Total Error Frameworks for Generic Datasets and Estimates

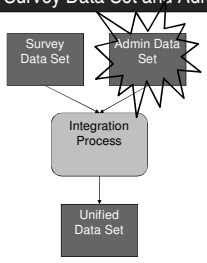
Paul P. Biemer
RTI International; University of North Carolina

AAPOR Webinar
October 18, 2018

Outline of this Webinar

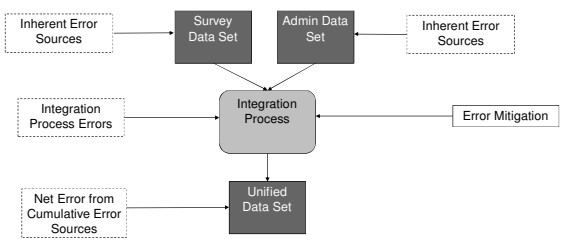
- Generic Data Sets
  - Single source data sets
  - Integrated data sets
- Estimates derived from generic data sets
  - Probability samples
  - Nonprobability samples

Integration of a Survey Data Set and Administrative Data Set



Definition: "data holdings containing information which is not primarily collected for statistical purposes." United Nations Economic Commission for Europe, 2011

Integration of a Survey Data Set and Administrative Data Set



A Simple Total Error Framework for Generic Data Sets

Typical File Structure

Record #	V_1	V_2	···	V_K
	← variables or features →			
Population units ↕				

Total error = column error
+ row error
+ cell error

Column and Cell Errors

Typical File Structure

Record #	V_1	V_2	···	V_K
	← variables or features →			
Population units ↕		X		
		?		

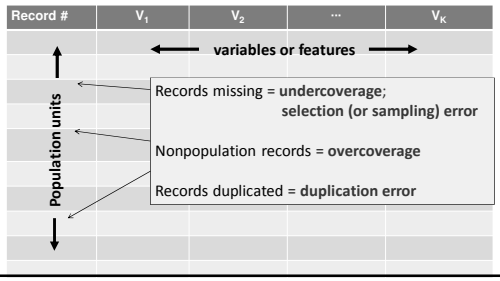
Misspecified variables = **specification error** (misaligned construct, outdated value, etc.)

Variable values in error = **content error** (measurement error, linking errors, data processing errors, etc.)

Variable values missing = **item missing data** (nonresponse, unavailable, etc.)

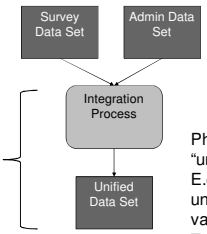
Row errors

Typical File Structure



Zhang's (2012) Two-phase Framework for a Generic Data Set

Phase I



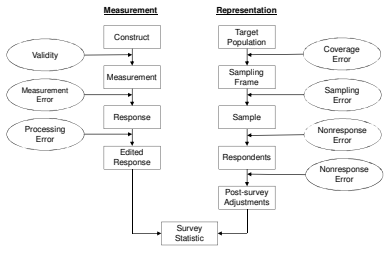
Phase II

Phase I creates a primary data set consisting of "objects" and "variables" E.g. "objects" may be Tweets and "variables" might be positive/negative attributes of these Tweets.

Phase II integrates two data sets of "units" and "harmonized variables" E.g. Tweets might be combined by unique individual and a composite variable created to summarize the Tweets' attributes.

Zhang's framework renames/redefines error sources in the TSE framework and identifies new root error sources in the integration process

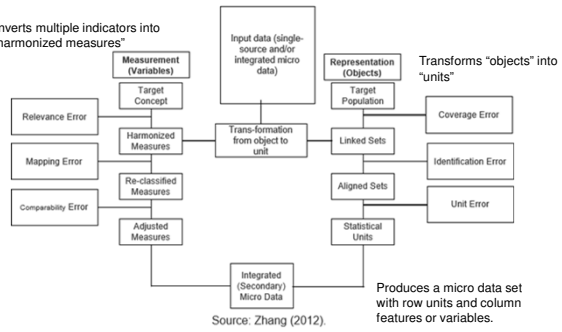
Groves et al. (2009) Error Framework for Survey Estimates



Source: Groves et al. (2009).

Zhang's Phase II Framework

Converts multiple indicators into a "harmonized measures"



Transforms "objects" into "units"

Produces a micro data set with row units and column features or variables.

Source: Zhang (2012).

Correspondence of Traditional Error Sources and Zhang's Error Sources

	Traditional Terminology	Zhang's Terminology
Column Error	Specification Error Invalidity	Validity Error (Phase I) Relevance Error (Phase II)
Cell Error	Content Error Measurement Error	Measurement Error (Phase I) Mapping Error (Phase II) Comparability Error (Phase II)
	Item Missing Data	Missing Redundancy (Phase I) Identification Error (Phase II) Unit Error (Phase II)
	Data Processing Error	Processing Error (Phase I)
Row Error	Sampling Error	Selection error (Phase I)
	Unit Nonresponse	Missing Redundancy (Phase I)
	Undercoverage Overcoverage Duplication	Frame Error (Phase I)

Error Risk Profile of the Error Sources

Types of Error Risks

- Intrinsic risk – risk that an error source poses if no steps are taken to reduce the error; error risk of "doing nothing."
  - Example: The intrinsic risk of nonresponse bias in a linear estimator is

$$B_I = \frac{\mathrm{cov}(y_i, \rho_i)}{\bar{\rho}}$$

- Residual risk – risk of error for a source that remains after mitigation strategies have been applied.
  - Example: After nonresponse weighting adjustments have been applied, the residual risk of bias is

$$B_R \leq B_I$$

Risk Profile Comparing Survey, Administrative and Unified Datasets: Either Intrinsic or Residual Risks

Error Sources	*Survey Dataset*	*Administrative Dataset*	*Unified Dataset*
Specification	Risk level (1, 2, 3)	Risk level (1, 2, 3)	Risk level (1, 2, 3)
Coverage: Undercoverage	Risk level (1, 2, 3)	Risk level (1, 2, 3)	Risk level (1, 2, 3)
Coverage: Overcoverage	Risk level (1, 2, 3)	Risk level (1, 2, 3)	Risk level (1, 2, 3)
Coverage: Duplication	Risk level (1, 2, 3)	Risk level (1, 2, 3)	Risk level (1, 2, 3)
Selection	Risk level (1, 2, 3)	Risk level (1, 2, 3)	Risk level (1, 2, 3)
Content	Risk level (1, 2, 3)	Risk level (1, 2, 3)	Risk level (1, 2, 3)
Missing Data	Risk level (1, 2, 3)	Risk level (1, 2, 3)	Risk level (1, 2, 3)

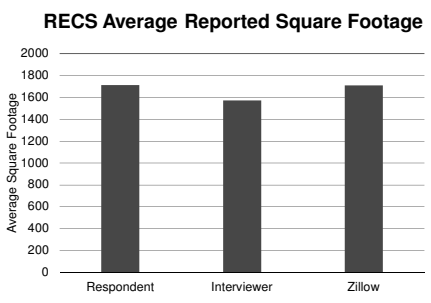
Case Study: Error Mitigation for Energy Use Survey Square Footage Data using Unified Data

- Data sources
  - Survey data: 2015 Residential Energy Consumption Survey (RECS)
    - n ≈ 2,400 completed cases
  - Big Data (data pulled from various sources)
    - Zillow
    - Acxiom
    - CoreLogic
- Variable of interest: housing unit square footage
- **Goal: Integrate the external data sources with the survey data to improve and/or evaluate the accuracy of survey square footage data**

Source: Amaya, A. (2017)

14

Evidence of Nonsampling Error from the RECS



More Evidence of Intrinsic Error Risks

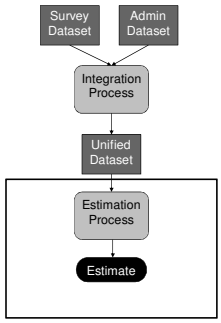
MSE Component	RECS	Zillow
Measurement Bias	-0.082	-0.14
Pop'n CV	0.64	0.64
Reliability	0.59	0.66
N	118,208,250	118,208,250
n	6,000	96,930,765
Response rate	55.4%	
Coverage rate	≈ 99%	82%
Selection rate	0.009%	

Intrinsic Error Risk Profile for the RECS, Zillow and Unified Datasets

Error Sources	*RECS*	*Zillow*	*RECS U Zillow*
Specification	2	2	2
Coverage: Undercoverage	1	2	1
Coverage: Overcoverage	2	1	1
Coverage: Duplication	2	1	1
Selection	3	1	3
Content	2	3	3
Missing Data	2	2	1
Average	**1.7**	**1.7**	**1.7**

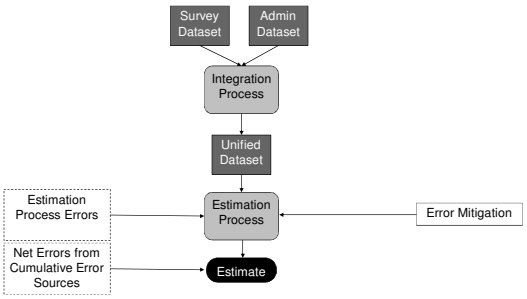
Unified data appears to offer no advantage to Zillow only dataset.

The Hybrid Estimation Process

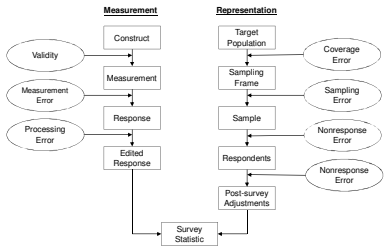


Reid et al. (2017) added a third phase for assessing the quality of the statistical estimates derived from the unified data set.

The Hybrid Estimation Process

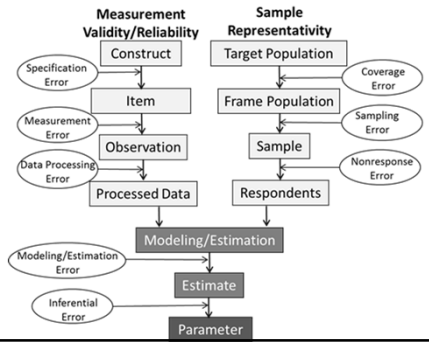


Groves et al. (2009) Total Survey Error Framework for Estimates

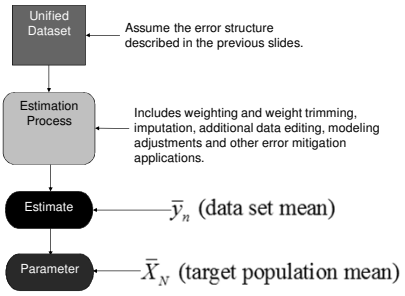


Source: Groves et al. (2009).

Total Survey Error Decomposition



Errors Associated with the Hybrid Estimation Process



Total Error Identity for the Mean of the a Generic Data Set

Total Error = Errors of Observation + Errors of Nonobservation

$$\bar{y}_n - \bar{X}_N = (\bar{y}_n - \bar{x}_n) + (\bar{x}_n - \bar{X}_N)$$

$$\bar{y}_n = \sum_{i=1}^{n} y_i / n \qquad \bar{x}_n = \sum_{i=1}^{n} x_i / n \qquad \bar{X}_N = \sum_{j=1}^{N} X_i / N$$

$X_i =$ true construct for pop'n unit j

$x_i =$ true construct for sample unit i

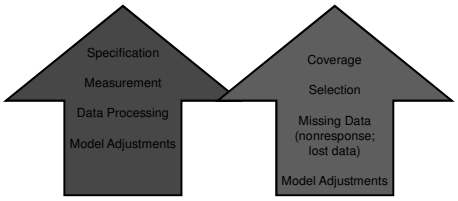
$y_i = x_i +$ error (the observation on sample unit i)

23

Total Error Identity for the Mean of the a Generic Data Set

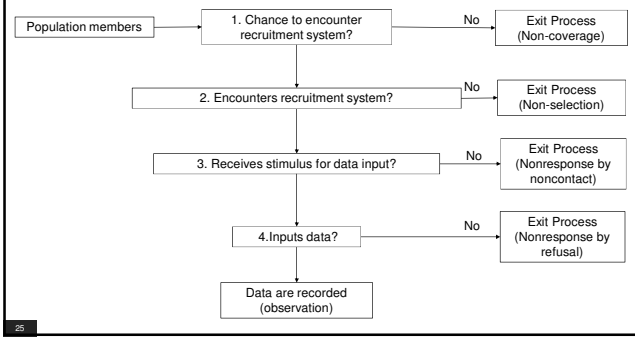
Total Error = Errors of Observation + Errors of Nonobservation

$$\bar{y}_n - \bar{X}_N = (\bar{y}_n - \bar{x}_n) + (\bar{x}_n - \bar{X}_N)$$



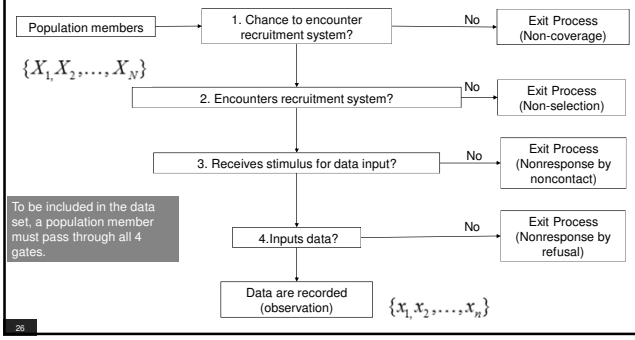
24

Framework for Total Error of Probability and Nonprobability Samples



25

Framework for Total Error of Probability and Nonprobability Samples



26

Total Error Assuming Sample Recruitment Error is Sole Error Source

- X_i denotes the characteristic measured for the ith person in the Recruitment Process
- $R_i = 1$ if population unit i is in the sample; $R_i = 0$ otherwise
  - i.e., $R_i = 1$ if unit i passes through all 4 sample recruitment gates
- $\rho_{RX} = \mathrm{Corr}(R_i, X_i \mid R)$ is a measure of sample recruitment bias

$$\mathrm{E}_R(\bar{x}_n - \bar{X})^2 = \sigma_X^2 \frac{N-n}{n} E_R(\rho_{RX}^2)$$

Population variance

Expectation with respect to the selection mechanism

Meng, 2017; Bethlehem, 1988

27

Total Error Assuming Sample Recruitment Error is Sole Error Source

- X_i denotes the characteristic measured for the ith person in the Recruitment Process
- $R_i = 1$ if population unit i is in the sample; $R_i = 0$ otherwise
  - i.e., $R_i = 1$ if unit i passes through all 4 sample recruitment gates
- $\rho_{RX} = \mathrm{Corr}(R_i, X_i \mid R)$ is a measure of sample recruitment bias

$$\mathrm{E}_R(\bar{x}_n - \bar{X})^2 = \sigma_X^2 \frac{N-n}{n} E_R(\rho_{RX}^2)$$

Example: For SRS sampling and no nonresponse, $\mathrm{E}_R(\rho_{RX}^2) = \frac{1}{N-1}$

$$\frac{N-n}{n}\sigma_X^2 \mathrm{E}_R(\rho_{RX}^2) = \left(1 - \frac{n}{N}\right)\frac{S_X^2}{n}$$

28

Interpretation of ρ_{RX}

- Not much is known about ρ_{RX} for nonprobability samples.
- However, ρ_{RX} has been studied extensively for surveys (through the estimation of nonresponse bias).
- ρ_{RX} will be smaller for nonprobability samples when gates 1, 2 and 3 are entered for all members of the population.
- Ability to adjust for sample recruitment bias is better for surveys because
  - We have more control over who enters gates 1-3 and thus more control over ρ_{RX}
  - We often know a lot about sample recruitment failures and how to adjust for them through weighting and imputation.

29

Useful Conversion Formula

It can be shown that, for unweighted data,

$$\rho_{RX} = \frac{\mathrm{RB}}{\mathrm{CV}_X\sqrt{(1-f)/f}}$$

Example: Let RB (Relative Bias) = 0.05

CV_X (Population CV) = 0.6

f (sampling fraction) = 0.8

Then

$$\rho_{RX} = \frac{\mathrm{RB}}{\mathrm{CV}_X\sqrt{(1-f)/f}} = \frac{0.05}{0.6\sqrt{(1-.8)/.8}} = 0.2$$

Suppose $f = 0.005$, then

$$\rho_{RX} = \frac{0.05}{0.6\sqrt{(1-.005)/.005}} = 0.007$$

Example 1. Revisiting the RECS Illustration

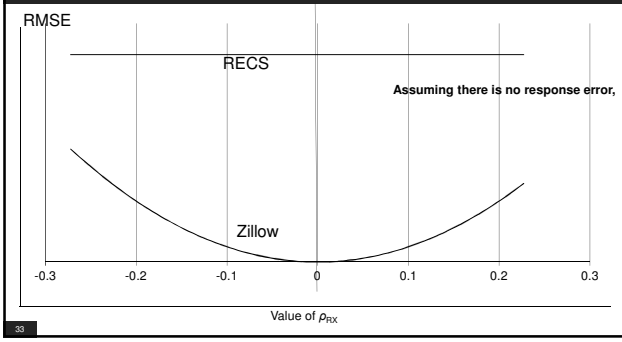
MSE Component	RECS	Zillow
Relative Bias	-0.082	-0.14
Pop'n CV	0.64	0.64
Reliability	0.59	0.66
ρ_{RX}	-0.000295	[-0.27,0.22]
N	118,208,250	118,208,250
n	6,000	96,930,765
Response rate	55.4%	
Coverage rate	≈ 99%	82%
Selection rate	0.009%	

Feasible range

Example 1. Revisiting the RECS Illustration

MSE Component	RECS	Zillow
Relative Bias	-0.082	-0.14 ← IGNORE
Pop'n CV	0.64	0.64
Reliability	0.59	0.66 ← IGNORE
ρ_{RX}	-0.000295	**[-0.27,0.22]**
N	118,208,250	118,208,250
n	6,000	96,930,765
Response rate	55.4%	
Coverage rate	≈ 99%	82%
Selection rate	0.009%	

RMSEs as a Function of ρ_{RX} for Zillow and RECS



33

Example 2. Bias in the 2016 U.S. Presidential Election (from Meng, 2017)

Examines the "missing data" bias in the 2016 Presidential Election

- Total sample size of combined state- and national-level polls

$$n = 2,315,570$$
$$f = 0.01$$

- Bias due to nonresponse can be computed from election results
- Simplification of ρ_{RX} for binary data (X=1→ Trump, X=0→Clinton)

$$\rho_{RX} = [\Pr(\text{Response}|X=1) - \Pr(\text{Response}|X=0)]\sqrt{\frac{p_X(1-p_X)}{f(1-f)}}$$
$$= -0.00502$$

Example 2. Bias in the 2016 U.S. Presidential Election (from Meng, 2017)

- While the actual sample size is 2,315,570, in terms of MSE, it is equivalent to an SRS sample of n = **400**!
- The error in the estimate might be reduced through weighting methods (see, for example, Haziza and Lesage, 2016).

Key Takeaways

- Total error models and frameworks have been developed for generic data sets and the estimates derived from them.
- These frameworks can help understand the root causes of error and how to address them to increase accuracy.
- When it comes to Big Data, i.e., massive data sets with unknown ρ_{RX} , much more work has been done on the "representation" than on the "measurement" side.
- However, more work is need on both gauge the magnitude of ρ_{RX} as well as how to reduce its effects on the final estimates.

References

- Amaya, A. (2017). "Appendix A. Research on Alternative Methods for Reporting Square Footage Estimates," downloaded on October 15, 2018 from https://www.eia.gov/consumption/residential/reports/2015/squarefootage/
- Biemer, P. and Amaya, A. (2018). "Some Tools for Assessing and Improving the Accuracy of Hybrid Estimators," paper presented at ITSEW 2018, Durham, NC
- Biemer, P., D. Trewin, H. Bergdahl, and L. Japec (2014). "A System for Managing the Quality of Official Statistics," *Journal of Official Statistics*, Vol. 30, No. 3, 2014, pp. 381–415.
- Groves, Robert M., Floyd J. Fowler, Jr., Mick Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourrangeau. *Survey Methodology. Revised edition*. New York: Wiley, 2009.
- Haziza, D. and E. Lasage (2016). "A Discussion of Weighting Procedures for Unit Nonresponse," *Journal of Official Statistics*, Vol. 32, No. 1, pp. 129-145.
- Meng, X. (2018). "Statistical Paradises and Paradoxes in Big Data (I): Law of Large Populations, Big Data Paradox, and the 2016 U.S. Presidential Election," *Annals of Applied Statistics*, Vol. 12, No. 2, 685–726.
- Reid, Giles, Felipa Zabala, and Anders Holmberg. "Extending TSE to Administrative Data: A Quality Framework and Case Studies from Stats NZ." *Journal of Official Statistics*, vol. 33, no. 2 (2017), pp. 477-511.
- Zhang, Li-Chun. "A Unit-Error Theory for Register-Based Household Statistics." *Journal of Official Statistics*, vol. 27, no. 3 (September 2011), pp. 415-432.