Anchoring Vignettes for Interpersonally Incomparable Survey Responses

Gary King
Institute for Quantitative Social Science
Harvard University

(talk at Graduate Methods and Models Class, Harvard University, 9/18/09)

Readings

- Gary King and Jonathan Wand. "Comparing Incomparable Survey Responses: Evaluating and Selecting Anchoring Vignettes," *Political Analysis*, 15, 1 (Winter, 2007): 46–66.
- Gary King; Christopher J.L. Murray; Joshua A. Salomon; and Ajay Tandon. "Enhancing the Validity and Cross-cultural Comparability of Measurement in Survey Research," *American Political Science Review*, Vol. 98, No. 1 (February, 2004): 191–207.
- Papers, FAQ, examples, software, conferences, videos: http://GKing.Harvard.edu/vign

Two Problems

1. How to measure "big" concepts we can define only by example
   - E.g., freedom, political efficacy, pornography, health, etc.
   - The usual advice: You do not have a methodological problem. Get a theory and it will produce a more concrete question. [Go away!]
   - The result of more concreteness: more reliability, no more validity
2. How to ensure interpersonal and cross-population comparability
   - Chinese report having more political efficacy than Americans
   - The most common measure of health — "How healthy are you? (Excellent, Good, Fair, Poor)" — often correlates negatively with actual health
   - Amartya Sen (2002): "The state of Kerala has the highest levels of literacy. . . and longevity. . . in India. But it also has, by a very wide margin, the highest rate of reported morbidity among all Indian states. . . . At the other extreme, states with low longevity, with woeful medical and educational facilities, such as Bihar, have the lowest rates of reported morbidity in India."

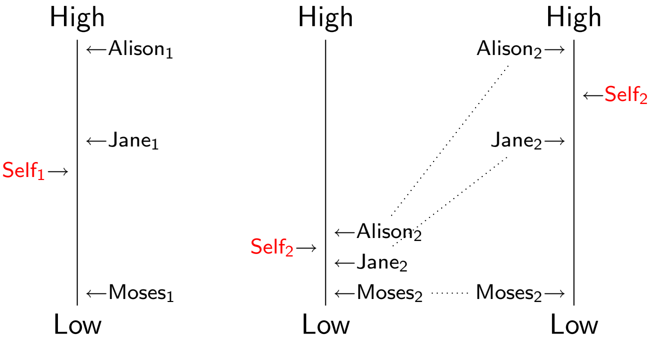
Anchoring Vignettes & Self-Assessments: Political Efficacy (about voting)

- "[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

Does R_1 or R_2 have More Political Efficacy?



- 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)


RUSSIA
KAZAKHSTAN
MONGOLIA
KYRGYZSTAN
NORTH KOREA
SOUTH KOREA
XINJIANG
NEI MONGOL
QINGHAI
XIZANG (TIBET)
SICHUAN
GANSU
HEILONGJIANG
JILIN
Beijing
Shanghai
Hong Kong
Taiwan
East China Sea
South China Sea
Yellow Sea
NEPAL
BHUTAN
INDIA
BANGLADESH
BURMA
LAOS
THAILAND
VIETNAM
PHILIPPINES
CAMBODIA
China
International boundary
Province-level boundary
National capital
Province-level capital
Railroad
Road
Autonomous regions and municipalities in italics.
500 Kilometers
500 Miles
Lambert Conformal Conic Projection, SP 23N/45N
Boundary representation is not necessarily authoritative.
Base 802714AI (R00152) 3-01


UNITED STATES
Gulf of Mexico
NORTH PACIFIC OCEAN
SONORA
CHIHUAHUA
BAJA CALIFORNIA
BAJA CALIFORNIA SUR
COAHUILA DE ZARAGOZA
NUEVO LEÓN
DURANGO
SINALOA
TAMAULIPAS
ZACATECAS
JALISCO
VERACRUZ-LLAVE
YUCATÁN
QUINTANA ROO
CAMPECHE
TABASCO
CHIAPAS
OAXACA
GUERRERO
BELIZE
GUATEMALA
HONDURAS
EL SALVADOR
Mexico
International boundary
State (estado) boundary
National capital
State (estado) capital
Railroad
Road
1. AGUASCALIENTES
2. DISTRITO FEDERAL
3. GUANAJUATO
4. HIDALGO
5. MÉXICO
6. MORELOS
7. QUERÉTARO DE ARTEAGA
8. TLAXCALA
Boundary representation is not necessarily authoritative.

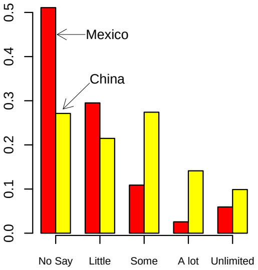
Mexico

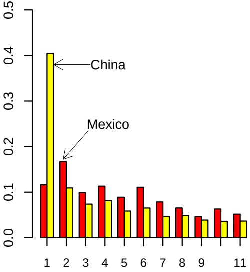
Opposition leader Vicente Fox elected President.
71-year rule of PRI party ends.
Peaceful transition of power begins.

Plenty of political efficacy

China: How much say do you have in getting the government to address issues that interest you?

Nonparametric Estimates of Political Efficacy





- 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*

Example	Survey Responses	1: $y < z_1$	2: $y = z_1$	3: $z_1 < y < z_2$	4: $y = z_2$	5: $y > z_2$	C
1	$y < z_1 < z_2$	T					$\{1\}$
2	$y = z_1 < z_2$		T				$\{2\}$
3	$z_1 < y < z_2$			T			$\{3\}$
4	$z_1 < y = z_2$				T		$\{4\}$
5	$z_1 < z_2 < y$					T	$\{5\}$
Ties:							
6	$y < z_1 = z_2$	T					$\{1\}$
7	$y = z_1 = z_2$		T		T		$\{2,3,4\}$
8	$z_1 = z_2 < y$					T	$\{5\}$
Inconsistencies:							
9	$y < z_2 < z_1$	T					$\{1\}$
10	$y = z_2 < z_1$	T			T		$\{1,2,3,4\}$
11	$z_2 < y < z_1$	T				T	$\{1,2,3,4,5\}$
12	$z_2 < y = z_1$		T			T	$\{2,3,4,5\}$
13	$z_2 < z_1 < y$					T	$\{5\}$

Analyzing the DIF-Free Variable: More Efficiencies

- How to analyze a variable with scalar and vector responses?
- Define an unobserved variable: $Y_i \sim \text{Normal}(x_i\beta, 1)$
- With observation mechanism, for scalar C, the same as ordered probit:

$$C_i = c \qquad \text{if } \tau_{c-1} \leq Y_i < \tau_c$$

- Probability of observing category c, for $X = x_0$:

$$\Pr(C = c|x_0) = \int_{\tau_{c-1}}^{\tau_c} \text{Normal}(y|x_0\beta, 1)dy$$

- Observation mechanism for vector valued C:

$$C_i = c \qquad \text{if } \tau_{\min(c)-1} \leq Y_i < \tau_{\max(c)}$$

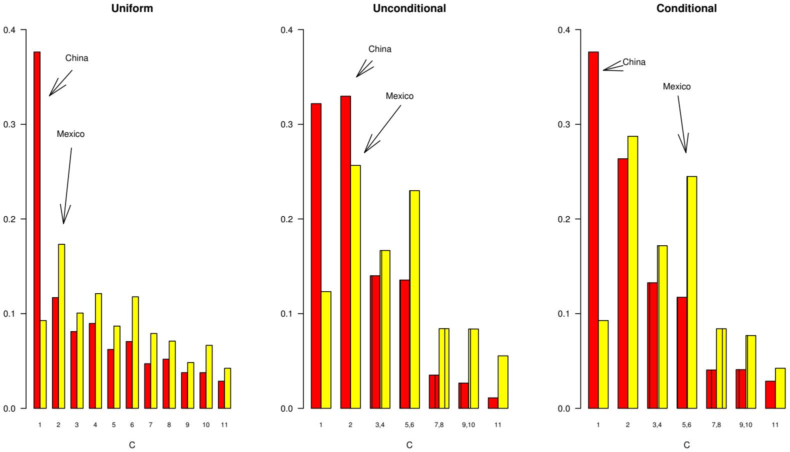
Robust Analysis via Conditional Model

- 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.

Improved Efficiency in Practice



Optimally Choosing Vignettes

- Ultimate Goal: Learn about a continuous unobserved variable (health, efficacy).
  - Observed: Proportions of the mass of the continuous variable (and hence observations) falling in each discrete category defined by the vignettes
  - Worst choice: All in one category; i.e., information = discriminatory power (E.g., "Bob ran two marathons last week. . ." does not discriminate among respondents)
  - Best choice: Largest number of categories, with mass of the unobserved variable spread uniformly over categories
- Immediate Goal: Measure information in a categorization scheme (defined by the choice of vignettes)
  - Formalization of the goal: Define a function $H(C)$ measuring information.
  - Operational use:
    - Run a pretest with lots of vignettes
    - Compute C and $H(C)$ for each possible subset,
    - Choose a subset for the main survey based on values of H and cost of survey questions.

Step 1: Criteria for Defining $H(C)$ for scalar C

Summarize C with a histogram, so $H(C) = H(p_1, \ldots, p_{2J+1})$. Add 3 criteria:

1. $H(0, 1, 0, 0, 0) = 0$, i.e., when all mass is in (any) one category and at a maximum when $p_1 = p_2 = \cdots = p_{2J+1}$
2. H is a monotonically increasing function of the number of vignettes J (and hence $2J + 1$, the number categories of C).
3. Assume consistent decomposition:
   - With one vignette, C has 3 categories (below, equal to, or above the vignette) and proportions $p_1 + p_2 + q = 1$
   - Add a new vignette and we can decompose the "above" category (into between the two vignettes, equal to the second, or above the second).
   - We now have 5 categories, with proportions $p_1 + p_2 + p_3 + p_4 + p_5 = 1$.
   - The information in the union of the smaller bins (3,4,5) should equal that in the original undecomposed bin since $q = p_3 + p_4 + p_5$.
   - The information in the unaffected bins (1, 2) should remain the same with the addition of the new vignette.
   - More formally: $H(p_1, p_2, p_3, p_4, p_5) = H(p_1, p_2, q) + qH(p_3, p_4, p_5)$

What Satisfies the Criteria for $H(C)$?

- Lots of candidates exist: Gini index, variance, absolute deviations 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 vector valued, and thus the p's are unknown?

Step 2: Defining $H(C)$ for scalar and vector C

- Without ties or inconsistencies, we simply compute entropy
- With ties and inconsistencies, we somehow estimate p's and then compute entropy, H.
- Rules for estimating the p's, and thus types of entropy:
  - Estimated entropy: using the multiple response ordered probit model
  - Known (minimum) entropy: information in the data we know exists for certain.

Estimated Entropy

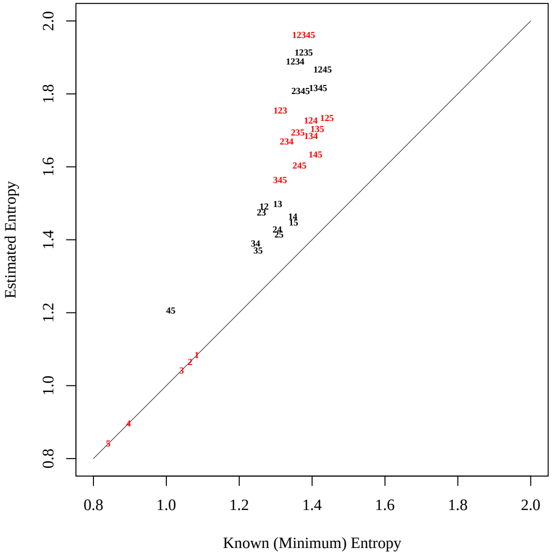
- Measures the informativeness of the vignettes,
- as supplemented by the predictive information in the covariates
- A reasonable approach, uses a modification of a standard statistical model, and robust to misspecification.
- *But* it assumes the probit specification is correct. Normally this is ok, 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".

Computing Known Entropy (no assumptions required)

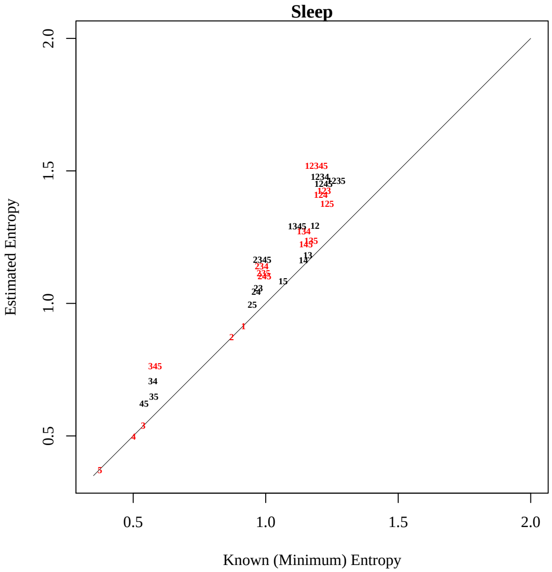
- Scalar-valued C_i observations are set to observed values.
- Vector-valued C_i:
  - Elements of all possible vector responses are parameterized: (e.g., p_1, p_2, p_3 for $C_i = \{2, 3, 4\}$)
  - All mass is restricted to within the vector (e.g., $p_1 + p_2 + p_3 = 1$)
  - Choose all p's to minimize 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.

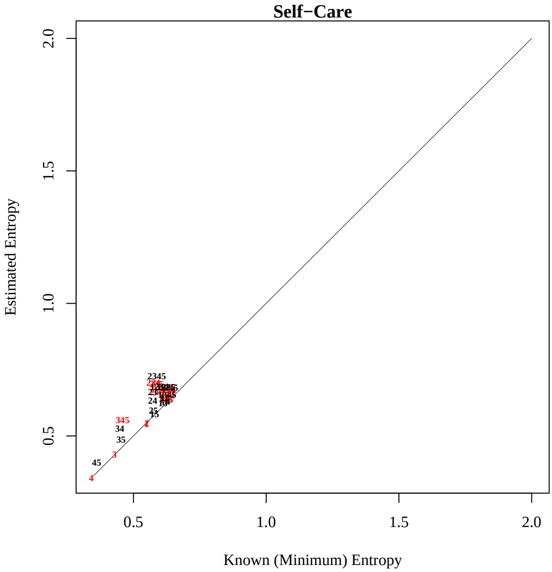
Political Efficacy (Mex & China)



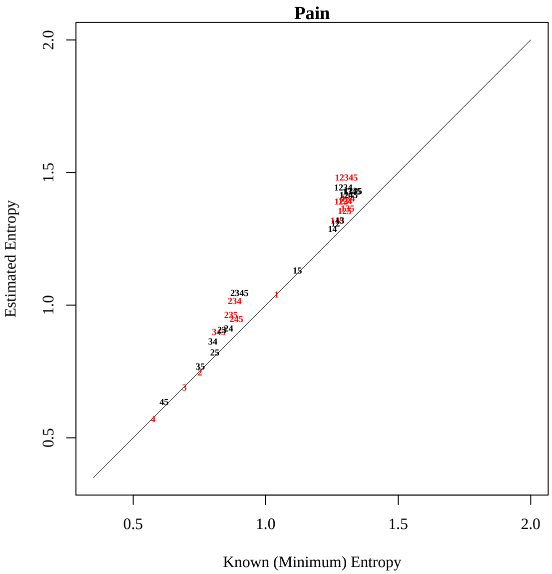
One vignette can be better than three: Sleep (China)



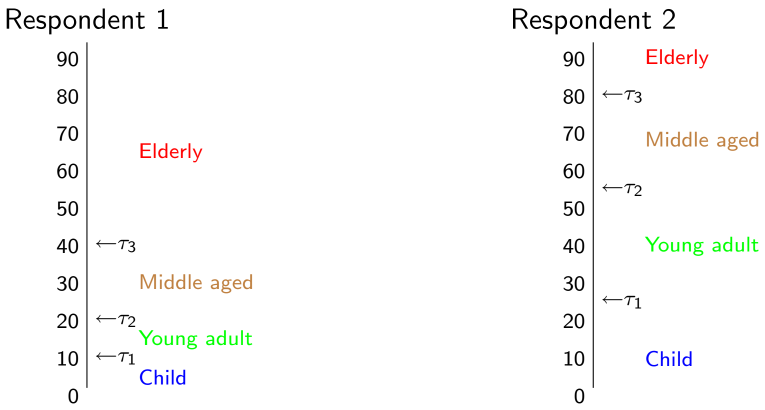
Some vignette sets are uninformative: Self-Care (China)



Some covariates are unhelpful: Pain (China)

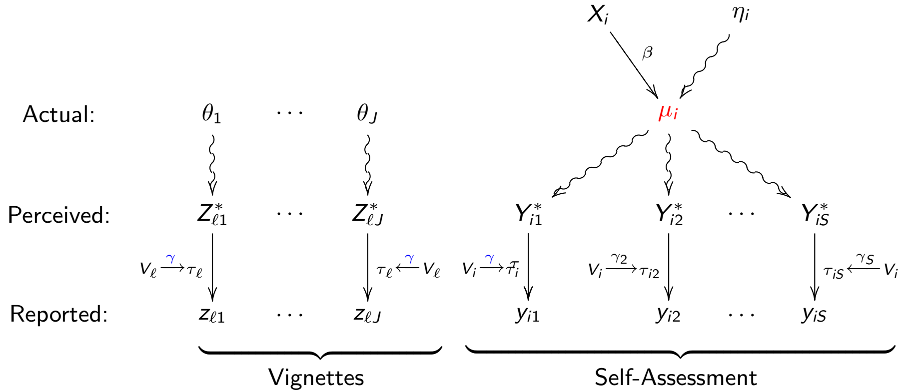


Categorizing Years of Age



- If thresholds vary, categorical answers are meaningless.
- Our parametric model works by estimating the thresholds.
- Vignettes provide identifying information for the τ's.

Model Summary

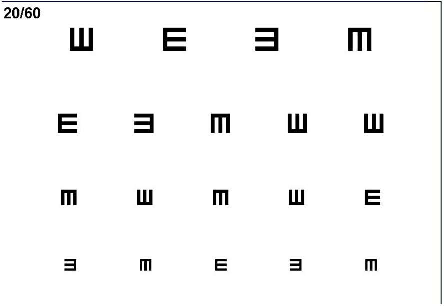


An ordinal probit model.
with varying thresholds,
a vignette for identification,
more vignettes for better discrimination,
optional multiple self-assessment questions,
and an optional random effect.

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

	Snellen Eye Chart		Ordinal Probit		Chopit	
	Mean	(s.e.)	μ	(s.e.)	μ	(s.e.)
Slovakia	8.006	(.272)	.660	(.127)	.286	(.129)
China	10.780	(.148)	.673	(.073)	.749	(.081)
Difference	-2.774	(.452)	$-.013$	(.053)	$-.463$	(.053)

- The medical test shows Slovakians 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.
- If you think you have DIF-free questions, you now have the first real opportunity to test that hypothesis.
- Whether or not you have DIF, vignettes can help us follow the usual survey advice of making questions concrete. (Compare "say in government" with that question plus the vignettes)
- 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 μ 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 $i = 1, \dots, n$

- Actual level: $\mu_i = X_i\beta + \eta_i$, with random effect $\eta_i \sim N(0, \omega^2)$
- Perceived level: $Y^*_{i1} \sim N(\mu_i, 1) \quad \dots \quad Y^*_{iS} \sim N(\mu_i, 1)$
- Reported Level:

$$
\begin{aligned}
y_{i1} &= k \quad \text{if } \tau_{i1}^{k-1} \le Y^*_{i1} < \tau_{i1}^{k} \\
&\vdots \\
y_{iS} &= k \quad \text{if } \tau_{is}^{k-1} \le Y^*_{is} < \tau_{is}^{k}
\end{aligned}
$$

where

$$
\begin{aligned}
\tau_{is}^{1} &= \gamma_1 V_i \\
\tau_{is}^{k} &= \tau_{is}^{k-1} + e^{\gamma_k V_i} \qquad (k = 2, \dots, K_s)
\end{aligned}
$$

Vignette Component: for $\ell = 1, \ldots, N$

- Actual level: $\theta_1, \ldots, \theta_J$
- Perceived level: $Z^*_{\ell 1} \sim N(\theta_1, \sigma^2) \quad \ldots \quad Z^*_{\ell J} \sim N(\theta_J, \sigma^2)$
- Reported Level: $z_{\ell j} = k \quad$ if $\tau^{k-1}_{\ell 1} \leq Z^*_{\ell j} < \tau^k_{\ell 1}$
  where

$$\begin{aligned} \tau^1_{\ell s} &= \gamma_1 V_\ell \\ \tau^k_{\ell s} &= \tau^{k-1}_{\ell s} + e^{\gamma_k V_\ell} \quad (k = 2, \ldots, K_s) \end{aligned}$$

The Likelihood Function: Self-Assessment Component

If η_i were observed:

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

(S ordered probits with varying thresholds). Since η_i is unobserved,

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

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

$$L_s(\beta,\omega^2,\gamma|y) = \prod_{i=1}^{n}\prod_{k=1}^{K_1}\left[F(\tau_{i1}^{k}|X_i\beta,\,1+\omega^2) - F(\tau_{i1}^{k-1}|X_i\beta,\,1+\omega^2)\right]^{\mathbf{1}(y_{i1}=k)}$$

The Likelihood Function: Adding the Vignette Component

The *vignette component* is a J-variate ordinal probit with varying thresholds:

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

The *joint likelihood* shares parameter γ:

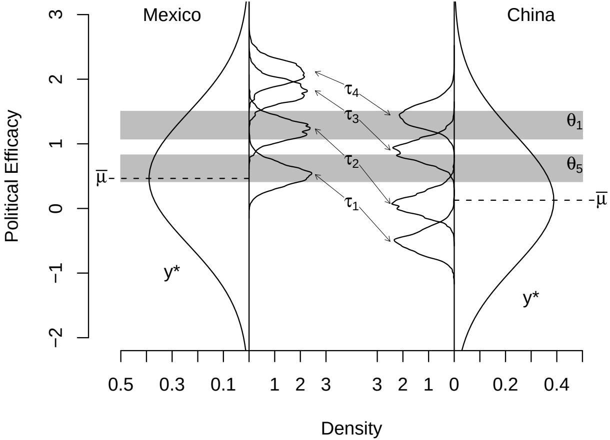
$$L(\beta, \sigma^2, \omega^2, \theta, \gamma|y, z) = L_s(\beta, \sigma^2, \omega^2, \gamma|y) \times L_v(\theta, \gamma|z).$$

and nests the ordinal probit model as a special case.

Fixing DIF in China and Mexico

		Ordinal Probit		Chopit	
Eqn.	Variable	Coeff.	(s.e.)	Coeff.	(s.e.)
μ	China	.670	(.081)	$-.362$	(.090)
	age	.004	(.003)	.006	(.003)
	male	.087	(.076)	.113	(.081)
	education	.020	(.008)	.019	(.008)
Vignettes	θ_1			1.393	(.190)
	θ_2			1.304	(.190)
	θ_3			.953	(.189)
	θ_4			.902	(.188)
	θ_5			.729	(.188)
	$\ln \sigma$			$-.238$	(.042)

The Source of DIF in China and Mexico: Threshold Variation



Computing Quantities of Interest

1. Effect Parameters
   The effect parameters β are interpreted as in a linear regression of actual levels μ_i on X_i and η_i.
2. Actual Levels, without a Self-Assessment
   - Choose hypothetical values of the explanatory variables, X_c
   - The posterior density of μ_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 $X = X_c$.

Estimating Actual Levels, with a Self-Assessment

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: $\mathrm{P}(\mu_J|y) = \mathrm{P}(\mu_E|y)$.
   - 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

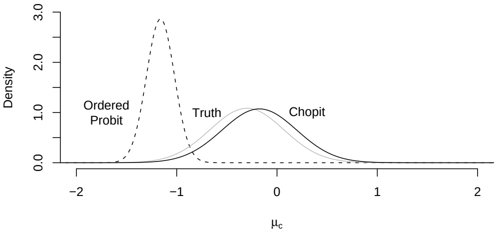
$$\mathrm{P}(\mu_i|y, y_i) \propto \mathrm{P}(y_i|\mu_i, y)\mathrm{P}(\mu_i|y),$$

the likelihood with η_i observed times the Method 1 posterior:

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

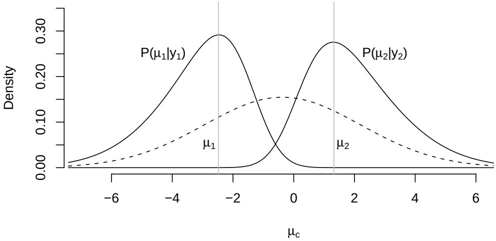
Key Difference:	$\mathrm{P}(\mu_i\|y)$	works for out-of-sample prediction
	$\mathrm{P}(\mu_i\|y, y_i)$	works better when y_i is available

Unconditional Posterior



Unconditional posterior for a hypothetical 65-year-old respondent in country 1, based on one simulated data set.

Conditional Posteriors



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.