Phone Surveys in LMICs through a Total Survey Error Framework: Focus on *Measurement*

Charles Lau, PhD (Gallup)

Co-author: Abigail Greenleaf, PhD, MPH (Columbia University)

GALLUP®

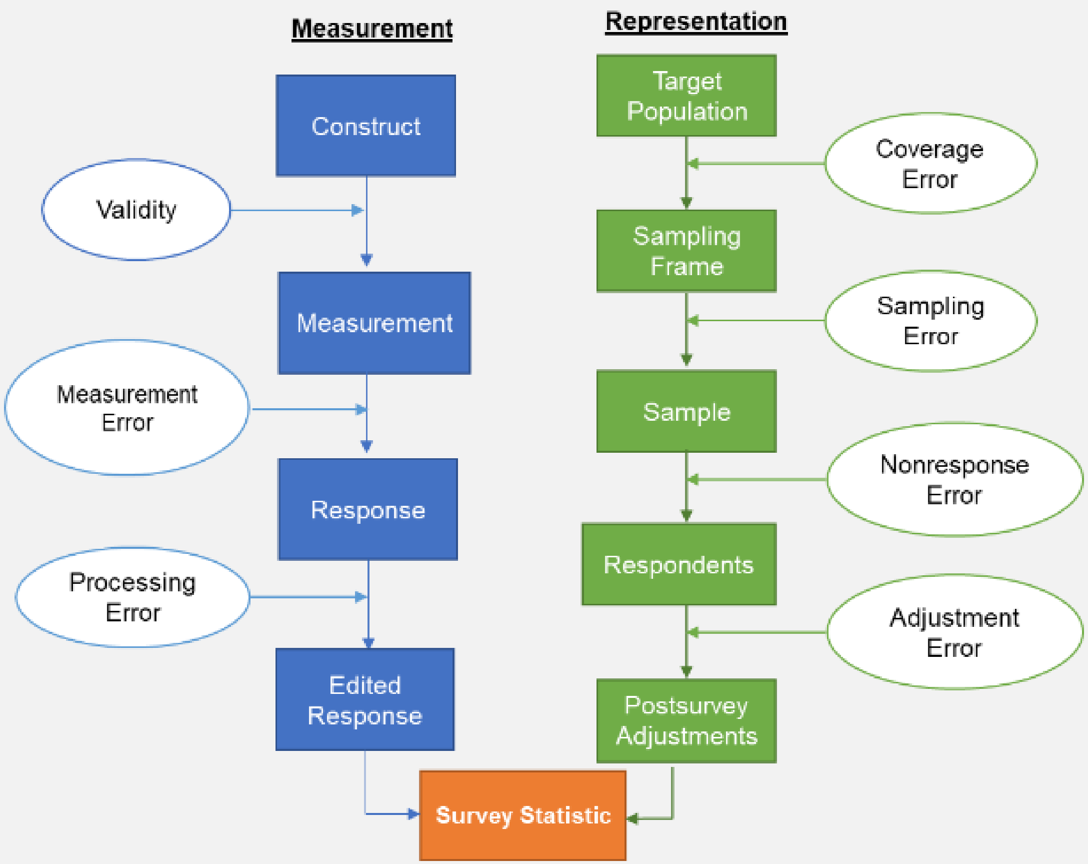
Measurement in a Total Survey Error (TSE) Framework

Construct: What we're studying

Measurement: How we're collecting the information (e.g., question wording, mode)

Respondent: How the respondent answers

Edited Response: What we do to the respondent's answer (e.g., data processing)



Copyright © 2025 Gallup, Inc. All rights reserved.
GALLUP®

Studies on Measurement in CATI Surveys

Types of Research

15 articles total

13 studied factual questions

2 studied knowledge questions

Note: No studies focused on attitudinal or opinion questions

Methods To Study Measurement

1. Comparison with reference standard (n = 1)
2. Repeated measures (n = 9)
3. Questionnaire design experiments (n = 2)
4. Comparison of independent CATI/CAPI samples (n = 2)
5. Psychometrics (n = 1)

Copyright © 2025 Gallup, Inc. All rights reserved.

Findings

Characteristics of …

1. Questions
2. Questionnaires
3. Respondent
4. Interviewer

→

Measurement Error
in CATI Surveys

Copyright © 2025 Gallup, Inc. All rights reserved.

1. Question Characteristics: The Good News

Questions about straightforward constructs using simple response options (e.g., yes/no) produce quality data

High levels of reliability when CATI surveys ask about alcohol, tobacco, diabetes (Lebanon, Tanzania/Bangladesh) (Mahfoud et al., 2014; Pariyo et al., 2019)

Eight-item scale using binary outcomes on food insecurity in the past three months produces reliable and valid data (Mexico) Gaitán-Rossi et al., 2020)

Scales assessing whether food item was consumed (and another assessing the number of days in the past 7 days) were consistent between separate CATI and FTF surveys (Ethiopia) (Abate et al., 2023)

A scale using binary outcomes to assess women’s consumption of ten different food groups produced similar data in both CATI and FTF modes conducted approximately nine days apart (Kenya) (Lamanna et al., 2019)

Copyright © 2025 Gallup, Inc. All rights reserved.

GALLUP

1. Question Characteristics: The Good News

Questions about straightforward constructs using simple response options (e.g., yes/no) produce quality data

High levels of agreement between FTF and CATI survey about simple monitoring indicators for a community health program (Mali)

High levels of agreement between two CATI calls to assess patient-reported post-delivery outcomes (Uttar Pradesh/India)

High levels of agreement between FTF and CATI survey about measles vaccine (Cameroon)

Copyright © 2025 Gallup, Inc. All rights reserved.

1. Question Characteristics: The Bad News

*Questions about **complex constructs** produced lower quality data*

Low reliability when assessing physical activity (Lebanon, Bangladesh/Tanzania) (Mahfoud et al., 2014; Pariyo et al., 2019)

Now I am going to ask you the time you spend being physically active in a typical week. Count the duration of how long you are physically active at work, at home at least for 10 minutes. First, I will ask you about your vigorous physical activity and then about moderate physical activity.

In a typical week, think about your moderate intensity activity. Moderate intensity activity (at least for 10 minutes) that causes small increase in breathing or heart rate e.g. carrying light loads, rowing a boat or riding a bicycle with a regular pace. Do not count walking in this activity.

Source: Pariyo et al. (2019)

Copyright © 2025 Gallup, Inc. All rights reserved.

1. Question Characteristics: The Bad News

Detailed questions *produced lower quality data*

Low levels of sensitivity (between CATI/FTF) when asking detailed questions to monitor community health program (Chen et al., 2021; Mali)

Example Indicator:
Proportion of ASCs reporting Amoxicillin tablets/syrup stock-out that lasted more than 1 consecutive week in the past 3 months

Copyright © 2025 Gallup, Inc. All rights reserved.
GALLUP

1. Question Characteristics: The Bad News

<u>Numeric questions</u> *produced lower quality data*

Wilson (2015) compared respondent self-reports from CATI surveys about cookstove usage with sensor data. Self-reports were approximately <u>double</u> the numbers recorded in the cookstove sensor (Darfur/Sudan).

Anderson et al. (2023) found that the FTF survey produced significantly lower crop production estimates (ranging from 14% for fava beans to 68% smaller for pigeon peas) compared to CATI (India)

When asked about 118 types of food consumed in the past seven days and expenditure on 25 items, CATI produced 23% lower estimates than the FTF survey – doubling the poverty rate (Abate et al., 2023) (Ethiopia)

Copyright © 2025 Gallup, Inc. All rights reserved.
GALLUP®

1. Question Characteristics: The Bad News

Knowledge questions *produced lower quality data*

Pregnant women (Ng et al., 2022): knowledge was low, resulting in poor reliability for both modes (India)

Among community health workers (Shah et al., 2020), reliability was low for complex questions (i.e. field-coded and count questions) (Mali)

Proxy reports *produced lower quality data*

Dietary diversity of female respondent is reliable, but not when caregivers are asked to report about diet of children (6-23 months) (Kenya)

A study about adverse birth outcomes in India finds greater inconsistency about adverse birth outcomes when report came from someone other than mother or mother’s husband (India)

Copyright © 2025 Gallup, Inc. All rights reserved.
GALLUP®

1. Question Characteristics: The Bad News

Proxy reports *produced lower quality data*

Dietary diversity of female respondent is reliable, but not when caregivers are asked to report about diet of children (6-23 months) (Kenya)

A study about adverse birth outcomes in India finds greater inconsistency about adverse birth outcomes when report came from someone other than mother or mother's husband (India)

Copyright © 2025 Gallup, Inc. All rights reserved.

2. Questionnaire Length

Questionnaire length varies:

- Some questionnaires were brief:
  - 7 minutes in Lamarange et al. (2016)
  - 9 minutes in Ramesh et al. (2023)

- Others exceeded 30 minutes
  - 36 minutes in Lambrecht et al.
  - 41 minutes in Abate et al. (2023)
  - 41-50 minutes in Glazerman et al. (2023)

Lots of rules of thumb, but little empirical evidence

Copyright © 2025 Gallup, Inc. All rights reserved.
GALLUP®

2. Questionnaire Length

Optimal questionnaire length may depend on complexity of questions

Torrisi et al. (2024) - Malawi

- Randomized respondents to receive a 10-, 20-, or 30-minute questionnaire.
- Completion and cooperation rates were over 94% in all questionnaire versions.
- Data quality indicators were the same across the 10-, 20-, or 30-minute questionnaires.

Abate et al. (2022) - Ethiopia

- Randomized the placement of a nutrition module within a CATI survey
- When nutrition questions are asked later in the interview (on average, 15-minutes later), respondents report consuming fewer food types
- More underreporting among less common food groups such as animal sources (40% decrease) and fruits and vegetables (11% decrease).
- Lower self-reports contribute to a 28% decrease in minimum dietary diversity.

Copyright © 2025 Gallup, Inc. All rights reserved.

3. Respondent Characteristics

- Few studies on impact of respondent characteristics on measurement

- One exception: Abate et al. (2023)
  - CATI underestimates consumption expenditures relative to FTF
  - But underestimation is greater among
    - Less educated (may experience greater cognitive burden)
    - Larger households (difficulty recalling or knowing expenditures from all household members).

But very limited evidence about how respondent characteristics affect the quality of measurement

Copyright © 2025 Gallup, Inc. All rights reserved.
GALLUP®

4. Interviewer Characteristics

- Few studies on interviewer effects
- Lammana et al. (2019)
  - Over 50% of the variance in self-reports about nutrition in Kenya was on the interviewer level.
  - When collecting data from caregivers about children's nutritional intake, male interviewers recorded fewer foods in CATI (but not FTF).

Abate et al. (2023) did not find any evidence of interviewer effects, however.

Copyright © 2025 Gallup, Inc. All rights reserved.

Summary: Measurement

Key Findings

1. CATI can produce quality data when asking about simple constructs and using simple response options (yes/no)
2. Concerns about measurement error when CATI surveys ask complex questions
3. Little evidence about optimal questionnaire length

Recommendations

- Keep it simple!
- Adopt respondent-centered design
- Read questionnaires out loud
- Conduct cognitive interviews
- Conduct field tests
- Don’t copy-paste CAPI questionnaire in CATI
- Conduct methods studies

Copyright © 2025 Gallup, Inc. All rights reserved.

Discussion

Key Question: When is CATI "Fit for Purpose?"

(1) When FTF surveys aren't feasible

- Remote or inaccessible areas
- Conflict zones
- High frequency data
- Longitudinal data
- During health emergencies
- Multi-country surveys with resource constraints

(2) Research not based on nationally representative samples

- Skilled professionals
- Special populations
- Recruited from locations
- Describing changes

(3) Mixed-mode *CATI as a complement for FTF surveys*

- CAPI → CATI
- CAPI + CATI

Copyright © 2025 Gallup, Inc. All rights reserved.
GALLUP®