What to do about Biases in Survey Research

Gary King


September 6, 2007
Readings


- Papers, FAQ, examples, software, conferences, videos: [http://GKing.Harvard.edu/vign](http://GKing.Harvard.edu/vign)
The Importance of Survey Research

- In political science: 1/2 of all quantitative articles
- Other social sciences and related professional areas: Widely used
- A large fraction of our information base over the last half century
- A multi-billion dollar industry
- Of widespread public interest
Examples of the Problem

Presidential Approval: the longest public opinion time series

- On 9/10/2001, 55% of Americans approved of the way George W. Bush was “handling his job as president”.
- The next day — which the president spent in hiding — 90% approved.
- Was this massive opinion change, or was the same question interpreted differently?
Examples of the Problem

The O.J. Simpson trial: most publicized murder trial in history

- The facts of the case seemed clear
- Did he do it? Whites: 62% say “yes”. Blacks: 14% say “yes”.
- Did black and white Americans have genuinely opposing views about whether Simpson committed murder, or did the two groups interpret the same survey question differently?

Gary King

Anchoring Vignettes
Examples of the Problem

The most common measure of the health of populations: “How healthy are you? Excellent, Good, Fair, or Poor”

- Suppose an otherwise healthy 25-year-old woman with a cold and a backache answers “fair” and a 90-year-old man just able to get out of bed says “excellent”
- Is the young woman less healthy or are the two interpreting the same question differently?
- In some countries, responses to this survey question correlate negatively with objective measures of health status (Sen, 2002).
Fig 1. Life expectancy among males and females in India compared with United States, mid-1990s.

Fig 2. Incidence of reported morbidity in India, mid-1970s, compared with United States, mid-1960s.
Anchoring Vignettes & Self-Assessments: Political Efficacy (about voting)

- “[Alison] lacks clean drinking water. She and her neighbors are supporting an opposition candidate in the forthcoming elections that has promised to address the issue. It appears that so many people in her area feel the same way that the opposition candidate will defeat the incumbent representative.”

- “[Jane] lacks clean drinking water because the government is pursuing an industrial development plan. In the campaign for an upcoming election, an opposition party has promised to address the issue, but she feels it would be futile to vote for the opposition since the government is certain to win.”

- “[Moses] lacks clean drinking water. He would like to change this, but he can’t vote, and feels that no one in the government cares about this issue. So he suffers in silence, hoping something will be done in the future.”

How much say [does ‘name’ / do you] have in getting the government to address issues that interest [him / her / you]?
(a) Unlimited say, (b) A lot of say, (c) Some say, (d) Little say, (e) No say at all
The only reason for vignette assessments to change over respondents is DIF.
Assumption holds because investigator creates the anchors (Alison, Jane, Moses).
Our simple (nonparametric) method works this way.
A Simple, Nonparametric Method

- Define self-assessment answers *relative* to vignettes answers.
- For respondents who rank vignettes, $z_{i1} < z_{i2} < \cdots < z_{iJ}$,

$$C_i = \begin{cases} 
1 & \text{if } y_i < z_{i1} \\
2 & \text{if } y_i = z_{i1} \\
3 & \text{if } z_{i1} < y_i < z_{i2} \\
\vdots & \vdots \\
2J + 1 & \text{if } y_i > z_{iJ} 
\end{cases}$$

- Apportion $C$ equally among tied vignette categories
- (This is wrong, but simple; we will improve shortly)
- Treat vignette ranking inconsistencies as ties
- Requires vignettes and self-assessments asked of all respondents
- (Our parametric method doesn’t)
Comparing China and Mexico
Mexico

Opposition leader Vicente Fox elected President.
71-year rule of PRI party ends.
Peaceful transition of power begins.

Plenty of political efficacy
China: How much say do you have in getting the government to address issues that interest you?
The left graph is a histogram of the observed categorical self-assessments.

The right graph is a histogram of $C$, our nonparametric DIF-corrected estimate of the same distribution.
Key Measurement Assumptions

1. **Response Consistency**: Each respondent uses the self-assessment and vignette categories in approximately the same way across questions. (DIF occurs across respondents, not across questions for any one respondent.)

2. **Vignette Equivalence**:
   (a) The actual level for any vignette is the same for all respondents.
   (b) The quantity being estimated exists.
   (c) The scale being tapped is perceived as unidimensional.

3. In other words: we allow response-category DIF but assume stem question equivalence.
Ties and Inconsistencies Produce *Ranges*

<table>
<thead>
<tr>
<th>Example</th>
<th>Survey Responses</th>
<th>1: $y &lt; z_1$</th>
<th>2: $y = z_1$</th>
<th>3: $z_1 &lt; y &lt; z_2$</th>
<th>4: $y = z_2$</th>
<th>5: $y &gt; z_2$</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$y &lt; z_1 &lt; z_2$</td>
<td>T</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>{1}</td>
</tr>
<tr>
<td>2</td>
<td>$y = z_1 &lt; z_2$</td>
<td></td>
<td>T</td>
<td></td>
<td></td>
<td></td>
<td>{2}</td>
</tr>
<tr>
<td>3</td>
<td>$z_1 &lt; y &lt; z_2$</td>
<td></td>
<td></td>
<td>T</td>
<td></td>
<td></td>
<td>{3}</td>
</tr>
<tr>
<td>4</td>
<td>$z_1 &lt; y = z_2$</td>
<td></td>
<td></td>
<td>T</td>
<td></td>
<td></td>
<td>{4}</td>
</tr>
<tr>
<td>5</td>
<td>$z_1 &lt; z_2 &lt; y$</td>
<td></td>
<td></td>
<td></td>
<td>T</td>
<td></td>
<td>{5}</td>
</tr>
</tbody>
</table>

**Ties:**

<table>
<thead>
<tr>
<th>Example</th>
<th>Survey Responses</th>
<th>1: $y &lt; z_1 = z_2$</th>
<th>2: $y = z_1 = z_2$</th>
<th>3: $z_1 = z_2 &lt; y$</th>
<th>4: $z_1 &lt; z_2$</th>
<th>5: $z_1 = z_2 &lt; z_1$</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>$y &lt; z_1 = z_2$</td>
<td>T</td>
<td></td>
<td>T</td>
<td></td>
<td>T</td>
<td>{1}</td>
</tr>
<tr>
<td>7</td>
<td>$y = z_1 = z_2$</td>
<td>T</td>
<td></td>
<td>T</td>
<td></td>
<td>T</td>
<td>{2,3,4}</td>
</tr>
<tr>
<td>8</td>
<td>$z_1 = z_2 &lt; y$</td>
<td>T</td>
<td></td>
<td></td>
<td></td>
<td>T</td>
<td>{5}</td>
</tr>
</tbody>
</table>

**Inconsistencies:**

<table>
<thead>
<tr>
<th>Example</th>
<th>Survey Responses</th>
<th>1: $y &lt; z_2 &lt; z_1$</th>
<th>2: $y = z_2 &lt; z_1$</th>
<th>3: $z_2 &lt; y &lt; z_1$</th>
<th>4: $z_2 &lt; y = z_1$</th>
<th>5: $z_2 &lt; z_1 &lt; y$</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>$y &lt; z_2 &lt; z_1$</td>
<td>T</td>
<td></td>
<td></td>
<td></td>
<td>T</td>
<td>{1}</td>
</tr>
<tr>
<td>10</td>
<td>$y = z_2 &lt; z_1$</td>
<td>T</td>
<td></td>
<td>T</td>
<td></td>
<td>T</td>
<td>{1,2,3,4}</td>
</tr>
<tr>
<td>11</td>
<td>$z_2 &lt; y &lt; z_1$</td>
<td>T</td>
<td></td>
<td>T</td>
<td></td>
<td>T</td>
<td>{1,2,3,4,5}</td>
</tr>
<tr>
<td>12</td>
<td>$z_2 &lt; y = z_1$</td>
<td>T</td>
<td></td>
<td>T</td>
<td></td>
<td>T</td>
<td>{2,3,4,5}</td>
</tr>
<tr>
<td>13</td>
<td>$z_2 &lt; z_1 &lt; y$</td>
<td>T</td>
<td></td>
<td></td>
<td></td>
<td>T</td>
<td>{5}</td>
</tr>
</tbody>
</table>
How to analyze a variable with scalar and vector responses?

We define a new method (censored ordered probit), a direct extension of ordinal probit allowing for ranges of responses

Useful for vignettes; also useful for survey questions that allow ranges of responses
Improved Efficiency in Practice

Uniform

Unconditional

Conditional

Gary King

Anchoring Vignettes
**Ultimate Goal:** Define categories with vignettes to learn about a continuous unobserved variable (health, efficacy).

- **Worst choice:** All in one category, no discriminatory power (E.g., “Bob ran two marathons last week...” does not discriminate among respondents)
- **Best choice:** Largest number of categories, equal proportions across categories

**Immediate Goal:** Measure information in a categorization scheme

- **Operational use:**
  - Run a pretest with lots of vignettes
  - Compute $C$ and $H(C)$ for each possible subset,
  - Choose vignettes for the main survey based on $H$ and cost of survey questions.
A measure of information $H(C)$?

- Three Criteria for a measure, $H(C)$:
  1. $H(C)$ should be 0 when all answers are in one category; at a maximum when proportions are equal across categories
  2. $H(C)$ should increase monotonically with the number of vignettes (and thus categories)
  3. Assume consistent decomposition as we add vignettes

- Lots of candidates exist (all inequality measures): Gini index, variance, absolute deviations, Herfindahl index, etc.
- Only one measure satisfies all three criteria: entropy.

- Thus, formally, we set:

$$H(p_1, \ldots, p_{2J+1}) = - \sum_{j=1}^{2J+1} p_j \ln(p_j)$$

- Only question remaining: How do we calculate entropy when $C$ is not a scalar?
Step 2: Defining $H(C)$ for scalar and vector $C$

- Without ties or inconsistencies, we simply compute entropy
- With ties and inconsistencies, 2 options:
  - Estimated entropy: using the censored ordinal probit model
  - Known (minimum) entropy: information in the data we know exists for certain (inferences do not depend on the model)
- Result is easy to use: one measure indicating information in survey question and vignettes
Political Efficacy (Mex & China)

Known (Minimum) Entropy

Estimated Entropy

12
3
4
5
123
124 125
134 135
145
234
235
245
345
12345
12 13
14 15
23
24 25
34 35
45
1234
1235
1245
1345 2345
One vignette can be better than three: Sleep (China)
Some vignette sets are uninformative: Self-Care (China)
Some covariates are unhelpful: Pain (China)
If thresholds vary, categorical answers are meaningless.

Our parametric model works by estimating the thresholds.

Vignettes provide identifying information for the $\tau$'s.
Self-Assessments v. Medical Tests

Self-Assessment:
In the last 30 days, how much difficulty did [you/name] have in seeing and recognizing a person you know across the road (i.e. from a distance of about 20 meters)? (A) none, (B) mild, (C) moderate, (D) severe, (E) extreme/cannot do

The Snellen Eye Chart Test:
Fixing DIF in Self-Assessments of Visual (Non)acuity

<table>
<thead>
<tr>
<th></th>
<th>Snellen Eye Chart</th>
<th>Ordinal Probit</th>
<th>Chopit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean (s.e.)</td>
<td>µ (s.e.)</td>
<td>µ (s.e.)</td>
</tr>
<tr>
<td>Slovakia</td>
<td>8.006 (.272)</td>
<td>.660 (.127)</td>
<td>.286 (.129)</td>
</tr>
<tr>
<td>China</td>
<td>10.780 (.148)</td>
<td>.673 (.073)</td>
<td>.749 (.081)</td>
</tr>
<tr>
<td>Difference</td>
<td>−2.774 (.452)</td>
<td>−.013 (.053)</td>
<td>−.463 (.053)</td>
</tr>
</tbody>
</table>

- The medical test shows Slovaksians see much better than the Chinese
- Ordinal probit finds no difference
- Chopit reproduces the same result as the medical test (though on different scale)
Conclusions

- Our approach can fix DIF, if response consistency and vignette equivalence hold — and the survey questions are good.

- Anchoring vignettes will not eliminate all DIF, but problems would have to occur at unrealistically extreme levels to make the unadjusted measures better than the adjusted ones.

- Expense can be held down to a minimum by assigning each vignette to a smaller subsample. E.g., 4 vignettes asked for 1/4 of the sample each adds only one question/respondent.

- Writing vignettes aids in the clarification and discovery of additional domains of the concept of interest — even if you do not do a survey.

- We do not provide a solution for other common survey problems: Question wording, Accurate translation, Question order, Sampling design, Interview length, Social backgrounds of interviewer and respondent, etc.
For More Information

http://GKing.Harvard.edu/vign

Includes:

- Academic papers
- Anchoring vignette examples by researchers in many fields,
- Frequently asked questions,
- Videos
- Conferences
- Statistical software
Anchoring Vignettes Measure DIF, not Vision: A Heuristic

Define $\mu$ as the quantity of interest; $D$ as DIF.

1. If model assumptions hold:
   - Self-assessments estimate: $(\mu + D)$.
   - Vignettes estimate: $D$ (they vary over $i$ only due to DIF)
   - Vignette-corrected self-assessments: $(\mu + D) - D = \mu$

2. If model assumptions do not hold:
   - Self-assessments estimate: $(\mu + D_s)$.
   - Vignettes estimate: $D_v$ (which may differ from $D_s$)
   - Vignette-corrected self-assessments: $(\mu + D_s) - D_v = \mu + (D_s - D_v)$
   - Which is larger?
     - (a) Self-assessment bias: $D_s$
     - (b) Vignette-corrected self-assessment bias: $(D_s - D_v)$
   - Since the same person generates both $D_s$ and $D_v$, (b) will usually be smaller.

3. Conclusion: Anchoring vignettes will usually help reduce bias. They will sometimes not make a difference. They will almost never exacerbate bias.
Self-Assessment Component: for \( i = 1, \ldots, n \)

- **Actual level:** \( \mu_i = X_i \beta + \eta_i \), with random effect \( \eta_i \sim N(0, \omega^2) \)
- **Perceived level:** \( Y_{i1}^* \sim N(\mu_i, 1) \) \( \ldots \) \( Y_{is}^* \sim N(\mu_i, 1) \)
- **Reported Level:**
  
  \[
  y_{i1} = k \quad \text{if} \quad \tau_{i1}^{k-1} \leq Y_{i1}^* < \tau_{i1}^k \\
  \vdots \\
  y_{is} = k \quad \text{if} \quad \tau_{is}^{k-1} \leq Y_{is}^* < \tau_{is}^k 
  \]

  where

  \[
  \tau_{is}^1 = \gamma_1 V_i \\
  \tau_{is}^k = \tau_{is}^{k-1} + e^{\gamma_k} V_i \quad (k = 2, \ldots, K_s)
  \]
Vignette Component: for $\ell = 1, \ldots, N$

- Actual level: $\theta_1, \ldots, \theta_J$
- Perceived level: $Z_{\ell1}^* \sim N(\theta_1, \sigma^2)$ \ldots $Z_{\ell J}^* \sim N(\theta_J, \sigma^2)$
- Reported Level: $z_{\ell j} = k$ if $\tau_{\ell 1}^{k-1} \leq Z_{\ell j}^* < \tau_{\ell 1}^k$

where

$$
\tau_{\ell s}^1 = \gamma_1 V_\ell \\
\tau_{\ell s}^k = \tau_{\ell s}^{k-1} + e^{\gamma_k V_\ell} \quad (k = 2, \ldots, K_s)
$$
The Likelihood Function: Self-Assessment Component

If $\eta_i$ were observed:

$$P(y_i|\eta_i) = \prod_{i=1}^{n} \prod_{s=1}^{S} \prod_{k=1}^{K_s} \left[ F(\tau_{is}^k|X_i\beta + \eta_i, 1) - F(\tau_{is}^{k-1}|X_i\beta + \eta_i, 1) \right]^{1(y_{is}=k)}$$

(S ordered probits with varying thresholds). Since $\eta_i$ is unobserved,

$$L_s(\beta, \omega^2, \gamma|y) \propto \prod_{i=1}^{n} \int_{-\infty}^{\infty} \prod_{s=1}^{S} \prod_{k=1}^{K_s} \left[ F(\tau_{is}^k|X_i\beta + \eta, 1) - F(\tau_{is}^{k-1}|X_i\beta + \eta, 1) \right]^{1(y_{is}=k)} N(\eta|0, \omega^2) d\eta$$

In the special case where $S = 1$, this simplifies to

$$L_s(\beta, \omega^2, \gamma|y) = \prod_{i=1}^{n} \prod_{k=1}^{K_1} \left[ F(\tau_{i1}^k|X_i\beta, 1 + \omega^2) - F(\tau_{i1}^{k-1}|X_i\beta, 1 + \omega^2) \right]^{1(y_{i1}=k)}$$
The vignette component is a $J$-variate ordinal probit with varying thresholds:

$$L_v(\theta, \sigma^2, \gamma|z) \propto \prod_{\ell=1}^{N} \prod_{j=1}^{J} \prod_{k=1}^{K_1} \left[ F(\tau_{\ell1}^k|\theta_j, 1) - F(\tau_{\ell1}^{k-1}|\theta_j, \sigma^2) \right] 1(z_{\ell j} = k)$$

The joint likelihood shares parameter $\gamma$:

$$L(\beta, \sigma^2, \omega^2, \theta, \gamma|y, z) = L_s(\beta, \sigma^2, \omega^2, \gamma|y) \times L_v(\theta, \gamma|z).$$

and nests the ordinal probit model as a special case.
## Fixing DIF in China and Mexico

<table>
<thead>
<tr>
<th>Eqn.</th>
<th>Variable</th>
<th>Ordinal Probit Coeff.</th>
<th>(s.e.)</th>
<th>Chopit Coeff.</th>
<th>(s.e.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu$</td>
<td>China</td>
<td>.670</td>
<td>(.081)</td>
<td>-.362</td>
<td>(.090)</td>
</tr>
<tr>
<td></td>
<td>age</td>
<td>.004</td>
<td>(.003)</td>
<td>.006</td>
<td>(.003)</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>.087</td>
<td>(.076)</td>
<td>.113</td>
<td>(.081)</td>
</tr>
<tr>
<td></td>
<td>education</td>
<td>.020</td>
<td>(.008)</td>
<td>.019</td>
<td>(.008)</td>
</tr>
<tr>
<td>Vignettes</td>
<td>$\theta_1$</td>
<td></td>
<td></td>
<td>1.393</td>
<td>(.190)</td>
</tr>
<tr>
<td></td>
<td>$\theta_2$</td>
<td></td>
<td></td>
<td>1.304</td>
<td>(.190)</td>
</tr>
<tr>
<td></td>
<td>$\theta_3$</td>
<td></td>
<td></td>
<td>.953</td>
<td>(.189)</td>
</tr>
<tr>
<td></td>
<td>$\theta_4$</td>
<td></td>
<td></td>
<td>.902</td>
<td>(.188)</td>
</tr>
<tr>
<td></td>
<td>$\theta_5$</td>
<td></td>
<td></td>
<td>.729</td>
<td>(.188)</td>
</tr>
<tr>
<td>ln $\sigma$</td>
<td></td>
<td></td>
<td></td>
<td>-.238</td>
<td>(.042)</td>
</tr>
</tbody>
</table>
The Source of DIF in China and Mexico: Threshold Variation

\[ \theta_1 \quad \theta_5 \]

\[ y^* \quad y^* \]

\[ \tau_1 \quad \tau_2 \quad \tau_3 \quad \tau_4 \]

\[ \mu \quad \bar{\mu} \]

\[ 0.5 \quad 0.3 \quad 0.1 \quad 1 \quad 2 \quad 3 \quad 3 \quad 2 \quad 1 \quad 0 \quad 0.2 \quad 0.4 \]

\[ \text{Political Efficacy} \]

\[ \text{Density} \]
1. Effect Parameters

The effect parameters $\beta$ are interpreted as in a linear regression of actual levels $\mu_i$ on $X_i$ and $\eta_i$.

2. Actual Levels, without a Self-Assessment

- Choose hypothetical values of the explanatory variables, $X_c$
- The posterior density of $\mu_c$ is similar to regression:

$$P(\mu_c | y) = N(\mu_c | X_c \hat{\beta}, X_c' \hat{V}(\hat{\beta}) X_c + \hat{\omega}^2)$$

- E.g., we can use the mean, $X_c \hat{\beta}$ as a point estimate of the actual level when $X = X_c$. 
1. If we know $y_i$, why not use it?

2. For example,
   - Suppose John and Esmeralda have the same $X$ values
   - By Method 1, they give the same inferences: $P(\mu_J|y) = P(\mu_E|y)$.
   - Suppose John’s $y_J$ value is near $\hat{\mu}_J$ and but Esmeralda’s is far away.
     - Under Method 1, nothing’s new. Predictions are unchanged.
     - Intuitively, John is average and Esmeralda is an outlier
     - We should adjust our prediction from $\hat{\mu}_E$ toward $y_E$.
   - So the new method takes roughly the weighted average of the model prediction $\hat{\mu}_E$ and the observed $y_E$, with weights determined by the how good a prediction it is.
More formally, we use Bayes theorem

\[ P(\mu_i | y, y_i) \propto P(y_i | \mu_i, y)P(\mu_i | y), \]

the likelihood with \( \eta_i \) observed times the Method 1 posterior:

\[
P(\mu_i | y, y_i) \propto \prod_{s=1}^{S} \prod_{k=1}^{K_s} \left[ F(\hat{\tau}_{is}^k | \mu_i, 1) - F(\hat{\tau}_{is}^{k-1} | \mu_i, 1) \right] 1(y_{is} = k) \times N(\mu_i | X_i \hat{\beta}, X_i \hat{V}(\hat{\beta})X_i' + \hat{\omega}^2)
\]

Key Difference:
- \( P(\mu_i | y) \) works for out-of-sample prediction
- \( P(\mu_i | y, y_i) \) works better when \( y_i \) is available
Unconditional posterior for a hypothetical 65-year-old respondent in country 1, based on one simulated data set.
Conditional posteriors for two different 21 year old respondents. Person 1 gave responses (1,1) on the two self-evaluation questions; Person 2 gave responses (4,3). The unconditional posterior, drawn with a dashed line, gives less specific predictions. Each curve was computed from one simulated data set.
Estimated Entropy

- Measures the informativeness of the vignettes,
- as supplemented by the predictive information in the covariates
- A reasonable approach, uses a modification of a standard statistical model, and robust to misspecification.
- But it assumes the probit specification is correct. Normally this is ok, but decisions here are more consequential since they affect data collection decisions and thus can preclude asking some questions
- Thus, we also want “known entropy”.

Gary King ()  Anchoring Vignettes
Scalar-valued $C_i$ observations are set to observed values.

Vector-valued $C_i$:

- Elements of all possible vector responses are parameterized: (e.g., $p_1, p_2, p_3$ for $C_i = \{2, 3, 4\}$)
- All mass is restricted to within the vector (e.g., $p_1 + p_2 + p_3 = 1$)
- Choose all $p$'s to minimize entropy (i.e., adjust the $p$'s to see how spiky the distribution can become)
- Some tricks make this easy with a genetic optimizer.

Then form the histogram (summing the $p$'s) and compute entropy.

We now compute *estimated entropy* and *known entropy* for all possible subsets of vignettes.
Robust Analysis via Conditional Model

- Condition on observed value of $c_i$:

$$\Pr(C = c|x_0, c_i) = \begin{cases} \frac{\Pr(C=c|x_0)}{\sum_{a \in c_i} \Pr(C=a|x_0)} & \text{for } c \in c_i \\ 0 & \text{otherwise} \end{cases}$$

- Advantages compared to unconditional probabilities:
  - Conditions on $c_i$ by normalizing the probability to sum to one within the set $c_i$ and zero outside that set.
  - For scalar values of $c_i$, this expression simply returns the observed category: $\Pr(C = c|x_i, c_i) = 1$ for category $c$ and 0 otherwise.
  - For vector valued $c_i$, it puts probability density over categories within $c_i$, which in total sum to one.
  - Probabilities can be interpreted for causal effects or summed to produce a histogram.
  - Result:
    - highly robust to model mispecification,
    - extracts considerably more information from anchoring vignette data.