

Surveys, Platforms, and Crowdsourcing

Jared Edgerton

Why Surveys Still Matter

Surveys remain a central tool for:

- Measuring attitudes and beliefs
- Capturing latent preferences
- Studying populations at scale

But modern surveys are rarely simple random samples.

The Shift to Platforms

Much contemporary survey data are collected via:

- Online labor markets
- Survey platforms
- Panel providers
- Hybrid recruitment strategies

Platforms reshape who answers and how.

Data as a Measurement Process

Survey data reflect:

- Question wording
- Response options
- Ordering and framing
- Mode of administration

Measurement decisions precede analysis.

From Instrument to Dataset

Surveys arrive as:

- Exported files (CSV, TSV, SPSS)
- Codebooks or questionnaires
- Metadata files
- Platform-generated diagnostics

Understanding structure comes first.

Reading Survey Exports (Python)

```
import pandas as pd

df = pd.read_csv('survey_export.csv')
df.head()
```

The raw export is not analysis-ready data.

Reading Survey Exports (R)

```
library(readr)
```

```
df ← read_csv('survey_export.csv')  
head(df)
```

Always inspect before cleaning.

Codebooks Are Data

Codebooks specify:

- Variable meanings
- Scales and categories
- Missing value codes
- Skip logic

Ignoring codebooks leads to silent errors.

Labeling and Recoding

Cleaning often requires:

- Relabeling categories
- Harmonizing scales
- Converting strings to factors
- Explicit missing-value handling

Recoding Example (Python)

```
df['gender'] = df['gender'].map({  
    1: 'Male',  
    2: 'Female',  
    3: 'Other'  
})
```

Recoding Example (R)

```
df$gender ← recode(  
  df$gender,  
  '1' = 'Male',  
  '2' = 'Female',  
  '3' = 'Other'  
)
```

Crowdsourcing and Human Subjects

Crowdsourced data involve:

- Heterogeneous respondents
- Variable attention and effort
- Strategic responding

Quality is uneven and must be assessed.

Quality Checks Are Essential

Common checks include:

- Completion time thresholds
- Attention checks
- Straight-lining detection
- Response consistency

These are measurement decisions.

Simple Quality Check (Python)

```
df = df[df['duration_seconds'] > 60]
```

Simple Quality Check (R)

```
df ← df ▷ dplyr::filter(duration_seconds > 60)
```

Missingness Is Informative

Missing data may reflect:

- Survey design
- Respondent fatigue
- Sensitivity of questions
- Platform defaults

Missingness is rarely random.

Platforms Shape Data

Platforms influence:

- Who participates
- Incentives and compensation
- Sampling frames
- Longitudinal availability

Platform choice is a design choice.

Documentation Is Non-Negotiable

Every survey pipeline should record:

- Recruitment method
- Platform used
- Field dates
- Exclusion criteria
- Known limitations

Survey data without documentation are not reusable.

What We Emphasize in Practice

- Treat surveys as constructed instruments
- Inspect raw exports carefully
- Make recoding explicit
- Document quality filters

Discussion

- Where do surveys introduce bias?
- How do platforms shape who responds?
- Which quality checks feel defensible?