Using non-traditional data sources to understand travel behavior

Virginie Lurkin

Laurie A. Garrow

Transport and Mobility Laboratory (TRANSP-OR),
École Polytechnique Fédérale de Lausanne (EPFL)

12th Workshop on Discrete Choice Models
June 22-24, 2017

Overview

Motivation

Research objectives

Survey of air travelers

Results

Conclusion

Overview

Motivation

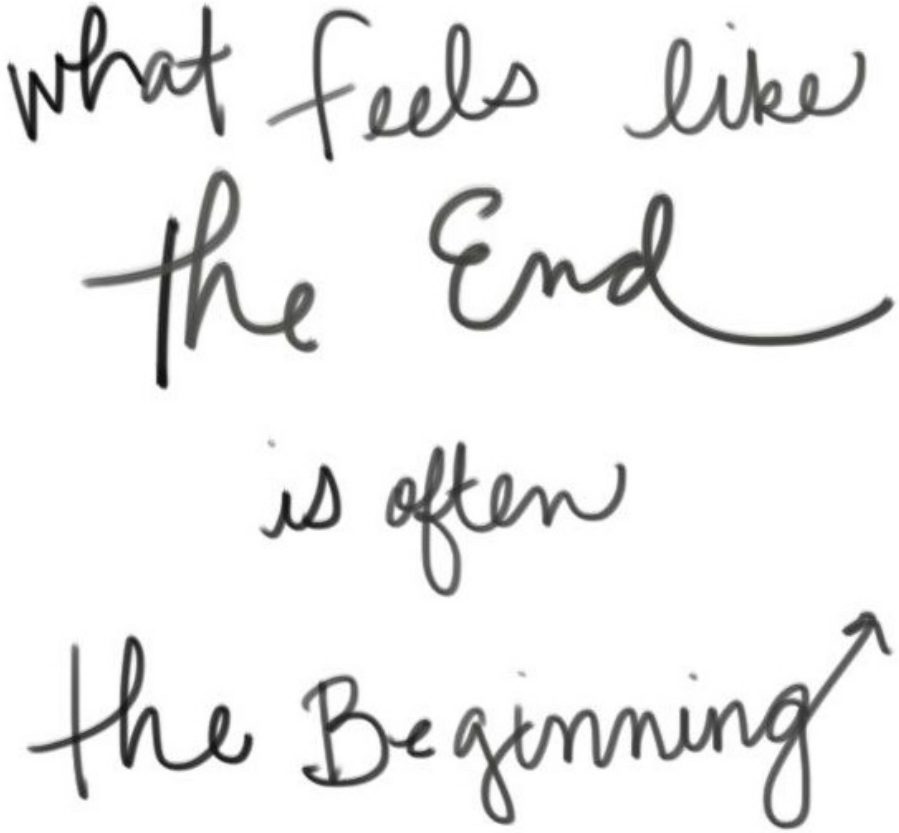
Research objectives

Survey of air travelers

Results

Conclusion

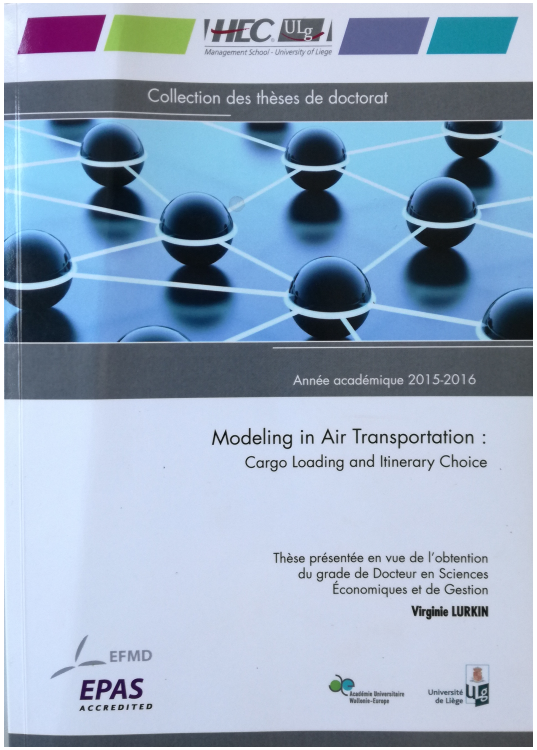
Motivation



Motivation

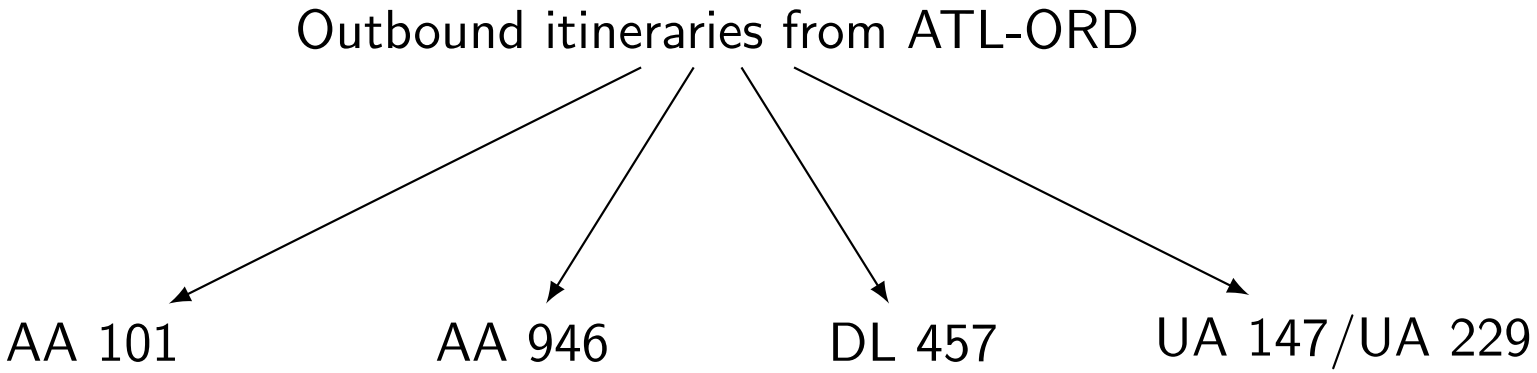
Motivation

Modeling in Air Transportation: Cargo Loading and **Itinerary Choice**



Itinerary choice model

$$y_{ni} = \begin{cases} 1 & \text{if individual } n \text{ chooses itinerary } i, \\ 0 & \text{otherwise} \end{cases}$$



$$U_i = V_i + \varepsilon_i$$

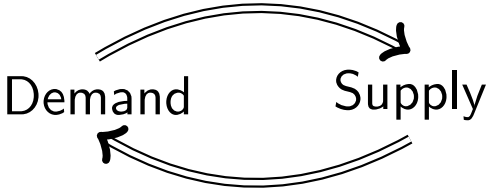
$$V_i = \alpha_i + \beta_1 Cost_i + \beta_2 Time_i + ...$$

$$P_i = \frac{e^{V_i}}{\sum_j e^{V_j}}$$

Factors influencing itinerary choice

The fundamental problem

$$demand = \beta \times price + ... + \varepsilon$$



Price endogeneity

Contributions

Main **contributions**:

- Estimate a baseline MNL model that controls for price endogeneity for high-yield and low-yield fare products using the control-function method
- Estimate more advanced DCM based on the GEV family that capture complex product substitution patterns

Contributions

Main **contributions**:

- Estimate a baseline MNL model that controls for price endogeneity for high-yield and low-yield fare products using the control-function method
- Estimate more advanced DCM based on the GEV family that capture complex product substitution patterns

Main **conclusions**:

- Importance to correct for price endogeneity
  - Over-estimation of customer's value of time and biased price elasticities
- Strong correlation across itineraries that share similar departure times

Choice set generation

- Construct **choice sets** for **each OD city pair that departs on day of week** d

Choice set generation

- Construct **choice sets** for **each OD city pair that departs on day of week** d

Segment	Choice sets	Choice sets Min Alts	Mean Alts	Max Alts
Same TZ, distance $\leq$ 600 miles	30,943	2	10.8	95
Same TZ, distance $>$ 600 miles	22,861	2	14.3	105
One TZ WB, distance $\leq$ 600 miles	5,617	2	10.6	64
One TZ WB, distance $>$ 600 miles	24,82	2	15.1	127
One TZ EB, distance $\leq$ 600 miles	5,630	2	10.3	63
One TZ EB, distance $>$ 600 miles.	25,062	2	14.5	137
Two TZ WB	11,505	2	17.1	133
Two TZ EB	11,267	2	15.3	93
Three TZ WB	6,732	2	21.3	156
Three TZ WB	6,619	2	19.2	138

Key: TZ = Time Zone, WB = Westbound, EB = Eastbound

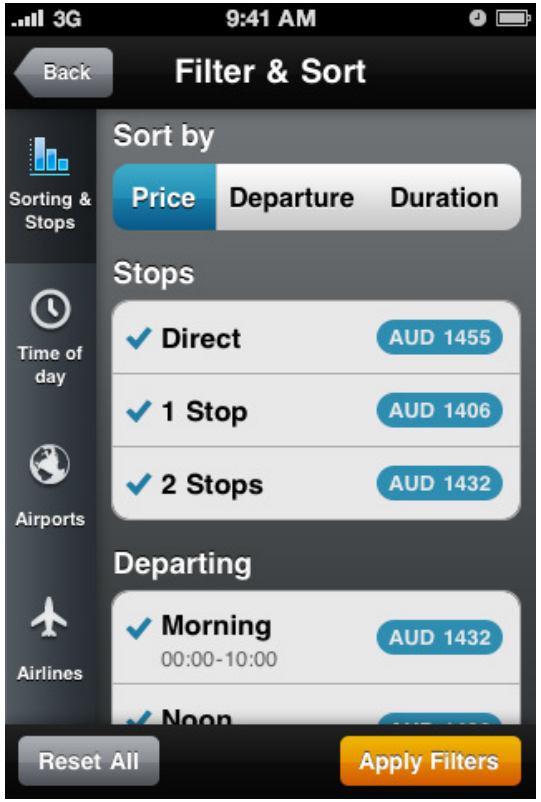
Choice set generation

- Construct **choice sets** for **each OD city pair that departs on day of week** d

Segment	**Choice sets**	**Choice sets**		
		Min Alts	**Mean Alts**	**Max Alts**
Same TZ, distance $\leq$ 600 miles	30,943	2	10.8	**95**
Same TZ, distance $>$ 600 miles	22,861	2	14.3	**105**
One TZ WB, distance $\leq$ 600 miles	5,617	2	10.6	**64**
One TZ WB, distance $>$ 600 miles	24,82	2	15.1	**127**
One TZ EB, distance $\leq$ 600 miles	5,630	2	10.3	**63**
One TZ EB, distance $>$ 600 miles.	25,062	2	14.5	**137**
Two TZ WB	11,505	2	17.1	**133**
Two TZ EB	11,267	2	15.3	**93**
Three TZ WB	6,732	2	21.3	**156**
Three TZ WB	6,619	2	19.2	**138**

Key: TZ = Time Zone, WB = Westbound, EB = Eastbound

Choice set generation



Overview

Motivation

Research objectives

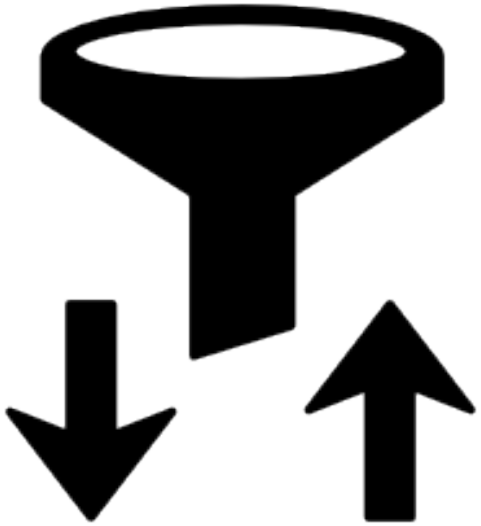
Survey of air travelers

Results

Conclusion

Ultimate objective

Develop a choice set generation model for itinerary choice models that incorporates sorting and filtering actions using an interactive online survey



Intermediate objective

Determine if lower-cost crowdsourcing worksites, such as Amazon Mechanical Turk provide similar results as more traditional survey panels

Overview

Motivation

Research objectives

Survey of air travelers

Results

Conclusion

Online survey

Background Questions

Georgia Institute of Technology

4. How long before your trip did you purchase your ticket?

- ○ **Zero to three days before**
- ○ **Four to six days before**
- ○ **One to two weeks before**
- ○ **Two to three weeks before**
- ○ **One month before**
- ○ **Two months before**
- ○ **Three or more months before**

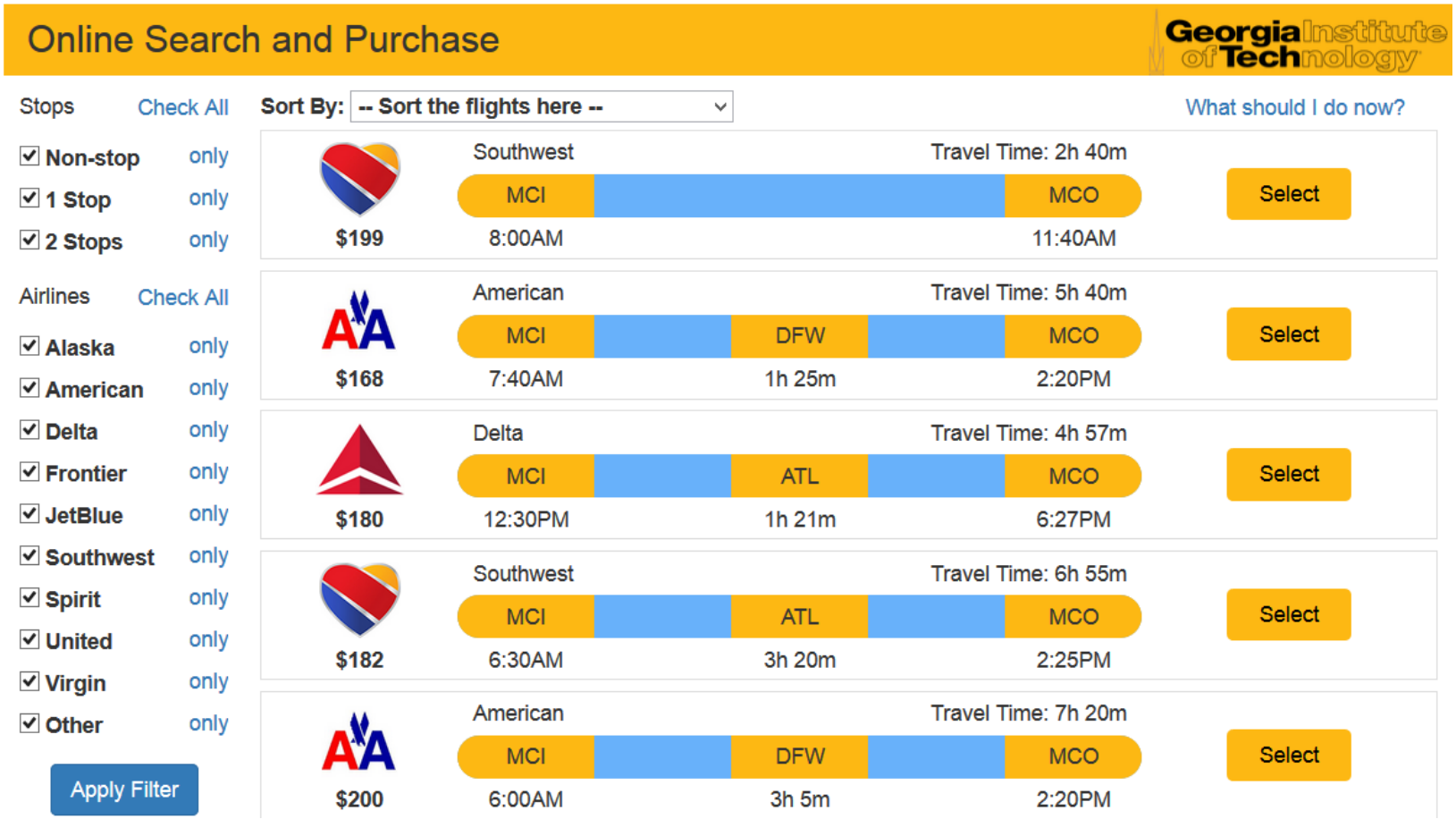
5. At what airport did you begin the outbound air portion of your trip? Enter the airport by typing the city, airport name or airport code in the box below. Then select the airport from the menu. If your airport is not one of the options, just leave its name and city name in the box.

My air trip began at:

6. At what airport did you conclude the outbound air portion of your trip? Please do not include any connecting airports you may have passed through during the air portion of your trip. Enter the airport by typing the city, airport name or airport code in the box below. Then select the airport from the menu. If your airport is not one of the options, just leave its name and city name in the box.

My air trip ended at:

Online survey

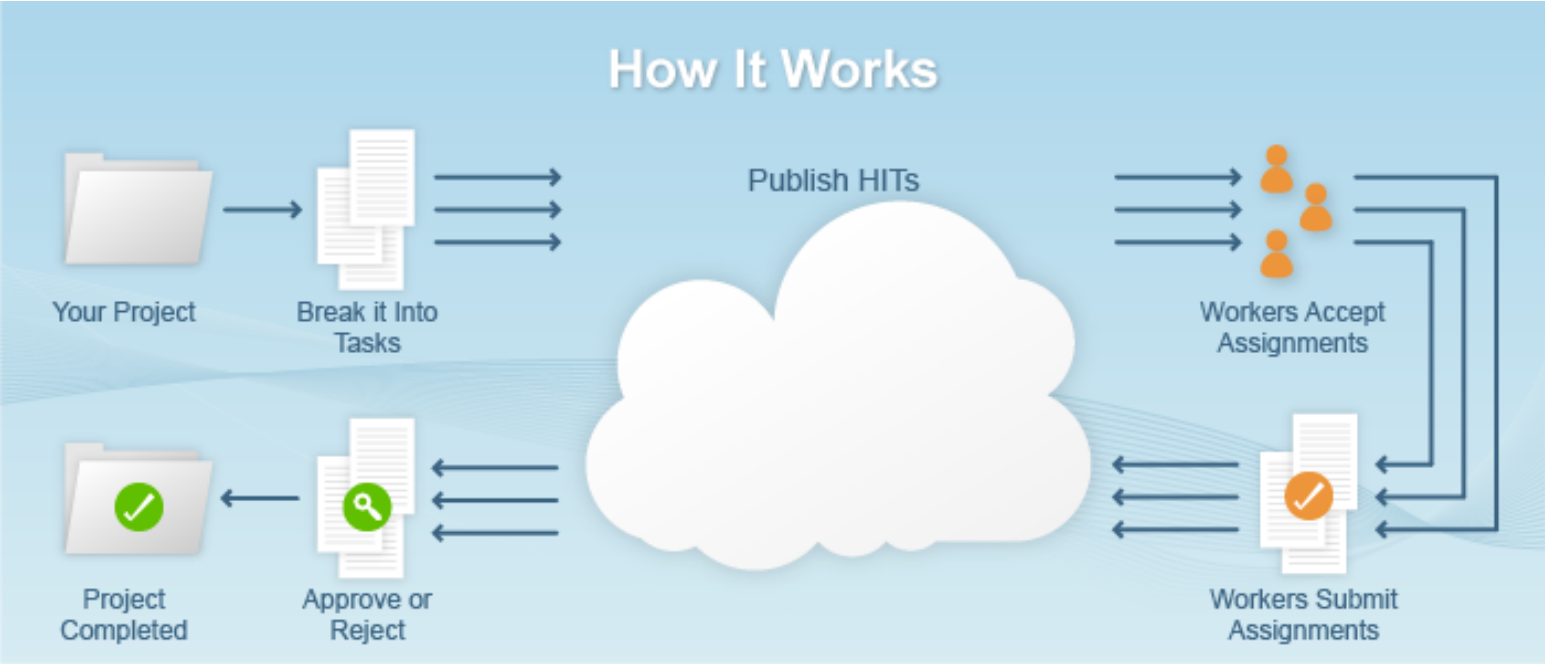


Data - AMT or Qualtrics?

1. Amazon Mechanical Turk (AMT) is an online outsourcing platform with more than 500,000 workers in 190+ countries that perform microtasks , typically for $0.10 USD or less

2. Qualtrics is a more traditional marketing firm that maintains a panel of respondents that complete surveys for a variety of clients

Amazon Mechanical Turk

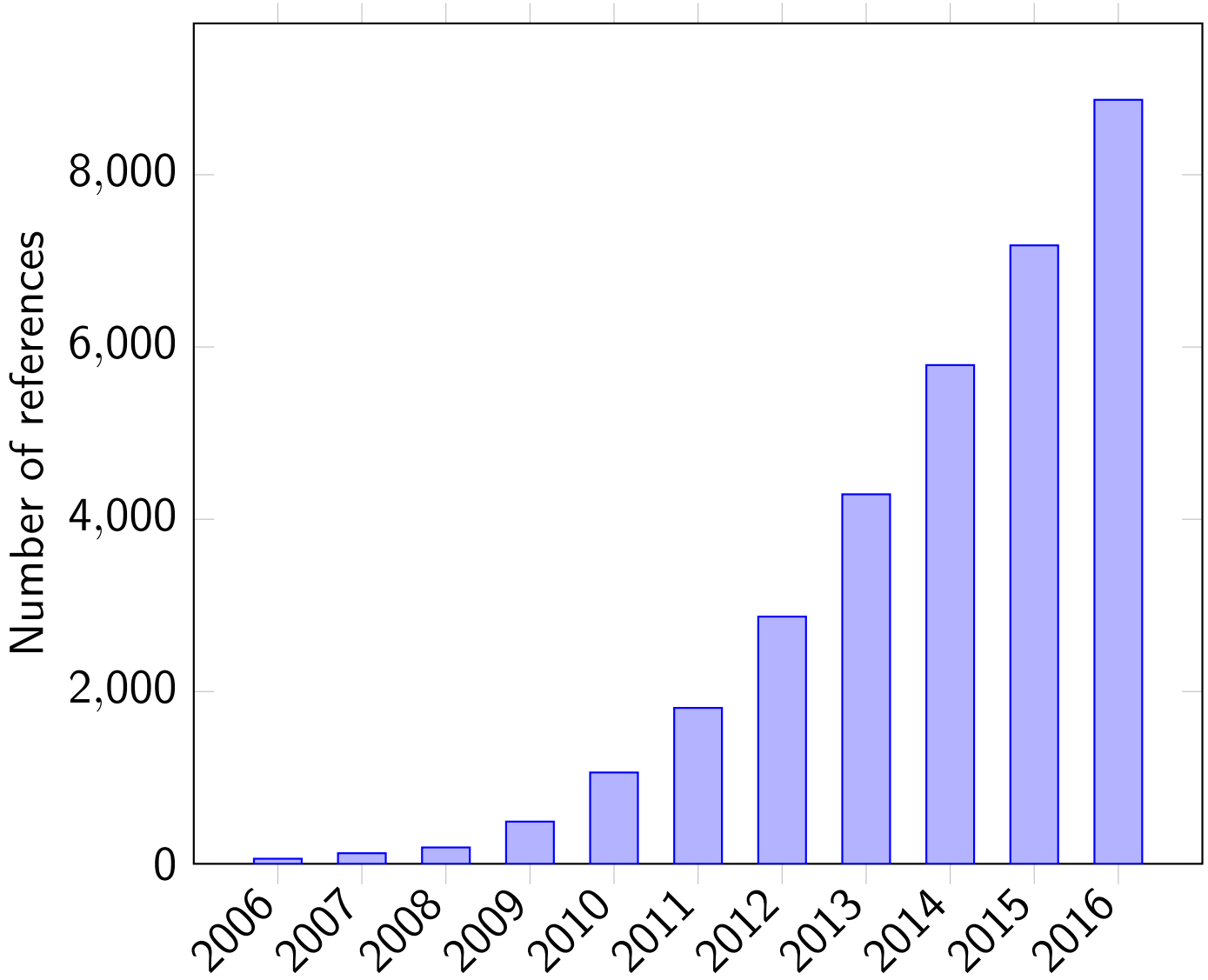


Comparison of AMT and Qualtrics

	AMT	Qualtrics
Number of respondents*	690	553
-High yield respondents	62	62
-Low yield respondents	628	491
Data collection period	Oct-Nov 2016	March 2017
Total survey cost	$305.25	$3,835
Participant reimbursement	$0.25 regular workers $1.00 master workers	$0.65

*after cleaning

Google Scholar Search of "Mechanical Turk"



Methodology

AMT $\times$ **Qualtrics**

1. Use Chi-Square test of homogeneity to determine if survey respondents and their responses to individual questions are similar

2. Estimate itinerary choice models from AMT and Qualtrics data and determine if results are similar

Chi-Square test of homogeneity

H_0: AMT and Qualtrics respondents are homogeneous with respect to the proposed categories

Chi-Square test of homogeneity

H_0: AMT and Qualtrics respondents are homogeneous with respect to the proposed categories

$$\chi^2 \text{ statistic} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \text{ where}$$

O_{ij} is the **observed** frequency for category i and population j

E_{ij} is the **expected** frequency for category i and population j

Chi-Square test of homogeneity

H_0: AMT and Qualtrics respondents are homogeneous with respect to the proposed categories

$$\chi^2 \text{ statistic} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \text{ where}$$

O_{ij} is the **observed** frequency for category i and population j

E_{ij} is the **expected** frequency for category i and population j

Categorical question	AMT	Qualtrics	All
$category_1$	O_{11}	O_{12}	c_1
$category_2$	O_{21}	O_{22}	c_2
...	...	...	
$category_r$	O_{r1}	O_{r2}	c_r
	n_1	n_2	n

Chi-Square test of homogeneity

H_0: AMT and Qualtrics respondents are homogeneous with respect to the proposed categories

$$\chi^2 \text{ statistic} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \text{ where}$$

O_{ij} is the **observed** frequency for category i and population j

E_{ij} is the **expected** frequency for category i and population j

Categorical question	AMT	Qualtrics	All
$category_1$	O_{11}	O_{12}	c_1
$category_2$	O_{21}	O_{22}	c_2
...	...	...	
$category_r$	O_{r1}	O_{r2}	c_r
	n_1	n_2	n

$$E_{ij} = \frac{c_i \times n_j}{n}$$

Chi-Square test of homogeneity

H_0: AMT and Qualtrics respondents are homogeneous with respect to the proposed categories

$$\chi^2 \text{ statistic} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \text{ where}$$

O_{ij} is the **observed** frequency for category i and population j

E_{ij} is the **expected** frequency for category i and population j

Categorical question	AMT	Qualtrics	All
$category_1$	O_{11}	O_{12}	c_1
$category_2$	O_{21}	O_{22}	c_2
...	...	...	
$category_r$	O_{r1}	O_{r2}	c_r
	n_1	n_2	n

$$E_{ij} = \frac{c_i \times n_j}{n}$$

Reject H_0 if χ^2 statistic $> \chi^2_{\alpha,df}$, where $df = (r-1) \times (2-1)$

Overview

Motivation

Research objectives

Survey of air travelers

Results

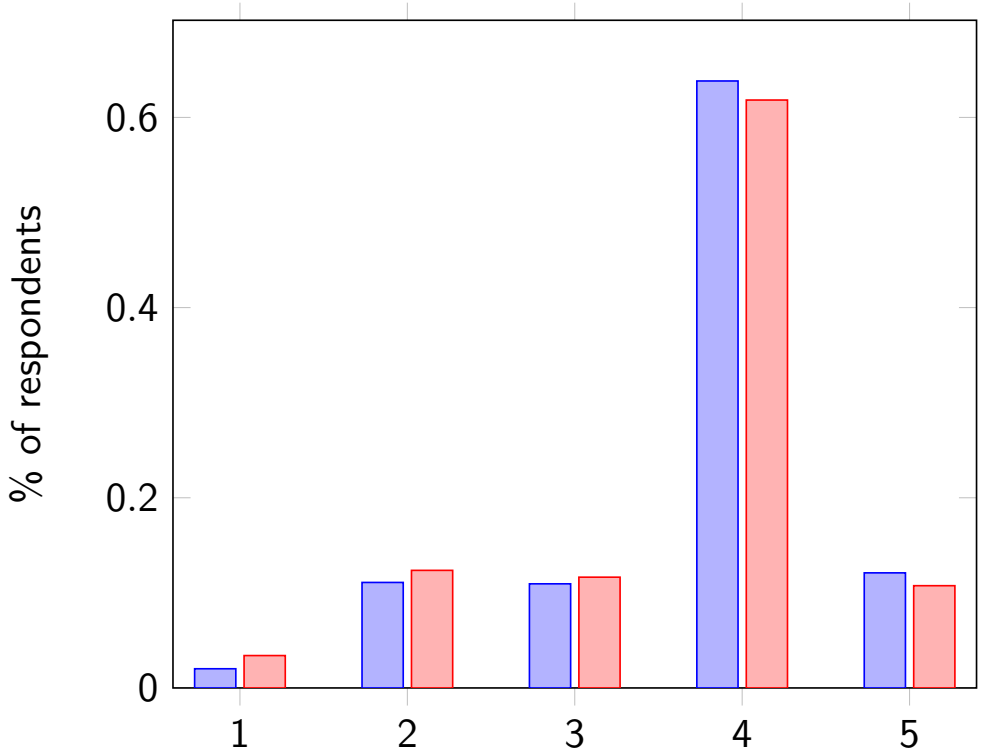
Conclusion

Audience poll

Who thinks AMT and Qualtrics respondents have statistically equivalent **trip characteristics** ?

Comparison AMT and Qualtrics

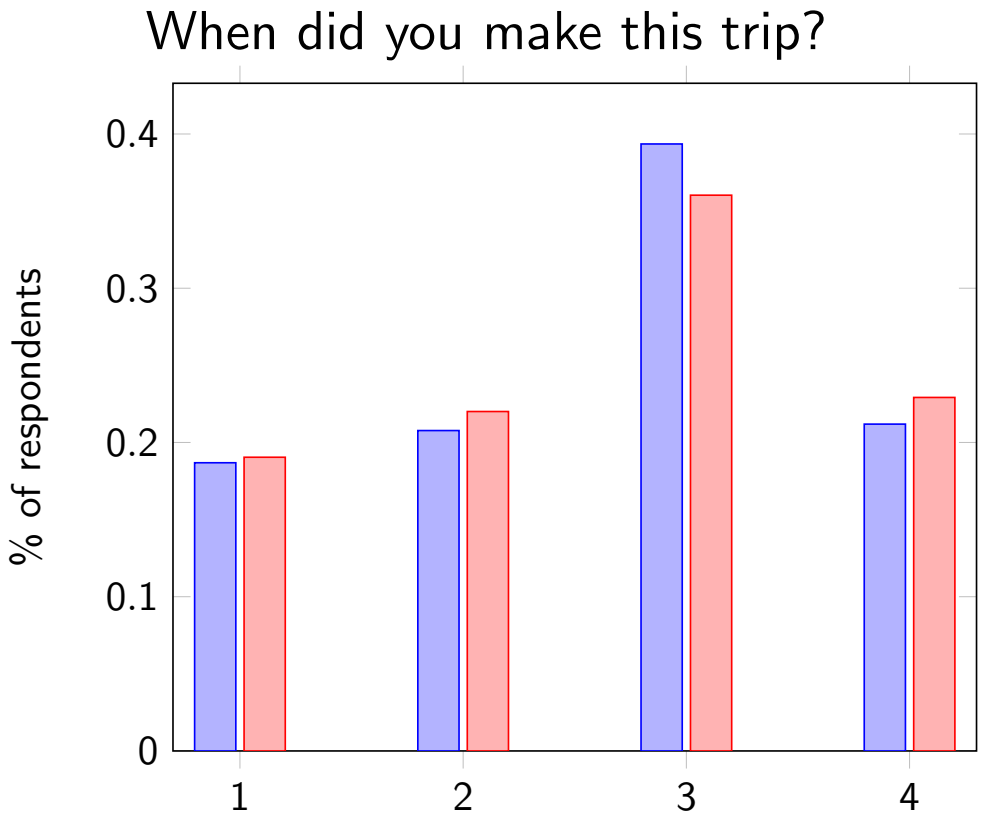
Approximately how often do you make air trips?



(1) 1 RT/week or more (2) 1-3 RT/month (3) 7-12 RT/year
(4) 1-6 RT/year (5) $<$ 1 RT/year

χ^2 statistic $= 3.5 < \chi^2_{0.05,4} = 9.5 \rightarrow$ **not reject** H_0

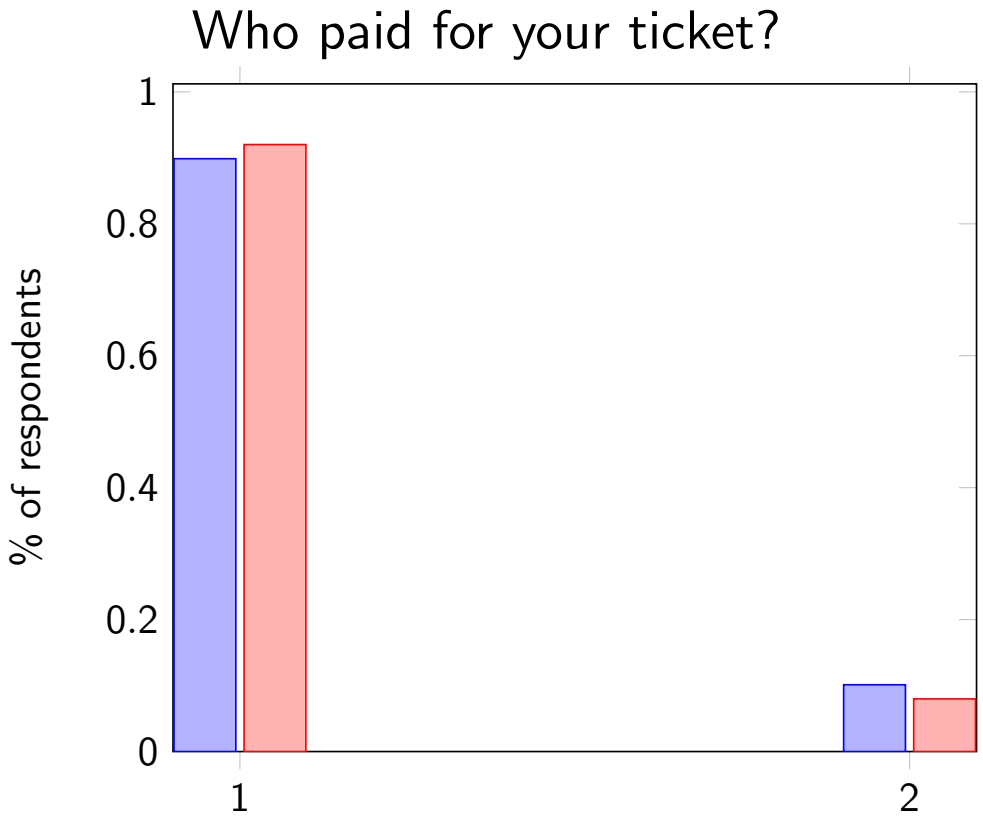
Comparison AMT and Qualtrics



(1) In the past month (2) 1-2 months ago
(3) 3-6 months ago (4) 7-12 months ago

χ^2 statistic $= 2.3 < \chi^2_{0.05,3} = 7.1 \rightarrow$ **not reject** H_0

Comparison AMT and Qualtrics

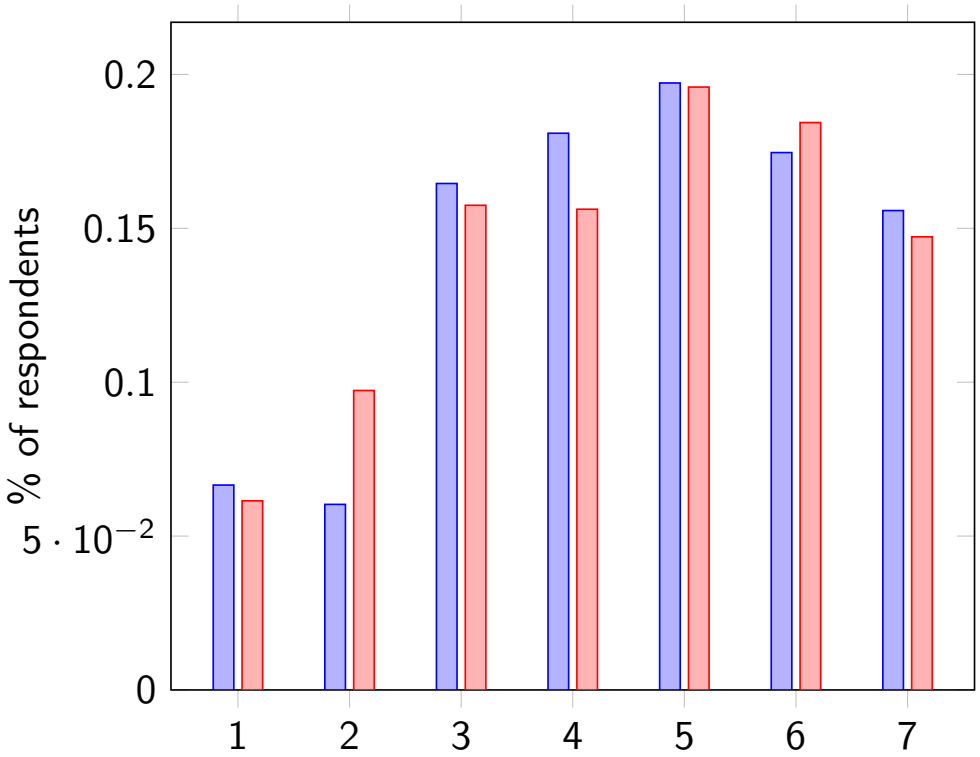


(1) I paid, personally (2) My company paid or reimbursed me

χ^2 statistic $= 2.5 < \chi^2_{0.05,3} = 3.8 \rightarrow$ **not reject** H_0

Comparison AMT and Qualtrics

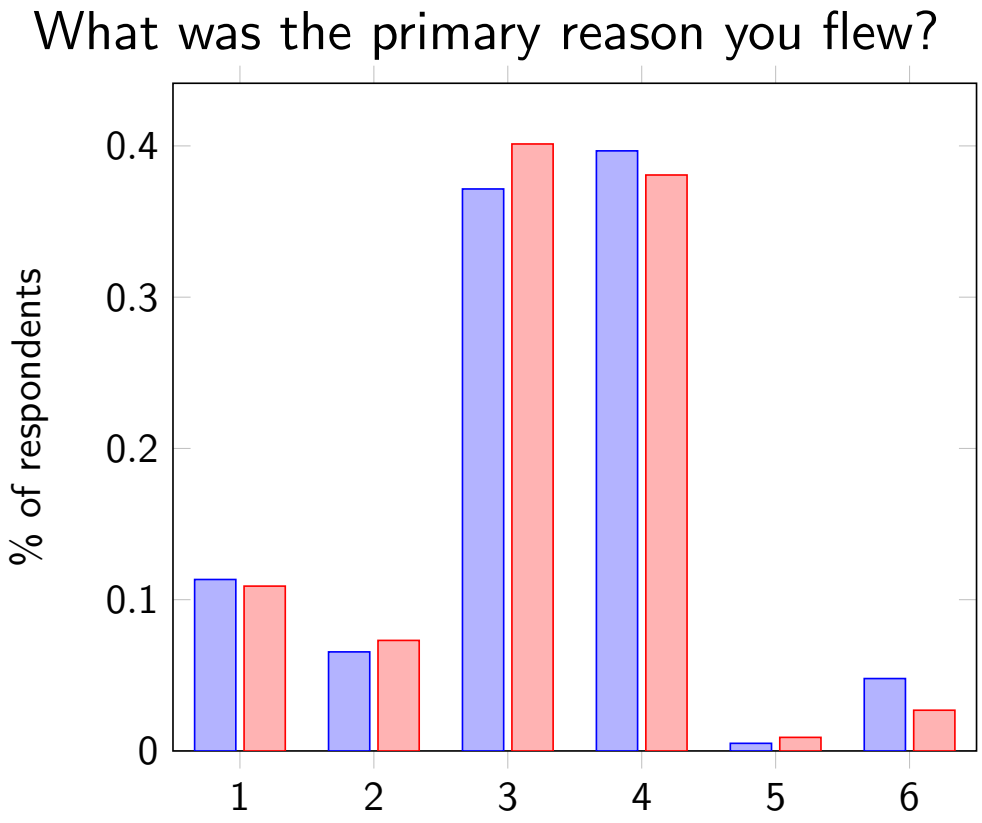
How long before your trip did you purchase your ticket?



(1) 0-3 days (2) 4-6 days (3) 1-2 weeks (4) 2-3 weeks
(5) 1 months (6) 2 months (7) 3+ months

χ^2 statistic $= 9.0 < \chi^2_{0.05,3} = 12.6 \rightarrow$ **not reject** H_0

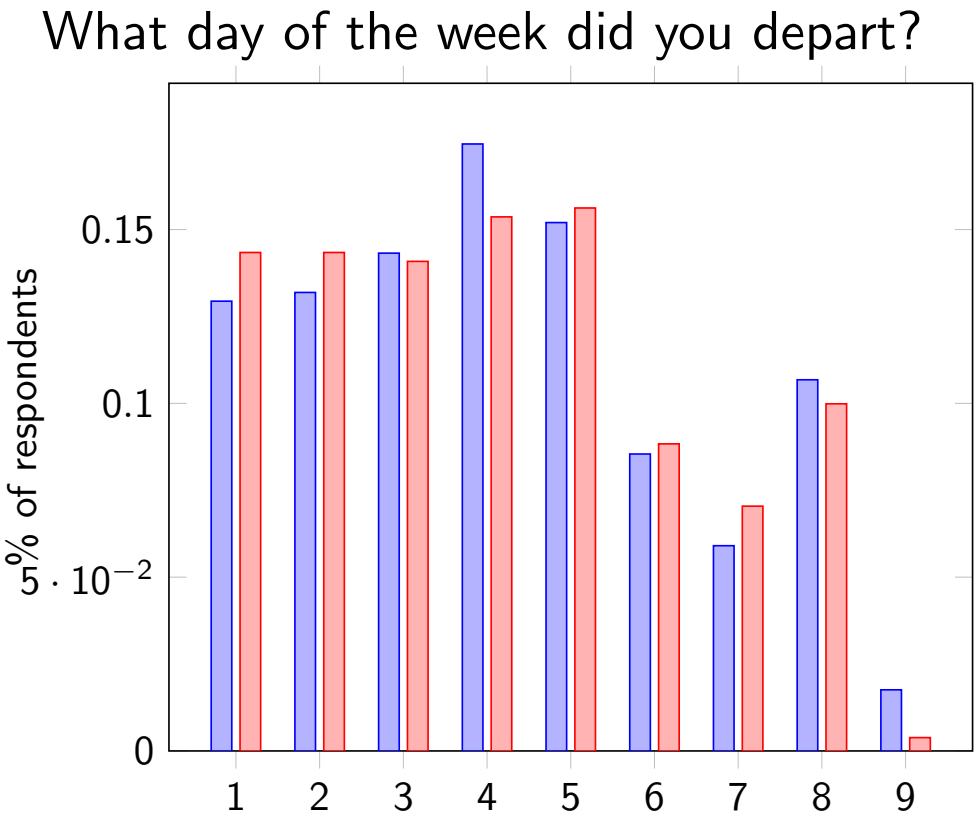
Comparison AMT and Qualtrics



(1) Business (2) Conference (3) Vacation
(4)VFR (5) School (6) Other

$$\chi^2 \text{ statistic} = 7.0 < \chi^2_{0.05,3} = 11.1 \rightarrow \textbf{not reject } H_0$$

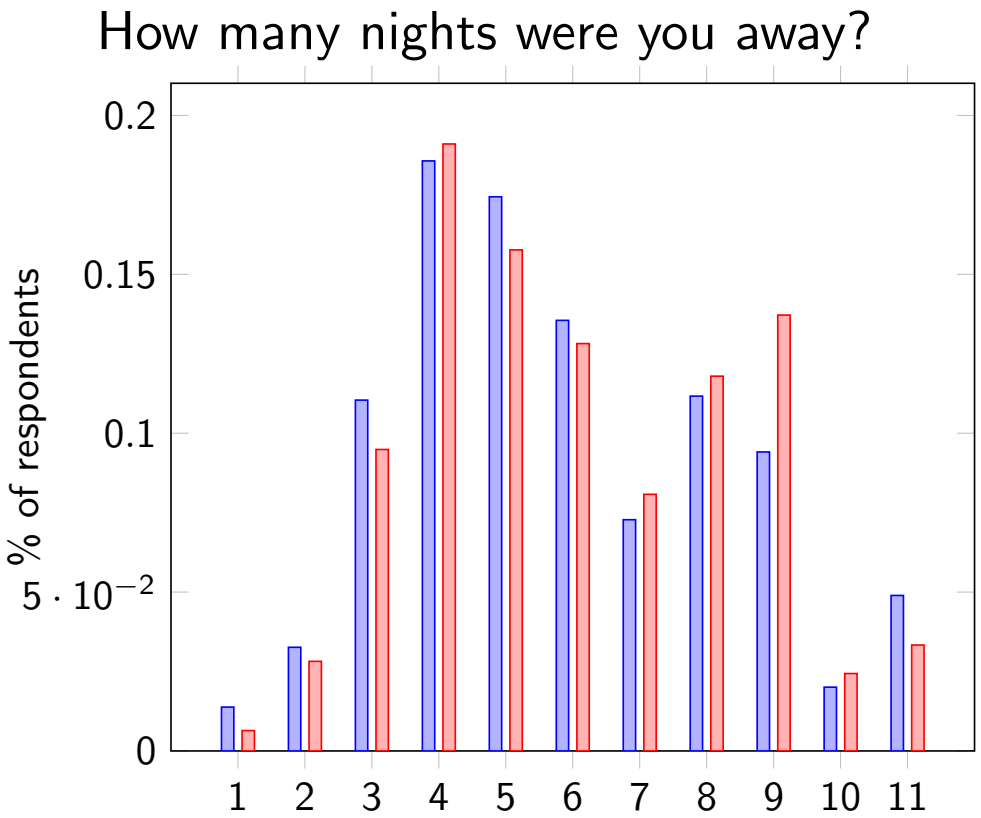
Comparison AMT and Qualtrics



(1) Monday (2) Tuesday (3) Wednesday (4) Thursday (5) Friday (6) Saturday
(7) Sunday (8) Don't remember but a weekday (9) Don't remember but a weekend

$$\chi^2 \text{ statistic} = 10.0 < \chi^2_{0.05,3} = 15.5 \rightarrow \textbf{not reject } H_0$$

Comparison AMT and Qualtrics

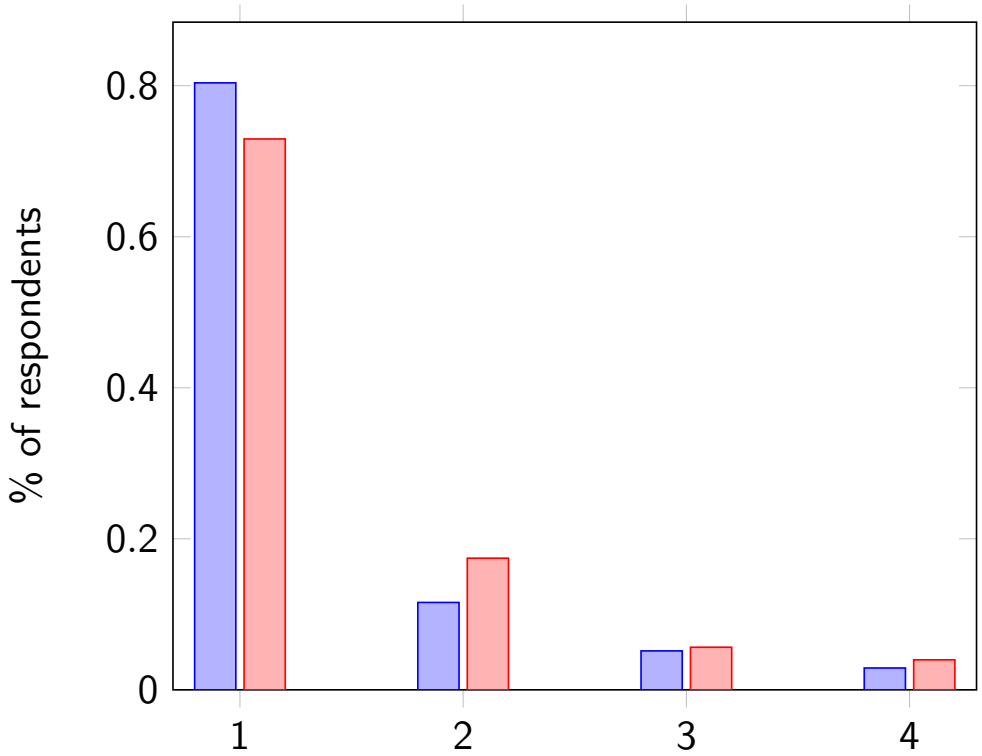


(1) 0 night (2) 1 night (3) 2 nights (4) 3 nights (5) 4 nights (6) 5 nights
(7) 6 nights (8) 7 nights (9) 8-14 nights (10) 15-20 nights (11) 3 weeks or more

$$\chi^2 \text{ statistic} = 13.6 < \chi^2_{0.05,3} = 18.3 \rightarrow \textbf{not reject } H_0$$

Comparison AMT and Qualtrics

What class of service did you use on your trip?

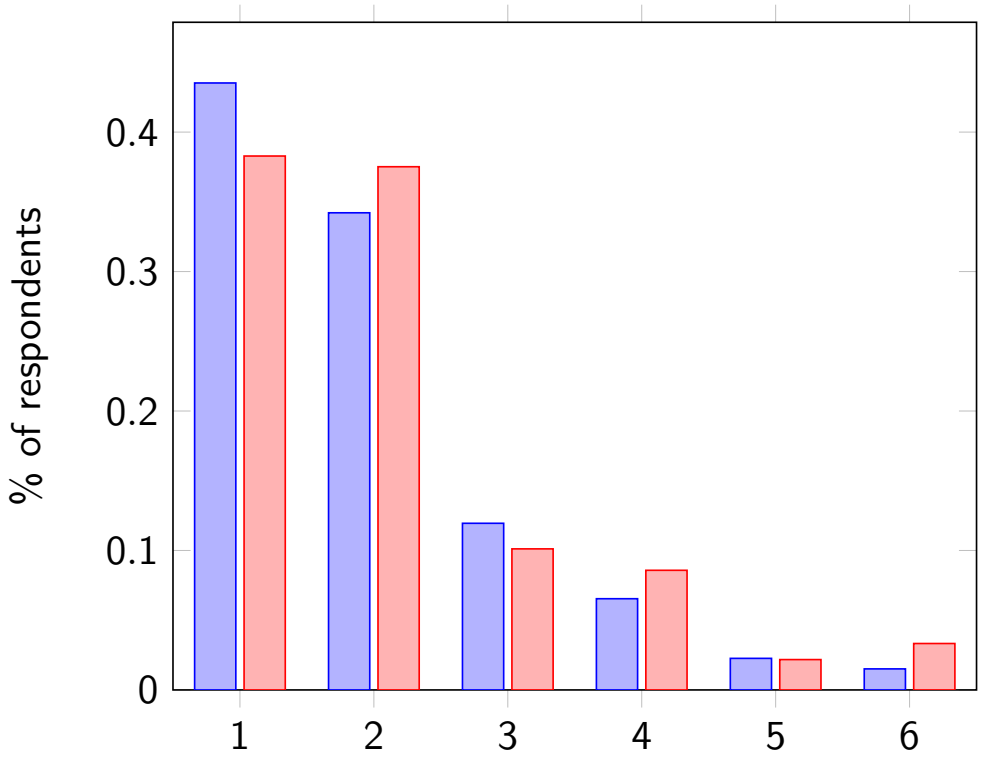


(1) Basic economy (2) Premium economy (3) Business (4) First class

$$\chi^2 \text{ statistic} = 13.7 > \chi^2_{0.05,3} = 7.8 \rightarrow \textbf{reject } H_0$$

Comparison AMT and Qualtrics

How many associates, friends, or family members travelled together?



(1) 1 (alone) (2) 2 person (3) 3 persons
(4) 4 persons (5) 5 persons (6) 6+ persons

$$\chi^2 \text{ statistic} = 12.6 > \chi^2_{0.05,3} = 11.1 \rightarrow \textbf{reject } H_0$$

Audience poll

Who thinks AMT and Qualtrics respondents have statistically equivalent **airline memberships** and **itinerary preferences**?

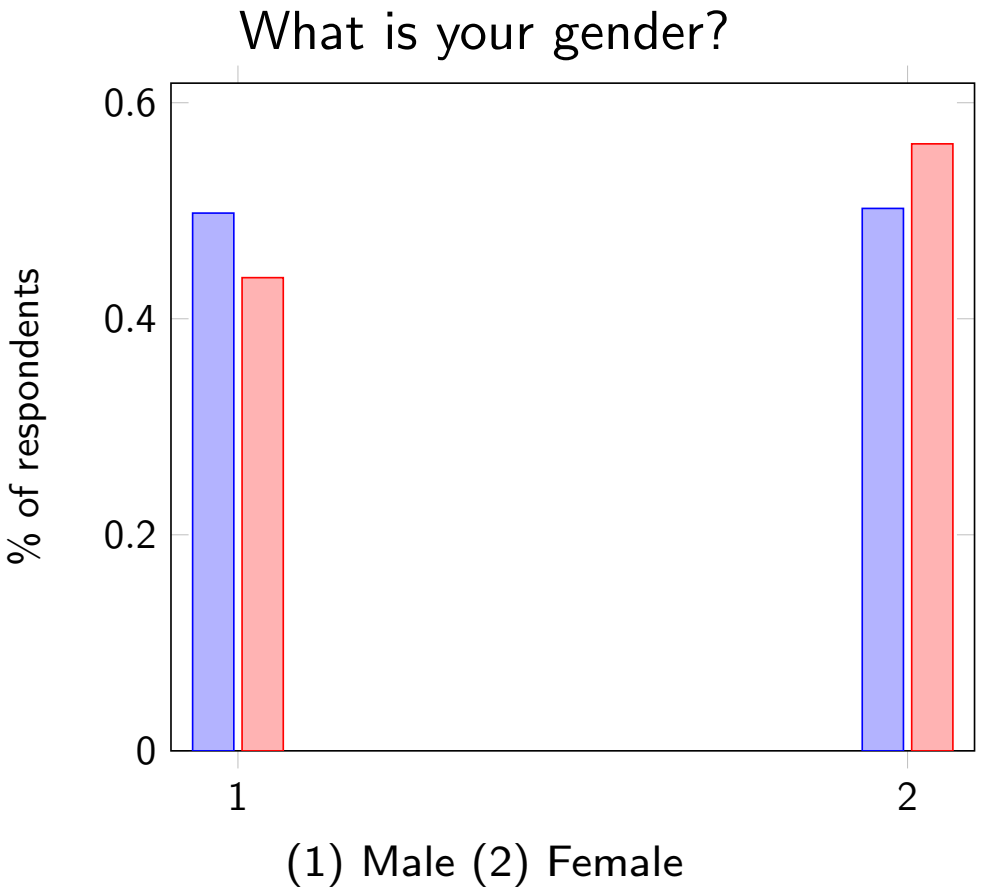
Comparison AMT and Qualtrics

Categorical question	Conclusion
Please indicate the airlines that you have previously flown on?	**not reject** H_0
I only fly certain airlines	**reject** H_0
I generally shop for the cheapest flights and do not consider other factors	**reject** H_0
I avoid small propeller and regional jet aircraft	**reject** H_0
Travel times are more important to me than price	**reject** H_0
Travel times are more important to me than the carrier	**reject** H_0
Price is more important to me than carrier	**reject** H_0

Audience poll

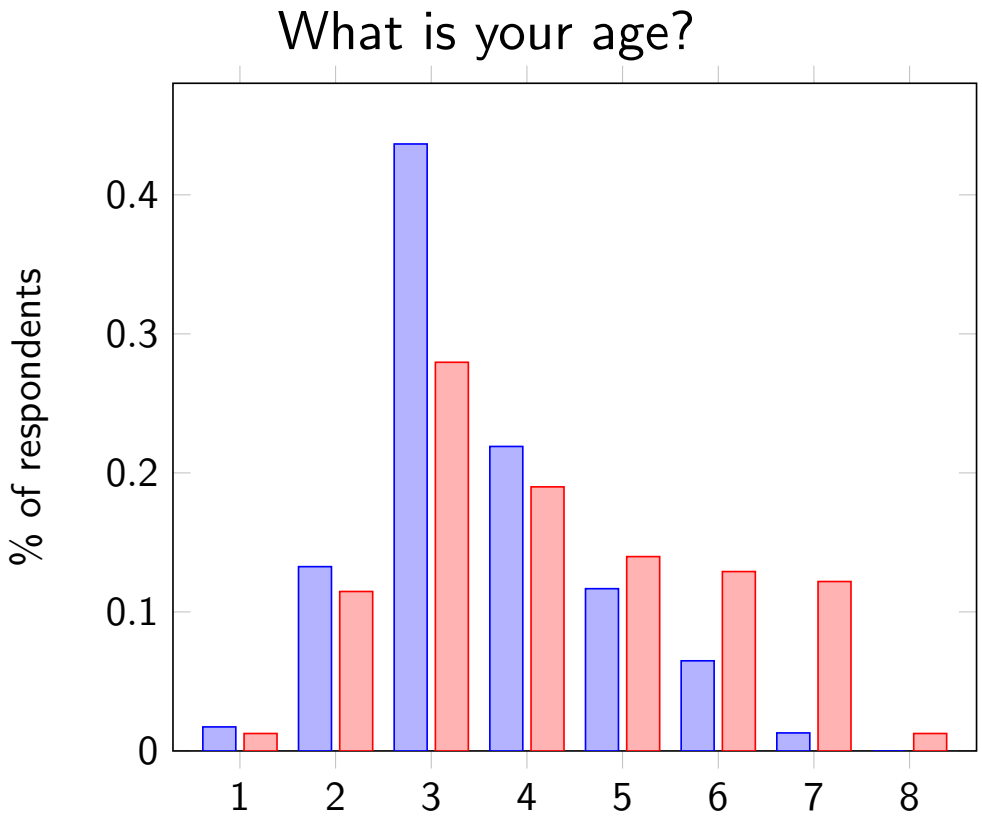
Who thinks AMT and Qualtrics respondents have statistically equivalent **sociodemographic characteristics**?

Comparison AMT and Qualtrics



$$\chi^2 \text{ statistic} = 4.4 > \chi^2_{0.05,3} = 3.8 \rightarrow \textbf{reject } H_0$$

Comparison AMT and Qualtrics

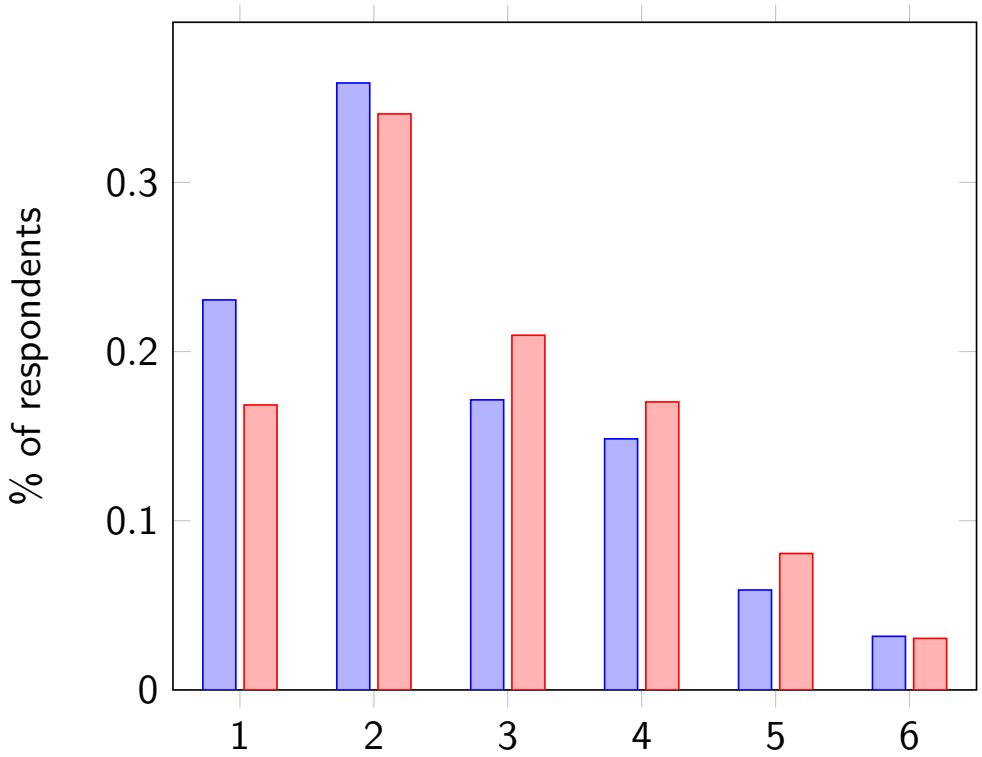


(1) 18-19 (2) 20-24 (3) 25-34 (4) 35-44
(5) 45-54 (6) 55-64 (7) 65-74 (8) 75+

$$\chi^2 \text{ statistic} = 106.6 > \chi^2_{0.05,3} = 14.1 \rightarrow \textbf{reject } H_0$$

Comparison AMT and Qualtrics

How many people are in your household?

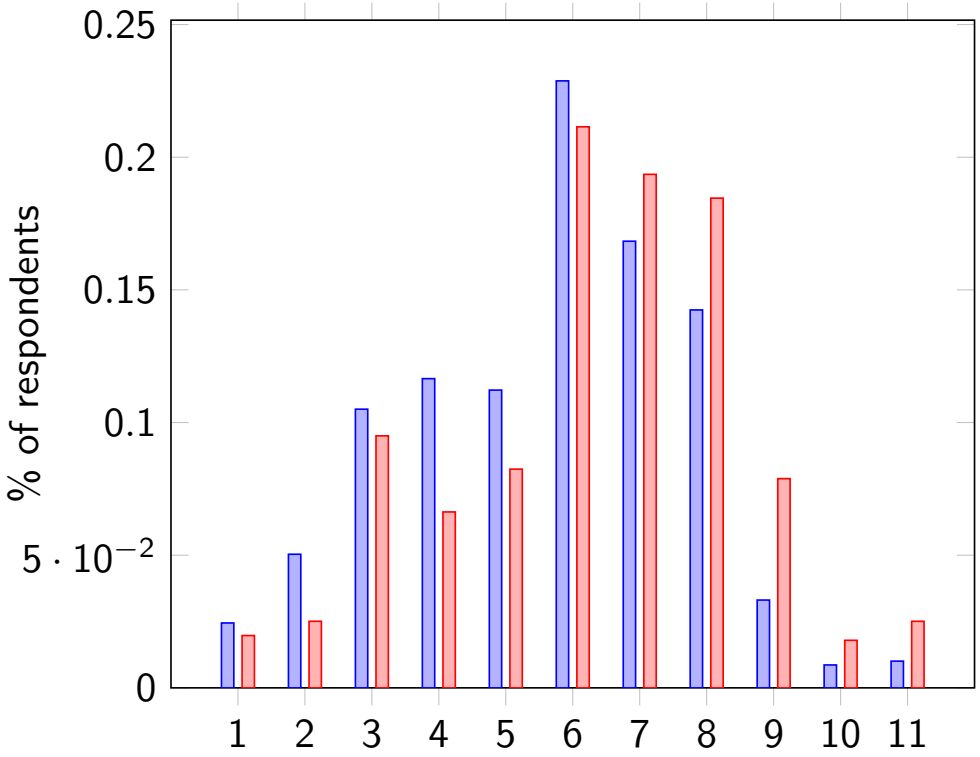


(1) 1 (alone) (2) 2 people (3) 3 people
(4) 4 people (5) 5 people (6) 6+ people

$$\chi^2 \text{ statistic} = 11.6 > \chi^2_{0.05,3} = 11.1 \rightarrow \textbf{reject } H_0$$

Comparison AMT and Qualtrics

What was your annual household income before taxes last year?



(1) <10K (2) 10-19K (3) 20-29K (4) 30-39K (5) 40-49K (6) 50-74K
(7) 75-99K (8) 100-149K (9) 150-199K (10) 200-249K (11) 250+K

$$\chi^2 \text{ statistic} = 40.0 > \chi^2_{0.05,3} = 18.3 \rightarrow \textbf{reject } H_0$$

Findings

AMT respondents are younger, lower-income, and more likely to live alone than Qualtrics

Suggests that **weighted sampling approaches** that weight as a function of socio-demographic characteristics (income, age) may provide similar results with respect to itinerary-choice model estimation

MNL Model - Model 1: Restricted Model

Variables	Estimates (t-stat)
American MTurk	-0.322 (-2.1)
Delta MTurk	-0.624 (-3.9)
United MTurk	-0.904 (-5.0)
American Qualtrics	-0.475 (-2.6)
Delta Qualtrics	-0.740 (-3.9)
United Qualtrics	-0.161 (-0.8)
Other (ref.)	0
Morning 12:00 AM-9:59 AM (ref.)	0
Afternoon 10 AM-3:59 PM	-0.324 (-4.7)
Evening 4 PM-11:59 PM	-0.895 (-9.5)
Elapsed Time	-0.008 (-10.3)
Number of Connections	-1.489 (-10.7)
Price	-0.017 (-20.9)
LL(0)	-3481.15
LL(model)	-2846.13
ρ_0^2	0.182
VOT ($/hr)	26,49

MNL Model - Model 2: Departure Time

Variables	Estimates (t-stat)
Morning 12:00 AM-9:59 AM (ref.)	0
Afternoon MTurk 10 AM-3:59 PM	-0.234 (-2.5)
Afternoon Qualtrics 10 AM-3:59 PM	-0.438 (-4.2)
Evening MTurk 4 PM-11:59 PM	-0.773 (-6.2)
Evening Qualtrics 4 PM-11:59 PM	-1.058 (-7.2)
LL(0)	-3481.15
LL(model)	-2844.48
ρ_0^2	0.183
VOT ($/hr)	26.59

Likelihood ratio statistic:

$$-2(LL_R - LL_U) = 3.3 < \chi^2_{0.05,2} = 5.99 \rightarrow \textbf{not reject } H_0$$

MNL Model - Model 3: Number of connections

Variables	Estimates (t-stat)
Number of connections Mturk	-1.600 (-9.5)
Number of connections Qualtrics	-1.365 (-7.9)
LL(0)	-3481.15
LL(model)	-2845.42
ρ_0^2	0.183
VOT ($/hr)	26.34

Likelihood ratio statistic:

$$-2(LL_R - LL_U) = 1.4 < \chi^2_{0.05,1} = 3.84 \rightarrow \textbf{not reject } H_0$$

MNL Model - Model 4: Elapsed Time

Variables	Estimates (t-stat)
Elapsed Time Mturk	-0.009 (-9.7)
Elapsed Time Qualtrics	-0.006 (-6.8)
LL(0)	-3481.15
LL(model)	-2843.85
ρ_0^2	0.183
VOT Mturk ($/hr)	29.96
VOT Qualtrics ($/hr)	22.03

Likelihood ratio statistic:

$$-2(LL_R - LL_U) = 4.6 > \chi^2_{0.05,1} = 3.84 \rightarrow \textbf{reject } H_0$$

MNL Model - Model 5: Price

Variables	Estimates (t-stat)
Price Mturk	-0.019 (-17.3)
Price Qualtrics	-0.015 (-13.0)
LL(0)	-3481.15
LL(model)	-2843.47
ρ_0^2	0.183
VOT Mturk ($/hr)	24.21
VOT Qualtrics ($/hr)	29.83

Likelihood ratio statistic:

$$-2(LL_R - LL_U) = 5.3 > \chi^2_{0.05,1} = 3.84 \rightarrow \textbf{reject } H_0$$

MNL Model - Model 6: Elapsed Time and Price

Variables	Estimates (t-stat)
Elapsed Time Mturk	-0.009 (-9.9)
Elapsed Time Qualtrics	-0.006 (-6.0)
Price Mturk	-0.019 (-17.3)
Price Qualtrics	-0.014 (-12.1)
LL(0)	-3481.15
LL(model)	-2838.69
ρ_0^2	0.185
VOT Mturk ($/hr)	28.06
VOT Qualtrics ($/hr)	23.62

Likelihood ratio statistic:

$$-2(LL_R - LL_U) = 14.9 > \chi^2_{0.05,2} = 5.99 \rightarrow \textbf{reject } H_0$$

Next steps

- Compare these results to the ones obtained using revealed preference data (my PhD thesis)
- Determine if unweighted estimation lead to similar results
- Analyze if AMT and Qualtrics respondents have the same behavior regarding the use of search and filter tools
- Develop a choice set generation model and compare results for AMT and Qualtrics

Q&As


merci
thank you
gracias
obrigado
danke
go raibh maith agat
спасибо
grazie
謝謝
dziękuję
sukriya
bedankt
arigatô
ngiyabonga
dakujem
mochchakkeram
terima kasih
sagolun
chokrane
taiku
teşekkür ederim
mauruuru
hvala
leat
takk
tapadh
krap
dank je
chnorakaloutioun
kia ora
ありがとう
grazzi
мерси
barka
tau
mersi
diolch
shukriya
spas
dhanyavad
teşekkü ederim
welalin
enkosi
gratias ago
misaotra
matondo
dhanyavadagalu
kop khun
najis tuke
mamnun
paldies
faafetai lava
kam sah hamnida
nanni
trugarez
asante
Баярлалаа
vinaka
ευχαριστώ
xièxie
dankie
kiitos
sobodi
dankon
manana
murakoze
tenki
rahmet
kop khun krap
mercé
nandri
didi madloba
mahalo
köszönöm
хвала
blagodaram
grăcie
mësi
gracias
ačiū
akun
sulpáy
tack
bayarlalaa
djiere dieuf
tanemirt
dëkuji
rahmat