of Official Statistics* 18: 371–391.

Received March 2015
Revised August 2016
Accepted September 2016