TECHNICAL COMMENT

POLICY FORUM

Technical Comment on "Policy impacts of statistical uncertainty and privacy"

Yifan Cui[1], Ruobin Gong[2]*, Jan Hannig[3], Kentaro Hoffman[4]

Steed *et al*. (*1*) illustrates the crucial impact that the quality of official statistical data products may exert on the accuracy, stability, and equity of policy decisions on which they are based. The authors remind us that data, however responsibly curated, can be fallible. With this comment, we underscore the importance of conducting principled quality assessment of official statistical data products. We observe that the quality assessment procedure employed by Steed *et al*. needs improvement, due to (i) the inadmissibility of the estimator used, and (ii) the inconsistent probability model it induces on the joint space of the estimator and the observed data. We discuss the design of alternative statistical methods to conduct principled quality assessments for official statistical data products, showcasing two simulation-based methods for admissible minimax shrinkage estimation via multilevel empirical Bayesian modeling. For policymakers and stakeholders to accurately gauge the context-specific usability of data, the assessment should take into account both uncertainty sources inherent to the data and the downstream use cases, such as policy decisions based on those data products.

We motivate the proposed assessment framework by considering Title I funding allocation by the U.S. Department of Education using the U.S. Census Bureau's Small Area Income and Poverty Estimates (SAIPE) dataset studied by Steed *et al.* (*1*). Let $\mu = (\mu_1, ..., \mu_k)$ be the true population counts for children under poverty in districts $i = 1, ..., k$, and n$x = (x_1, ..., x_k)$ be the official SAIPE poverty estimates. Denote by $y : \mathbb{N}^k \rightarrow (\mathbb{R}^+)^k$ the entitlement function, that is, $y(x) = (y_1(x), ..., y_k(x))$ are the districts' official entitlements (in USD) based on x, and $y(\mu)$ the *true* entitlements were the true poverty population μ known. Finally, let $\mathcal{L}(\cdot;\cdot)$ be a loss function that measures the misallocation of funding between $y(x)$ and $y(\mu)$. The assessment estimates the average loss between the ideal and the realized allocations:

$$\mathbb{E}(\mathcal{L}(y(\mu); y(x))|x) \tag{1}$$

with expectation taken over what we denote as $p_c(\mu|x)$, the available distributional information about the true poverty counts μ given the observed estimate x and any auxiliary parameter c. The parameter c may encode known information about the variability in the observed estimates, such as their sampling or model-based variance. When $p_c(\mu|x)$ is given, Eq. 1 can be approximated via simulation:

$$\frac{1}{T}\sum_{t=1}^{T}\mathcal{L}\left(y\left(\mu^{(t)}\right)y(x)\right) \tag{2}$$

where $\mu^{(t)} \sim p_c(\mu|x)$ i.i.d. for some T large. This assessment is uncertainty- and policy-aware by the specifications of $p_c(\mu|x)$ and $\mathcal{L}$, respectively.

Typically, the loss function $\mathcal{L}$ is chosen by the assessor depending on the policy context, whereas $p_c(\mu|x)$ relies on information available to the assessor. Following Steed *et al.* (*1*), the available information are (i) the coefficients of variation upper bounds $c = (c_1, ..., c_k)$ suggested by the Census Bureau; and (ii) that x is approximately normally distributed around μ. That is,

$$x|\mu \sim N(\mu, \text{diag}(v)) \tag{3}$$

where $v = (v_1, ..., v_k), v_i = (c_i x_i)^2$ are the sampling variances of x.

Steed *et al.* employ a simulation procedure [section 2 of the supplementary materials in (*1*)] that approximates Eq. 1 by using x as a plug-in estimate for μ, and producing replicates of x using Eq. 3 based on this plug-in estimate. Understood within our proposal, this procedure amounts to simulating $\mu^{(t)}$ replicates ($T = 1000$) under the following choice of $p_c(\mu|x)$:

$$\mu|x{\sim}N(x, \text{diag}(v)) \tag{4}$$

This is not ideal for two reasons. First, each $\mu^{(t)}$ generated through (Eq. 4) is *inadmissible* for the true poverty count μ, a classic observation from Charles Stein (*2*, *3*). Second, Eq. 3 and Eq. 4 together do not admit a consistent joint probability distribution for (μ, x) (*4*), exposing the procedure to potential paradoxical conclusions [e.g., (*5*)].

How should the assessor construct $p_c(\mu|x)$? There is unlikely a unique "best" approach for all contexts, but reasonable starting points exist. Here, we discuss a class of distributions derived via multi-level empirical Bayesian modeling that accord to admissible and minimax shrinkage estimates for μ. The class follows the general form

$$\mu_i|x \overset{ind}{\sim} N((1-B_i)x_i + B_i\beta, (1-B_i)v_i), \quad i = 1, ..., k \tag{5}$$

where β and all $B_i \in [0, 1]$ are functions of x and the auxiliary c. The method is called *shrinkage* because compared to Eq. 4, it adjusts each poverty estimate μ_i based on the observed x_i to account for a common baseline β with a $100B_i\%$ variance reduction. This restores consistency on the joint specification of (μ, x) whenever $B_i \neq 0$.

Two possible constructions of Eq. 5 are (i) The Hudson-Berger (HB) construction (*6*, *7*), for which $\beta = 0$ and

$$B_i^{HB} = \min\left(1, \frac{(k-2)/v_i}{\sum_{j=1}^{k}/(x_j/v_j)^2}\right) \tag{6}$$

and (ii) the Morris-Lysy (ML) construction (*8*), for which $\beta = \bar{x}$,

$$B_i^{ML} = \frac{v_i}{v_i + \bar{v}_h\left(1-\hat{B}\right)/\hat{B}} \tag{7}$$

where $\bar{v}_h = k/\left(\sum_{i=1}^{k} v_i^{-1}\right)$ is the harmonic mean of the v_i's, $\hat{B} = (k-3)/(k-4)\hat{\sigma}^2$, and $\hat{\sigma}^2 = (k-1)^{-1}\sum_{i=1}^{k}(x_i - \bar{x})^2/v_i$ is the mean square error in the observed poverty counts.

Both constructions cater to unequal sampling variances v_i. They differ in that Hudson-Berger exerts stronger shrinkage for larger v_i whereas Morris-Lysy for smaller v_i. Due to the heavy tail of the SAIPE poverty estimates x and increasing c_i for larger x_i, we apply the Morris-Lysy method on the observed poverty proportion x_i/n_i (rather than x_i), where n_i is the total population of district i in order to mitigate overly strong shrinkage effects.

We compare the proposed approaches with the evaluation of Steed *et al.* (*1*). The top panel of Fig. 1 compares the quantiles of the expected Hudson-Berger and Morris-Lysy poverty estimates with the SAIPE estimates. The bottom panel displays poverty estimate replicates generated through the constructions of Hudson-Berger, Morris-Lysy, and Eq. 2 for four districts with different population sizes (at 1, 5, 50, and 100% quantiles). For small counts, Hudson-Berger shrinks strongly resulting in nearly constant $\mu^{(t)}$ replicates, whereas its replicates are comparable to that of Steed *et al.* (*1*) for larger counts. On the other hand, the Morris-Lysy method exhibits a varying and moderate shrinkage effect at all count levels.

Table 1 displays the estimated lost entitlement based on the three approaches, with data error alone and with differential privacy

[1]Center for Data Science, Zhejiang University, China. [2]Dept of Statistics, Rutgers University, New Brunswick, NJ. [3]Dept of Statistics & Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC. [4]Johns Hopkins Institute for NanoBioTechnology, Baltimore, MD.
*Corresponding author. Email: ruobin.gong@rutgers.edu

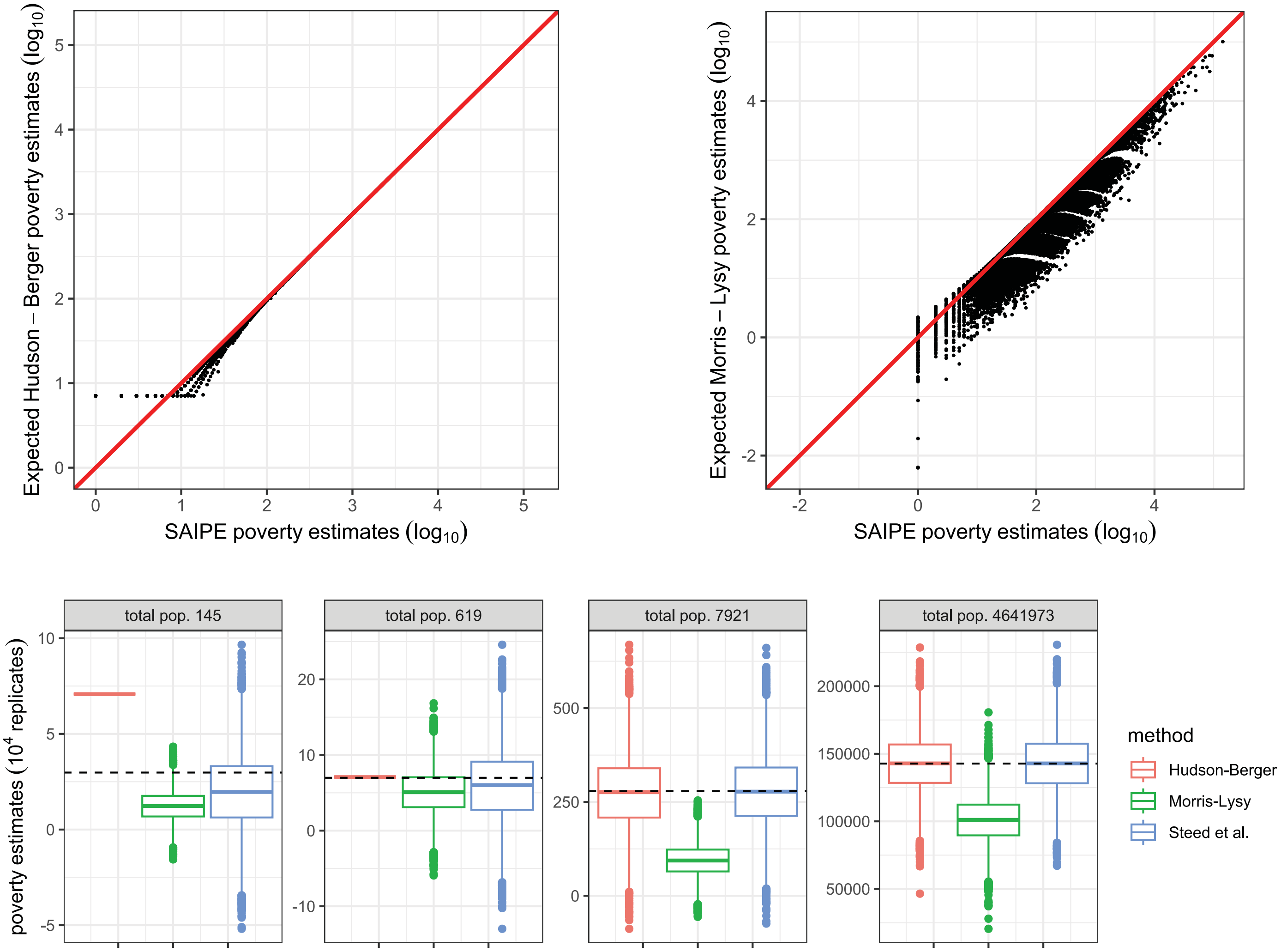

Fig. 1. (Top) quantile-quantile comparisons of expected Hudson-Berger (left) and Morris-Lysy (right) poverty estimates with SAIPE estimates ($\log_{10}$). (Bottom) Boxplots of 10^4 poverty estimate replicates from the Hudson-Berger, Morris-Lysy, and Steed *et al.* (1) constructions for four districts with total population sizes at 1, 5, 50, and 100% quantiles.

Table 1. Estimated lost entitlements (in USD, billions) due to data error (left) and due to data and privacy error (middle; $\epsilon = 0.1$) according to each assessment construction. Additional loss due to privacy (percent) is shown on the right.

	data error (s.e.)	data + privacy error (s.e.)	diff. (%)
Steed *et al.* (1)	1.058 (0.031)	1.109 (0.031)	4.756
Hudson-Berger	1.060 (0.032)	1.110 (0.033)	4.650
Morris-Lysy	2.385 (0.044)	2.429 (0.044)	1.840

protection ($\epsilon = 0.1$) applied to the observed SAIPE estimates first. These results are reproduced and/or implemented using the code provided by Steed *et al.* (*1*). The Hudson-Berger assessment agrees closely with the assessment by Steed *et al.* (*1*), putting the expected lost entitlement at $1.06 billion due to data error and an additional 4.65% due to privacy protection. The Morris-Lysy assessment estimates the lost entitlement at $2.385 billion due to data error, and an additional 1.84% due to privacy protection. The code we used to conduct these experiments relies in part on the public codebase that accompanies Steed *et al.* (*1*) and can be found at https://github.com/khoffm4/dp-policy-shrink.

The analysis by Steed *et al.* (*1*) is a timely companion to the rapid emergence of differential privacy as the new formal privacy standard for statistical disclosure limitation (SDL), anticipating possible adoption in complex survey programs at the Census Bureau (*9*, *10*) and at the IRS (*11*). The privacy revamp has been met with critical feedback from data users (*12*), who question the usability of differentially private data products after deliberate noise injection which instills distrust both in the data product and in the competence of the curator. The privacy innovation inadvertently ruptured, in the words of (*13*), a "statistical imaginary" that official statistics are somehow pristine. Steed *et al.* (*1*) point out that data users' distrust may be misplaced, as the impact of errors and uncertainty stemming from sampling, response, measurement, reporting, and editing may dominate that of errors from privacy. It exposes the need to examine, accurately and often, the extent to which every error source in an official statistical product affects policy decisions. The development of quality assessment tools that are theoretically sound, substantively relevant, and practically deployable calls for quantitative research.

REFERENCES AND NOTES

1. R. Steed, T. Liu, Z. S. Wu, A. Acquisti, *Science* **377**, 928–931 (2022).
2. C. M. Stein, "Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution" in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics* (Univ. California Press, 1956), pp. 197–206.
3. The inadmissibility of $\mu^{(t)}$ generated through Eq. 4 means its estimation risk for μ based on the ℓ_2 loss can be uniformly dominated by an alternative and admissible estimate.
4. The inconsistency can be seen via a simple argument by the *law of iterated variances*. For each district *i*, the variance of the poverty estimate x_i may be written as a sum of two components,$\mathbb{V}(x_i) = \mathbb{E}(\mathbb{V}(x_i|\mu_i)) + \mathbb{V}(\mathbb{E}(x_i|\mu_i))$, where the first component is equal to v_i and the second is equal to $\mathbb{V}(\mu_i)$. Applying the same argument to $\mathbb{V}(\mu_i)$ we see that it is the sum of v_i and $\mathbb{V}(x_i)$. Since $\mathbb{V}(\mu_i)$, $\mathbb{V}(x_i)$, and v_i are all non-negative, both results can hold only when all v_i's are uniformly zero, leading to a contradiction.
5. A. P. Dawid, M. Stone, J. V. Zidek, *J. R. Stat. Soc. B* **35**, 189–213 (1973).
6. H. Hudson, Empirical Bayes estimation (technical report no. 58, Stanford Univ, 1974).
7. J. O. Berger, *Ann. Stat.* **4**, 223–226 (1976). http://www.jstor.org/stable/43590231.
8. C. N. Morris, M. Lysy, *Stat. Sci.* **27**, 115 (2012).
9. M. H. Freiman, R. A. Rodríguez, J. P. Reiter, A. Lauger, Formal Privacy and Synthetic Data for the American Community Survey (U.S. Census Bureau, 2018); https://www.census.gov/library/working-papers/2018/adrm/formal-privacy-synthetic-data-acs.html
10. A Roadmap for Disclosure Avoidance in the Survey of Income and Program Participation (SIPP) (National Academies of Sciences, Engineering, and Medicine, 2022); Https://www.nationalacademies.org/our-work/a-roadmap-for-disclosure-avoidance-in-the-survey-of-income-and-program-participation-sipp.
11. A. F. Barrientos, A. R. Williams, J. Snoke, C. M. Bowen, Differentially Private Methods for Validation Servers, (Urban Institute research report, 2021); https://www.urban.org/research/publication/differentially-private-methods-validation-servers.
12. V. J. Hotz, J. Salvo, A Chronicle of the Application of Differential Privacy to the 2020 Census. *HDSR* **2**, ff891fe5 (2022).
13. D. Boyd, J. Sarathy, *HDSR* **2**, 66882f0e (2022).

ACKNOWLEDGMENTS

The authors thank the authors of Steed *et al*. (*1*) for valuable comments and discussions. Y. C. was supported in part by the National Natural Science Foundation of China. J. H. was supported in part by the National Science Foundation under grant DMS-1916115, 2113404, and 2210337.

Submitted 25 November 2022; accepted 5 April 2023
10.1126/science.adf9724

Science
AAAS

Technical Comment on “Policy impacts of statistical uncertainty and privacy”

Yifan Cui, Ruobin Gong, Jan Hannig, and Kentaro Hoffman

Science, **380** (6648), eadf9724.
DOI: 10.1126/science.adf9724

Science (ISSN) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science* is a registered trademark of AAAS.