What are the impacts of using a geo-sampling design on the precision of study outcomes?
The Study

**Study:** Barriers to internet adoption

**Where:** Low/middle income countries

**Outcomes:**
- Internet use in past 30 days
- Number mobile phones used in last 30 days
Key Parameters:

- Nationally representative probability-based sample of 15 to 64 year-olds
- 3,000 respondents
- Urban/rural and various poverty levels
- Logistically feasible
- Time and cost effective
  - No listing process
- No random walk
- Adaptable across countries
- Electronic data collection
Geo-sampling Methodology

Geo-sampling is:

- An innovative, multi-stage, probability-based sampling methodology

- Integrates Geographic Information Systems (GIS) at several stages of the sampling process

- Divides target areas into logistically manageable units (Amer, AAPOR 2015)
Geo-sampling Methodology

[Map images showing Municipio Boundary, Secondary Grid Cell, Primary Grid Cell, and Main Sample/Muestra principal]
Standardized Steps of Geo-sampling

- Select sample up to “district” level
  - “District” varies by country
Standardized Steps of Geo-sampling

- Select sample up to “district” level
  - “District” varies by country

- Divide districts into 1-km$^2$ grid cells; (“primary grid units”)
  - Randomly sample residential PGUs
Standardized Steps of Geo-sampling

- Select sample up to “district” level
  - “District” varies by country

- Divide districts into 1-km² grid cells (“primary grid units”)
  - Randomly sample residential PGUs

- Divide selected 1-km² grid cells into smaller sub-grid cells (“secondary grid units”)
  - Randomly sample residential SGUs
  - 100 m² in rural areas / 50 m² in urban areas
Standardized Steps of Geo-sampling

- Select sample up to “district” level
  - “District” varies by country

- Divide districts into 1-km² grid cells (“primary grid units”)
  - Randomly sample residential PGUs

- Divide selected 1-km² grid cells into smaller sub-grid cells (“secondary grid units”)
  - Randomly sample residential SGUs
  - 100 m² in rural areas / 50 m² in urban areas

- All households within selected SGUs

- One adult 15-64 per household
Standardization/Tailoring of Sample Design

- **Countries**
  - **Asia**: Thailand, Bangladesh, India
  - **Africa**: Ghana, Kenya, Nigeria, Rwanda, Uganda
  - **South America**: Brazil, Colombia, Guatemala

- **Variability in sample design resources across countries**
  - Different geographic boundaries ("district")
  - Availability of administrative unit frames
  - Availability/level of Census data for stratification

- **Country-specific tailoring**
  - Number of stages prior to geo-sampling
  - Size of area units selected before grid units
  - Sampling method (e.g., SRS, probability proportional to size)

*Most tailoring occurred at higher stages before geo-sampling stages*
What are the impacts of using a geo-sampling design on the precision of study outcomes?
Design Effect (DEFF)

- “Penalty” to the precision of study outcomes from using a complex sample design (e.g., unequal selection probabilities, clustering, stratification) instead of simple random sampling

- Determines effective sample size

- Decomposed conceptually as product of unequal weighting effect ($UWE$) and clustering effects ($DEFF_{clust}$)

\[
DEFF = UWE \times DEFF_{clust}
\]
### Internet Used in Past 30 Days

#### Design Effects

<table>
<thead>
<tr>
<th></th>
<th>Africa</th>
<th>Asia</th>
<th>South America</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Nigeria</td>
<td>Uganda</td>
<td>Thailand</td>
</tr>
<tr>
<td>Sample Size</td>
<td>3,042</td>
<td>3,071</td>
<td>3,134</td>
</tr>
<tr>
<td>DEFF</td>
<td>15.0</td>
<td>2.9</td>
<td>2.0</td>
</tr>
<tr>
<td>UWE (% of DEFF)</td>
<td>2.2 (15%)</td>
<td>2.2 (76%)</td>
<td>2.4 (120%)</td>
</tr>
</tbody>
</table>
## Number Mobile Phones Used in Past 30 Days

### Design Effects

<table>
<thead>
<tr>
<th>Country</th>
<th>Sample Size</th>
<th>DEFF</th>
<th>UWE (% of DEFF)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nigeria</td>
<td>2,844</td>
<td>27.0</td>
<td>2.2 (8%)</td>
</tr>
<tr>
<td>Uganda</td>
<td>2,635</td>
<td>4.1</td>
<td>2.3 (56%)</td>
</tr>
<tr>
<td>Asia</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thailand</td>
<td>3,111</td>
<td>18.6</td>
<td>2.4 (13%)</td>
</tr>
<tr>
<td>Brazil</td>
<td>3,022</td>
<td>8.3</td>
<td>2.3 (28%)</td>
</tr>
<tr>
<td>South America</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Colombia</td>
<td>2,577</td>
<td>7.6</td>
<td>3.4 (45%)</td>
</tr>
</tbody>
</table>
Components of the Clustering Effect

- **Average cluster size**
  - Weighted average is appropriate when cluster sizes vary (Holt 1980; Chromy 2014)

- **Intracluster correlation coefficient** (*ICC*)
  - Measure of homogeneity within clusters
  - Can vary by outcome
  - Our method: ANOVA model with fully nested random effects for clusters and stratification variables as covariates (Ganninger 2010; Delong and Lokhnygina 2014)
<table>
<thead>
<tr>
<th>Sampling Stage</th>
<th>Cluster Unit</th>
<th>ICC</th>
<th>Average Cluster Size</th>
<th>Clustering Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>State</td>
<td>.04</td>
<td>268.4</td>
<td>12.7</td>
</tr>
<tr>
<td>2</td>
<td>Local Government Area</td>
<td>.03</td>
<td>62.2</td>
<td>2.8</td>
</tr>
<tr>
<td>3</td>
<td>Primary Grid Unit (1-km² cell)</td>
<td>.04</td>
<td>29.6</td>
<td>2.3</td>
</tr>
<tr>
<td>4</td>
<td>Secondary Grid Unit</td>
<td>.06</td>
<td>14.6</td>
<td>1.9</td>
</tr>
<tr>
<td>Sampling Stage</td>
<td>Cluster Unit</td>
<td>ICC</td>
<td>Average Cluster Size</td>
<td>Clustering Effect</td>
</tr>
<tr>
<td>----------------</td>
<td>-------------------------------------</td>
<td>------</td>
<td>----------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>1</td>
<td>State</td>
<td>.15</td>
<td>245.8</td>
<td>36.9</td>
</tr>
<tr>
<td>2</td>
<td>Local Government Area</td>
<td>.02</td>
<td>57.7</td>
<td>2.1</td>
</tr>
<tr>
<td>3</td>
<td>Primary Grid Unit (1-km² cell)</td>
<td>.02</td>
<td>27.4</td>
<td>1.6</td>
</tr>
<tr>
<td>4</td>
<td>Secondary Grid Unit</td>
<td>.06</td>
<td>13.5</td>
<td>1.7</td>
</tr>
</tbody>
</table>
Conclusions and future analysis

Benefits of geo-sampling methodology
- Innovative design, adaptable across countries
- No listing/No random walk
- Cuts costs and implementation timeline

High design effects not due to geo-sampling methodology
- Large cluster sizes from higher stages of sampling contributing to clustering effects

Future analysis
- Why do the clustering effects vary by outcome?
- Decompose clustering effects across additional countries
Thank you

Jennifer Unangst  
Research Statistician  
junangst@rti.org

Jeniffer Iriondo-Perez  
Research Statistician  
perez@rti.org

Safaa Amer, PhD  
Senior Research Statistician  
samer@rti.org