THE EFFECT OF MEASUREMENT ERROR ON BINARY RANDOMIZED RESPONSE TECHNIQUE (RRT) MODELS

B. Meche[1] V. Poruri[2] Sat Gupta[3] Sadia Khalil[3,4]

[1]University of Louisiana at Lafayette & University of Chicago

[2]University of California, Berkeley

[3]University of North Carolina Greensboro

[4]Lahore College for Women University

This research was funded by the National Science Foundation Grant No. DMS-2244160.

THE UNIVERSITY of NORTH CAROLINA GREENSBORO
Department of Mathematics & Statistics

- Introduction
- Measurement error in RRT models
- Simulation results
- Discussion of Measurement Error
- Proposed Measure of Privacy
- Future Work

THE UNIVERSITY of NORTH CAROLINA GREENSBORO
Department of Mathematics & Statistics

Introduction

Measurement error:

- Difference between the true value and the recorded value
- Quantitative models: recorded response Z, true response Y, and measurement error U

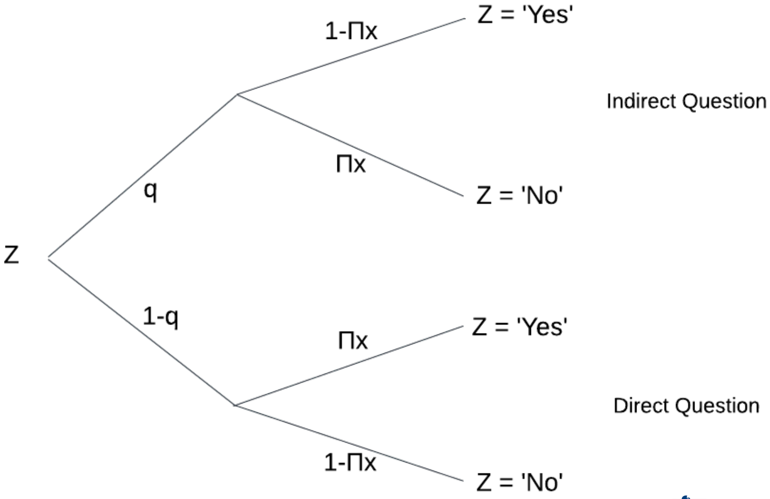
$$Z = Y + U$$

- Binary models: recorded response Z, true response Y, and measurement error C

$$Z = Y + C \mod 2$$

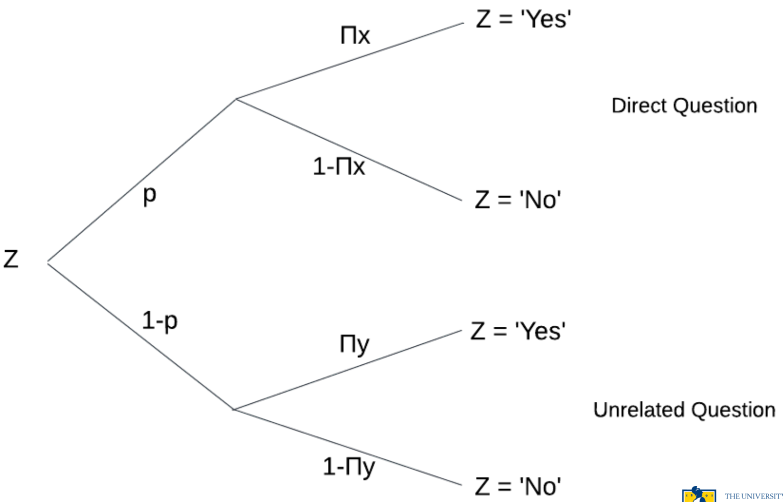
Goal: Observe the effect that measurement error has on binary RRT models

Warner model

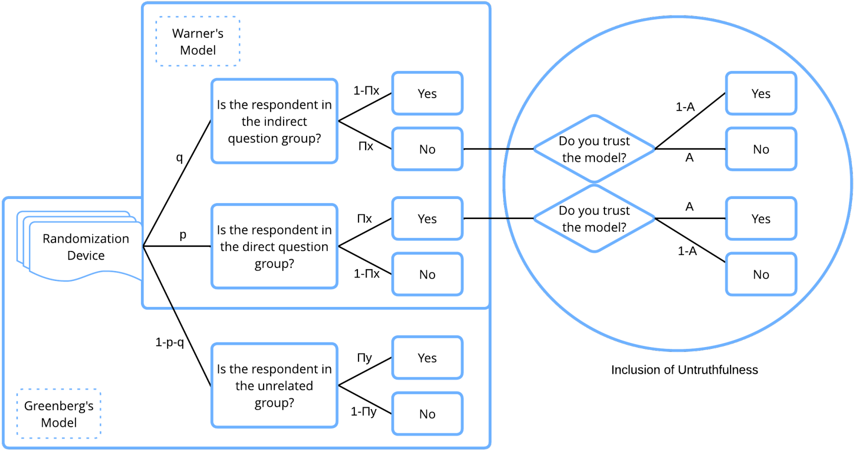


THE UNIVERSITY of NORTH CAROLINA GREENSBORO
Department of Mathematics & Statistics

Greenberg model



Lovig Model (2021)



Research question: Which model is best?

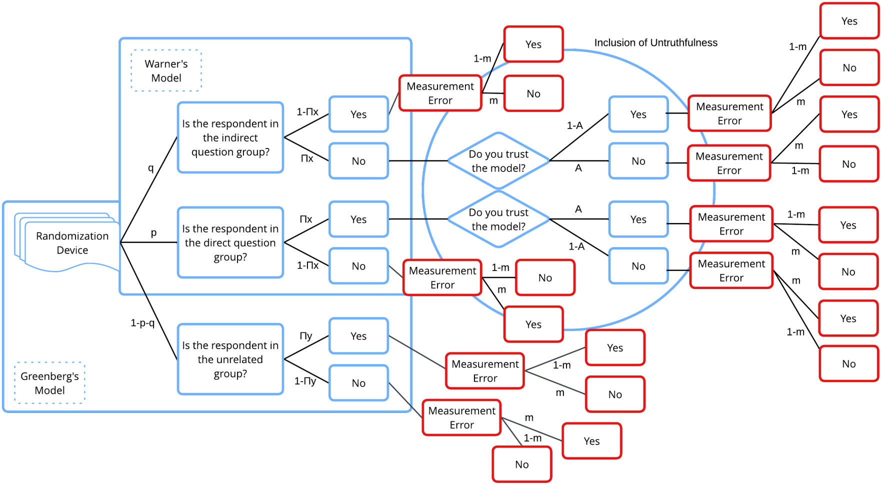
- A unified measure of efficiency and privacy proposed by Lovig et al. (2021)

$$\mathbb{M} = \frac{PP^a}{MSE^b}$$

  - PP: privacy protection $= \dfrac{1 - \text{privacy loss}}{1 - \pi_x}$
  - a, b are weights for the importance of efficiency and privacy respectively
- Lovig model is the best model by $\mathbb{M}$
- However, we suggest that accounting for measurement error will affect a researcher's model choice
- We will apply measurement error to the three models and compare the effect that measurement error has on each

Measurement error in RRT models

Meche et al. (2024) Model: Accounting for Measurement Error



B. Meche, V. Poruri, S. Gupta, and S. Khalil

The University of North Carolina Greensboro
Department of Mathematics & Statistics

Lovig model for measurement error:

Question 3: (With Mixture Model) *"Do you have the sensitive trait?"*

$$P_y = (A\pi_x (p-q) + \pi_y (1-p-q))(1-2m) + m(1-2q) + q$$

$$\hat{\pi}_x = \frac{\widehat{P_y} - \hat{m}(1-2q) - q - \pi_y (1-p-q)(1-2\hat{m})}{\hat{A}(1-2\hat{m})(p-q)}$$

$$\mathbb{E}[\widehat{P_y}] = P_y, \ \mathrm{Var}[\widehat{P_y}] = \frac{P_y(1-P_y)}{n}$$

THE UNIVERSITY *of* NORTH CAROLINA GREENSBORO
Department *of* Mathematics & Statistics

Lovig model for measurement error: estimating measurement error m

Question 2: (Rigged question where $\pi_x = 0$) *"Are you a robot?"*

$$P_{y_0} = m\left(1 - 2q - 2\pi_y\left(1 - p - q\right)\right) + q + \left(1 - p - q\right)\pi_y$$

$$\widehat{m} = \frac{\widehat{P_{y0}} - q - \pi_y(1 - p - q)}{1 - 2q - 2\pi_y(1 - p - q)}$$

$$\begin{aligned}\mathrm{Var}[\widehat{m}] &= \frac{\mathrm{Var}[\widehat{P_{y0}}]}{(1 - 2q - 2\pi_y(1 - p - q))^2} \\ &= \frac{P_{y_0}\left(1 - P_{y_0}\right)}{n\left(1 - 2q - 2\pi_y(1 - p - q)\right)^2}\end{aligned}$$

THE UNIVERSITY of NORTH CAROLINA GREENSBORO
Department of Mathematics & Statistics

Lovig model for measurement error:

Question 1: (With Greenberg Model) *"Do you trust the model?"*

$$P_{yg} = p_g A + (1 - p_g)\pi_{yg}$$

$$\hat{A} = \frac{\widehat{P_{yg}} - (1 - p_g)\pi_{yg}}{p_g}$$

$$\mathbb{E}[\widehat{P_{yg}}] = P_{yg}\ ;\ \mathrm{Var}[\widehat{P_{yg}}] = \frac{P_{yg}(1 - P_{yg})}{n}$$

$$\mathbb{E}[\hat{A}] = A\ ;\ \mathrm{Var}[\hat{A}] = \frac{P_{yg}(1 - P_{yg})}{n p_g^2}$$

THE UNIVERSITY of NORTH CAROLINA GREENSBORO
Department of Mathematics & Statistics

Lovig model for measurement error: estimating prevalence of sensitive trait $\hat{\pi}_x$ using a first order Taylor's approximation

$$f(x,y,z) \approx f(x_0,y_0,z_0)+(x-x_0)f_x(x_0,y_0,z_0)+(y-y_0)f_y(x_0,y_0,z_0)$$
$$+(z-z_0)f_z(x_0,y_0,z_0),$$

$$\hat{\pi}_x = \frac{\widehat{P_y} - \hat{m}(1-2q) - q - \pi_y(1-p-q)(1-2\hat{m})}{\hat{A}(1-2\hat{m})(p-q)}$$

$$\begin{aligned}\hat{\pi}_x \approx\ & \frac{P_y - m(1-2q) - q - \pi_y(1-p-q)(1-2m)}{A(1-2m)(p-q)} \\ & + \frac{\widehat{P_y} - P_y}{A(1-2m)(p-q)} \\ & - (\hat{A} - A)\left[\frac{P_y - m(1-2q) - q - \pi_y(1-p-q)(1-2m)}{A^2(1-2m)(p-q)}\right] \\ & + (\hat{m} - m)\left[\frac{2P_y - 1}{A(p-q)(1-2m)^2}\right]\end{aligned}$$

THE UNIVERSITY of NORTH CAROLINA GREENSBORO
Department of Mathematics & Statistics

Lovig model for measurement error: MSE of estimated prevalence of sensitive trait $\hat{\pi}_x$ using a Taylor's approximation

$$\mathbb{E}[\hat{\pi}_x] = \pi_x$$

$$\begin{aligned}\mathrm{MSE}[\hat{\pi}_x] = {} & \frac{\mathrm{Var}[\widehat{P_y}]}{[A\,(1-2m)\,(p-q)]^2} \\ & + \mathrm{Var}[\widehat{m}]\left[\frac{2P_y - 1}{A(p-q)(1-2m)^2}\right]^2 \\ & + \mathrm{Var}[\widehat{A}]\left[\frac{P_y - m\,(1-2q) - q - \pi_y\,(1-p-q)\,(1-2m)}{A^2\,(1-2m)\,(p-q)}\right]^2\end{aligned}$$

THE UNIVERSITY *of* NORTH CAROLINA GREENSBORO
Department *of* Mathematics & Statistics

Approach: We may treat the Warner and Greenberg models as special cases of the Lovig model

- Warner model: $q = 1 - p$
- Greenberg model: $q = 0$

Now, we may use the results from the Lovig model to solve for measurement error m, trust A, and sensitive trait π_X statistics.

The University *of* North Carolina Greensboro
Department *of* Mathematics & Statistics

Simulation results

Table 1. Estimates of $\hat{\pi}_{xu}$ (*not* accounting for measurement error) and $\hat{\pi}_x$ (accounting for measurement error). Results aggregated over 10000 trials with $n = 500$, $p = 0.7$, $q = 0.15$, $\pi_x = 0.4$, and $\pi_y = \frac{1}{12}$.

q	A	m	$\hat{\pi}_{xu}$	$\hat{\pi}_x$
0.15	1	0	0.3998	0.4003
0.15	1	0.01	0.4041	0.3996
0.15	1	0.05	0.4208	0.4003
0.15	1	0.1	0.4432	0.4014
0.15	0.95	0	0.3991	0.3998
0.15	0.95	0.01	0.4048	0.3995
0.15	0.95	0.05	0.4240	0.3995
0.15	0.95	0.1	0.4489	0.3997
0.15	0.9	0	0.4002	0.4006
0.15	0.9	0.01	0.4061	0.3993
0.15	0.9	0.05	0.4284	0.3996
0.15	0.9	0.1	0.4563	0.4004

A	$\hat{A}$	m	$\hat{m}$	MSE[$\hat{m}$]	$\hat{\pi}_x$	MSE[$\hat{\pi}_x$]	PP	$\mathbb{M}$
1	0.9997	0.01	0.0099	0.0001 **0.0001**	0.4001	0.0009 **0.0010**	0.1119 **0.1117**	123.2866 **107.6471**
1	1.0002	0.05	0.0499	0.0001 **0.0001**	0.3997	0.0011 **0.0012**	0.2234 **0.2234**	199.7414 **180.9270**
1	0.9995	0.1	0.1003	0.0002 **0.0002**	0.3994	0.0014 **0.0016**	0.3494 **0.3488**	245.2008 **222.2470**
0.95	0.9503	0.01	0.0101	0.0001 **0.0001**	0.3997	0.0010 **0.0011**	0.1168 **0.1169**	120.1149 **103.0169**
0.95	0.9497	0.05	0.0501	0.0002 **0.0001**	0.4000	0.0012 **0.0014**	0.2320 **0.2324**	194.7648 **171.7727**
0.95	0.9499	0.1	0.0999	0.0002 **0.0002**	0.4001	0.0016 **0.0017**	0.3604 **0.3606**	224.7503 **209.1361**
0.9	0.9002	0.01	0.0101	0.0001 **0.0001**	0.3996	0.0011 **0.0012**	0.1227 **0.1226**	113.0005 **98.5821**
0.9	0.9001	0.05	0.0502	0.0001 **0.0001**	0.4005	0.0013 **0.0015**	0.2436 **0.2422**	189.5389 **162.8959**
0.9	0.9002	0.1	0.0999	0.0002 **0.0002**	0.3992	0.0017 **0.0019**	0.3729 **0.3731**	220.4197 **196.3539**

Table 2. Theoretical (**bold**) and empirical values based on 10000 simulations with $n = 500$, $p = 0.7$, $q = 0.0$, $\pi_x = 0.4$, and $\pi_y = \frac{1}{12}$.

THE UNIVERSITY of NORTH CAROLINA GREENSBORO
Department of Mathematics & Statistics

A	$\hat{A}$	m	$\hat{m}$	MSE[$\hat{m}$]	$\hat{\pi}_x$	MSE[$\hat{\pi}_x$]	PP	$\mathbb{M}$
1	1.0000	0.01	0.0100	0.0006 **0.0006**	0.3996	0.0016 **0.0018**	0.4395 **0.4398**	271.3308 **249.8652**
1	0.9993	0.05	0.0501	0.0007 **0.0007**	0.4003	0.0019 **0.0021**	0.4975 **0.4978**	257.3392 **239.3552**
1	0.9998	0.1	0.0996	0.0008 **0.0008**	0.4014	0.0025 **0.0026**	0.5670 **0.5665**	228.5641 **216.0764**
0.95	0.9500	0.01	0.0103	0.0006 **0.0006**	0.3995	0.0018 **0.0019**	0.4516 **0.4525**	253.8097 **233.1124**
0.95	0.9501	0.05	0.0499	0.0007 **0.0007**	0.3995	0.0021 **0.0023**	0.5111 **0.5106**	242.4145 **222.4784**
0.95	0.9499	0.1	0.1001	0.0008 **0.0008**	0.3997	0.0028 **0.0029**	0.5792 **0.5791**	208.7430 **199.9674**
0.9	0.8995	0.01	0.0102	0.0006 **0.0006**	0.3993	0.0020 **0.0021**	0.4645 **0.4659**	235.4783 **216.8621**
0.9	0.9002	0.05	0.0502	0.0007 **0.0007**	0.3996	0.0023 **0.0025**	0.5225 **0.5241**	223.8847 **206.1068**
0.9	0.8999	0.1	0.0996	0.0008 **0.0008**	0.4004	0.0030 **0.0032**	0.5923 **0.5922**	195.9438 **184.3541**

Table 3. Theoretical (**bold**) and empirical values based on 10000 simulations with $n = 500$, $p = 0.7$, $q = 0.15$, $\pi_x = 0.4$, and $\pi_y = \frac{1}{12}$.

THE UNIVERSITY of NORTH CAROLINA GREENSBORO
Department of Mathematics & Statistics

Simulation results: Warner model

A	$\hat{A}$	m	$\hat{m}$	MSE$[\hat{m}]$	$\hat{\pi}_x$	MSE$[\hat{\pi}_x]$	PP	$\mathbb{M}$
1	1.0000	0.01	0.0104	0.0027 **0.0026**	0.3993	0.0032 **0.0034**	0.6596 **0.6597**	204.8066 **195.7034**
1	1.0003	0.05	0.0520	0.0028 **0.0027**	0.4001	0.0038 **0.0040**	0.6898 **0.6897**	179.6031 **173.4890**
1	1.0003	0.1	0.0995	0.0027 **0.0028**	0.3998	0.0048 **0.0050**	0.7270 **0.7265**	150.4105 **145.2988**
0.95	0.9500	0.01	0.0107	0.0026 **0.0026**	0.4012	0.0036 **0.0037**	0.6727 **0.6711**	186.7708 **179.7029**
0.95	0.9498	0.05	0.0510	0.0027 **0.0027**	0.4006	0.0042 **0.0044**	0.7016 **0.7005**	167.1141 **159.0592**
0.95	0.9503	0.1	0.1001	0.0028 **0.0028**	0.3996	0.0054 **0.0055**	0.7351 **0.7366**	135.4669 **132.9618**
0.9	0.9002	0.01	0.0099	0.0026 **0.0026**	0.3987	0.0040 **0.0042**	0.6817 **0.6830**	171.1752 **164.3027**
0.9	0.8998	0.05	0.0507	0.0027 **0.0027**	0.4006	0.0046 **0.0049**	0.7139 **0.7117**	153.5604 **145.1756**
0.9	0.8996	0.1	0.1000	0.0028 **0.0028**	0.3992	0.0060 **0.0062**	0.7456 **0.7469**	125.1110 **121.0998**

Table 4. Theoretical (**bold**) and empirical values based on 10000 simulations with $n = 500$, $p = 0.7$, $q = 0.3$, $\pi_x = 0.4$, and $\pi_y = \frac{1}{12}$.

Discussion of Simulation Results:

- Estimates provided by $\hat{\pi}_{xu}$ are severely biased, but $\hat{\pi}_x$ provides accurate estimates of π_x
- Estimators $\hat{A}$ and $\hat{m}$ are good estimators of m and A
- Error estimators $\widehat{\text{MSE}[\hat{m}]}$ and $\widehat{\text{MSE}[\hat{\pi}_x]}$ are good estimators of $\text{MSE}[\hat{m}]$ and $\text{MSE}[\hat{\pi}_x]$, respectively
  - The Warner model has the most measurement error, as estimated by the metric $\text{MSE}[\hat{m}]$
  - The Greenberg model has the least $\text{MSE}[\hat{\pi}_x]$
- The Warner model offers the highest privacy protection PP and the Greenberg model offers the least
- The Lovig model offers the highest unified measure $\mathbb{M}$
  - (Lovig, et al. 2021) arrived at the same conclusions for efficiency, privacy, and unified measure when measurement error was not accounted for
  - However, the Lovig mixture model has more measurement error than the Greenberg model

Discussion of Measurement Error

Recall the formula for and a few values of measurement error:

Table 5. Theoretical and estimated values of MSE of measurement error estimators. Estimates aggregated over 10000 trials, with $n = 500$, $\pi_x = 0.4$, $p = 0.7$, and $\pi_y = \frac{1}{12}$.

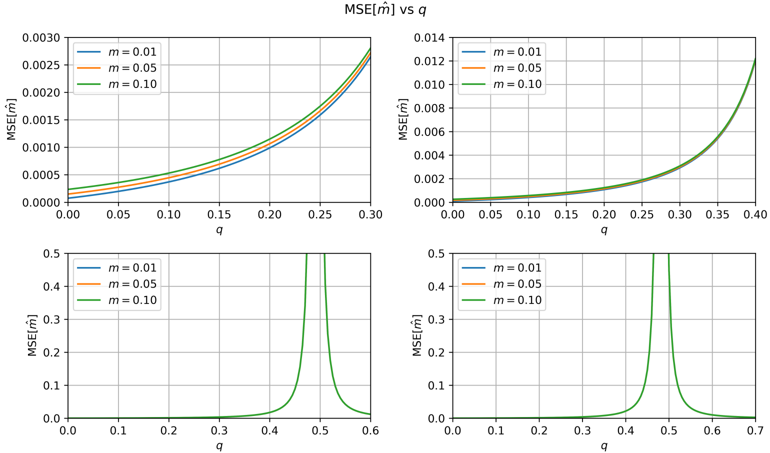
Model	p	q	m	$\hat{m}$	MSE[$\hat{m}$]	$\widehat{\text{MSE}[\hat{m}]}$
Greenberg	0.7	0	0.01	0.0099	0.0001	0.0001
	0.7	0	0.05	0.0499	0.0001	0.0001
	0.7	0	0.1	0.1003	0.0002	0.0002
Lovig	0.7	0.15	0.01	0.0100	0.0006	0.0006
	0.7	0.15	0.05	0.0501	0.0007	0.0007
	0.7	0.15	0.1	0.0996	0.0008	0.0008
Warner	0.7	0.3	0.01	0.0104	0.0026	0.0027
	0.7	0.3	0.05	0.0520	0.0027	0.0028
	0.7	0.3	0.1	0.0995	0.0028	0.0027

$$\text{MSE}[\hat{m}] = \frac{P_{y_0}(1 - P_{y_0})}{n(1 - 2q - 2\pi_y(1 - p - q))^2}$$

The effect that measurement error has on the model not only depends on the choice of q but also p.

THE UNIVERSITY of NORTH CAROLINA GREENSBORO
Department of Mathematics & Statistics

Figure 1. Impact of varying q on MSE$[\hat{m}]$ with $n = 500$ and $\pi_y = \frac{1}{12}$. Values of $q = 0, 0.3$ represent the Greenberg and Warner models, respectively. Values of $q \in (0, 0.3)$ represent the Lovig model.



Top Left: $p = 0.7$, **Top Right:** $p = 0.6$

Bottom Left: $p = 0.4$, **Bottom Right:** $p = 0.3$.

THE UNIVERSITY of NORTH CAROLINA GREENSBORO Department of Mathematics & Statistics

Proposed Measure of Privacy

Proposed Measure of Privacy: **Odds Ratio**

- We propose using the Odds Ratio as a measure of the predictability of participants' true responses from model responses.
- The O.R. quantifies the change in odds of the sensitive trait as the reported response changes from a 0 ("No") to a 1 ("Yes").
- To compute the Odds Ratio, we first compute a binary logistic regression model.
  - The odds ratio is calculated using the coefficients estimated from the model:

$$\log(\text{odds of } 1) = \beta_0 + \beta_1 X$$

  - Odds Ratio $(X) = e^{\beta_1}$
- Recall the traditional method of privacy *PP* states that the Greenberg model offers the least privacy and Warner offers the most.

Table 6. Odds ratio values from estimated true and model responses. Estimates aggregated over 10000 trials, with $n = 500$, $\pi_x = 0.4$, $p = 0.7$, and $\pi_y = \frac{1}{12}$

q	$p+q$	A	m	OR
Greenberg Model				
0	0.7	1.00	0.00	124.93
0	0.7	1.00	0.05	124.62
0	0.7	1.00	0.10	124.62
0	0.7	0.95	0.00	105.53
0	0.7	0.95	0.05	105.74
0	0.7	0.95	0.10	105.74
0	0.7	0.90	0.00	89.93
0	0.7	0.90	0.05	90.05
0	0.7	0.90	0.10	90.05
Lovig Model				
0.15	0.85	1.00	0.00	13.26
0.15	0.85	1.00	0.05	13.27
0.15	0.85	1.00	0.10	13.27
0.15	0.85	0.95	0.00	11.61
0.15	0.85	0.95	0.05	11.63

q	$p+q$	A	m	OR
Lovig Model				
0.15	0.85	0.95	0.10	11.63
0.15	0.85	0.90	0.00	10.24
0.15	0.85	0.90	0.05	10.25
0.15	0.85	0.90	0.10	10.25
Warner Model				
0.3	1.00	1.00	0.00	5.60
0.3	1.00	1.00	0.05	5.60
0.3	1.00	1.00	0.10	5.60
0.3	1.00	0.95	0.00	5.09
0.3	1.00	0.95	0.05	5.11
0.3	1.00	0.95	0.10	5.11
0.3	1.00	0.90	0.00	4.65
0.3	1.00	0.90	0.05	4.66
0.3	1.00	0.90	0.10	4.6

Future Work

- Applying different types of measurement error to the mixture model to account for unequal probability of error occurrence
- Measurement error in Optional Mixture Model by Sapra et al. (2022)

THE UNIVERSITY *of* NORTH CAROLINA
GREENSBORO
Department *of* Mathematics & Statistics

- WARNER, S. L., *Randomized response: A survey technique for eliminating evasive answer bias.*, Journal of the American Statistical Association, 60(309):63-69. 1965

- GREENBERG, B. G., ABUL-ELA, A.-L. A., SIMMONS, W. R., AND HORVITZ, D. G. , *The unrelated question randomized response model: Theoretical framework.*, Journal of the American Statistical Association, 64(326):520-539. 1969.

- GUPTA S., MEHTA S., SHABBIR J. AND KHALIL S. , *A unified measure of respondent privacy and model efficiency in quantitative RRT models*, Journal of Statistical Theory and Practice, 12:3, 506-511. 2018.

- LOVIG M., KHALIL S., RAHMAN S., SAPRA P. AND GUPTA S., *A mixture binary RRT model with a unified measure of privacy and efficiency*, Communications in Statistics - Simulation and Computation, DOI: 10.1080/03610918.2021.1914092. 2021.

- SAPRA, P., KHALIL, S. AND GUPTA, S. (2022): *Accounting for Lack of Trust in Optional Binary RRT Models Using a Unified Measure of Privacy and Efficiency*, Journal of Statistical Theory and Practice, Vol. 16, Article 51

THE UNIVERSITY of NORTH CAROLINA GREENSBORO
Department of Mathematics & Statistics

Thank you!
Any questions?

THE UNIVERSITY of NORTH CAROLINA GREENSBORO
Department of Mathematics & Statistics