Benefits and Issues in the Use of Internet-Based Surveys- Experience from Israel

Danny Pfeffermann,

National Statistician and CBS Director, Israel

Professor of Statistics,

Hebrew University of Jerusalem and Southampton University

Conference on future of online data collection in social surveys

University of Southampton, June 2019

Digital Israel

Government decision 2017:

"Israel will leverage the opportunities inherent in the **digital revolution** and in the advancements in **information and communication technologies** to reduce social and geographic disparities, to accelerate economic growth, and to promote smart and friendly government services for its citizens, with the aim of becoming a global leader in the digital sector."

And what about the Israel Central Bureau of Statistics (ICBS)?

Among other things,

- Allow (encourage) response by **Internet** in our surveys,
- Integrate response by Internet with other modes of response, within a generic management system.

Reasons for encouraging response by internet are obvious:

No interviewers, Generally no prior contacts with sampled units, possible **increase in response rates** by letting sampled units respond when convenient, **much cheaper** than other data collection methods, enabling **much larger samples,…**

Possible disadvantages: **No direct interaction with interviewer, low response** (in Israel ☺), possible **mode effects…**

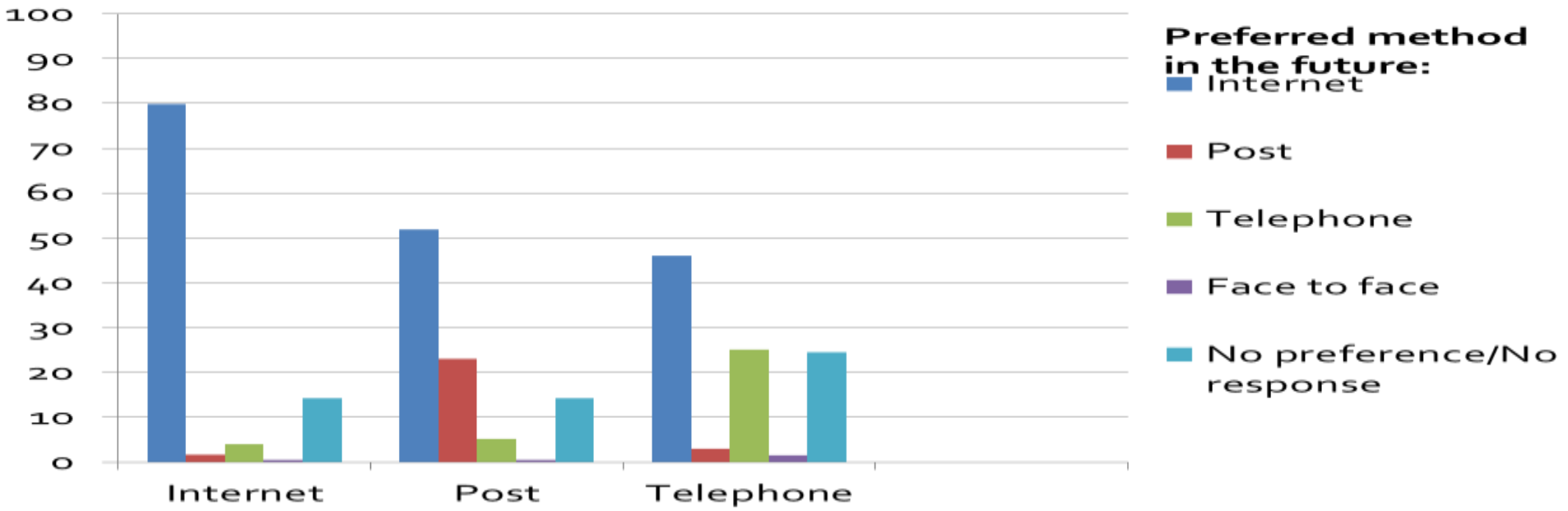
Integration of internet mode in surveys at ICBS so far

- Survey of higher education
- Survey of personal security
- Census of agricultural
- Pilot (rehearsal) census

In what follows I shall highlight some interesting outcomes from these four surveys.

Survey of higher education 2012 - preferred mode of response

Assigned mode of response and preferred mode in the future



- Sample of students divided into the 3 groups. Telephone only offered to those who didn't respond by internet or post. Sampled units asked how they would like to respond in the future.

Survey of higher education (cont.)

	Sample size	Response		Response by internet	
YEAR	**N**	**N**	**%**	**N**	**%**
2012	10,223	8,809	86.2	2,510	28.7
2014	10,207	8,244	80.6	4,744	57.5
2017	9,655	8,139	84.3	3,388	41.6

2012 sample - 2nd year of undergraduates.

2014 sample - Same sample as 2012, mail sent with address and password for use of internet.

2017 sample - new sample.

Lower response in 2017- no longer students and hence lower accuracy of mail addresses and lower motivation and interest.

Personal Security Survey- response rates by Internet

- In 2016- 2018- similar to 2015.
- Low response by orthodox Jews and Arab population.

Mode effects

Mixed mode surveys: different modes of response**; telephone**, **personal interview**, **email, internet…** different modes often offered **sequentially** to non-respondents with a previous mode.

Mode-effects encompass two confounded effects:

Selection effect; different characteristics of respondents using different modes ⇒ **possible differences in study variables;**

Measurement effect; effect of **responding differently by same person**, depending on mode of response.

Big differences often observed in answers with different modes.

Example of mode effects- Agriculture Census, Israel, 2018

- **210** farmers responded both by internet and by telephone**!!**
- Ideal for assessing existence of **measurement effects**.

Study variables	# Farmers T=I	# Farmers T>I	# Farmers T<I
# of workers	131	39	40
Cultivated area	139	38	33

Study variables	Mean Internet (I)	Mean Telephone(T)	Mean for T>I	Mean for T<I
# of workers	5.9	5.8	T=15.5 I= 7.0	T= 7.5 I=17.0
Cultivated area	108.5	105.9	T= 318.4 I= 192.0	T= 88.3 I= 144.5

Another example- pilot (rehearsal) census, Israel 2017

Po – Internet	Po – telephone	Response rate	
	0.75	80	**Telephone**
0.92	0.745	87	**Internet, if not telephone**

Internet respondents differ from telephone respondents

Use of internet increases response rate

P_O - Percentage of correct addresses in Israel Population Register.

How to account for mode effects?

A common approach to deal with mode effects**:** assume that one of the modes has **no measurement effect** $\Rightarrow$ by restricting to this mode, the estimate of the population parameter is **unbiased**.

Uses **observational study** theory**;** requires knowledge of covariates satisfying **strong ignorability conditions**.

- No such mode guaranteed - not clear how to test its existence.

New- Adjusting for mode effects by use of Bayes theorem

Suppose $M \geq 2$ modes and denote $M_i \rightarrow$mode used by unit $i \in S$.

Denote by $\mathbf{x}_i$ covariates explaining Y.

Assumption- for every $j \in U$ exists a **true** value Y_j with ***pdf*** $f_p(y_j \mid \mathbf{x}_j)$.

❖ **Not assumed** that Y is measured accurately under any mode**!!**

Adjusting for mode effects by Bayes theorem (cont.)

By **Bayes theorem**,

$$f_M(y_i \mid \mathrm{x}_i, M_i = m) = \frac{\Pr(M_i = m \mid \mathrm{y}_i, \mathrm{x}_i)\, f_p(y_i \mid \mathrm{x}_i)}{\Pr(M_i = m \mid \mathrm{x}_i)}$$

$f_p(y_i \mid \mathrm{x}_i) \rightarrow$ **Target distribution in population**.

$f_M(y_i \mid \mathrm{x}_i, M_i = m) \rightarrow$ accounts for **Selection** effects from using mode m.

- Requires modelling $\Pr(M_i = m \mid \mathrm{y}_i, \mathrm{x}_i)$ **e.g., multinomial logistic**
- Covariates explaining chosen mode not necessarily the same as covariates explaining the outcome. (For model identification, the two sets of covariates need to **differ** in at least one variable.)

A small simulation study

Y_i binary, $\mathrm{x}_i = (x_{1i}, x_{2i})$ generated independently from *Beta*(2,5).

$$\Pr(Y_i = 1 \mid \mathrm{x}_i) = \frac{1}{1 + \exp(\alpha_0 + \alpha_1 x_{1i})};$$

$$\Pr(M_i = m \mid Y_i, \mathrm{x}_i) = \frac{\exp(\beta_{0m} + \beta_{1m} Y_i + \gamma_1 x_{2i})}{1 + \sum_{m=1}^{3} \exp(\beta_{0m} + \beta_{1m} Y_i + \gamma_1 x_{2i})}, \; m = 1, 2, 3.$$

- **Four** possible modes. ***m*=4→ non-ignorable non response**.
- Simple random sample without replacement of size ***n*=1,000**, out of population of size ***N*=10,000**.
- **Single** population, **250** independent samples (**design-based**).
- Model parameters re-estimated from each sample by **MLE**.

	$m=1$	$m=2$	$m=3$	$m=4$
N_m	9474	24436	40608	25482
$\hat{N}_{m,\text{Model}}$	10181 (692)	24439 (1071)	40382 (990)	24998 (863)
$\hat{N}_{m,\text{HT}}$	10503 (795)	24321 (1074)	39778 (977)	24753 (776)
$\bar{Y}_m$	0.40	0.56	0.62	0.85
$\hat{\bar{Y}}_{m,\text{Model}}$	0.42 (0.036)	0.55 (0.024)	0.62 (0.009)	0.84 (0.029)
$\hat{\bar{Y}}_{m,\text{HT}}$	0.43 (0.065)	0.54 (0.023)	0.63 (0.011)	0.80 (0.049)

$$\hat{N}_{m,\text{Model}} = \sum_{i=1}^{N} \hat{Pr}(M_i = m \mid \mathrm{x}_i); \quad \hat{N}_{m,HT} = \sum_{i=1}^{n} \frac{\hat{Pr}(M_i = m \mid \mathrm{x}_i)}{Pr(i \in S)},$$

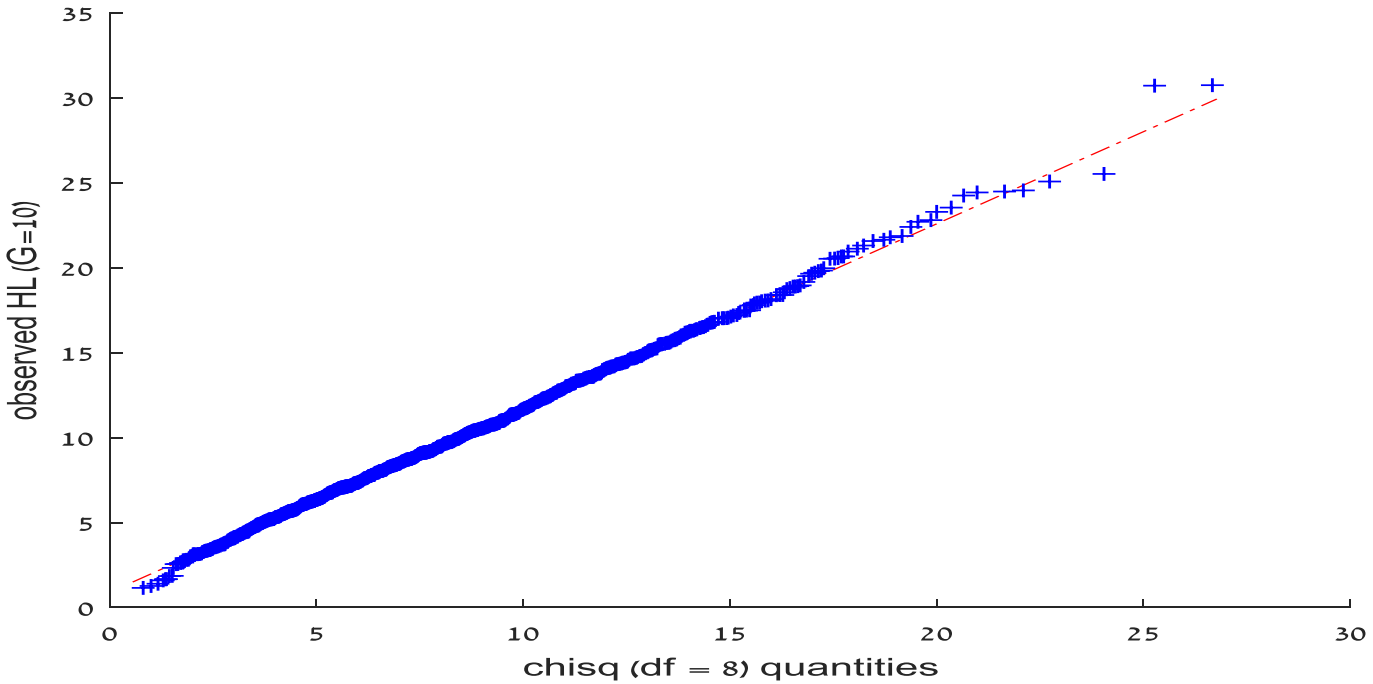
$$\hat{\bar{Y}}_{m,\text{Model}} = (\hat{N}_{m,\text{Model}})^{-1} \sum_{i=1}^{N} \hat{Pr}(Y_i = 1 \mid m, \mathrm{x}_i); \quad \hat{\bar{Y}}_{m,HT} = \hat{\mathrm{N}}_{m,HT}^{-1} \frac{\sum_{i=1}^{n} \hat{Pr}(Y_i = 1 \mid m, \mathrm{x}_i)}{Pr(i \in S)}.$$

$\bar{Y}$	$\hat{\bar{Y}}_{\text{Model}} = \mathrm{N}^{-1} \sum_{i=1}^{N} \hat{Pr}(Y_i = 1 \mid \mathrm{x}_i)$	$\hat{\bar{Y}}_{\text{HT}}$	$\bar{Y}_s$
0.64	0.64 (0.015)	0.62 (0.016)	0.52 (0.022)

Model testing (Hosmer & Lemashov, 1980)

Divide the sample into G groups based on probabilities $\hat{p}_r(Y_i = 1 | \mathrm{x}_i)$

$$\mathbf{H\text{-}L} = \sum_{g=1}^{G} \frac{(\bar{y}_g - \bar{\hat{q}}_g)^2}{\bar{\hat{q}}_g(1-\bar{\hat{q}}_g)/r_g} \overset{H_0}{\sim} \chi^2_{(G-2)}; \bar{y}_g = \frac{1}{r_g}\sum_{i\in g} y_i, \bar{\hat{q}}_g = \frac{1}{r_g}\sum_{i\in g} \hat{p}_r(Y_i = 1 | \mathrm{x}_i).$$



Further remarks

1- The proposed approach does not require the existence of covariates satisfying strong ignorability conditions.

2- The approach does not assume that the responses obtained by one of the modes are **correct**.

3- **Nonignorable nonresponse** accounted for.

4- Knowledge of covariates for outside the sample **not required**

5- **Requires** modelling $\Pr(M_i = m \mid \mathrm{y}_i, \mathrm{x}_i)$ and $f_p(y_i \mid \mathrm{x}_i)$, but the Models $\Pr(M_i = m \mid \mathrm{y}_i, \mathrm{x}_i)$ and $f_M(y_i \mid \mathrm{x}_i, M_i = m)$, and hence $f_p(y_i \mid \mathrm{x}_i)$ **can be tested** using standard test procedures.

6- The approach can be extended to adjust also for measurement errors (**in work**).

ICBS project to Integrate Internet Response – 2019

Target: **Enable online response in all surveys with a generic management system.**

CAXL	❖ **Call – Computer Assisted Internet Interview,** ❖ **CAPI – Computer Assisted Personal Interview,** ❖ **CATI - Computer Assisted Telephone Interview.**

❖ **Initiated by Israel's Finance Ministry as part of a project aimed at digitation of the CBS.**

Adapting surveys to the use of the internet

- **Management:** allow simultaneous use of different response modes, make sure that a person responding by the internet is no longer asked to respond by other modes;

- **Questionnaire:** **"Taylor made"**- don't ask a person questions for which the answer is known from administrative files; Format, length, structure, graphics, of questionnaire…

- **Interaction with sampled units:** reference/Introduction letter, Number, frequency and gaps between reminders, find ways of encouraging (attracting) response by the internet.

Cost Savings

Annual saving of integrating internet response in all our surveys is assessed at about **8 million NIS**, **~16%** of the current cost of our data collection. Assumes **15%** response by internet in the household surveys, and **50%** response in the business surveys.

Conclusions

- Use of internet has big operational, logistic and cost Advantages with the potential of increasing response rates.
- However, in many countries, the use of internet not found **yet** to be the preferred mode by respondents.
- Answers in internet surveys may differ from answers obtained with other modes- not clear which mode provides more accurate answers.
- By appropriate modelling, it is possible to adjust for possible mode effects. **The models can be tested**.