Learning More from Conjoint Experiments through a Doubly Randomized Design

Jens Hainmueller Daniel J. Hopkins Teppei Yamamoto

Stanford University University of Pennsylvania MIT

International Methods Colloquium, October 23, 2015

Ben Carson vs. Donald Trump

What is Conjoint Analysis?

- Multi-dimensional choice: central problem in social science
- Conjoint analysis: experimental technique for analyzing preferences/choices about multi-dimensional objects
- Introduced in 1971 (Green and Rao 1971); “factorial surveys” in sociology (Rossi et al. 1974)
- Widely used in marketing/business (e.g. Courtyard Marriott)

What is Conjoint Analysis?

- Multi-dimensional choice: central problem in social science
- Conjoint analysis: experimental technique for analyzing preferences/choices about multi-dimensional objects
- Introduced in 1971 (Green and Rao 1971); "factorial surveys" in sociology (Rossi et al. 1974)
- Widely used in marketing/business (e.g. Courtyard Marriott)

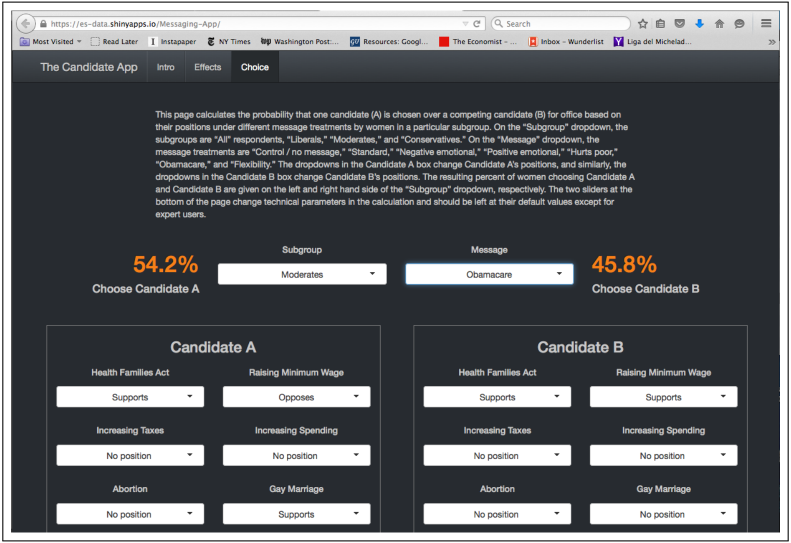
- Respondents read profiles which vary across multiple attributes
- Respondents rank or rate profiles
- Repeat the exercise

Growing Use of Conjoint Designs

- $\uparrow$ computing power $\rightarrow$ increasing use of conjoint experiments in public opinion research

  Recent examples: immigration attitudes, housing preferences, support for climate change agreements, bailout agreements, choices among media outlets, vote choice in various contexts (Loewen et al. 2012; Bechtel and Scheve 2013; Nall and Mummolo 2013; Franchino and Zucchini 2014; Wright et al. 2014; Abrajano et al. 2015; Carlson 2015; Carnes and Lupu 2015; Crowder-Meyer et al. 2015; Goggin et al. 2015; Hankinson 2015; Sen 2015; Teele et al. 2015)

Conjoint designs in practice



Quantities of Interest in Conjoint Analysis

- Treatments are *composites* of multiple causal components of interest
- Researchers want to identify:
  1. average marginal effect of each component (AMCE)
  2. how marginal effects compare with one another
  3. whether the components interact with one another
  4. whether respondent characteristics moderate the marginal effects
- Standard survey experiments cannot identify these quantities

Advantages of Conjoint Analysis

- Can identify relative weights among multiple attributes
- Cost-effective: various quantities of interest via single experiment
- Test competing hypotheses
- Decompose composite treatment effects (Hainmueller, Hopkins & Yamamoto 2014)
- Can mirror real-world choice contexts $\longrightarrow$ better external validity
    - e.g. Hainmueller, Hangartner & Yamamoto (2015) recover determinants of voting in Swiss citizenship elections at least a decade after they ended
- May suppress social desirability bias
- Can estimate Average Marginal Component Effects (AMCEs) via regression (difference-in-means estimators)

Open Questions for Conjoint Analysis

- We do not know much about optimal design for conjoint studies
- Choices when implementing a conjoint experiment:
  - Which/how many attributes per profile?
  - How many profiles per task?
  - How many tasks per respondent?
  - Forced choice or rating?
- Our research agenda: how do these choices affect the inferences we draw from conjoint designs? Which choices are optimal?

Open Questions for Conjoint Analysis

- We do not know much about optimal design for conjoint studies
- Choices when implementing a conjoint experiment:
  - Which/how many attributes per profile?
  - How many profiles per task?
  - How many tasks per respondent?
  - Forced choice or rating?
- Our research agenda: how do these choices affect the inferences we draw from conjoint designs? Which choices are optimal?
- In this paper we focus on one key question: How many attributes should a conjoint profile include?
- Results → may yield insights about online survey administration generally

The Masking-Satisficing Tradeoff

Stanford

Please carefully review the two candidates for President detailed below.

Which of these two candidates would you prefer to see as President of the United States?

	Candidate A	Candidate B
Highest education	graduated from high school	graduated from college
Largest campaign contributor	auto workers' unions	wall street firms
State of residence	Alabama	Ohio
Party affiliation	Republican	Republican
Your Choice:	○	○

NEXT

The Masking-Satisficing Tradeoff

Stanford

Please carefully review the two candidates for President detailed below.

Which of these two candidates would you prefer to see as President of the United States?

	Candidate A	Candidate B
Highest education	graduated from high school	graduated from college
Largest campaign contributor	auto workers' unions	wall street firms
State of residence	Alabama	Ohio
Annual income	$75k	$32k
Race/Ethnicity	Asian American	Black
Profession	lawyer	farmer
Car	Ford pick-up truck	Toyota Sedan
Favorite professional sport	football	basketball
Military service	served in U.S. military	served in U.S. military
Age	72	63
Marital status	single	single

Marital status	single	single
Position on abortion	pro-life	neutral
Position on health care	government should do more	government should do more
Religion	Evangelical Protestant	Evangelical Protestant
Prior elected office	state attorney general	state attorney general
Favorite music	hip hop	country
Religious activity	occasionally attends church	attends church weekly
Gender	female	female
Position on gay marriage	opposes gay marriage	favors gay marriage
Party affiliation	Republican	Republican
Your Choice:	○	○

NEXT

Masking

- One key problem when including **too few** attributes is masking
  - Respondents use one attribute because of perceived correlation with unobserved attribute (see also Dafoe et al. 2015)
  - E.g. If conjoint table does not include issue positions, the effect of partisanship $\rightarrow$ inflated
- Masking does not invalidate the identification of the AMCE, but muddies interpretation (effect of party or inferred issue position?)
- Expect masking to decline as more correlated attributes are included

Survey Satisficing

- However, including **too many** attributes causes excessive survey satisficing

Survey Satisficing

- However, including **too many** attributes causes excessive survey satisficing
- Limits of working memory (Miller 1994)
- The more attributes you add, the more difficult the task becomes $\rightarrow$ respondents might stop paying attention or change cognitive strategies
- Threat is especially pronounced in conjoint experiments, which often ask respondents to sort through many separate pieces of information in a single task
- Expect satisficing to rise as we include more attributes and the conjoint task becomes more challenging (Krosnick 1999)
- Task difficulty also depends on subject matter

Notation

- Respondent: $i \in \{1, ..., N\}$
- Attribute: $l \in \{1, ..., L\}$, each discrete with D_l levels

Notation

- Respondent: $i \in \{1, ..., N\}$
- Attribute: $l \in \{1, ..., L\}$, each discrete with D_l levels
- Assume a single-task, single-profile conjoint for simplicity (result fully generalizable for multi-task, multi-profile designs)
- Observed stated preferences: Y_i
  - Can be any real-valued random variable
  - Typically a binary choice ($Y_i \in \{0, 1\}$)

Doubly Randomized Design

Suppose that...

- Researcher is interested in all L attributes, but unwilling to include them all at once because of expected excessive satisficing
- Researcher believes satisficing will be tolerable if $L^* < L$ attributes are included

Doubly Randomized Design

Suppose that...

- Researcher is interested in all L attributes, but unwilling to include them all at once because of expected excessive satisficing
- Researcher believes satisficing will be tolerable if $L^* < L$ attributes are included

The doubly randomized design can then be defined as:

Definition (Doubly Randomized Design (DRD))

1. For each respondent i, randomly choose L^* out of the population of L attributes.
2. For each attribute l in the set of L^* chosen attributes, randomly sample a level out of the D_l possible levels.

Doubly Randomized Design

Suppose that...

- Researcher is interested in all L attributes, but unwilling to include them all at once because of expected excessive satisficing
- Researcher believes satisficing will be tolerable if $L^* < L$ attributes are included

The doubly randomized design can then be defined as:

Definition (Doubly Randomized Design (DRD))

1. For each respondent i, randomly choose L^* out of the population of L attributes.
2. For each attribute l in the set of L^* chosen attributes, randomly sample a level out of the D_l possible levels.

- Profile (or treatment) vector under DRD: $\boldsymbol{T}_i = [T_{i1}, ..., T_{iL}]^\top$ where $T_{il} \in \{0, 1, ..., D_l\}$
- $0 =$ *missingness* of the attribute due to the first-stage randomization

Satisficing and Potential Outcomes

Formalize satisficing using the potential outcomes framework:

Assumption (Satisficing)

$$Y_i = \begin{cases} Y_i(\boldsymbol{T}_i) & \text{if } \sum_{l=1}^{L} \boldsymbol{I}\{T_{il} = 0\} \geq L - L^*, \\ Y_i^*(\boldsymbol{T}_i) & \text{if } T_{il} > 0 \text{ for all } l \in \{1, ..., L\}, \end{cases}$$

where $\boldsymbol{I}\{\cdot\}$ represents the indicator function.

- $Y_i(\boldsymbol{t})$: "true" potential outcomes, indicating response to profile $\boldsymbol{t}$ without excessive satisficing
- $Y_i^*(\boldsymbol{t})$: potential outcomes "contaminated" by satisficing
- Assumption says potential outcomes of interest are observable under DRD, but not when all L attributes are shown (called the global design, GD).

Idealized Average Marginal Component Effect (IAMCE)

- Researcher may seek to identify the idealized average marginal component effect (IAMCE):

$$\pi^I(p) \equiv \sum_{t_2 \in \{1,\ldots,D_2\}} \cdots \sum_{t_L \in \{1,\ldots,D_L\}} \mathbb{E}[Y_i(2, t_2, ..., t_L) - Y_i(1, t_2, ..., t_L)] \times p(T_{i2} = t_2, ..., T_{iL} = t_L)$$

- Average marginal effect of the attribute of interest (T_1) when all attributes are shown and yet no excessive satisficing

Idealized Average Marginal Component Effect (IAMCE)

- Researcher may seek to identify the idealized average marginal component effect (IAMCE):

$$\pi^I(p) \equiv \sum_{t_2 \in \{1,\ldots,D_2\}} \cdots \sum_{t_L \in \{1,\ldots,D_L\}} \mathbb{E}[Y_i(2, t_2, ..., t_L) - Y_i(1, t_2, ..., t_L)] \times p(T_{i2} = t_2, ..., T_{iL} = t_L)$$

- Average marginal effect of the attribute of interest (T_1) when all attributes are shown and yet no excessive satisficing
- The global design cannot identify IAMCE because of satisficing; it instead identifies the naïve AMCE:

$$\begin{aligned}
\pi^N(p) &\equiv \sum_{t_2 \in \{1,\ldots,D_2\}} \cdots \sum_{t_L \in \{1,\ldots,D_L\}} \mathbb{E}[Y_i^*(2, t_2, ..., t_L) - Y_i^*(1, t_2, ..., t_L)] \times p(T_{i2} = t_2, ..., T_{iL} = t_L) \\
&= \sum_{t_2 \in \{1,\ldots,D_2\}} \cdots \sum_{t_L \in \{1,\ldots,D_L\}} \{\mathbb{E}[Y_i \mid T_1 = 2, T_2 = t_2, ..., T_L = t_L] \\
&\qquad - \mathbb{E}[Y_i \mid T_1 = 1, T_2 = t_2, ..., T_L = t_L]\} \times p(T_{i2} = t_2, ..., T_{iL} = t_L) \\
&\neq \pi^I(p)
\end{aligned}$$

Feasible Average Marginal Component Effect (FAMCE)

- What about DRD? It identifies the feasible AMCE (FAMCE):

$$
\begin{aligned}
\pi^F(p) &= \sum_{(t_2,\ldots,t_L)\in\mathcal{T}^F} \mathbb{E}[Y_i(2,t_2,\ldots,t_L) - Y_i(1,t_2,\ldots,t_L)] \\
&\qquad\qquad \times p(T_{i2}=t_2,\ldots,T_{iL}=t_L) \\
&= \sum_{(t_2,\ldots,t_L)\in\mathcal{T}^F} \{\mathbb{E}[Y_i \mid T_1=2, T_2=t_2,\ldots,T_L=t_L] \\
&\qquad - \mathbb{E}[Y_i \mid T_1=1, T_2=t_2,\ldots,T_L=t_L]\} \\
&\qquad\qquad \times p(T_{i2}=t_2,\ldots,T_{iL}=t_L),
\end{aligned}
$$

where $\mathcal{T}^F = \left\{(t_2,\ldots,t_L) : \sum_{l=2}^{L} \boldsymbol{I}\{t_l=0\} = L - L^*\right\}$

- Correct potential outcomes, but wrong range of summation
- In fact, treatment assignment distribution has *no common support* between DRD and GD
- Identification of IAMCE is impossible without assumptions that allow extrapolation

Masking and the Substitution Assumption

One possible assumption is the substitution assumption:

Assumption (Substitution)

For any vector of attributes $[t_2, ..., t_L]$ *and its possible permutations, it is assumed:*

$$Y_i(t, t_2, ..., t_{L^*}, 0, ..., 0) = \sum_{t_{L^*+1} \in \{1,...,D_{L^*+1}\}} \cdots \sum_{t_L \in \{1,...,D_L\}} Y_i(t, t_2, ..., t_L) f_i(t_{L^*+1}, ..., t_L \mid t, t_2, ..., t_{L^*}),$$

where $0 \leq f_i(t_{L^*+1}, ..., t_L \mid t, t_2, ..., t_{L^*})$ *and*
$\sum_{t_{L^*+1} \in \{1,...,D_{L^*+1}\}} \cdots \sum_{t_L \in \{1,...,D_L\}} f_i(t_{L^*+1}, ..., t_L \mid t, t_2, ..., t_{L^*}) = 1,\ \forall\ t \in \{1, ..., D_1\}$.

- Assumption: potential outcome under DRD = weighted average of true potential outcomes with missing attributes substituted
- Weights (f_i) can be interpreted as perceived (subjective) conditional probability of missing attribute levels given observed attribute levels
- e.g. Respondent sees a Democratic candidate with no info about policy position given, so she makes a guess and responds based on the guess
- Formalizes a psychological mechanism that generates masking

Unidentifiability of IAMCE under DRD

- Does the substitution assumption allow any inference about IAMCE based on FAMCE?

Unidentifiability of IAMCE under DRD

- Does the substitution assumption allow any inference about IAMCE based on FAMCE?
- Result: Excluding knife-edge scenarios, IAMCE = FAMCE if

$$f_i(t_{L^*+1}, ..., t_L \mid 2, t_2, ..., t_{L^*}) \ = \ f_i(t_{L^*+1}, ..., t_L \mid 1, t_2, ..., t_{L^*}),$$

  for all indices $[2, ..., L]$ and their possible permutations
- This is when *there is no masking* for the attribute of interest!

Unidentifiability of IAMCE under DRD

- Does the substitution assumption allow any inference about IAMCE based on FAMCE?
- Result: Excluding knife-edge scenarios, IAMCE = FAMCE if

$$f_i(t_{L^*+1}, ..., t_L \mid 2, t_2, ..., t_{L^*}) \;=\; f_i(t_{L^*+1}, ..., t_L \mid 1, t_2, ..., t_{L^*}),$$

  for all indices $[2, ..., L]$ and their possible permutations
- This is when *there is no masking* for the attribute of interest!
- More generally, the substitution assumption does **not** identify IAMCE under DRD

Summary: Analytical results

Analytical results:

- The overall average effect is a variant of the average marginal component effect (AMCE) and identified under DRD
- We call this quantity the feasible AMCE (FAMCE)
- Conditional and interactive FAMCEs can also be identified under DRD or TRD
- Estimation and inference are easily done via regression and difference-in-differences

What Questions Can DRD Answer?

- What is the overall average effect of party affiliation when respondents see various other subsets of candidate attributes?
- Robustness check: Is the effect of party affiliation stable even when other, correlated attributes are also presented?
- Effect decomposition: How much of the party effect is "explained away" by candidates' policy positions or personal attributes?

What Questions Can DRD *Not* Answer?

- What would the effect of party affiliation be if respondents saw *all* other candidate attributes at the same time?

What Questions Can DRD *Not* Answer?

- What would the effect of party affiliation be if respondents saw *all* other candidate attributes at the same time?

- Note this is a purely counterfactual question because of satisficing
- Maybe of theoretical interest, but empirically empty!

What Questions Can DRD *Not* Answer?

- What would the effect of party affiliation be if respondents saw *all* other candidate attributes at the same time?

- Note this is a purely counterfactual question because of satisficing
- Maybe of theoretical interest, but empirically empty!

- Result: The "idealized AMCE" cannot be identified unless we assume 1) no masking or 2) biases from masking cancel out exactly

What Questions Can DRD *Not* Answer?

- What would the effect of party affiliation be if respondents saw *all* other candidate attributes at the same time?

- Note this is a purely counterfactual question because of satisficing
- Maybe of theoretical interest, but empirically empty!

- Result: The "idealized AMCE" cannot be identified unless we assume 1) no masking or 2) biases from masking cancel out exactly

- Proposal: Focus on FACME, which is feasible, empirically meaningful and potentially more interesting

Data Collection

- Six conjoint experiments implemented on MTurk and SSI between February - May 2015.
- Focus here: MTurk, SSI experiments in May 2015 ($n \approx 1,600$)
- Running example: presidential vote choice
  - "This study is about voting and about your views on potential candidates for President. We are going to present pairs of hypothetical presidential candidates in the United States. For each pair, please indicate which of the two candidates you would prefer to see as President."
- Include 20 attributes that could define U.S. presidential candidates: education, income, partisanship, issue positions, favorite professional sport, etc.

Empirical Evidence: Experimental Design

- Respondents randomly assigned to see 4, 5, 6, 7, 8, 9, 10, 15 or all of the 20 attributes (i.e. TRD with 8 numbers of attributes)
- Goals:
  - Illustrate the proposed conjoint design
  - Examine tradeoff between masking and satisficing empirically

Empirical Evidence: Experimental Design

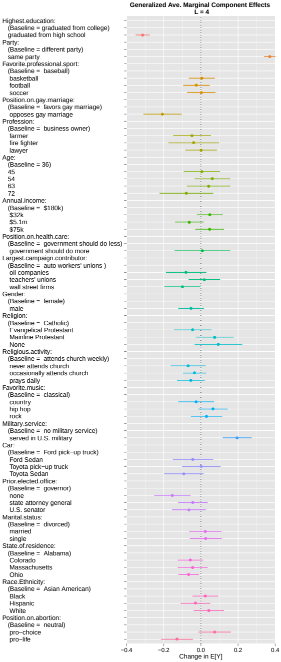
- Respondents randomly assigned to see 4, 5, 6, 7, 8, 9, 10, 15 or all of the 20 attributes (i.e. TRD with 8 numbers of attributes)
- Goals:
  - Illustrate the proposed conjoint design
  - Examine tradeoff between masking and satisficing empirically
- All respondents complete 30 choice tasks
  - How do response patterns change as the number of tasks increases? (for another paper)

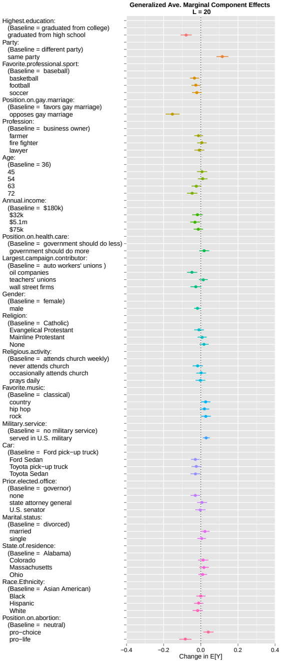
Empirical Evidence: Experimental Design

- Respondents randomly assigned to see 4, 5, 6, 7, 8, 9, 10, 15 or all of the 20 attributes (i.e. TRD with 8 numbers of attributes)
- Goals:
  - Illustrate the proposed conjoint design
  - Examine tradeoff between masking and satisficing empirically
- All respondents complete 30 choice tasks
  - How do response patterns change as the number of tasks increases? (for another paper)
- Attributes shown are randomly selected from the pool for each respondent
- Partisanship and education (our attributes of interest) always included to maximize power

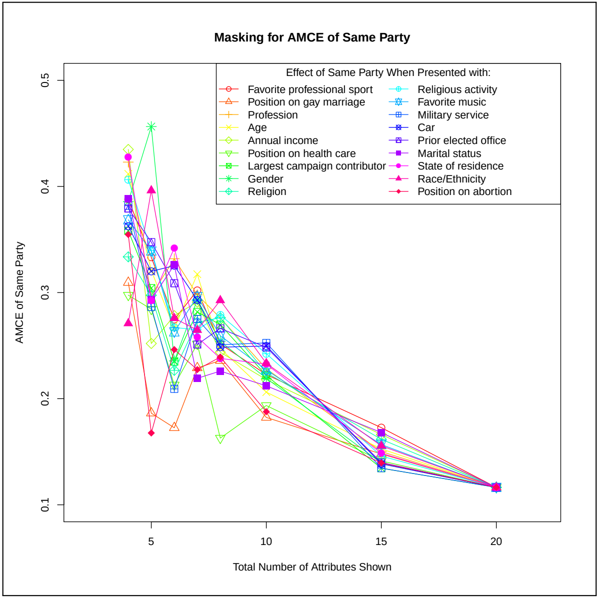
Candidate Attributes for Conjoint

Attribute	Levels
Age	36, 45, 54, 63, 72
Gender	male, female
Race/Ethnicity	Hispanic, White, Black, Asian American
Religion	Evangelical Protestant, Mainline Protestant, Catholic, None
Religious activity	prays daily, attends church weekly, occasionally attends church, never attends church
Military service	served in U.S. military, no military service
Profession	lawyer, business owner, farmer, fire fighter
Annual income	\$32k, \$75k, \$180k, \$5.1m
State of residence	Massachusetts, Ohio, Colorado, Alabama
Prior elected office	governor, U.S. senator, state attorney general, none
Car	Ford pick-up truck, Toyota pick-up truck, Ford Sedan, Toyota Sedan
Favorite music	country, rock, hip hop, classical
Favorite professional sport	baseball, football, basketball, soccer
Marital status	single, married, divorced
Position on health care	government should do more, government should do less
Position on abortion	pro-choice, pro-life, neutral
Position on gay marriage	favors gay marriage, opposes gay marriage
Largest campaign contributor	oil companies, teachers' unions, wall street firms, auto workers' unions
Party affiliation	Republican, Democrat
Highest education	graduated from high school, graduated from college

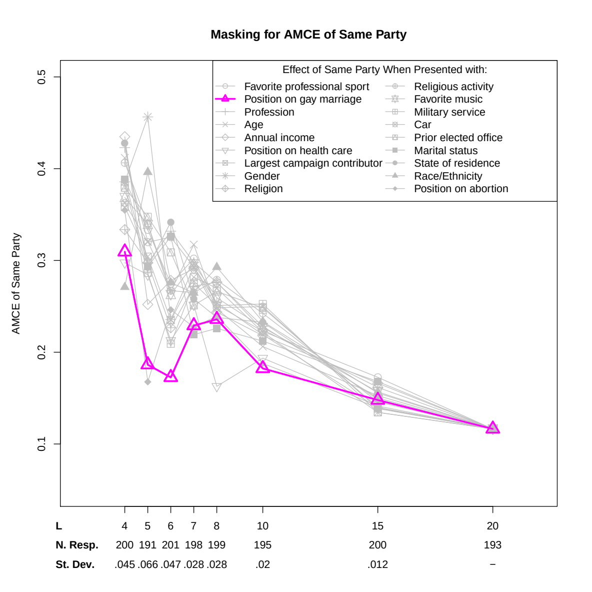

Generalized Ave. Marginal Component Effects
L = 4
Highest.education:
(Baseline = graduated from college)
graduated from high school
Party:
(Baseline = different party)
same party
Favorite.professional.sport:
(Baseline = baseball)
basketball
football
soccer
Position.on.gay.marriage:
(Baseline = favors gay marriage)
opposes gay marriage
Profession:
(Baseline = business owner)
farmer
fire fighter
lawyer
Age:
(Baseline = 36)
45
54
63
72
Annual.income:
(Baseline = $180k)
$32k
$5.1m
$75k
Position.on.health.care:
(Baseline = government should do less)
government should do more
Largest.campaign.contributor:
(Baseline = auto workers' unions)
oil companies
teachers' unions
wall street firms
Gender:
(Baseline = female)
male
Religion:
(Baseline = Catholic)
Evangelical Protestant
Mainline Protestant
None
Religious.activity:
(Baseline = attends church weekly)
never attends church
occasionally attends church
prays daily
Favorite.music:
(Baseline = classical)
country
hip hop
rock
Military.service:
(Baseline = no military service)
served in U.S. military
Car:
(Baseline = Ford pick−up truck)
Ford Sedan
Toyota pick−up truck
Toyota Sedan
Prior.elected.office:
(Baseline = governor)
none
state attorney general
U.S. senator
Marital.status:
(Baseline = divorced)
married
single
State.of.residence:
(Baseline = Alabama)
Colorado
Massachusetts
Ohio
Race.Ethnicity:
(Baseline = Asian American)
Black
Hispanic
White
Position.on.abortion:
(Baseline = neutral)
pro−choice
pro−life
−0.4
−0.2
0.0
0.2
0.4
Change in E[Y]


Generalized Ave. Marginal Component Effects
L = 20
−0.4
−0.2
0.0
0.2
0.4
Change in E[Y]

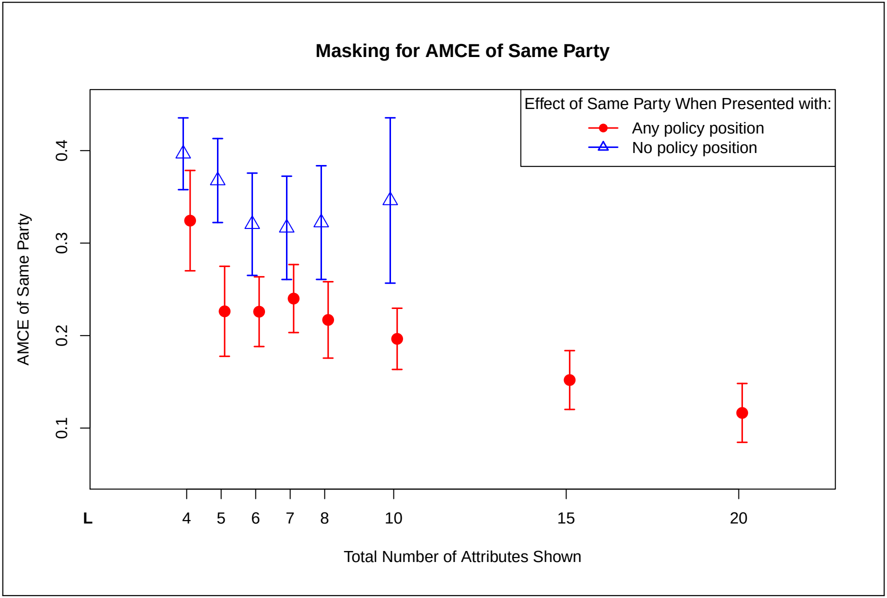
Conditional FAMCEs



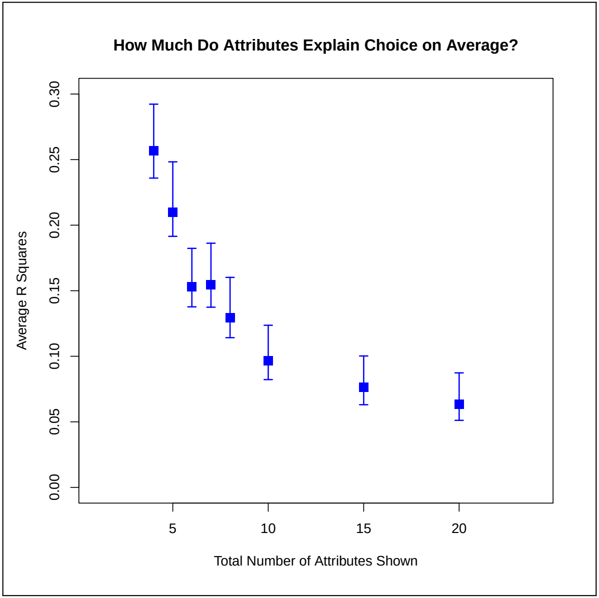
Masking by Issue Positions: Gay Marriage



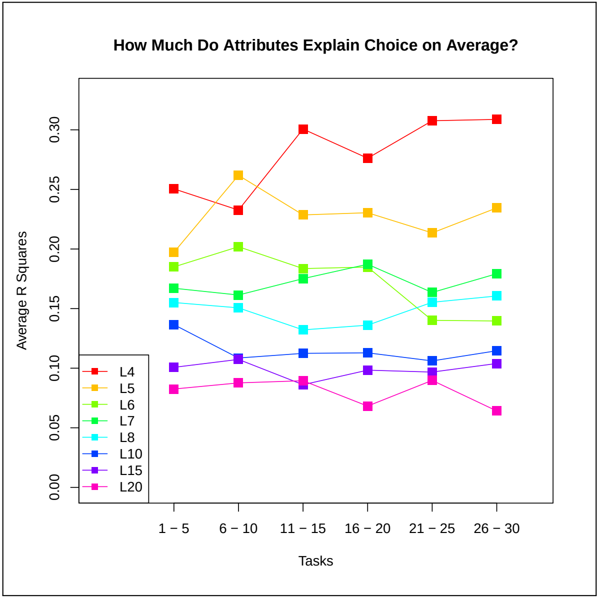
Masking by Issue Positions: Overall



Satisficing



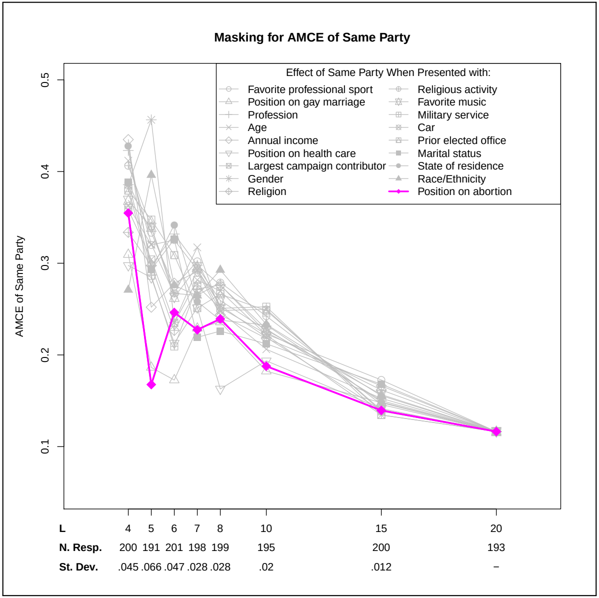
Satisficing



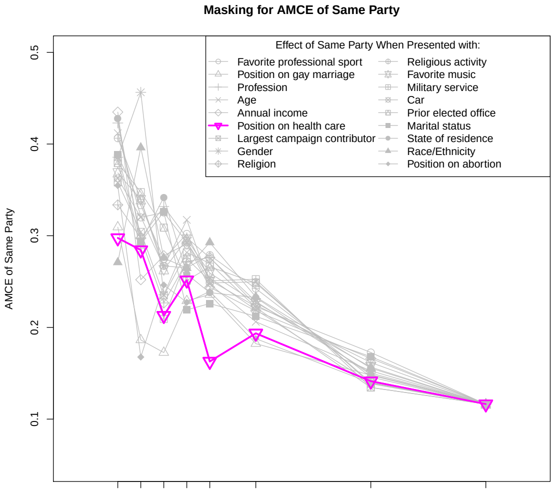
Summary and Conclusion

- Tradeoff between masking and satisficing is very real; choice of attributes matters!
  - Include too few and effects will be masked
  - Include too many and respondents will pay less attention
  - Traditional metrics of satisficing do not pick this up
- The tradeoff is an inherent feature of survey experimentation; "solving" it is impossible
- Goal: not to eliminate masking but to simulate real-world choice contexts
- The doubly (and triply) randomized design is an effective alternative
- Allows inference about interesting quantities in the face of the tradeoff
- Next step: Implementation in software (`Conjoint SDT` and `cjoint`)

Masking

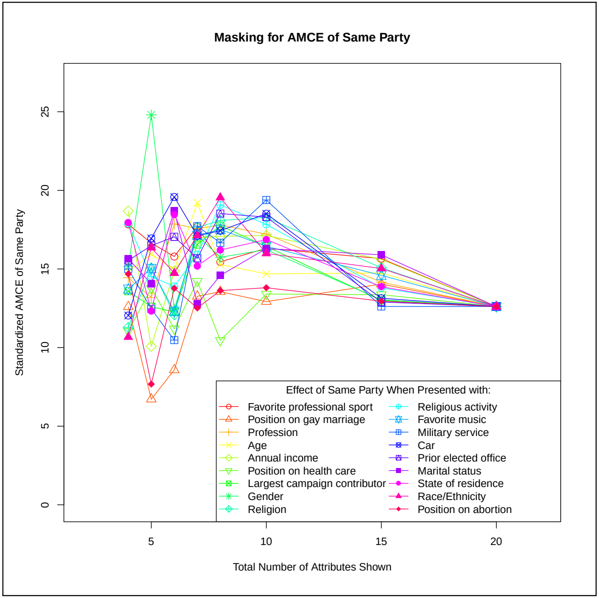

Masking for AMCE of Same Party
Effect of Same Party When Presented with:
Favorite professional sport
Position on gay marriage
Profession
Age
Annual income
Position on health care
Largest campaign contributor
Gender
Religion
Religious activity
Favorite music
Military service
Car
Prior elected office
Marital status
State of residence
Race/Ethnicity
Position on abortion
AMCE of Same Party
0.1
0.2
0.3
0.4
0.5
L 4 5 6 7 8 10 15 20
N. Resp. 200 191 201 198 199 195 200 193
St. Dev. .045 .066 .047 .028 .028 .02 .012 –

Masking

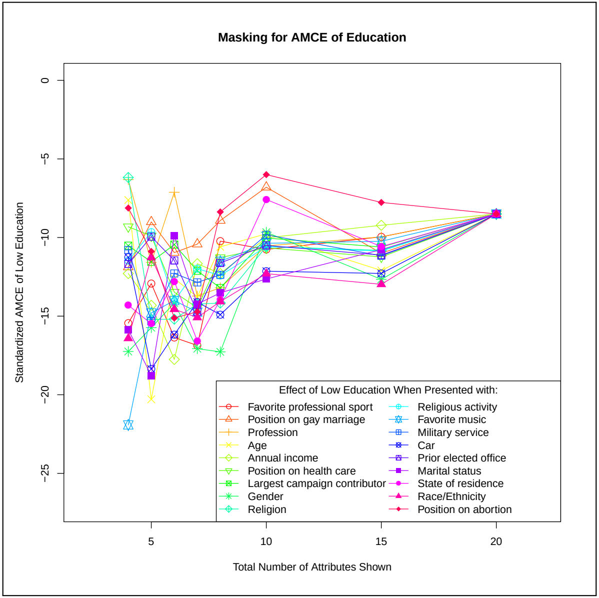


L	4	5	6	7	8	10	15	20
N. Resp.	200	191	201	198	199	195	200	193
St. Dev.	.045	.066	.047	.028	.028	.02	.012	–

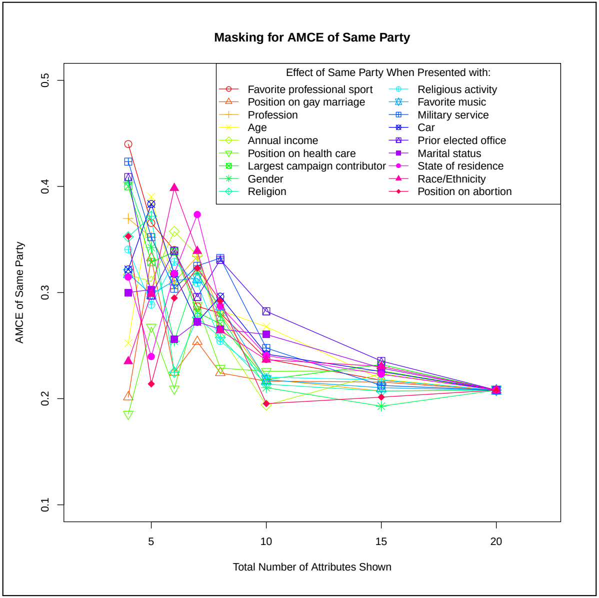
t-values: MTurk experiment



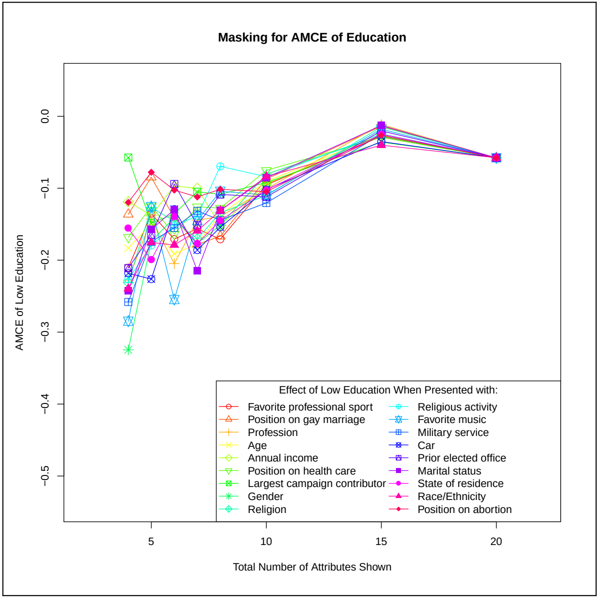
t-values: MTurk experiment



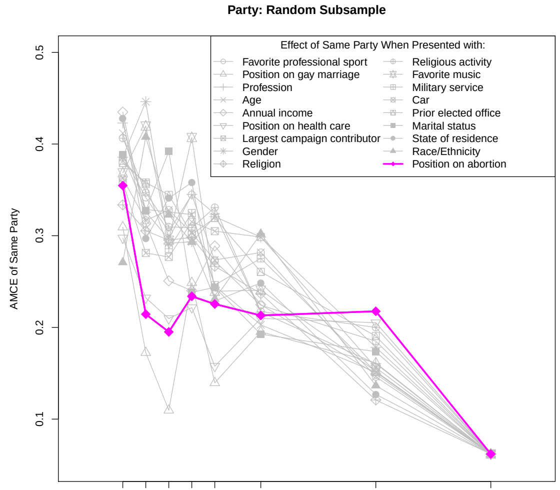
SSI experiment



SSI experiment

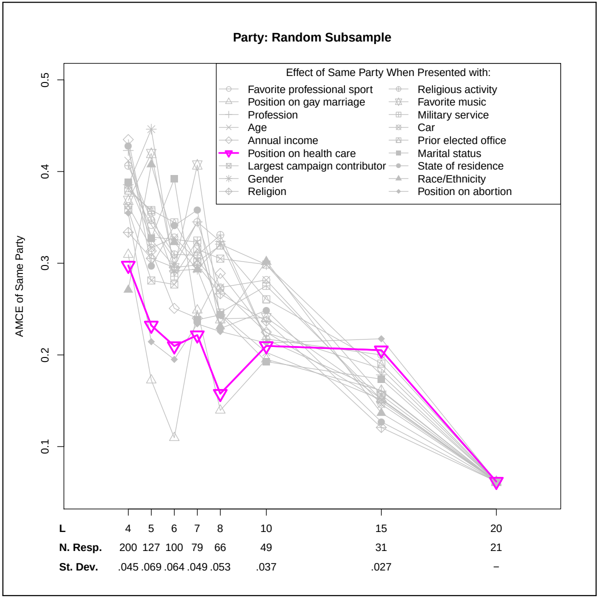


MTurk experiment

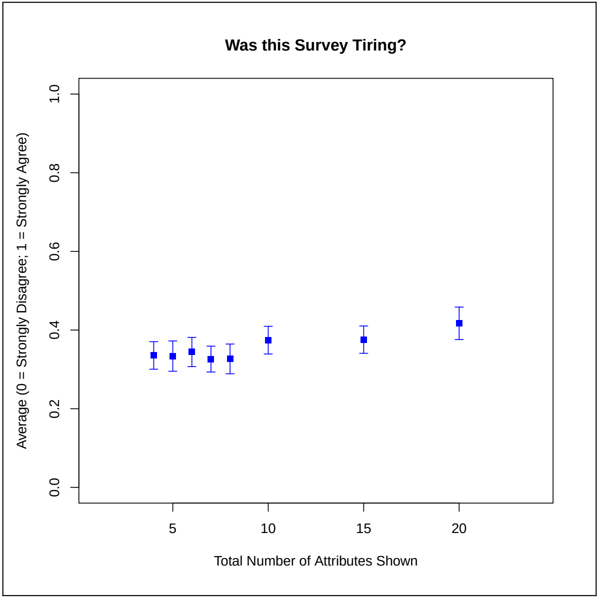


L	4	5	6	7	8	10	15	20
N. Resp.	200	127	100	79	66	49	31	21
St. Dev.	.045	.069	.064	.049	.053	.037	.027	–

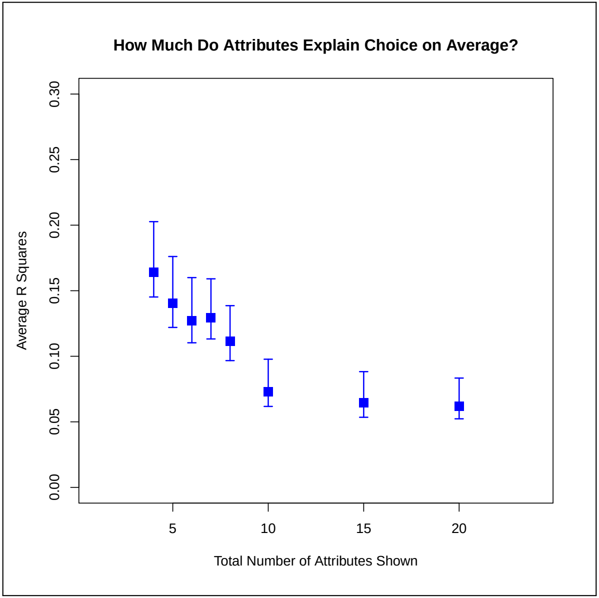
MTurk experiment



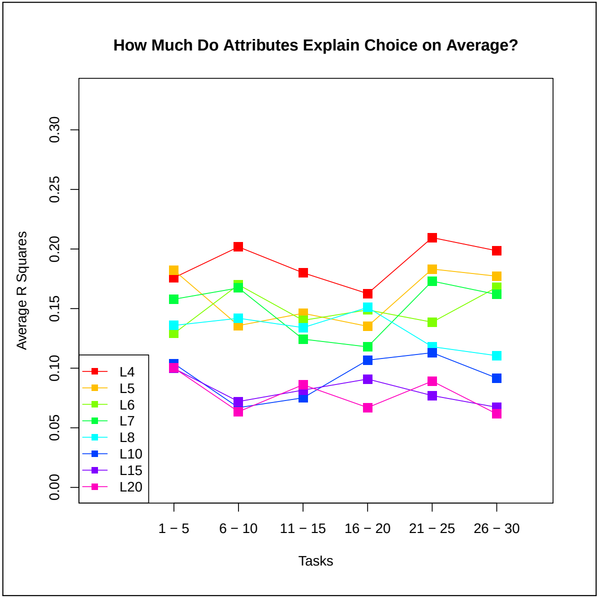
MTurk experiment



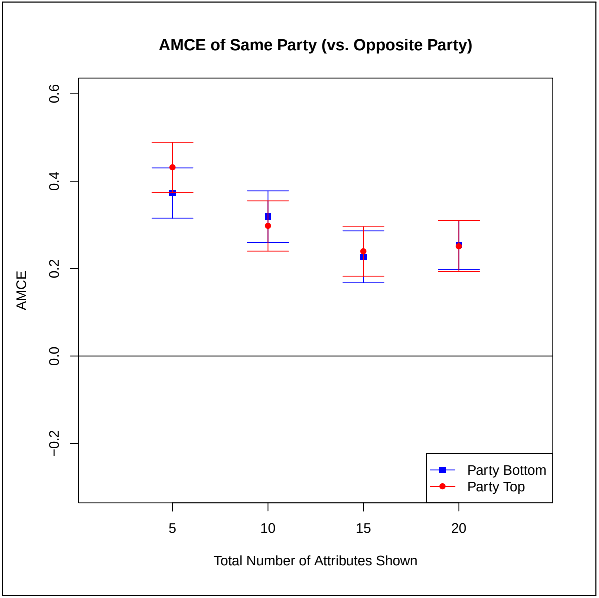
SSI experiment



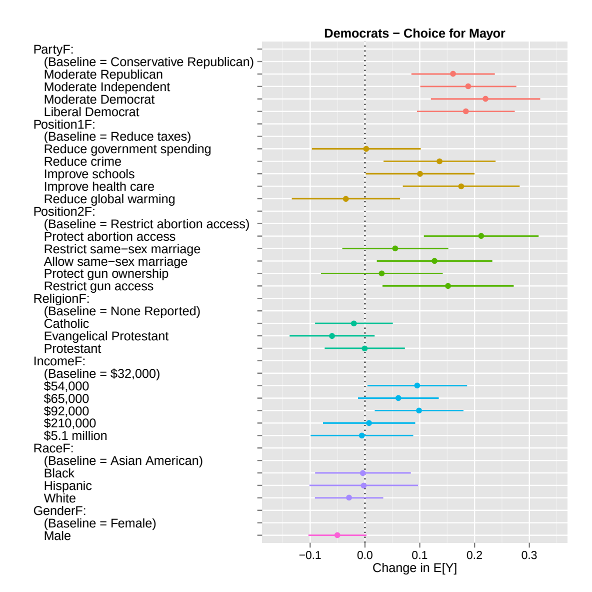
SSI experiment



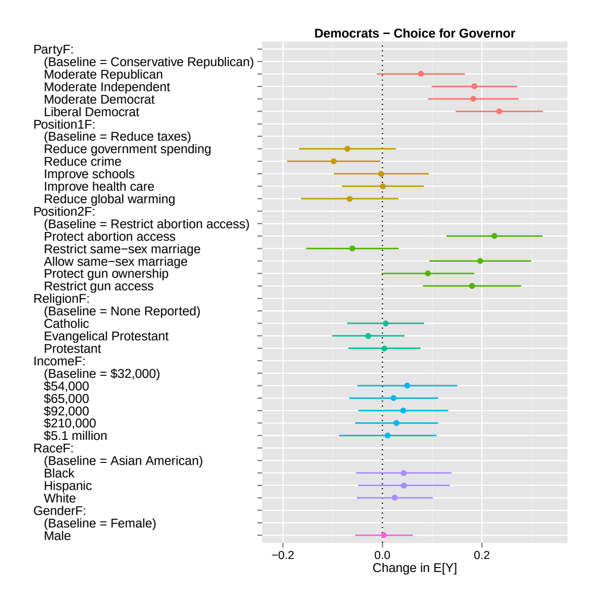
MTurk experiment



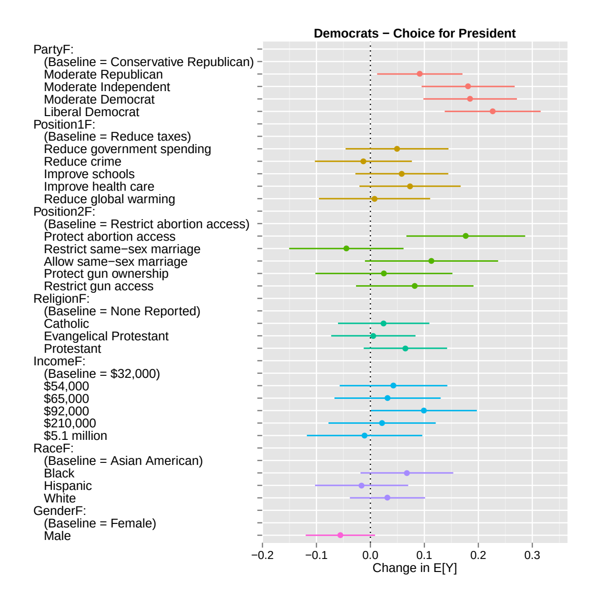
GfK/TESS Experiment



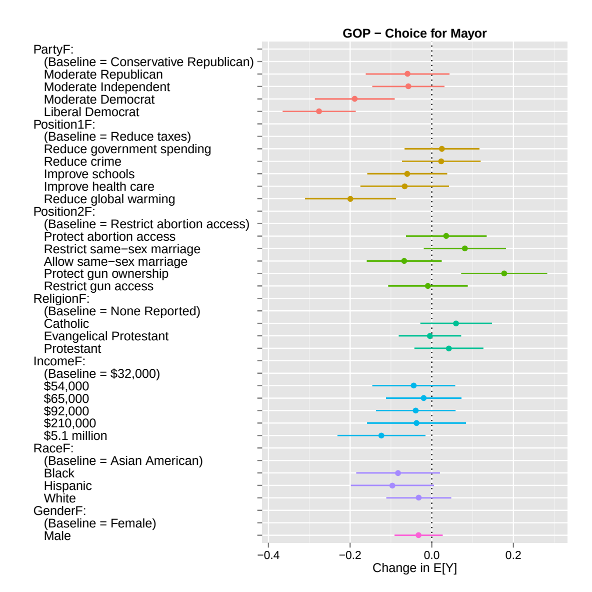
GfK/TESS Experiment



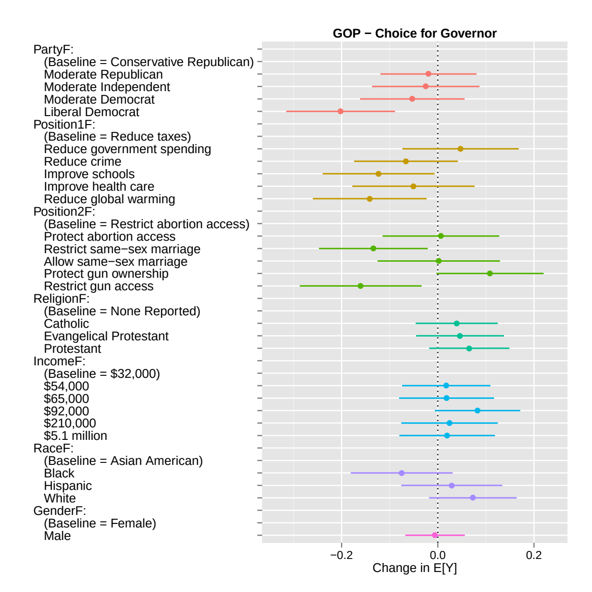
GfK/TESS Experiment



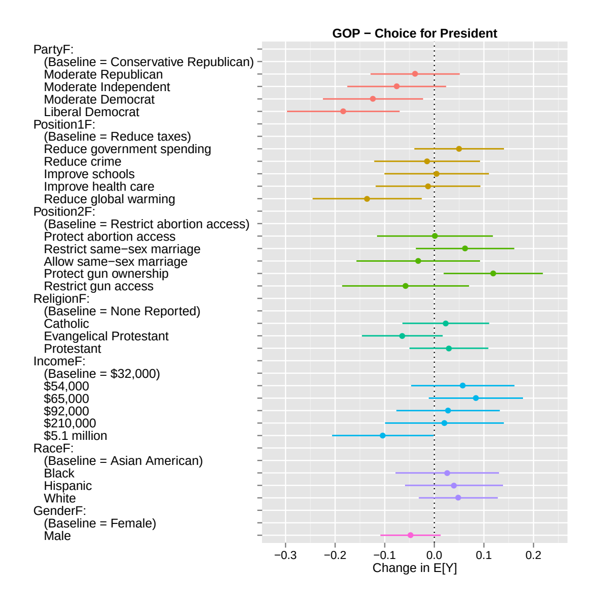
GfK/TESS Experiment



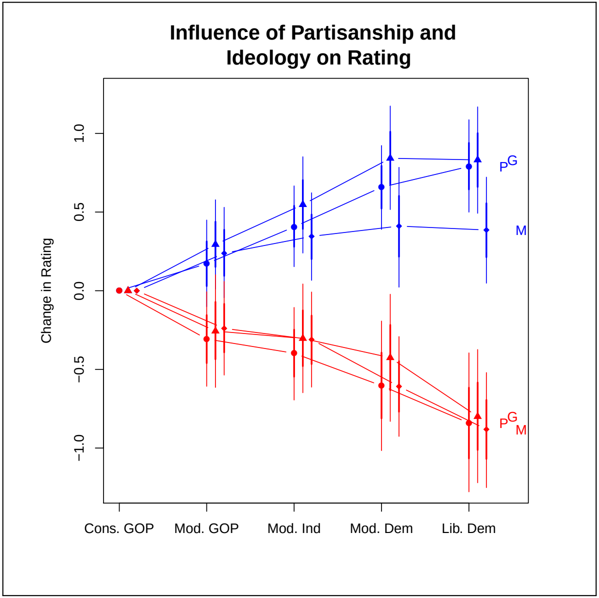
GfK/TESS Experiment



GfK/TESS Experiment



Gfk/TESS Experiment



Median Response Time