Alaska

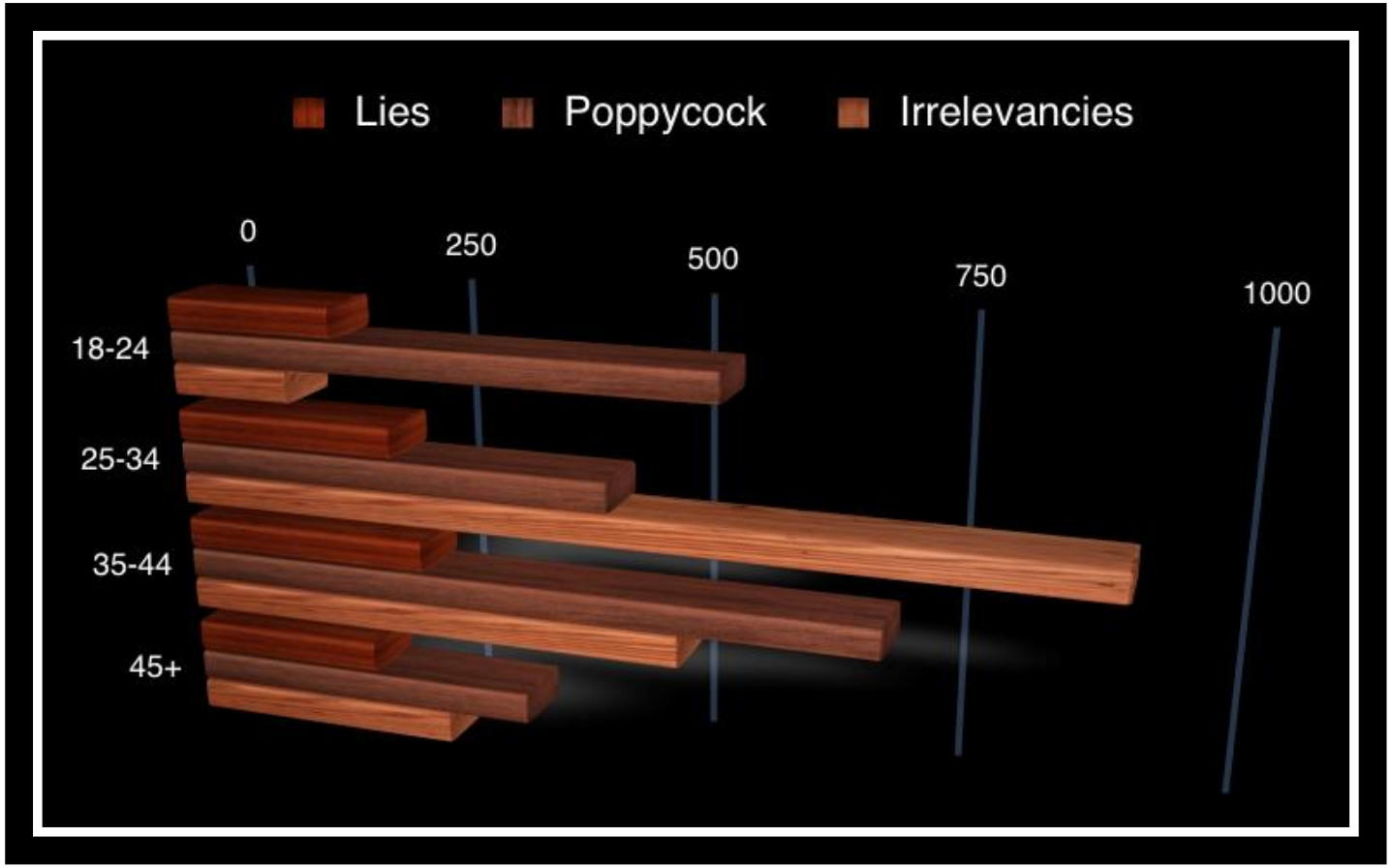
Survey Design

Jeff Sauro, PhD
Jim Lewis, PhD

Measuring U

CUSTOMERS · EMPLOYEES · SUPPLIERS
People power the internet.™
PEOPLE
Soft

Do Surveys Always Suck?



Source: Erika Hall: On Surveys

When a Survey is a Better Research Method

1. **Identifying your users**:
   age, gender, purchase frequency, product experience, geography.

2. **Identifying the most important features or content**:
   Top-task analysis.

3. **Benchmarking attitudes/Creating Standardized Metrics**:
   Do people trust us, think the experience is usable? Attitudes affect/predict/explain behavior.

4. **Identifying key drivers of a product or experience**:
   Why do people have low trust, satisfaction?

5. **Finding the likelihood to repurchase or recommend**:
   Intent is a good predictor of behavior.

What makes a good survey?

Surveys are artifacts designed for human use

- Consider human factors fundamentals
- Avoid excessive length
- Avoid unfamiliar terms
- Write clearly
- Use appropriate types of items
- Design for all platforms respondents might use
- Don’t forget usability and pilot testing

Ways to get a **terrible** response rate

1. Please rate how well you agree with the following statments

	Strongly Disagree 1	2	3	4	5	6	Strongly Agree 7
This website is easy							
This website is attractive							
This website meets my needs							
This website was fast							
This website was modern							
This website is what I need it to be							
This website makes me want to return							
This website makes me want to shop with them							
I would like to use this website on my mobile phone							
This website communicates with me							
I think this website would be easy to use on my mobile phone							
I had no questions while using this website							
I would not need any help when coming to this website for the first time							
I found everything I needed to find							
This website is clean and simple							
I will often return to this website							
I would recommend this website to a friend							
I would visit this website to kill time							
My experience on this website was positive							
This website has a great selection of items I would like to buy							
This website is for people who shop like me							
This website is for a brand I can trust							
This website meets my needs							
This website was fast							
This website was modern							
This website is what I need it to be							
This website makes me want to return							
This website makes me want to shop with them							
I would like to use this website on my mobile phone							
This website communicates with me							
I think this website would be easy to use on my mobile phone							
I had no questions while using this website							
I would not need any help when coming to this website for the first time							
I found everything I needed to find							
This website is clean and simple							
I will often return to this website							
I would recommend this website to a friend							
I would visit this website to kill time							
My experience on this website was positive							
This website has a great selection of items I would like to buy							
This website is for people who shop like me							
This website is for a brand I can trust							

- Make it feel like a test – lots of open-ended items
- Impossible screener
- Giant matrices
- No sense of progress
- Ask the really personal questions first
- Ask irrelevant questions without an N/A option
- Don’t think about respondent incentives

Writing survey questions

Good research questions for surveys

- How many?
- Demographics
- Attitudes
- Perceptions
- Expectations

Classes of survey questions

- **Open-ended:**
  Coded into variables
- **Close-ended (static):**
  Multiple choice, ratings
  – always same response options
- **Close-ended (dynamic):**
  Adaptive conjoint, MaxDiff, item response theory
  – items or response options differ depending on previous responses
- **Task based:**
  Unmoderated usability study

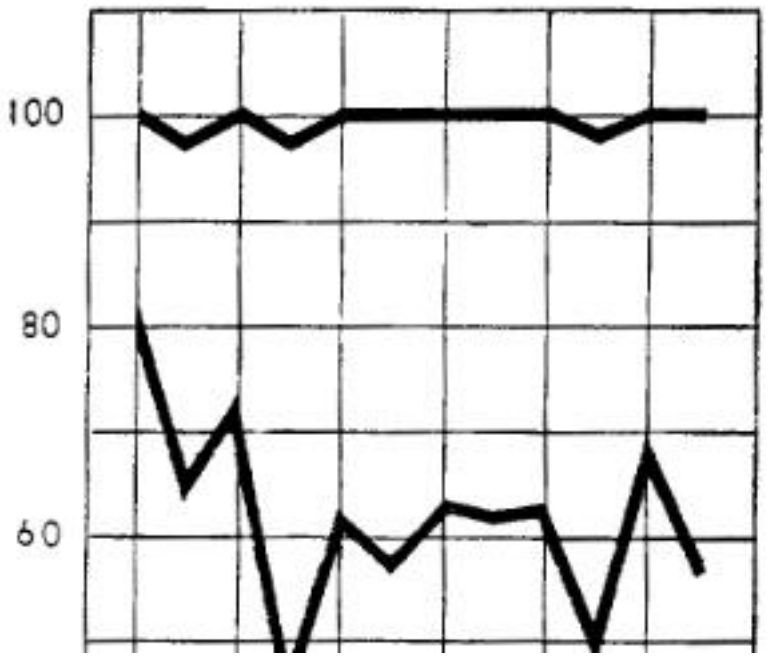
What **not** to include in a survey

- *Overall, how satisfied are you with the cost and quality of the product?*
  Double-barreled – what is rated?

- *Don't you agree that employers should be required to give paid time off?*
  Leading

- *Age range: 10-20; 20-30; 30-40; 40-50; 60+*
  Incomplete list of options, overlapping ranges

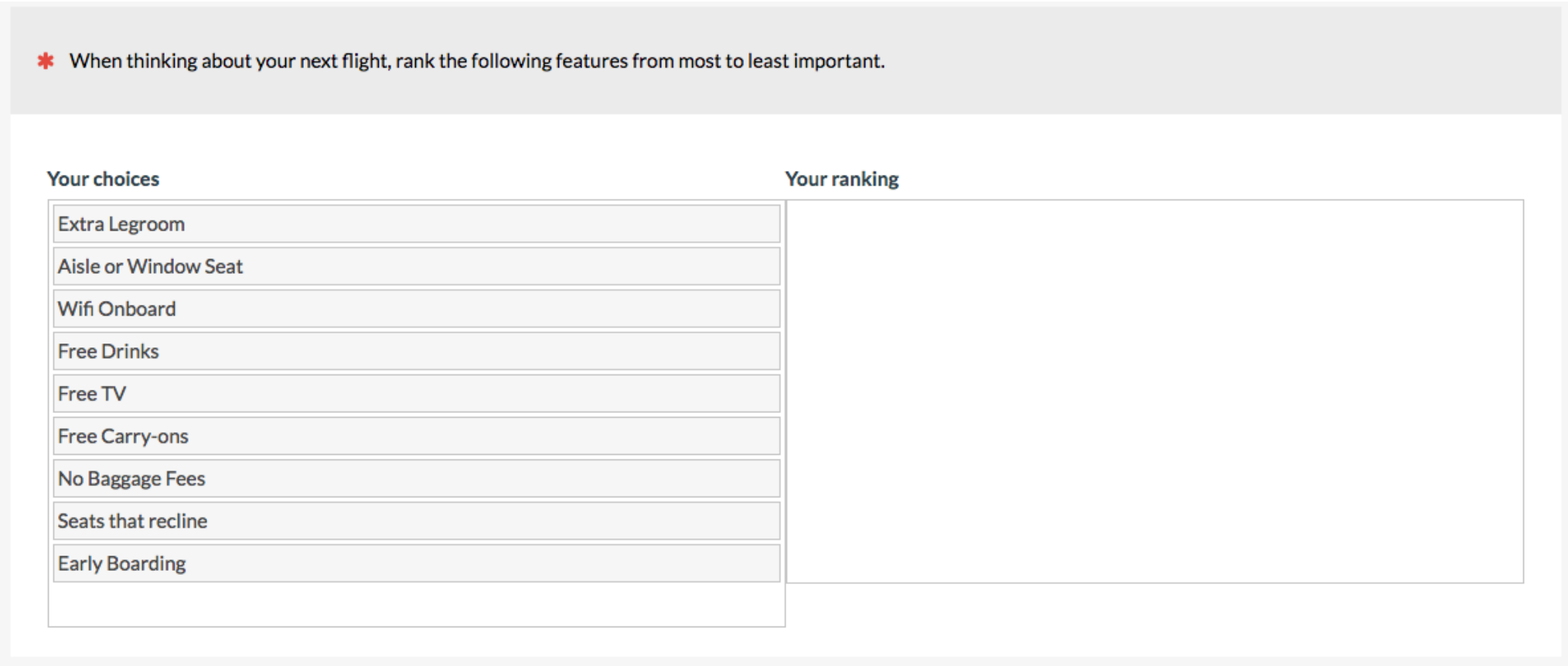
Identifying potential bias



- Social desirability and conformity
- Yea saying and acquiescence
- Order effects
- Prestige
- Threat and hostility
- Sponsorship
- Stereotype

Types of Questions/Items

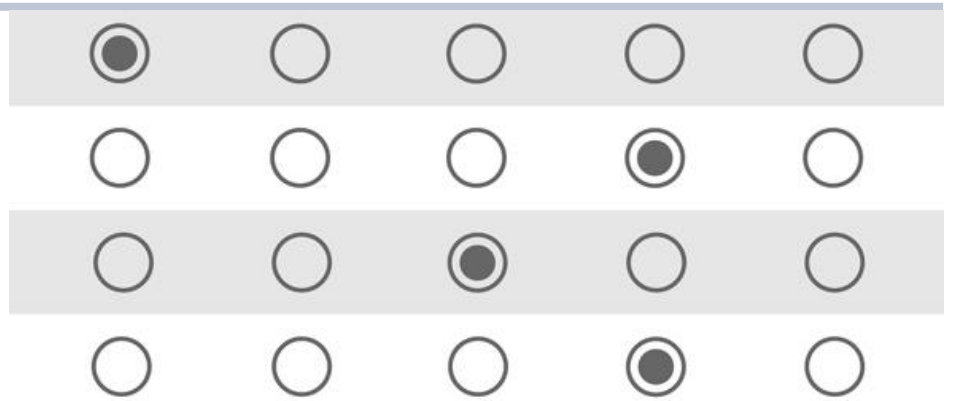
Multiple choice vs forced choice



Open-ended responses

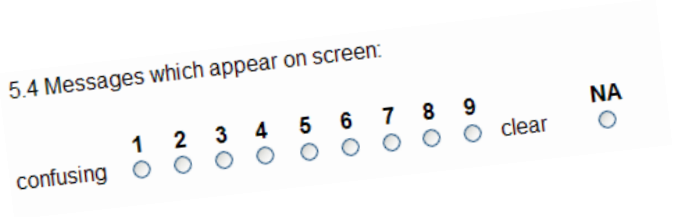
- Avoids limitations of closed-ended items
- No substitute for reading content
- Quantify using manual coding/classification
  - Current automation inadequate
- To code and classify:
  - Sort and clean
  - Group into themes
  - Split and combine as needed

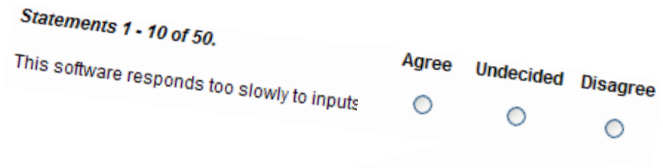
Grids and matrices

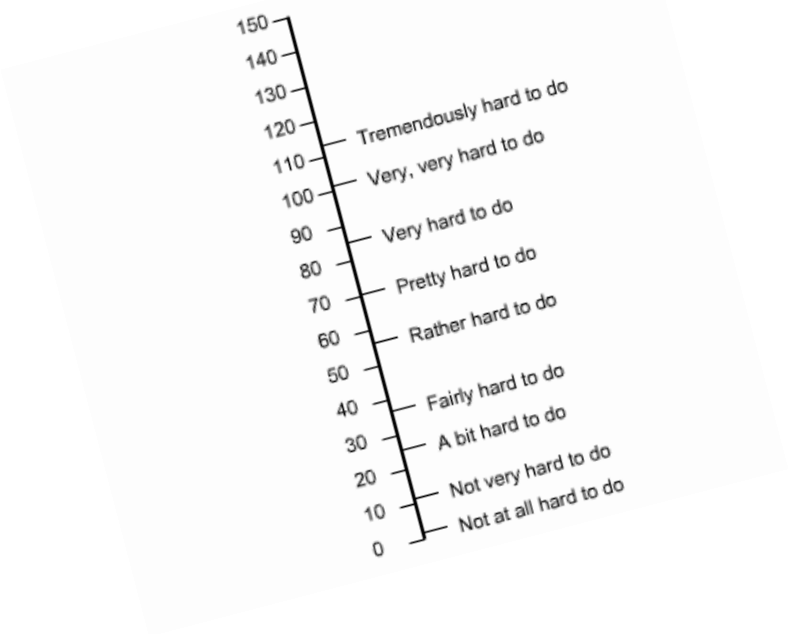


- Faster to complete than individual items
- More straightlining and nonresponse
- Little effect on scores and distributions
- May be some effect on correlation and loading
- Respondents dislike large grids but actually prefer smaller ones

Rating scales: Standardization > format issues



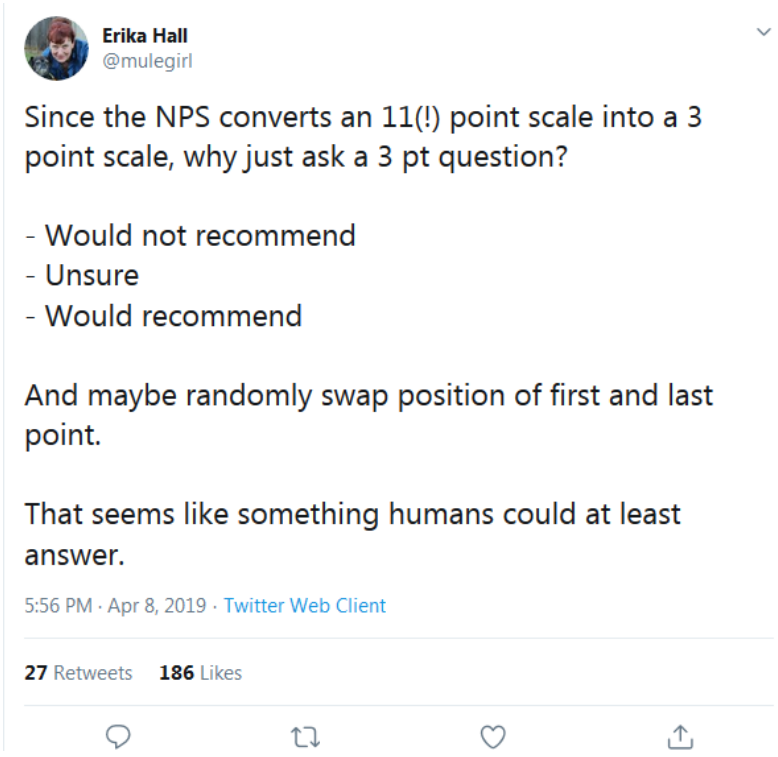




- Number of points: More than 3, usually 5-7, 11
- Labels and anchors: Little effect on responses using common formats (1-7, 0-10)
- Neutral point: Provide unless strong reason
- Positive/negative wording: Prefer positive
- Survey items in grids vs alone: Avoid large grids

Sauro corollary to Parkinson's law of triviality

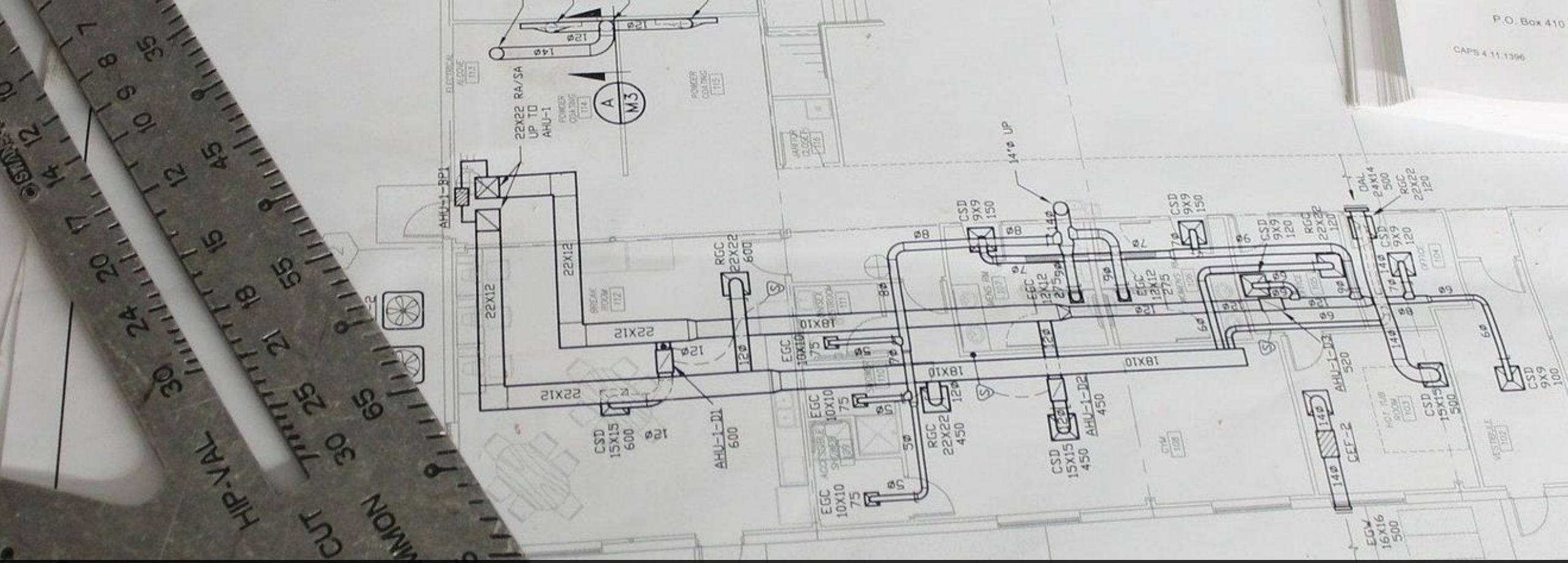
3 Point scale superiority is a myth.



"Thumbs up, thumbs down is simple. 3-point scales are simple. If you feel strongly about a 4 or 5-point scale, go for it. Not more. If you measure very specific experiences, it is very easy for a user to decide if they are happy or not." *Tomer Sharon*

- More scale points increase reliability (3 are unreliable)
- There's a loss of **intensity** and **validity** with three-point scales

See: *Is a Three-Point Scale Good Enough?*



Survey Setup

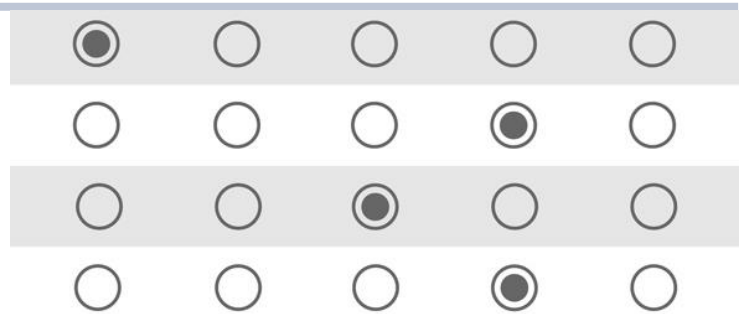
Introduction page

- Include title of survey
- Thanks for agreeing to participate
- Estimated completion time
- Purpose of survey
- Not a test of you – you're helping us
- Contact info in case there's a problem
- Collect any data needed to continue

Formatting

- Include progress bar (if possible)
- Limit number of questions per page
- One-per-page can increase completion time
- Biggest drop-out factor is length

More formatting



- **Counterbalancing/randomization**:

  Apply to response options, items, and tasks when appropriate

- **Use of bolding**:

  Don’t overuse, but consider bolding key instructions so they stand out from the rest of the text

- **Color and backgrounds**:

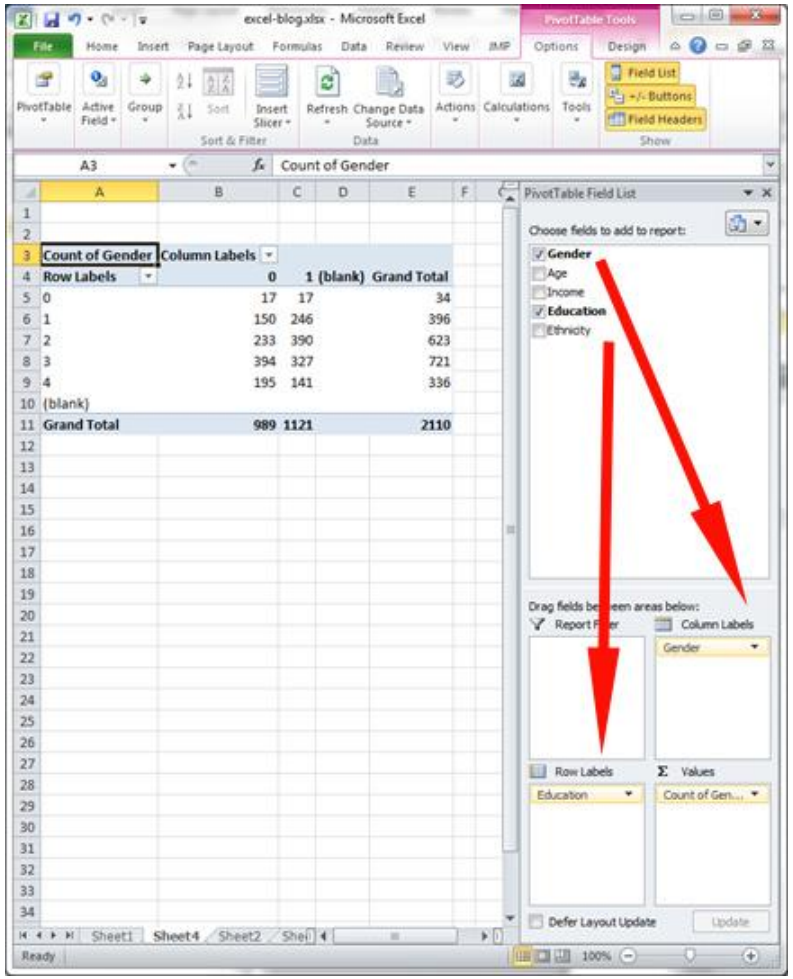
  Be legible and be consistent, and for periodic surveys it will be easier to be consistent if you avoid using nonstandard colors or backgrounds

Best practices in length of survey

- Best predictor of abandonment is length
- Acceptable length depends on delivery method
  - Shortest for intercept surveys
  - Longer for email surveys
- Median time for MeasuringU surveys ~5 min
- Multiple short surveys usually better than one long one

Data Analysis

Creating pivot tables

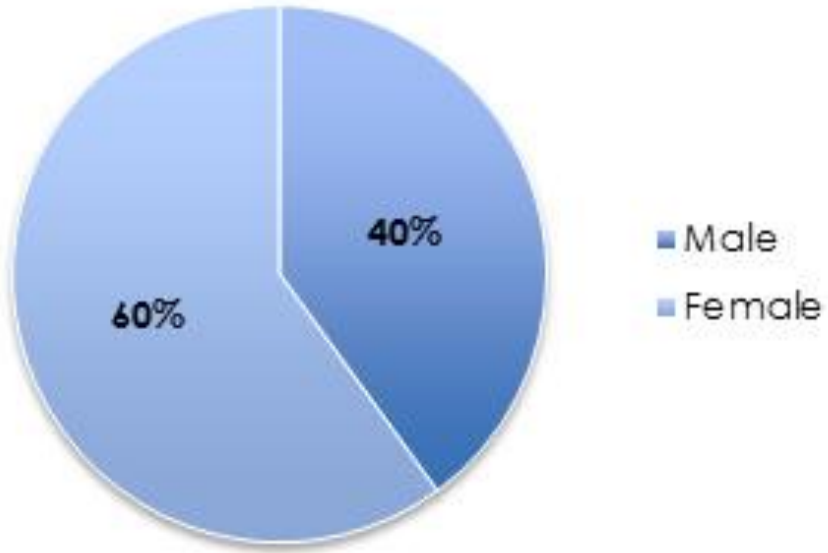


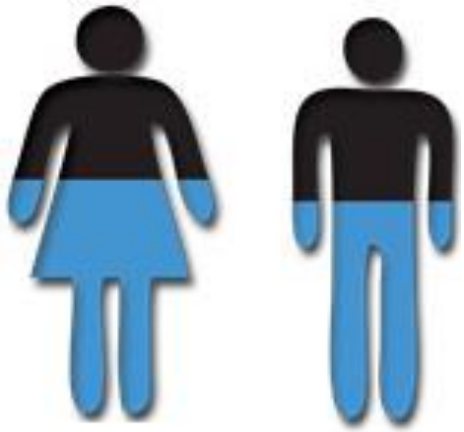
- Powerful tools for exploring data
- Easy way to set up cross tabulation
- Click Pivot Table on Excel ribbon
- Select columns that contain the data
- Labels at top, no spaces
- Drag/drop desired row/column labels
- Modify operations if desired (average, sum)

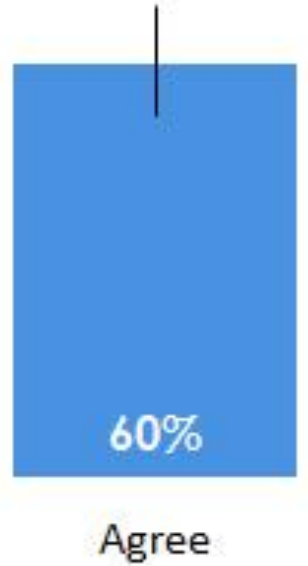
Displaying and summarizing survey data

Binary responses:

pie, USA Today, bars with CI



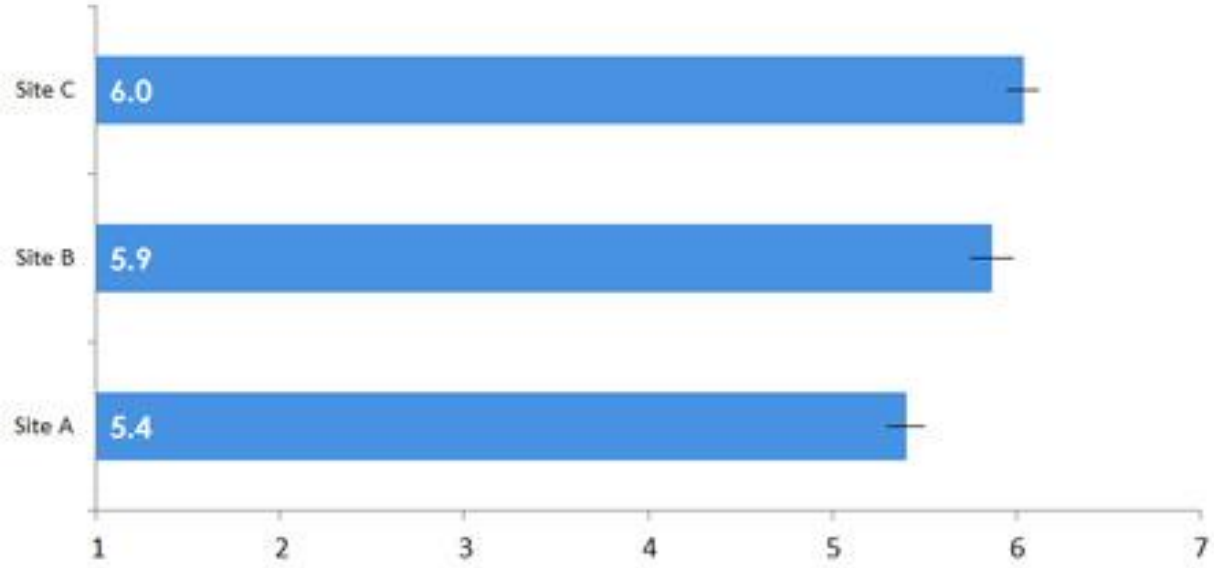




Displaying and summarizing survey data

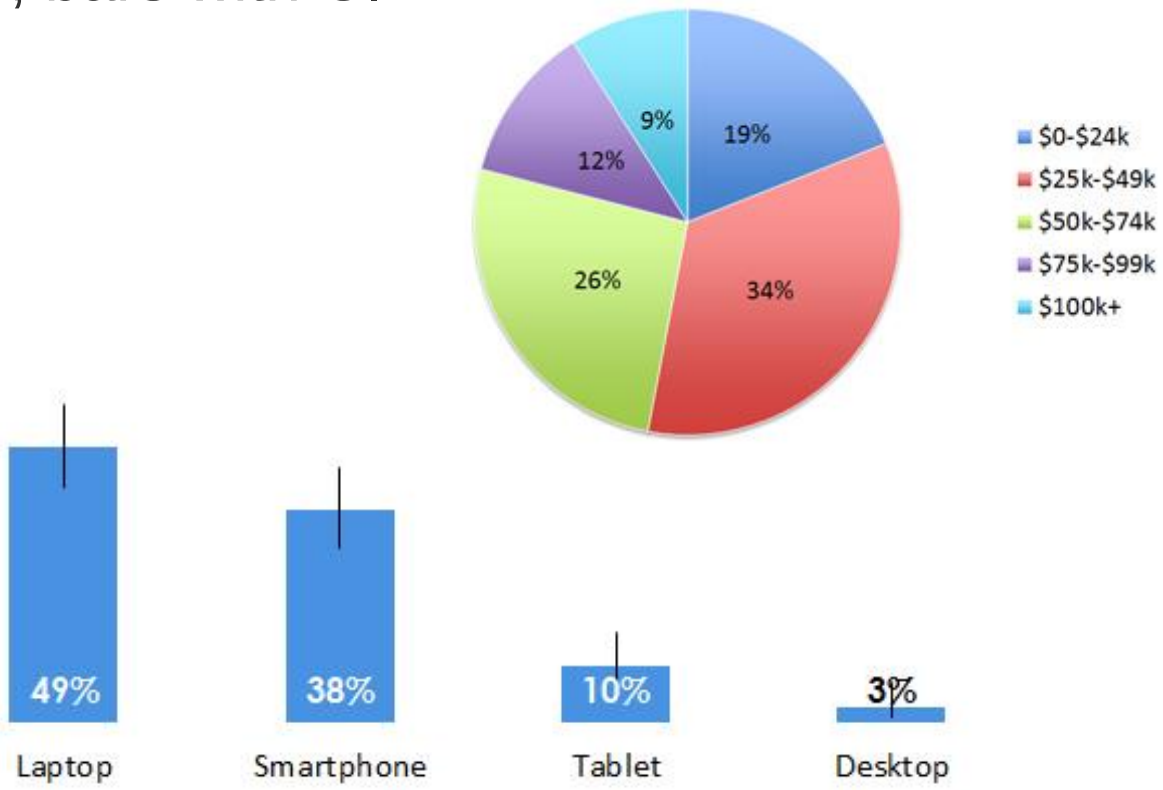
Rating scales:

mean, top box, net, bars with CI

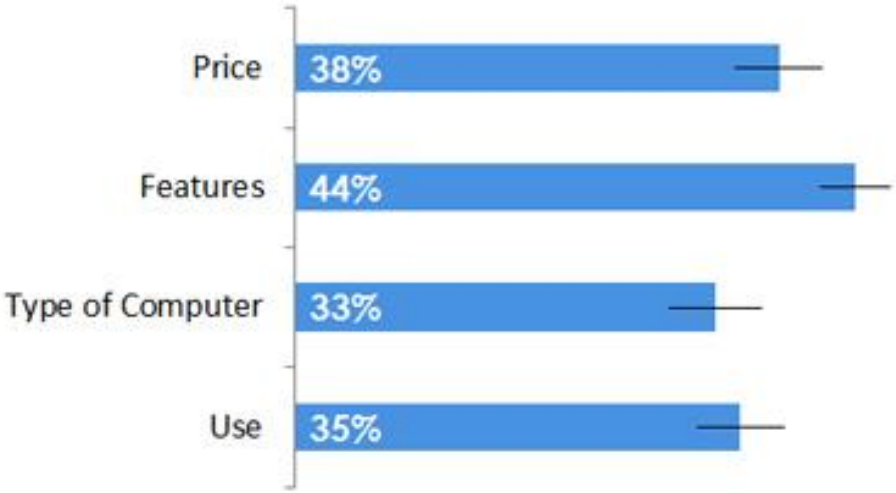


Displaying and summarizing survey data

Single select:
pie, bars with CI



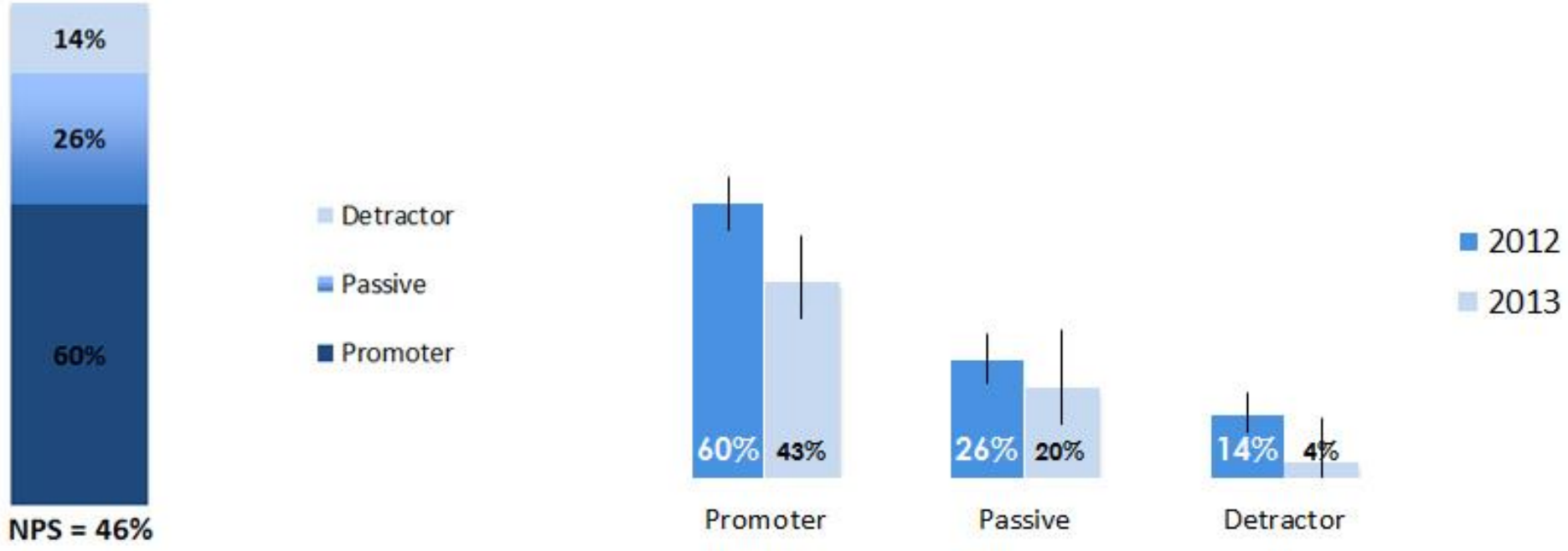
Multiple select:
bars with CI



Displaying and summarizing survey data

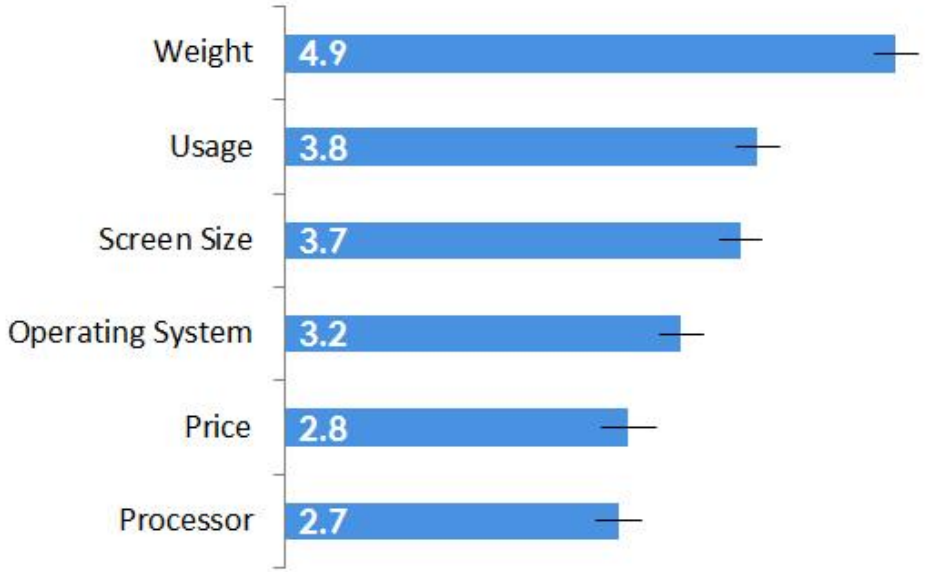
Likelihood-to-recommend:

mean, top box, net, bars with CI

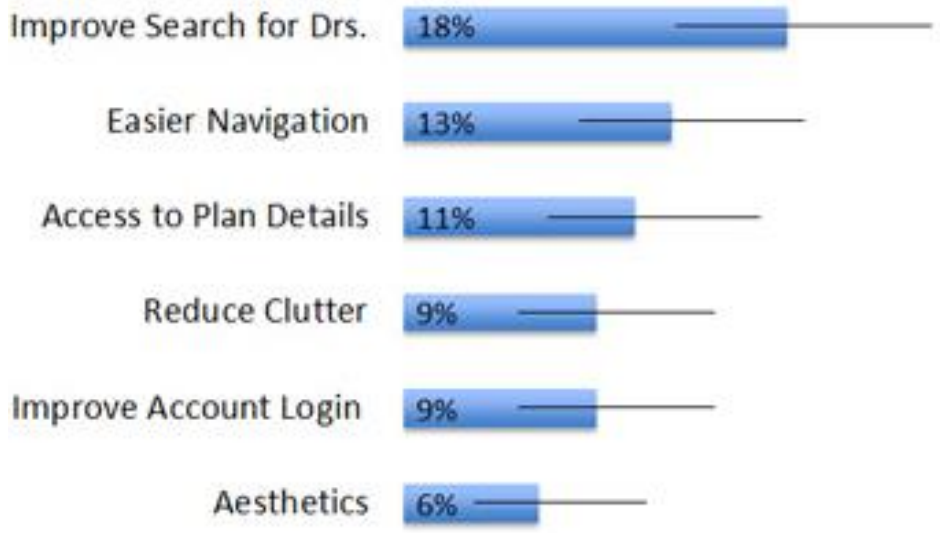


Displaying and summarizing survey data

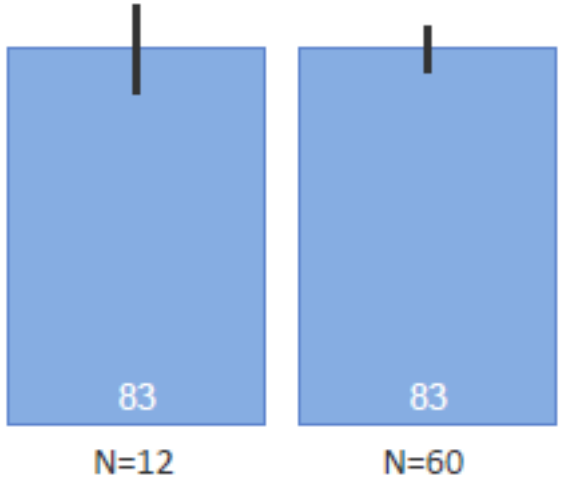
Forced rank:
mean, CI



Open-ended:
categorize, bars with CI



Using confidence intervals



- CI show precision and location
- 3 ingredients: confidence level, variability, n
- Produce plausible range given data in hand
- Intervals wider with higher confidence
- Intervals wider with higher variability
- Intervals narrower with higher n
- Different methods for means and percentages

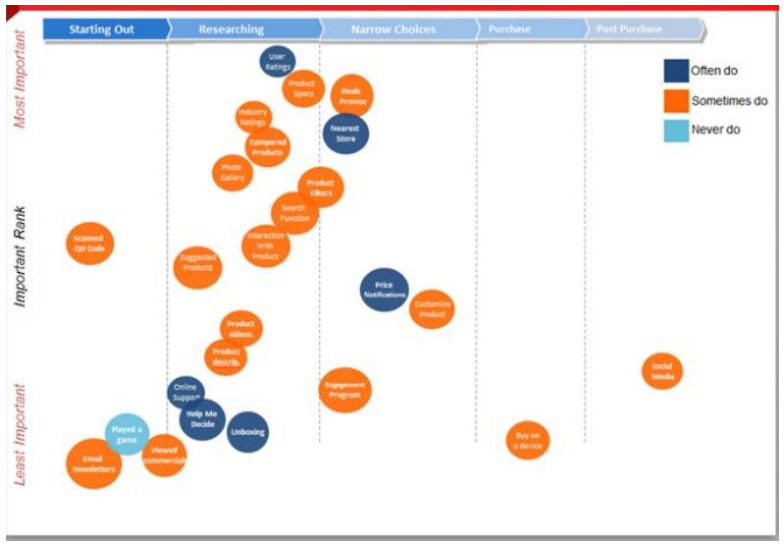
Analyzing preference data

- Binomial test with confidence interval
- Chi-square goodness of fit test
- Comparison of binomial confidence intervals
- McNemar exact test
- Rank tests (Friedman, Kruskal-Wallis)

Understanding brand lift and drag

- A way to understand how particular experiences impact attitudes
- Comparison of pre and post ratings
- Lift can be positive or negative
- When negative, also known as “drag”

Examining variables – cross tabbing



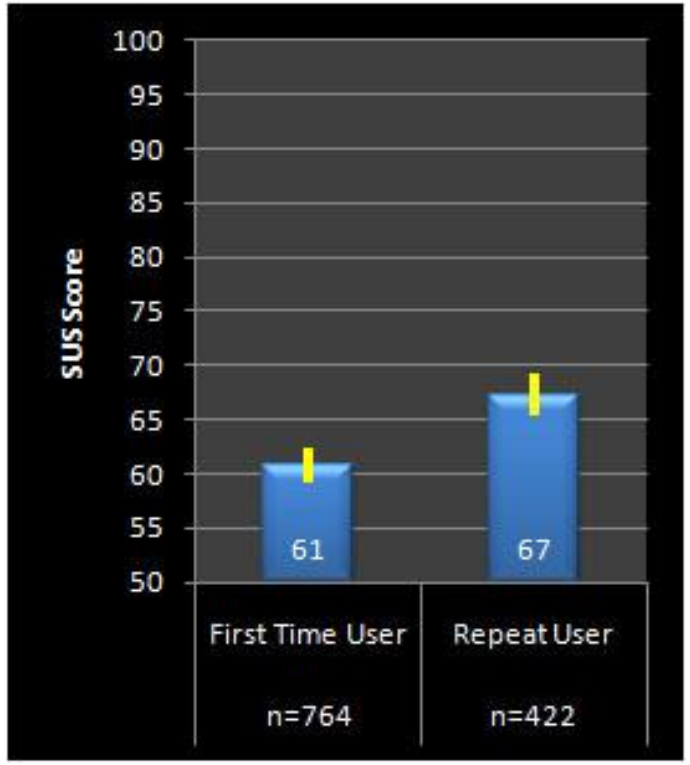
- Process of examining more than one variable in the same table or chart ("crossing" them)
- Used to identify interesting clusters
- Commonly done with observed variables to illustrate group similarities and differences
- Can be part of the analysis of open-ended responses after classification

Collapsing variables

	Range	Code	X	N	P	AdjX	Adj N	Ac
Much more likely	B3:B118	0	36	116	31%	37.3527717	118.705543	0.314
	D3:D118	0	35	116	30%	36.3527717	118.705543	0.306
	F3:F113	0	29	111	26%	30.3527717	113.705543	0.266
	H3:H109	0	33	107	31%	34.3527717	109.705543	0.313
	J3:J114	0	38	112	34%	39.3527717	114.705543	0.343
Somewhat more likely						27717	118.705543	0.323
						27717	118.705543	0.348
						27717	113.705543	0.416
						27717	109.705543	0.431
						27717	114.705543	0.43
Would have no effect						27717	118.705543	0.348
						27717	118.705543	0.323
						27717	113.705543	0.266
						27717	109.705543	0.203
						27717	114.705543	0.177
Somewhat less likely						77173	118.705543	0.028

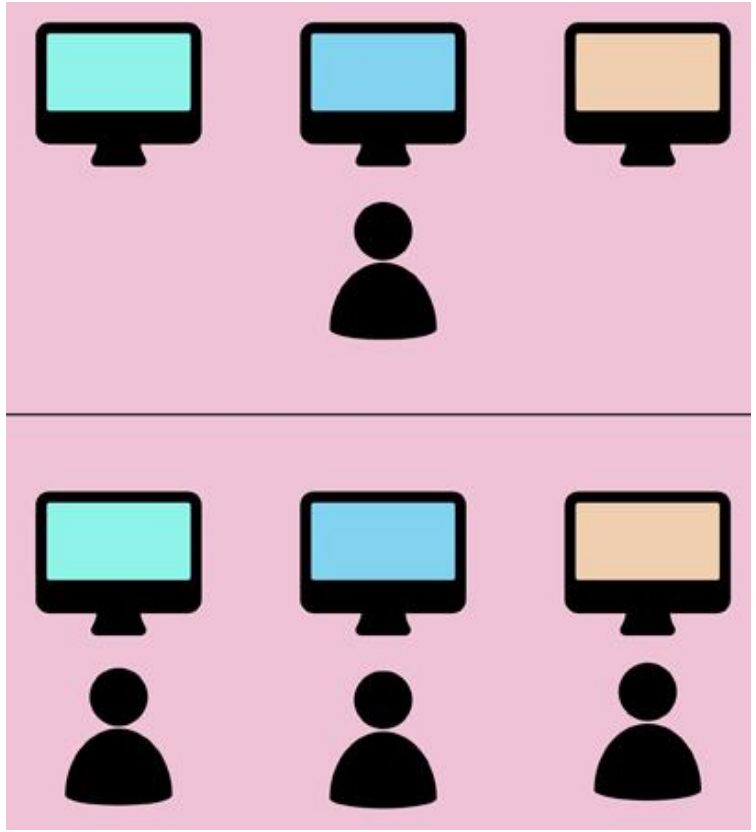
- Collapsing can simplify research questions
- Convert multiple frequency options into Low/High
- Convert multiple loyalty membership options into Member/Nonmember
- Convert 11-point scale to 3 categories (NPS)
- More advanced: principal components and factor analysis to combine multiple items into a single score

Controlling for prior experience



- Prior use tends to lift measures of perceived usability (replicated numerous times)
- With three categories of years of experience (<3, 4-5, >5) each level had 5% higher SUS
- Important to track prior experience
- Rule of thumb – if can't control for prior experience, use a 6-15% score adjustment
- More advanced: Analysis of covariance

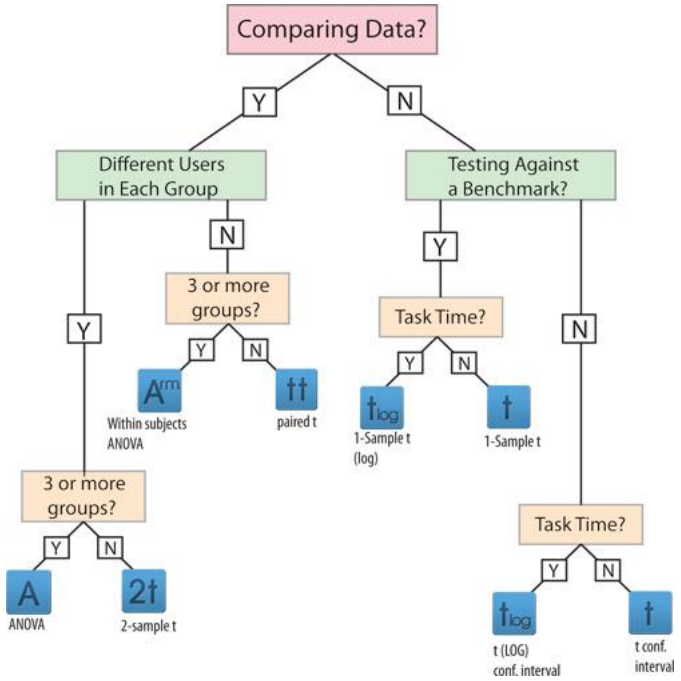
When to use between- vs within-subjects designs



Factor to Consider	Between	Within
Sample Size & Power	-	▼
Carryover Effects	▼	-
Impact on Attitudes	▼	-
Comparative Judgment	-	▼
Study Duration	▼	-

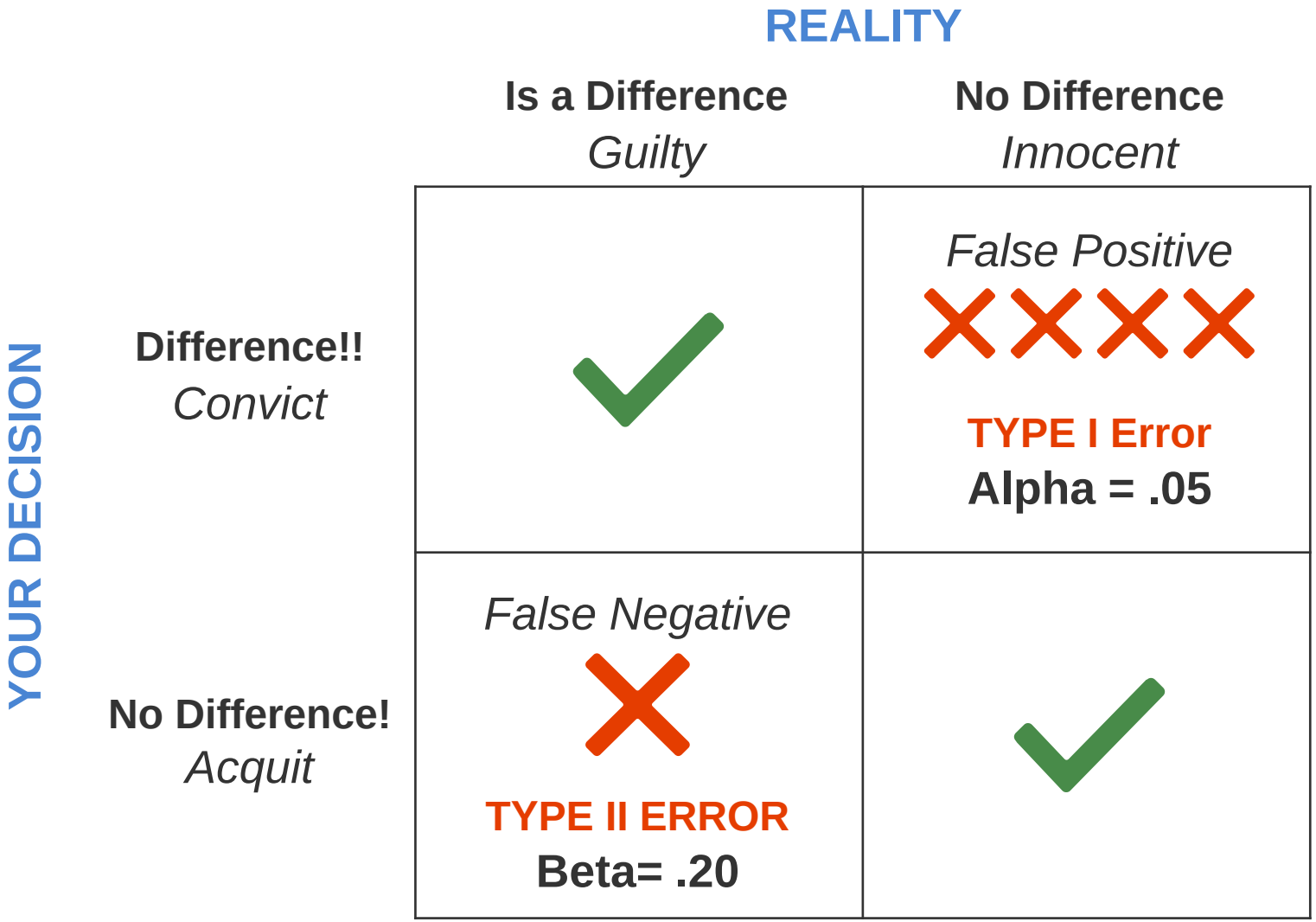
Compromise – all participants get baseline design with one alternate

Conducting statistical tests and interpreting *p*

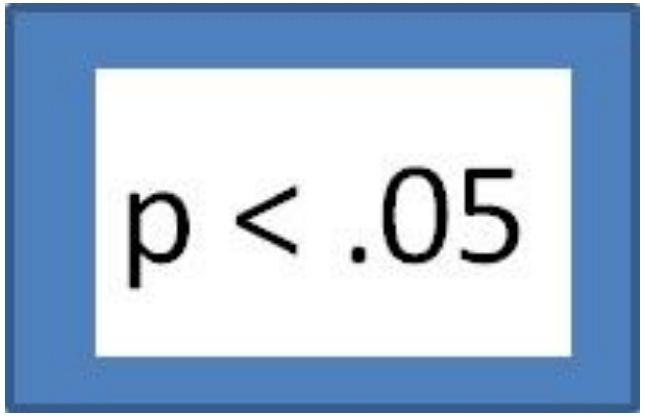


- Many different statistical tests
- Appropriateness depends on type of data and question being asked
- All produce a value for *p*
- The decision regarding statistical significance is binary and depends on whether the *p*-value is greater or lower than a criterion
- Most common criterion is $p < .05$

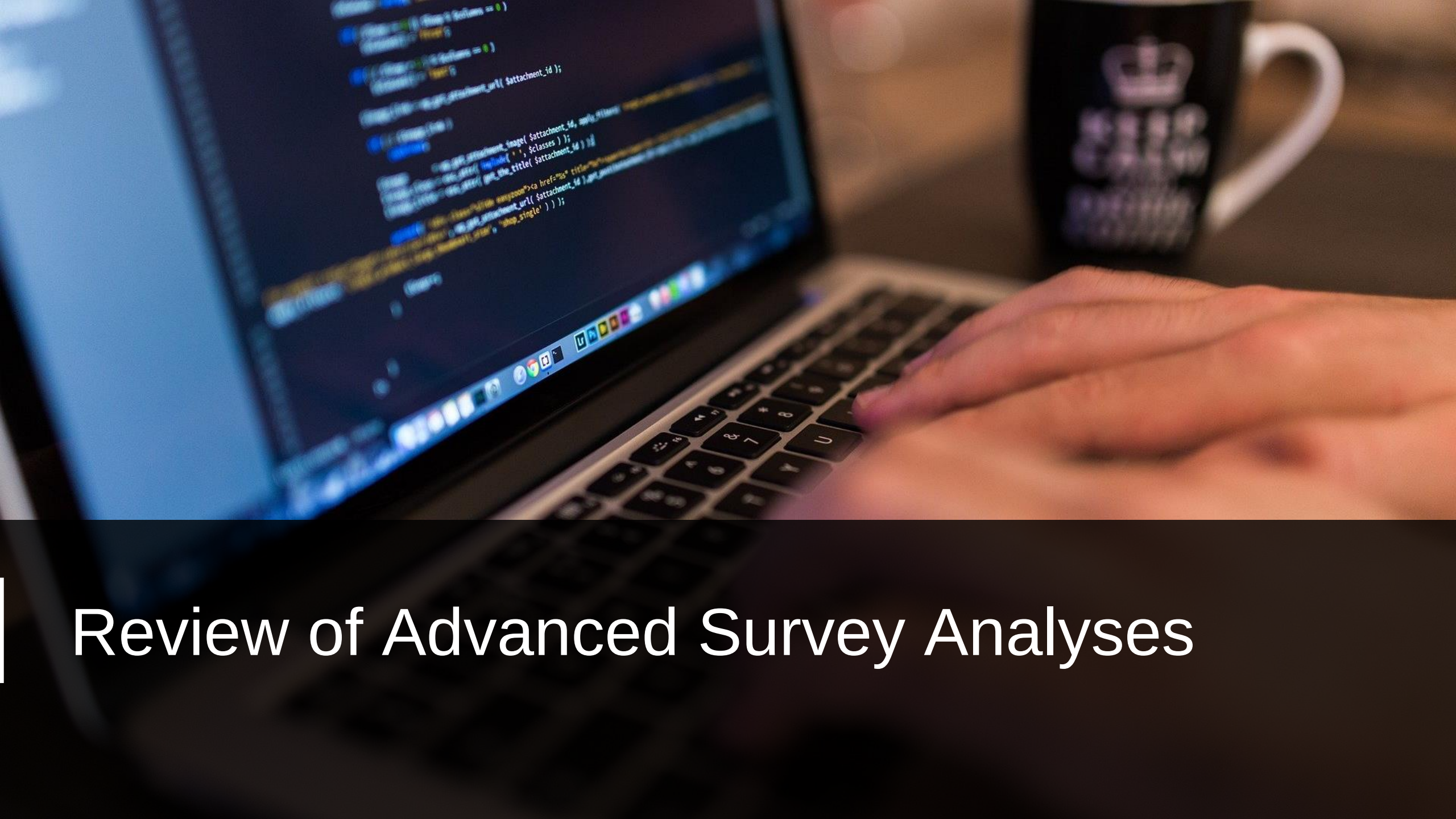
Hypothesis testing errors



What does statistically significant mean?

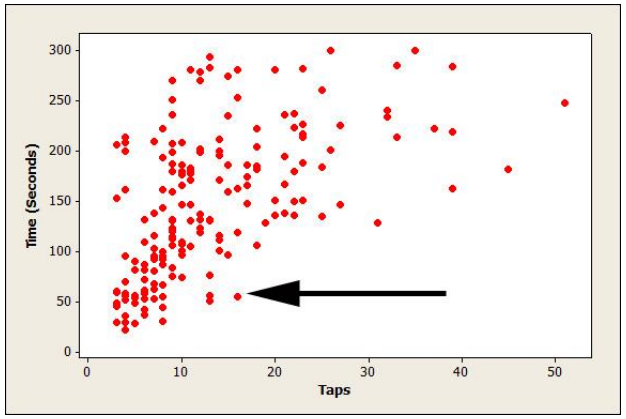


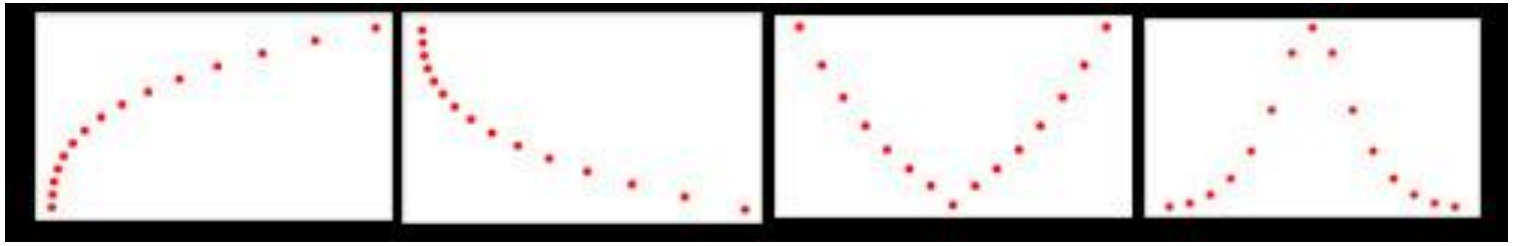
- Probably not due to chance
- Probability of getting this or a more extreme result if there really is no difference
- Sampling error = never absolutely certain
- Most exact way of saying maybe
- Setting p to .05 is a scientific convention, not necessarily best criterion in industrial studies
- Best practice to compute CI around differences to show plausible range – moves from just statistical to practical significance

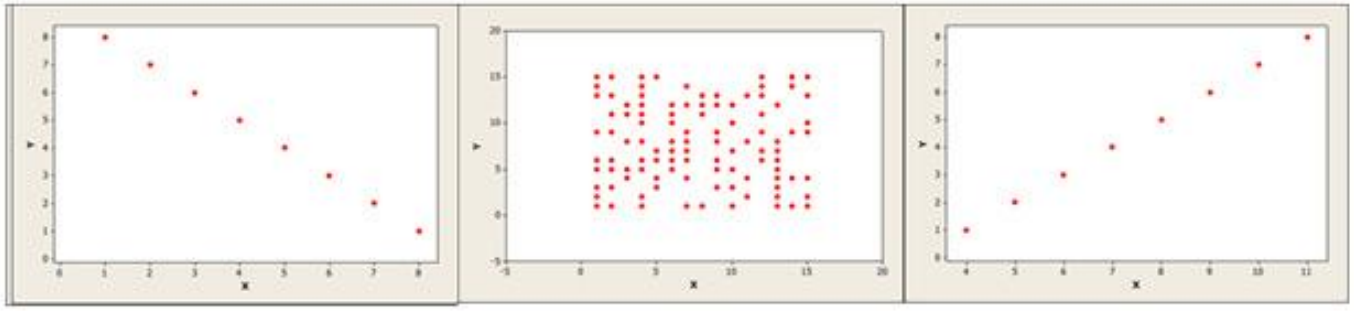


Review of Advanced Survey Analyses

Understanding Correlations







Interpreting Correlations

Description	Correlation
Aspirin and reduced risk of heart attack	0.02
Ever Smoking and Lung Cancer after 25 years	0.08
College Grades and Job Performance	0.16
Years of Experience & Job Performance	0.18
SAT Scores and Cumulative GPA at University of Pennsylvania for (White & Asian Students)	0.20
HS Class Rank and Cumulative GPA at University of Pennsylvania for (White & Asian Students)	0.26
Psychotherapy and Subsequent Well Being	0.32
Raw Net Promoter Scores and Future Firm Revenue Growth in 14 Industries	0.35
GRE Quantitative Reasoning and MBA GPA	0.37
Unstructured Job Interviews and Job Performance	0.38
Viagra and improved sexual functioning	0.38
Height and Weight from 639 Bangladeshi Students (Average of Men and Women)	0.38
Past Behavior as Predictor of Future Behavior	0.39
% of Adult Population that Smokes and Life Expectancy in Developing Countries	0.40
College Entrance Exam and College GPA in Yemen	0.41
SAT Scores and Cumulative GPA from Dartmouth Students	0.43
Height and Weight in US from 16,948 participants	0.44
NPS Ranks and Future Firm Revenue Growth in 14 Industries	0.44
Rorschach PRS scores and subsequent psychotherapy outcome	0.44

Description	Correlation
Intention to use technology and actual usage	0.50
General Mental Ability and Job Performance	0.51
Purchase Intention and Purchasing Meta Analysis (60 Studies)	0.53
Work Sample and Job Performance	0.54
PURE Scores From Expert and SUPR-Q Scores from Users	0.55
PURE Scores From Expert and SEQ Scores from Users	0.67
Likelihood to Recommend and Recommend Rate (Recent Recommendation)	0.69
SUS Scores and Future Software Revenue Growth (Selected Products)	0.74
Purchase Intent and Purchase Rate for New Products (n=18)	0.75
SUPR-Q quintiles and 90 Day purchase rates	0.78
Likelihood to Recommend and Recommend Rate (Recent Purchase)	0.79
PURE Scores From Expert and Task Time Scores from Users	0.88
Accuracy of Pulse Oximeter and Oxygen Saturation	0.89
Likelihood to Recommend and Reported Recommend Rate (Brands)	0.90

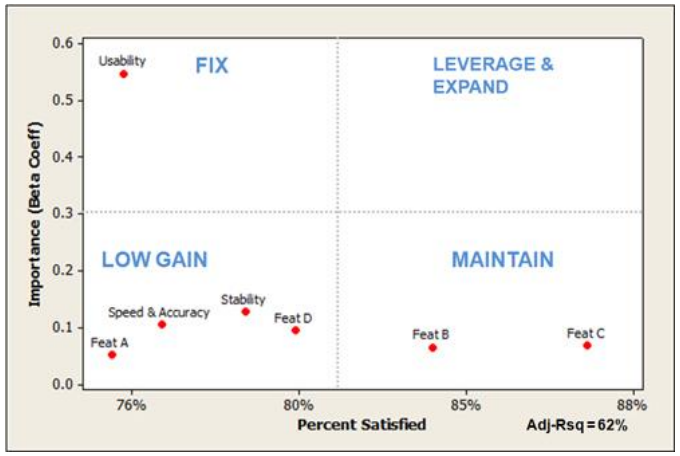
r = .10 small
r = .30 medium
r = .50 large

Context Matters Most

5 Advanced Stats Techniques & When to Use Them

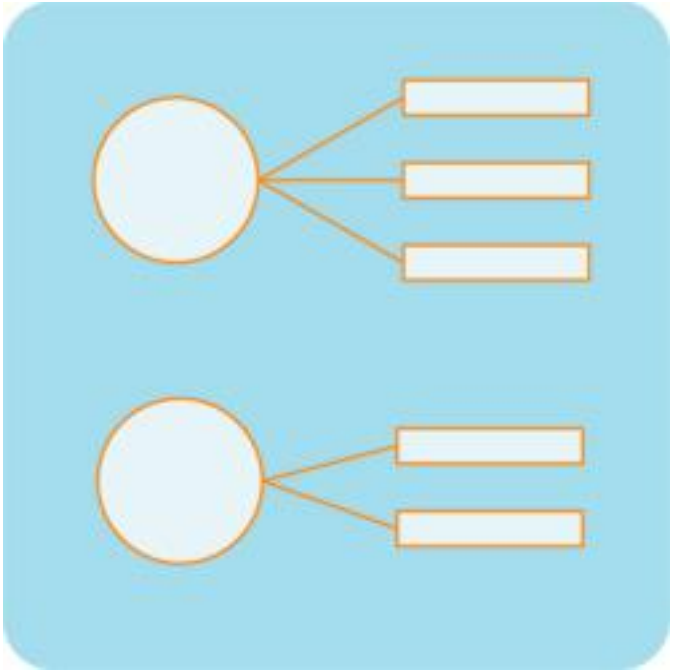
Technique	Focus	Research Example	Gotcha
Regression Analysis	Best combination of variables to predict continuous outcome variables	What features best predict likelihood to recommend?	Correlated independent variables and linearity
ANOVA	Comparison of multiple variables plus interactions	How does device type and form type affect task completion time?	Alpha inflation
Factor Analysis	Identify latent variables that form groups (factors)	What combination of items describes the constructs of appearance and trust?	Subjective factors & linearity
Cluster Analysis	Uncover how items form latent groups	Do users group two products together?	Subjective clusters
Logistic Regression	Best combination of variables to predict discrete outcome variables	How much does a service experience and tenure affect purchase rates?	Correlated independent variables and linearity

Regression/Key driver analysis



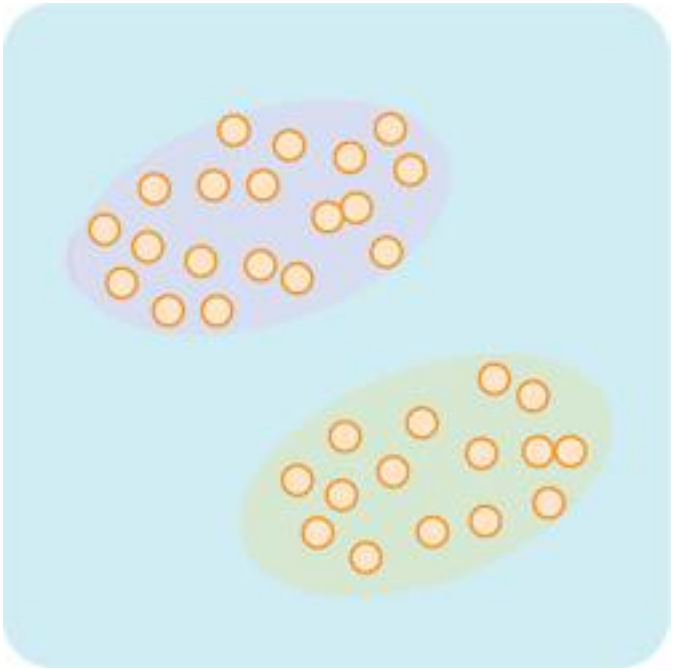
- Relative importance of predictors on outcomes
- Based on output from multiple linear regression
- Importance derived from standard beta weights
- Two key visualizations
- 2x2 matrix based on impact and intensity
- Rectangle chart of percentage explained variability

Factor analysis



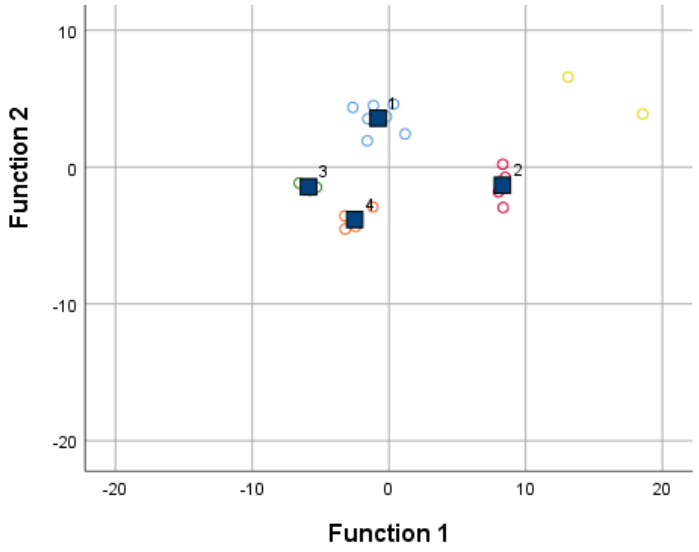
- Variety of methods for identifying latent structure in data
- Takes many observed variables and shows alignment of observed variables with factors
- Some subjectivity involved in determining number of factors
- Commonly used for development of standardized questionnaires
- Quantitative basis for combination of item ratings into scale scores

Cluster analysis

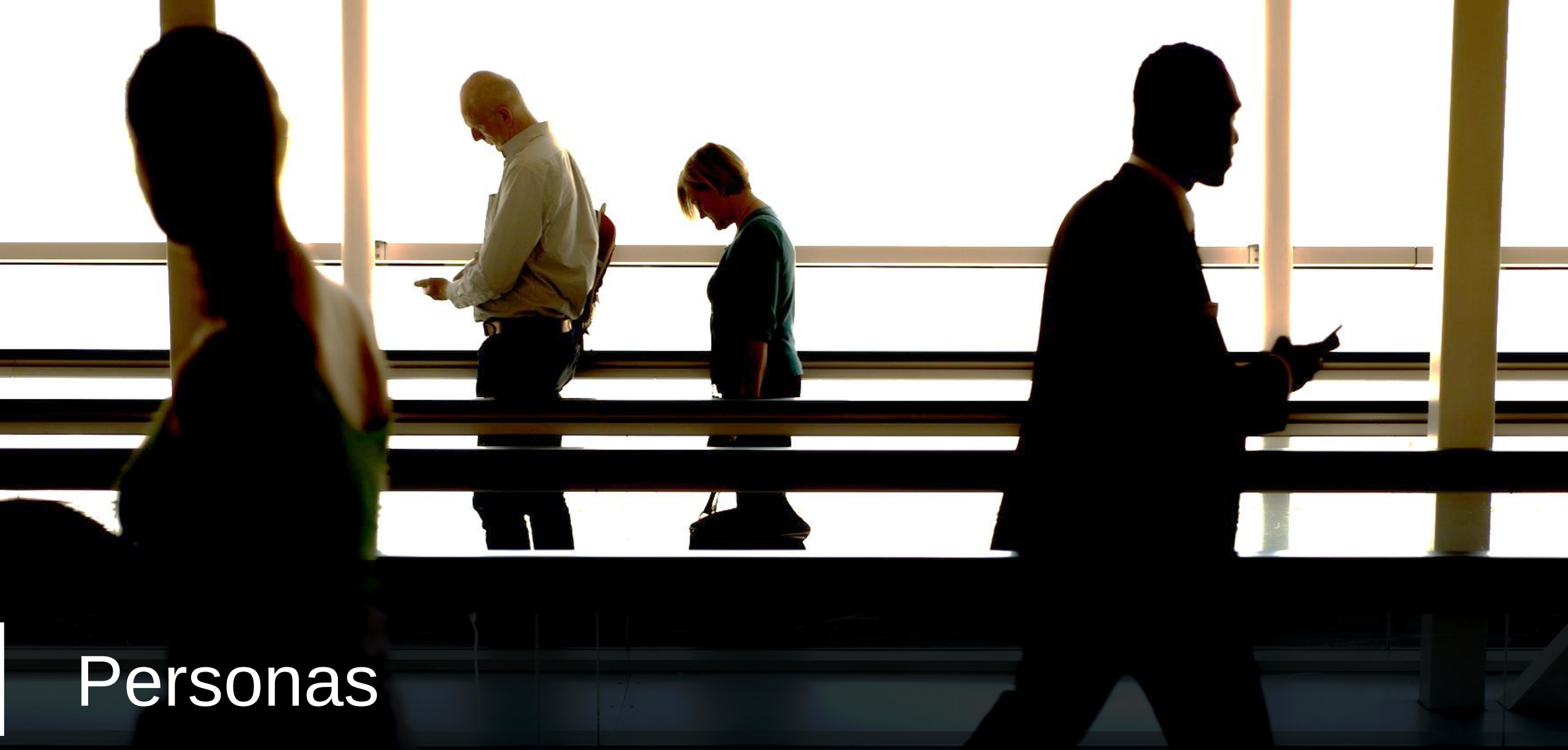


- Variety of methods for identifying groups of items
- Groups inferred from data
- Some subjectivity involved in determining number of clusters
- Commonly used for research in segmentation and personas

Discriminant analysis



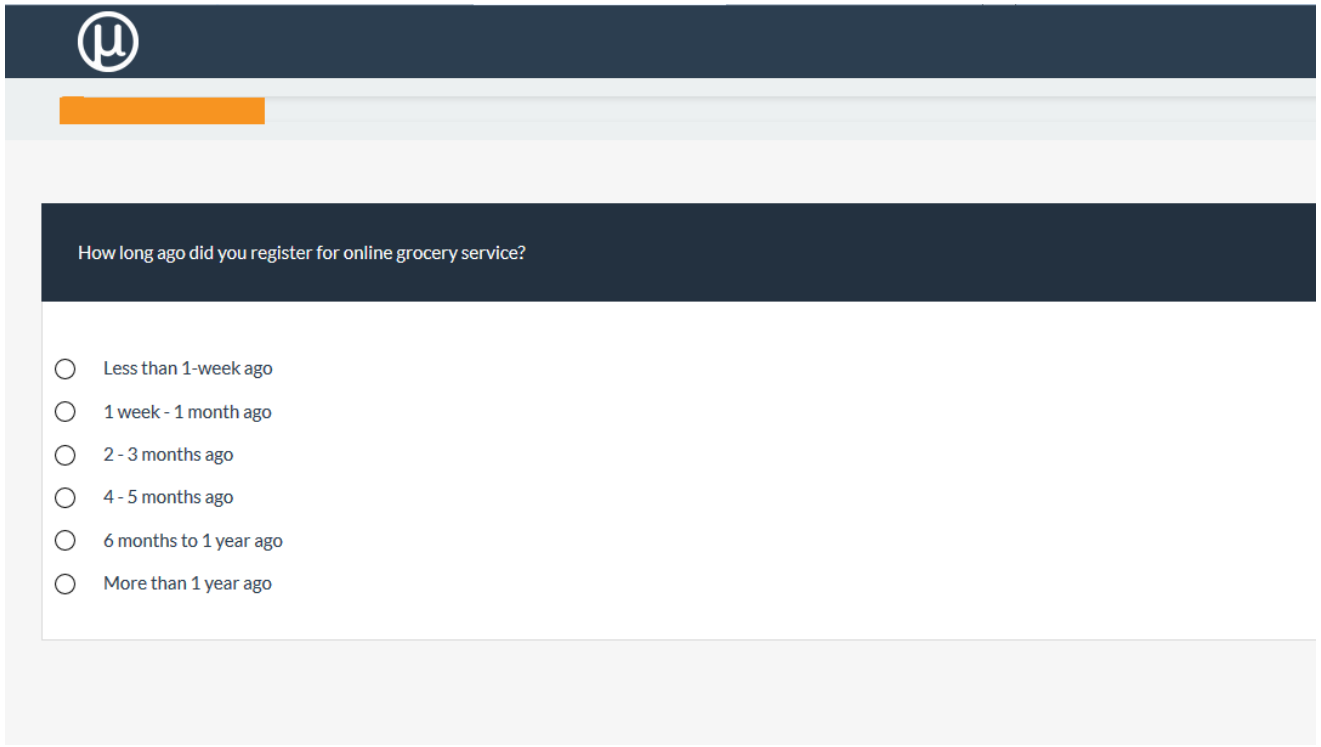
- Commonly used classification method
- Identify groups using cluster/class analysis
- Train classification model using demographic and rating variables
- Use classification model to make typing tool
- Use typing tool to predict segment membership



Personas

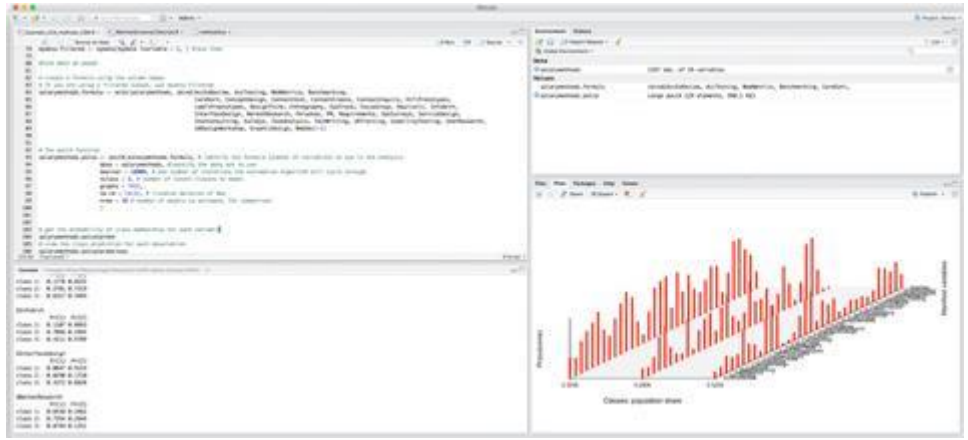
1. Conduct qualitative interviews and observation

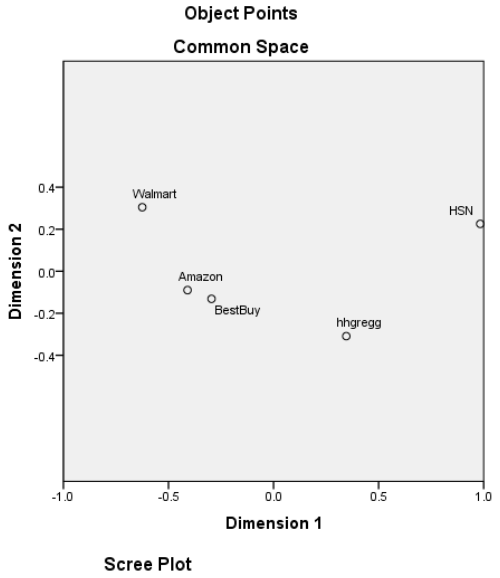
2. Survey a large sample of users or prospects

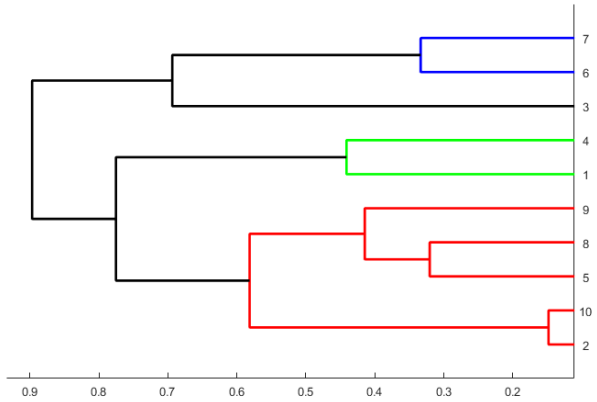


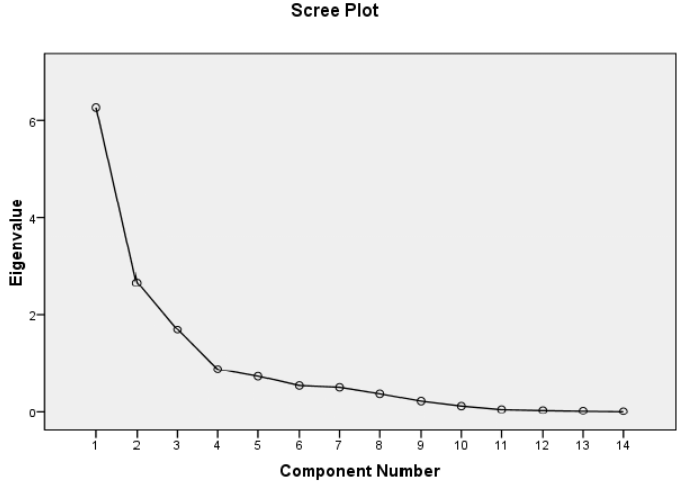
MUIQ

3. Identify the segments

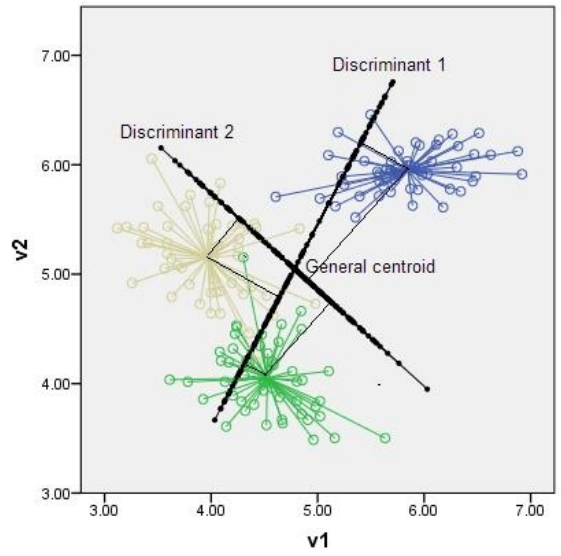


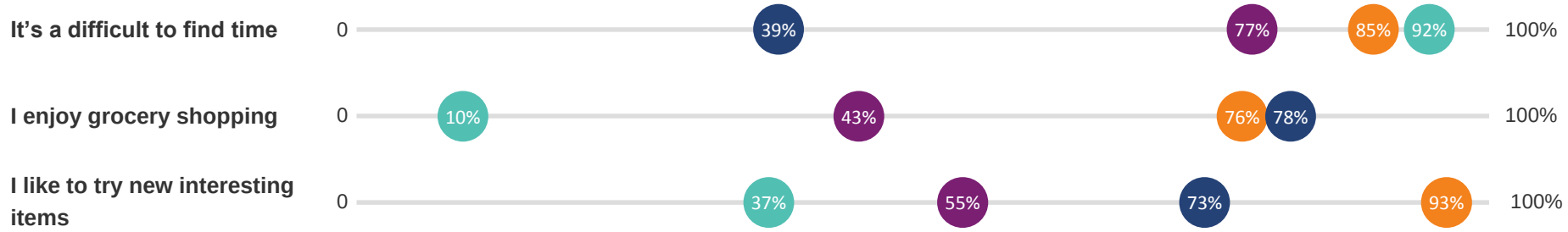




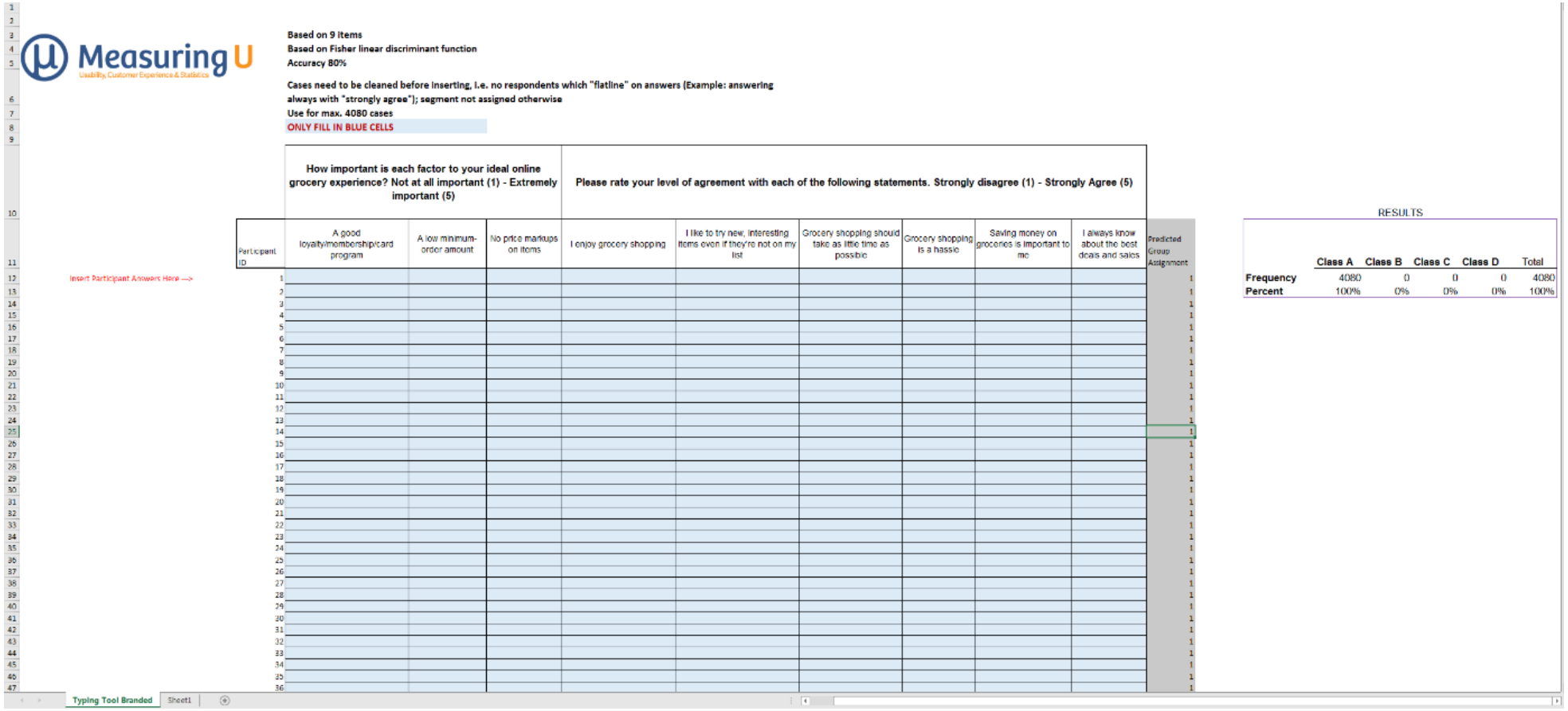


4. Determine key variables that differentiate segments





5. Predict segment membership using a typing tool



Measuring U
Usability, Customer Experience & Statistics

Based on 9 items
Based on Fisher linear discriminant function
Accuracy 80%

Cases need to be cleaned before inserting, i.e. no respondents which "flatline" on answers (Example: answering always with "strongly agree"); segment not assigned otherwise
Use for max. 4080 cases
ONLY FILL IN BLUE CELLS

Insert Participant Answers Here --->

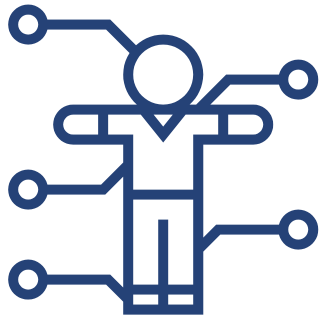
	How important is each factor to your ideal online grocery experience? Not at all important (1) - Extremely important (5)			Please rate your level of agreement with each of the following statements. Strongly disagree (1) - Strongly Agree (5)						
Participant ID	A good loyalty/membership/card program	A low minimum-order amount	No price markups on items	I enjoy grocery shopping	I like to try new, interesting items even if they're not on my list	Grocery shopping should take as little time as possible	Grocery shopping is a hassle	Saving money on groceries is important to me	I always know about the best deals and sales	Predicted Group Assignment
1										1
2										1
3										1
4										1
5										1
6										1
7										1
8										1
9										1
10										1
11										1
12										1
13										1
14										1
15										1
16										1
17										1
18										1
19										1
20										1
21										1
22										1
23										1
24										1
25										1
26										1
27										1
28										1
29										1
30										1
31										1
32										1
33										1
34										1
35										1
36										1

RESULTS

	Class A	Class B	Class C	Class D	Total
Frequency	4080	0	0	0	4080
Percent	100%	0%	0%	0%	100%

Typing Tool Branded | Sheet1

6. Personify or qualify your segments



Personify

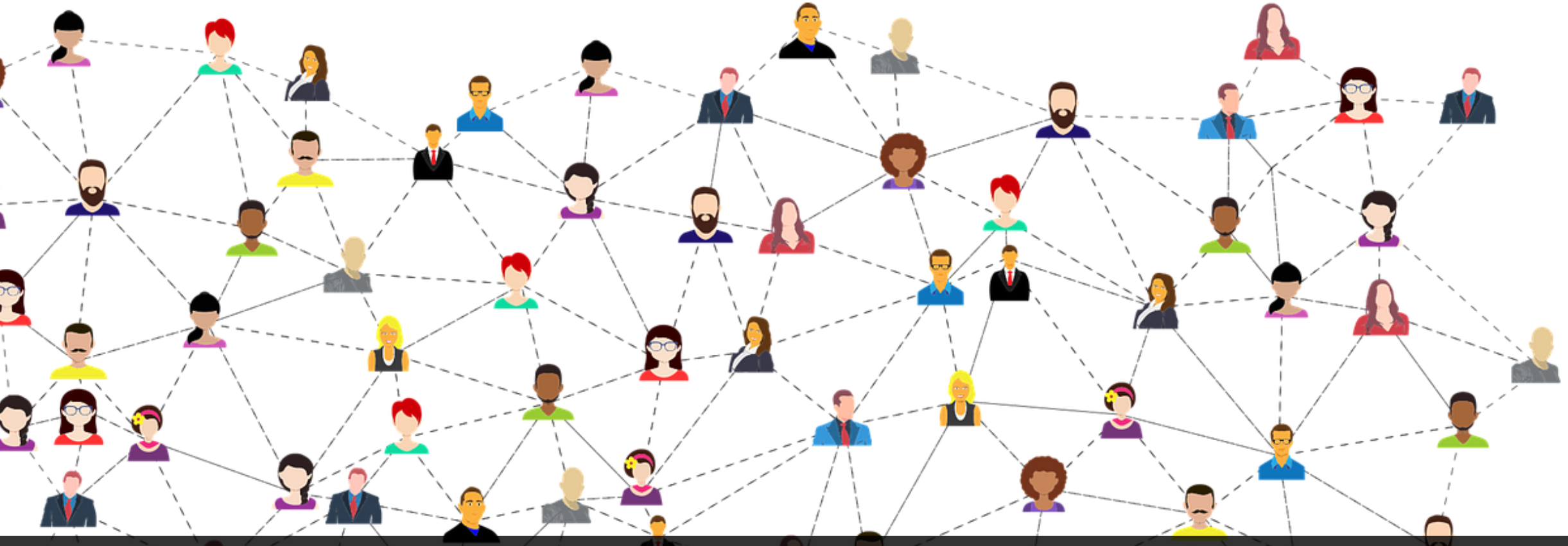
Go deep to get insights.

Conduct qualitative interviews & observations.

Qualify

What variables matter?

What percent of the population?



Scenarios of Interest

Persona identification and classification

- Quantitative cluster and classification requires large sample sizes to build precise models
- That said, you can do some quantitative work with small samples, keeping in mind this limits what you can claim
- Early trends often continue with collection of more data … except when they don't

With that in mind, here are some VERY preliminary results from the data you provided

Data reduction of ratings – **PCA**

	Component			
	1	2	3	4
1. Anx2Rel	-0.118	-0.538	0.045	-0.286
2. Spon2Plan	0.477	0.503	-0.411	0.481
3. Lux2Budg	0.892	-0.050	-0.170	0.193
4. Comf2Barg	0.882	0.207	0.052	-0.055
5. Money2Time	0.847	0.000	0.021	0.196
6. Discon2Con	0.357	-0.220	0.556	0.127
7. Exp2Inexp	0.135	0.490	-0.493	0.373
8. Solo2Group	-0.022	0.021	0.914	0.152
9. Thrill2Relax	-0.236	0.751	0.040	0.227
10. Bout2Chain	0.492	0.452	0.305	-0.306
11. Flow2Control	-0.223	0.321	-0.438	0.032
12. New2Favorite	0.036	0.717	-0.331	-0.092
13. Mem2Photo	0.056	-0.164	0.219	0.821
14. Self2CustSvc	-0.366	-0.049	0.725	-0.027
15. Intro2Extro	-0.239	-0.838	0.076	0.090
16. Prep2Wing	-0.183	-0.254	0.340	-0.692
17. Preplan2Imm	-0.051	-0.327	-0.437	-0.739

Not clearly aligned – 2, 7, 10

Component 1: 3, 4, 5

Component 2: 1, 9, 12, 15

Component 3: 6, 8, 11, 14

Component 4: 13, 16, 17

Spread of items across components good sign for discrimination/classification

Alignments seem reasonable?

Discriminant analysis – **all rating variables**

Classification Results[a,c]

			Predicted Group Membership				
		PropSegNum	1.00	2.00	3.00	4.00	Total
Cross-validated[b]	Count	1.00	2	2	3	0	7
		2.00	1	2	0	1	4
		3.00	0	1	2	0	3
		4.00	1	1	1	1	4
	%	1.00	28.6	28.6	42.9	.0	100.0
		2.00	25.0	50.0	.0	25.0	100.0
		3.00	.0	33.3	66.7	.0	100.0
		4.00	25.0	25.0	25.0	25.0	100.0

a. 100.0% of original grouped cases correctly classified.

b. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

c. 38.9% of cross-validated grouped cases correctly classified.

Overall classification accuracy was 38.9%

Not great, but better than chance expectation of 25%

Check diagonal for accuracy within class

Discriminant analysis – **best stepwise model**

Classification Results[a,c]

			Predicted Group Membership				
		PropSegNum	1.00	2.00	3.00	4.00	Total
Cross-validated[b]	Count	1.00	6	0	1	0	7
		2.00	1	3	0	0	4
		3.00	1	0	1	1	3
		4.00	0	1	1	2	4
	%	1.00	85.7	.0	14.3	.0	100.0
		2.00	25.0	75.0	.0	.0	100.0
		3.00	33.3	.0	33.3	33.3	100.0
		4.00	.0	25.0	25.0	50.0	100.0

a. 66.7% of original grouped cases correctly classified.

b. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

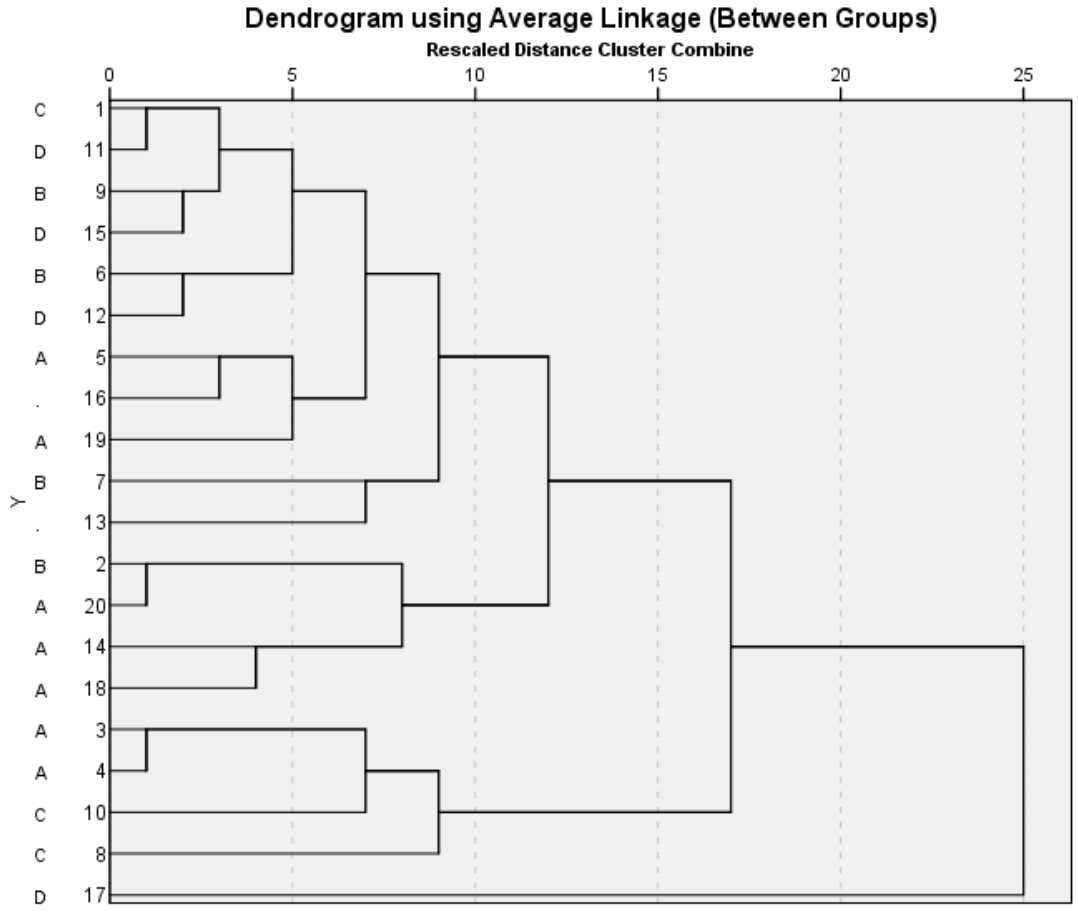
c. 66.7% of cross-validated grouped cases correctly classified.

Stepwise process retained two predictors: 5. Money2Time and 7. Exp2Inexp

Overall classification accuracy was 66.7%

Check diagonal for accuracy within class

Cluster analysis – **all rating variables**



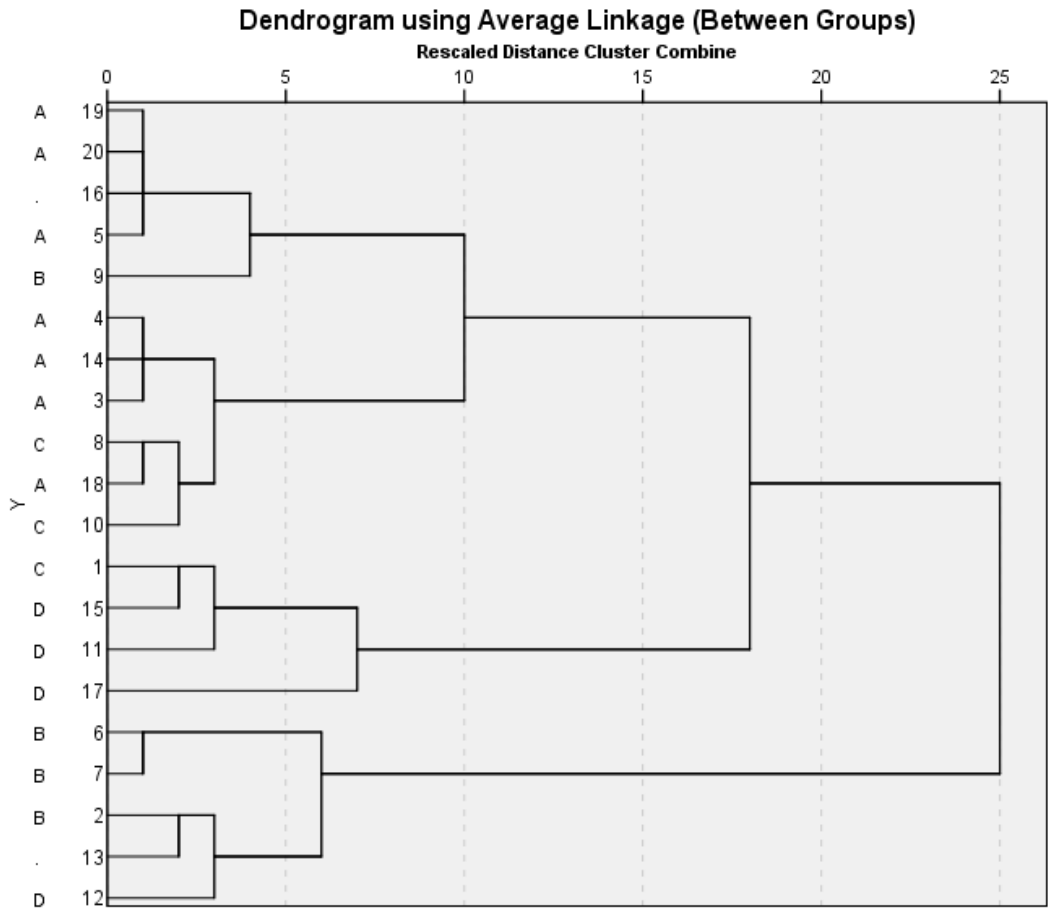
Training discriminant analysis depends on correct assignment of participants to groups

Preliminary hierarchical cluster analysis with all rating variables suggests this not the case

You would expect the letters to be more tightly grouped

Use this to discuss revising your preliminary assignments

Cluster analysis – **best two predictors**



Note more consistent grouping of letters in this graph

Consistent with finding from stepwise discriminant analysis

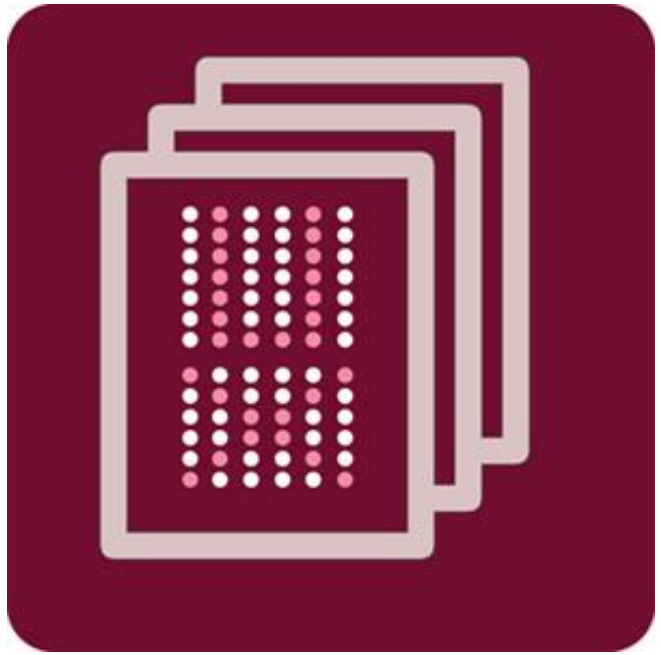
If current assignment of participant to persona is correct, only need two ratings for reasonably accurate prediction

Need more data!

Exploring your UX metrics data

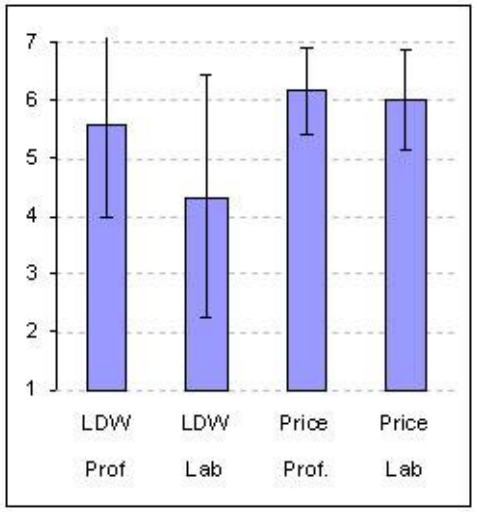
- Are the data ready for analysis?
- Which categorical variables drive CSAT?
- Which UX rating variables drive CSAT?
- Is there any latent structure in the ratings?
- Recommendations

Data preparation



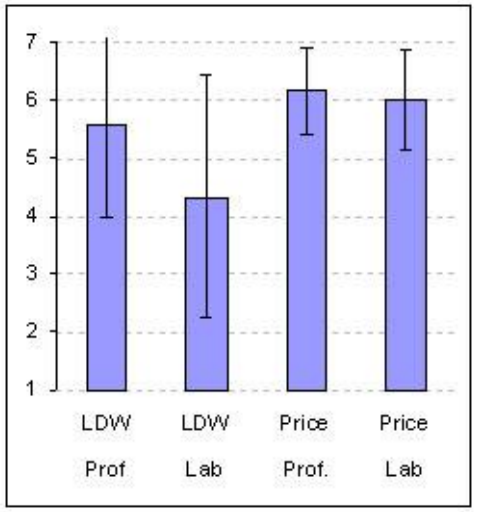
- Created numeric variable for CSAT
  (1: Extremely dissatisfied – 5: Extremely satisfied)
- Created Group variable for tracking which of six groups the respondent was a member
- Deleted thousands of incomplete responses
- Check coding of External Consistency
  – might be better to reverse it

Which categorical variables affect CSAT?



- Task - significant
- Task completion - significant
- Group - **not** significant
- Age range - significant
- Membership - significant
- Frequency of website use - significant
- New/repeat customer - significant
- MP status - significant

Which categorical variables affect CSAT?



With large samples expect statistical significance

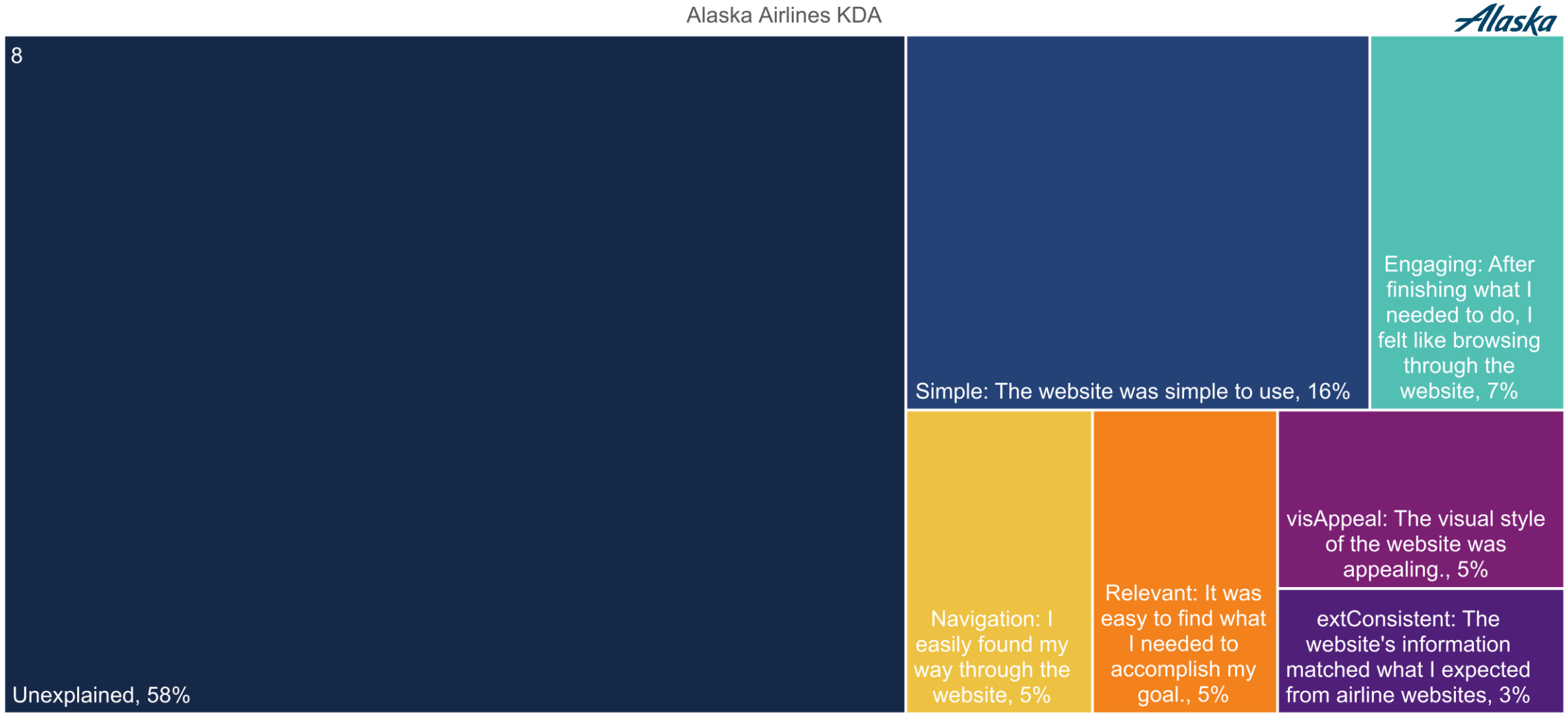
CSAT stability over groups surprising

CSAT differences for age groups kind of surprising

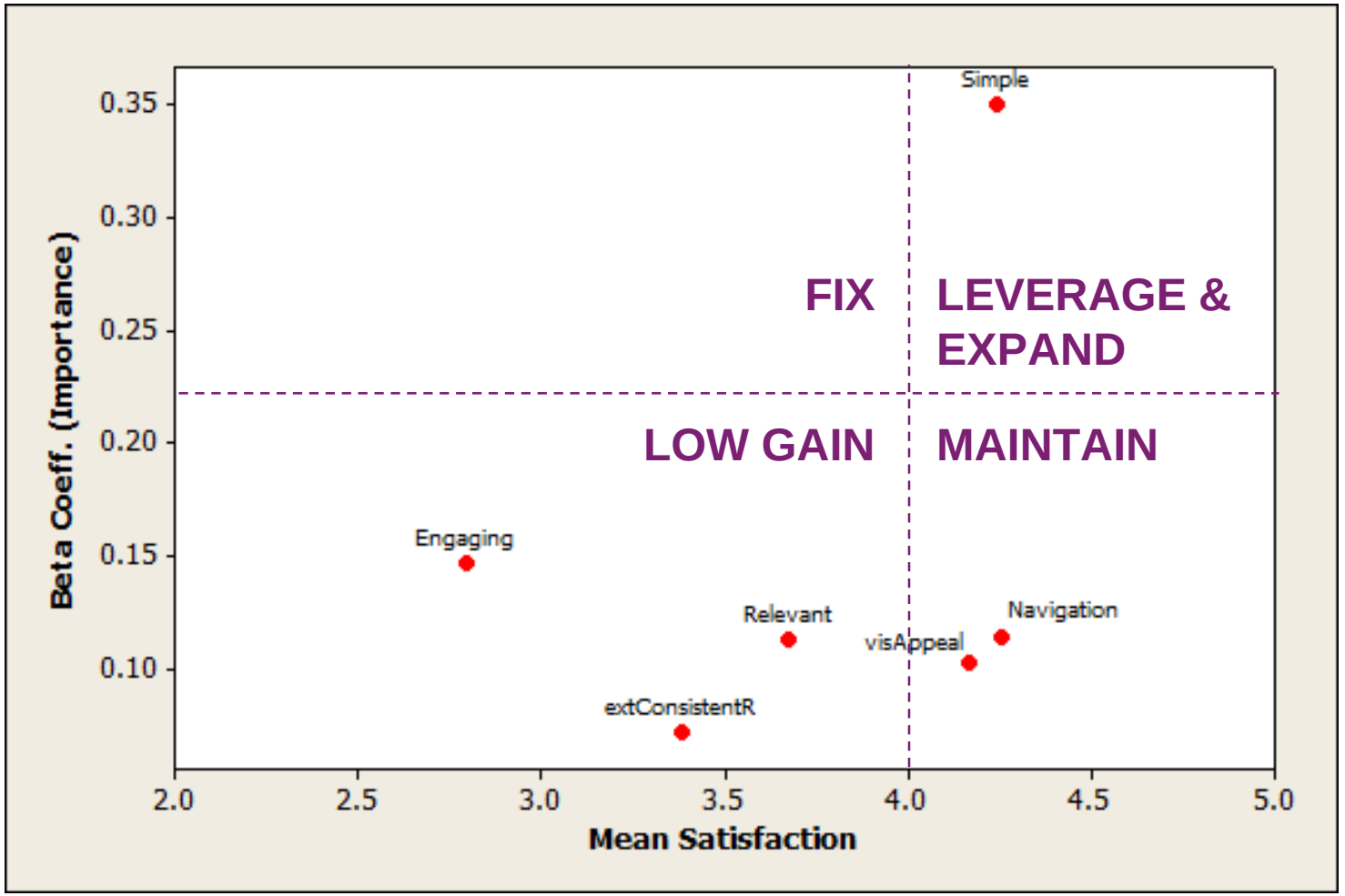
Not surprising that CSAT was:

- Different for different tasks
- Higher when task successful
- Higher for more frequent users
- Higher for loyalty members

Key driver analysis – **rectangle**



Key driver analysis – **2x2 matrix**

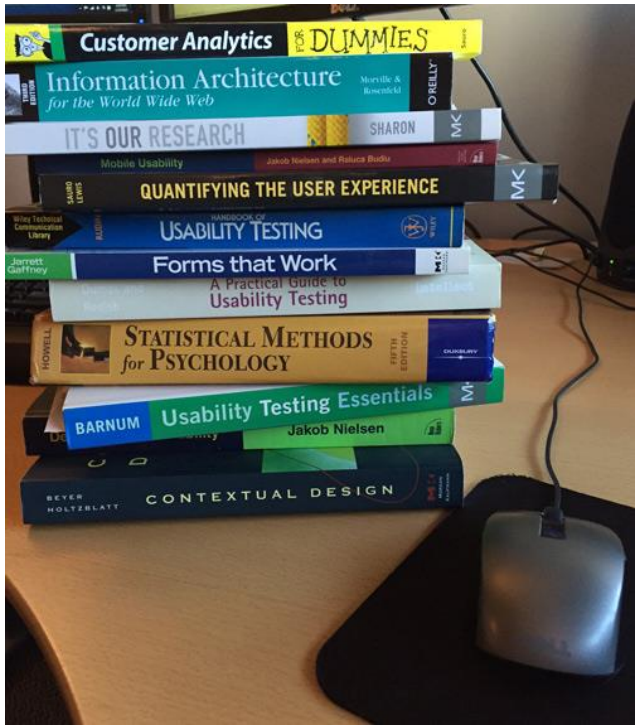


Is there any latent structure in the ratings?

Rotated Factor Matrix[a]		
	Factor	
	1	2
Simple	.862	.289
Navigation	.810	.238
intConsistent	.492	.349
extConsistent	-.100	-.753
Relevant	.429	-.145
Engaging	.353	.085
visAppeal	.613	.187
Extraction Method: Maximum Likelihood. Rotation Method: Varimax with Kaiser Normalization.		
a. Rotation converged in 3 iterations.		

- Parallel analysis indicated two factors
- Used MLFA and Varimax rotation
- 60% of variance explained
- All ratings except External Consistency aligned with first factor
- Remaining items, based on content, appear to be consistent with latent factor of perceived usability
- Reliability of Factor 1 items: $\alpha = .75$

UX metrics - recommendations



- Automate conversion of CSAT verbal responses to numeric (1: worst; 5 : best)
- Keep track of survey waves with a variable
- Consider reverse coding of External Consistency
- Consider collection of 11-point LTR rating
  – can't identify extreme ratings with five-point scale
  (> 10,000 respondents rated CSAT = 5)
- Consider adding UMUX-LITE to rating battery
- Consider temporary collection of SUS to check correspondence with custom global score

MUIQ

Remote UX Testing Platform
(Desktop & Mobile)

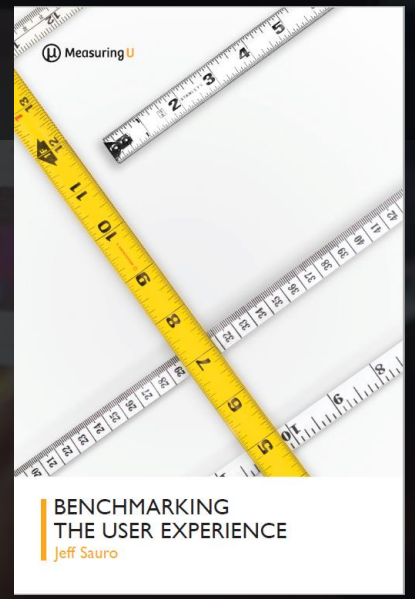
UX Research

SUPR-Q

Measurement
& Statistical Analysis

Eye Tracking &
Lab Based Testing

MeasuringU is a research firm based in Denver, Colorado focusing on quantifying the user experience.



MeasuringU
UX BOOTCAMP

measuringu.com
@MeasuringU

Exercises using “Subset for Exercises” sheet

- Confidence Interval Estimation
- Between-Subjects t-Test
- Within-Subjects t-Test
- Correlation
- 2 Proportion Test
- Between-Subjects ANOVA
- Within-Subjects ANOVA

CONFIDENCE INTERVAL ESTIMATION

Confidence Interval Estimation: CSAT

You want to know your overall mean CSAT with 90% confidence.

Copy CSATnum data and paste into statsUsabilityPak - 1 sample t.

What is the plausible range? Is it significantly greater than 4?

T-TESTS

Between-subjects t-test: Task completion

Is CSAT different for those who complete or fail to complete tasks?

Sort on “Task accomplished?” and paste CSATnum into statsUsabilityPak - 2 sample t.

What is the mean difference? Is it statistically significant? Practically significant?

Between-subjects t-test: New vs repeat visit

Is CSAT different for those who are new versus repeat visitors?

Sort on “NewVSRepeatVisit” and paste CSATnum into statsUsabilityPak - 2 sample t.

What is the mean difference? Is it statistically significant? Practically?

Within-subjects t-test: CSATper vs UXnum

Is there a difference in magnitude between CSAT expressed as a percentage and a custom UX metric based on the previously described factor analysis of the ratings, manipulated like the SUS to produce a 0-100-point measure?

Paste CSATper and UXnum into statsUsabilityPak - Paired t test.

What is the mean difference? Is it statistically significant? Practically?

CORRELATION

Correlation: CSATper with UXnum

Is there a significant correlation between CSAT expressed as a percentage and the custom UX metric?

Paste CSATper and UXnum into statsUsabilityPak - Regression.

What is the correlation? Is it statistically significant? Practically?

2 PROPORTION TEST

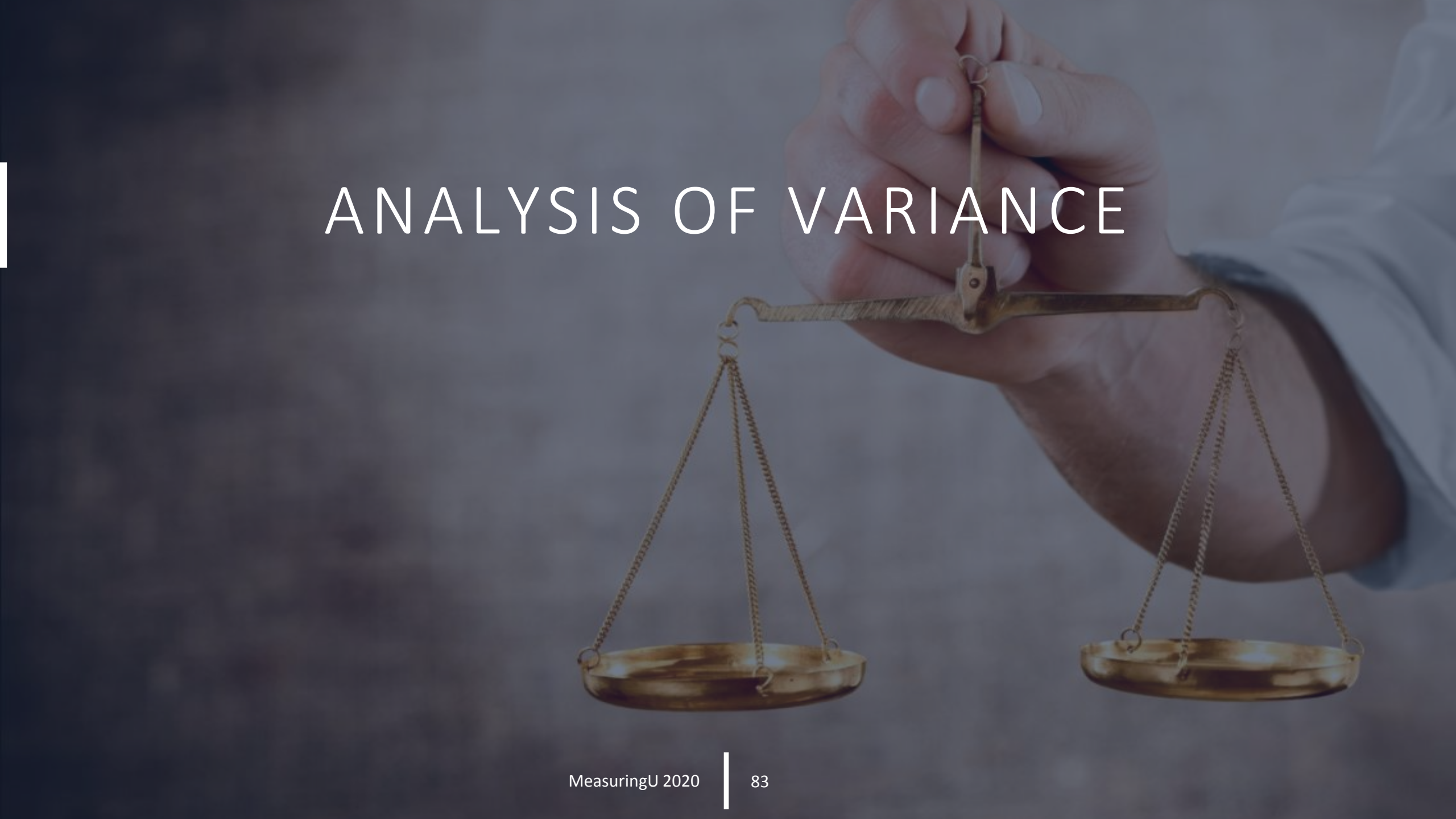
2 Proportion Test: Booking vs Find Info successes

Is there a difference in reported success completing booking versus looking for information?

Sort on Reason for Visit and Task Accomplished. Determine # successes and totals for both and enter into statsUsabilityPak - 2 comp rates.

What is the difference? Is it statistically significant? Practically?

ANALYSIS OF VARIANCE



ANOVA: Effect of fare on CSAT (between subjects)

Are there differences in CSAT for different fares in which customers are interested?

Sort by fares of interest and copy CSATnum values into statsUsabilityPak - ANOVA

What are the differences? Statistically significant? Practically?

ANOVA: Compare three ratings (within subjects)

Are there differences in the magnitudes of ratings for Simple, Engaging, and VisAppeal?

Copy Simple, Engaging, and VisAppeal into statsUsabilityPak – RM-ANOVA

What are the differences? Statistically significant? Practically?

Works Cited

10 Ways to Get a Horrible Survey Response Rate
https://measuringu.com/horrible-responserate/

Using Surveys to Measure the User Experience
https://measuringu.com/survey-ux/

10 Tips For Your Next Survey
https://measuringu.com/survey-tips/

What is a Representative Sample Size for a Survey?
https://measuringu.com/survey-sample-size/

How To Make Personas More Scientific
https://measuringu.com/scientific-personas/

9 Biases That Affect Survey Responses
https://measuringu.com/survey-biases/

How to Assess the Quality of a Measure
https://measuringu.com/measure-quality/

Do Survey Grids Affect Responses?
https://measuringu.com/grids-responses/

Very vs. Extremely Satisfied
https://measuringu.com/very-vs-extremely/

Is a Three-Point Scale Good Enough?
https://measuringu.com/three-points/

Does Coloring Response Categories Affect Responses?
https://measuringu.com/coloring-responses/

Can You Use a 3-Point Instead of an 11-Point Scale for the NPS?
https://measuringu.com/3-point-nps/

Effects of Labeling the Neutral Response in the NPS
https://measuringu.com/labeling-nps-effects/

What Motivates People to Take Free Surveys?
https://measuringu.com/free-surveys/

Does Changing the Order of the NPS Item Affect Results?
https://measuringu.com/nps-order/

4 Classes of Survey Questions
https://measuringu.com/survey-question-classes/

15 Common Rating Scales Explained
https://measuringu.com/rating-scales/

15 Metrics for UX Benchmarking
https://measuringu.com/ux-benchmark-metrics/

Changing the Net Promoter Scale: How Much Does It Matter?
https://measuringu.com/nps-scale-change/

How to Know Which Items to Remove in a Questionnaire https://measuringu.com/remove-items/

Cleaning Data From Surveys & Online Research
https://measuringu.com/cleaning-data/

5 Techniques to Identify Clusters In Your Data
https://measuringu.com/identify-clusters/

A Better Way To Segment Your Customers
https://measuringu.com/better-segmentation/

How Long Is the Typical Online Study?
https://measuringu.com/online-study-time/

5 Ways to Increase Study Participation Rates https://measuringu.com/study-participate/

10 Things to Know about a Key Driver Analysis
https://measuringu.com/key-drivers/

Picking the Right Data Collection Method
https://measuringu.com/data-collection/

12 Tips For Writing Better Survey Questions
https://measuringu.com/survey-questions/

Pros and Cons of Requiring Survey Responses
https://measuringu.com/requiring-responses/

13 Things to Consider When Offering Survey Incentives
https://measuringu.com/offering-incentives/

5 Steps for Better Customer Sampling
https://measuringu.com/sampling-customers/

7 Survey Types to Measure the Customer Experience
https://measuringu.com/cux-surveys/

Are Top Box Scores a Better Predictor of Behavior?
https://measuringu.com/top-box-behavior/

Answers to Exercises

- Confidence Interval Estimation
- Between-Subjects t-Test
- Within-Subjects t-Test
- Correlation
- 2 Proportion Test
- Between-Subjects ANOVA
- Within-Subjects ANOVA

CONFIDENCE INTERVAL ESTIMATION

Confidence Interval Estimation: CSAT

Copy CSATnum data and paste into statsUsabilityPak - 1 sample t.

What is the plausible range? Is it significantly greater than 4?

Contents measuringu.com

Confidence Interval and Test Around Raw Continuous data

Paste the Raw Data in the 1st Column--Remove Non Numeric Values

* *Required Fields* Enter Summary Data

For Task Times Use the Time CI & Test Tab

Clear Values

Raw Data*
5
5
5
1
5
5
5
5
4
5
4
4
5
5
5
5
4
4

Input

* Confidence Level 90%

Test Benchmark

Descriptive Stats	
Mean	4.22
Standard Deviation	1.22
n	99

Results

	Low	High
Confidence Interval	4.02	4.43

Margin of Error 5%

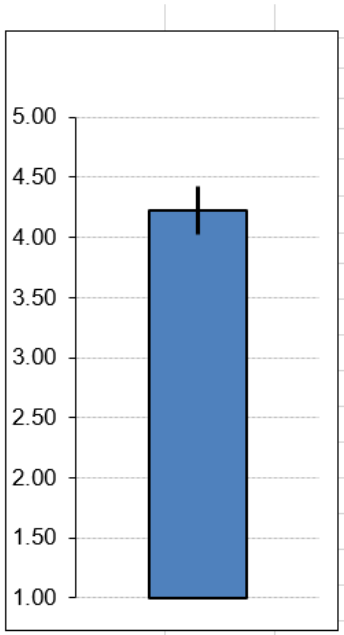
p-values

Population Mean = Test Mean

Population Mean >Test Mean

Population Mean < Test Mean

Power



How to Report

We can be 95.0% confident the population mean is above 4.02

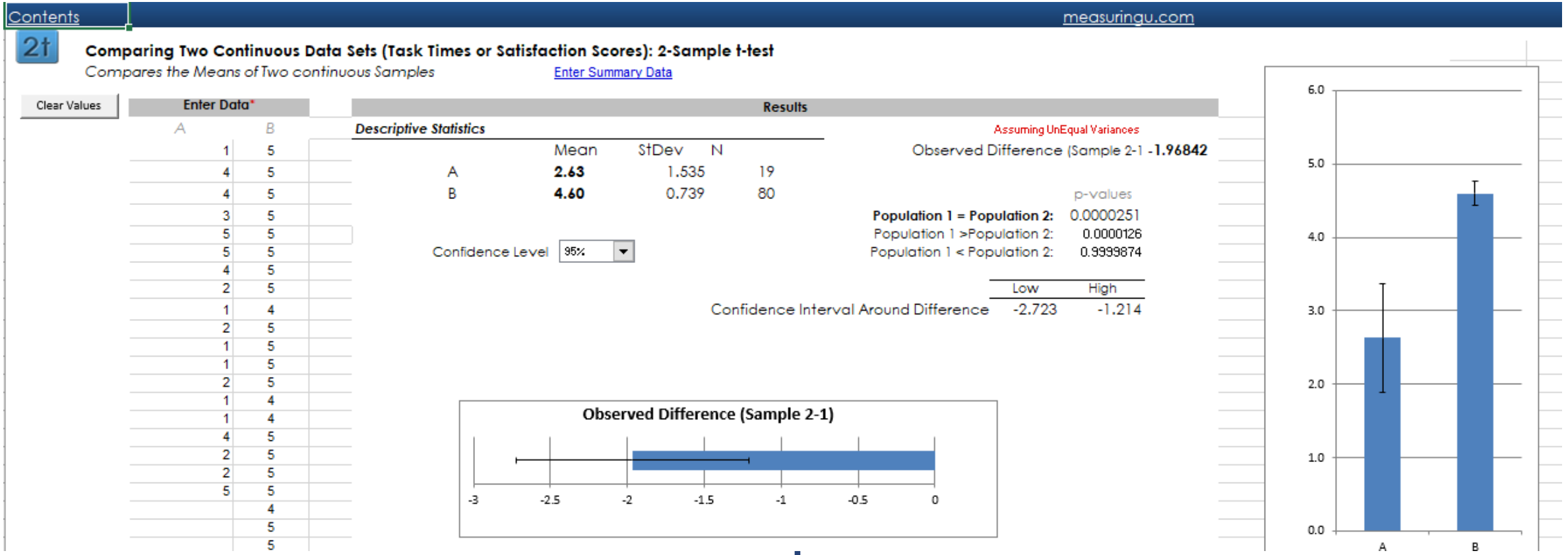
We can be 90.0% confident the population mean is **between** 4.02 and 4.43

T-TESTS

Between-subjects t-test: Task completion

Sort “Task accomplished?”; paste CSATnum in statsUsabilityPak - 2 sample t.

What is the mean difference? Is it statistically significant? Practically?



Between-subjects t-test: New vs repeat visit

Sort “NewVSRepeatVisit”; paste CSATnum in statsUsabilityPak - 2 sample t.

What is the mean difference? Is it statistically significant? Practically?

Comparing Two Continuous Data Sets (Task Times or Satisfaction Scores): 2-Sample t-test

Compares the Means of Two continuous Samples Enter Summary Data

Clear Values

Enter Data*	
A	B
1	4
5	4
2	3
1	2
1	1
1	2
2	4
5	5
5	5
5	5
4	5
5	5
4	5
4	5
5	5
5	5
4	5
5	5
5	5
5	4
5	4
5	5

Results

Descriptive Statistics

	Mean	StDev	N
A	**4.09**	1.330	46
B	**4.33**	1.060	39

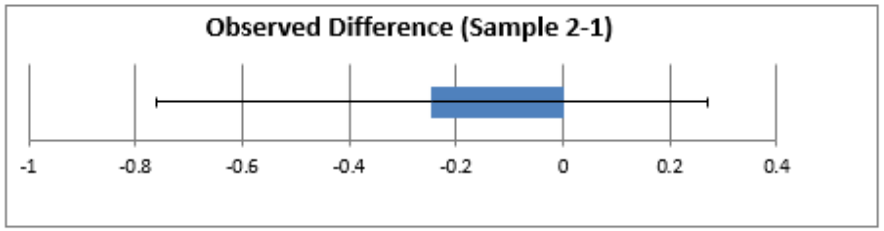
Confidence Level 95%

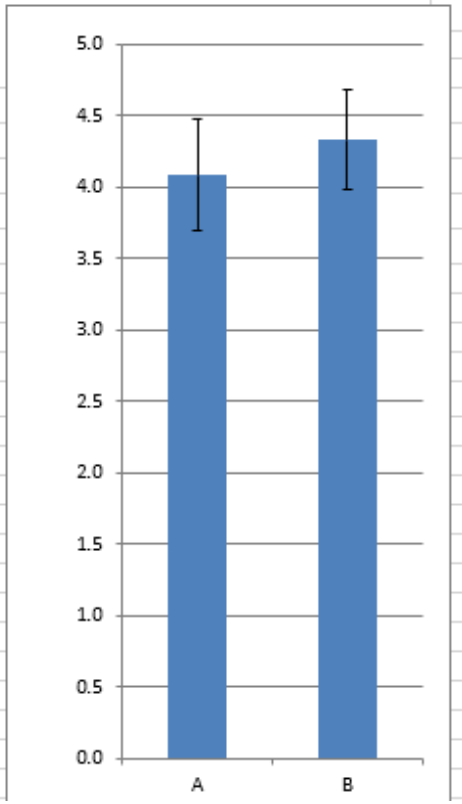
Assuming UnEqual Variances

Observed Difference (Sample 2-1 **-0.24638**

	p-values
Population 1 = Population 2:	0.3449392
Population 1 >Population 2:	0.1769217
Population 1 < Population 2:	0.8230783

	Low	High
Confidence Interval Around Difference	-0.762	0.270





Within-subjects t-test: CSATper vs UXnum

Paste CSATper and UXnum into statsUsabilityPak - Paired t test.

What is the mean difference? Is it statistically significant? Practically?

Paired t-test

Paste the Raw Values in the 1st two Columns. Each row should be the same person. Remove Non Numeric Values

* *Required Fields*

Clear Values

Enter Data*		
CSAT	UX	Difference
0	0.25	-0.25
1	0.75	0.25
0.25	0.7083333	-0.46
0	0.1666667	-0.17
0	0.2916667	-0.29
0	0	0.00
0.25	0.7083333	-0.46
1	0.875	0.13
1	0.75	0.25
1	0.8333333	0.17
0.75	0.5833333	0.17
1	0.7916667	0.21
0.75	0.7083333	0.04
0.75	0.6666667	0.08
1	0.9166667	0.08
1	0.7083333	0.29
0.75	0.625	0.13
1	0.9583333	0.04
1	0.9166667	0.08
1	0.75	0.25
1	0.875	0.13
1	0.5416667	0.46

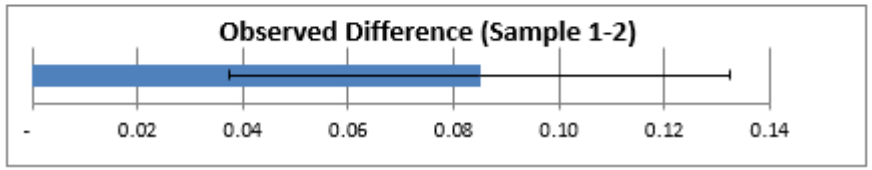
Input	
* Confidence Level	95%
Null Hypothesis	
Difference is	Equal To
	0

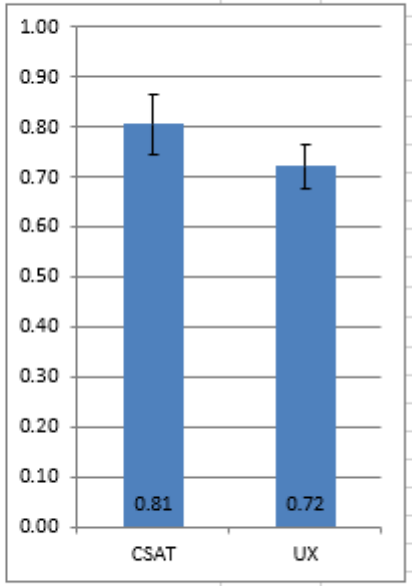
Descriptive Stats of the Difference	
Mean Difference	0.1
Median Difference	0.1
Standard Deviation	0.238
N (sample size)	99

Results	
Average Difference	0.09
Confidence Interval Low	0.04
Confidence Interval High	0.13
Margin of Error	0.0
p-value	0.000576963
Power	0.942813380

Descriptive Statistics for each Group

	Mean	StDev	N
CSAT	0.81	0.30	99.00
UX	0.72	0.23	99.00





CORRELATION

Correlation: CSATper with UXnum

Paste CSATper and UXnum into statsUsabilityPak - Regression

What is the correlation? Is it statistically significant? Practically?

Correlation & Simple Linear Regression

Paste the Raw X and Y values in the input below (up to 9750 values)

Input	
CSAT	UX
0.00	0.25
1.00	0.75
0.25	0.71
0.00	0.17
0.00	0.29
0.00	0.00
0.25	0.71
1.00	0.88
1.00	0.75
1.00	0.83
1	1
1	1
0.8	1
0.8	1
1.0	1
1	0.7083333
0.75	0.625
1	0.9583333
1	0.9166667
1	0.75
1	0.875
1	0.5416667
0.75	0.75

Results

		90% CI Around r	
		Low	High
Correlation *r*	0.633931	0.523	0.724
p-value (2-sided)	0.00000		
R-Squared	40.19%		

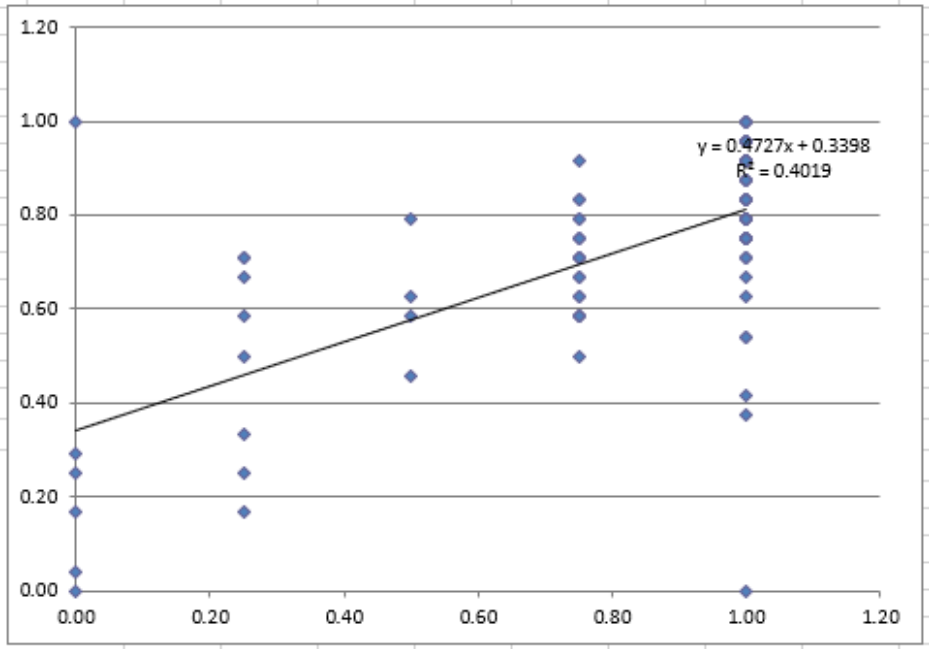
Regression Line	Slope		Y-Intercept
Y-hat =	0.472690148 x	+	0.33976

90% CI Around Slope		90% CI Around Intercept	
Low	High	Low	High
0.375	0.570	0.26	0.42

Predicted Values

Enter an X Value for a prediction	Predicted Y Value	90% CI Around Prediction	
		Low	High
80	38.15	30.45	45.86

Alpha 0.1



2 PROPORTION TEST

2 Proportion Test: Booking vs Find Info successes

Paste # successes and totals into statsUsabilityPak - 2 comp rates.

What is the difference? Is it statistically significant? Practically?

Comparing 2 Proportions (N-1 Two Proportion Test)

For comparing two independent sample proportions for all sample sizes. When the Expected Cell counts is below 1, the Fisher Exact Test is Used.

* *Required Fields*

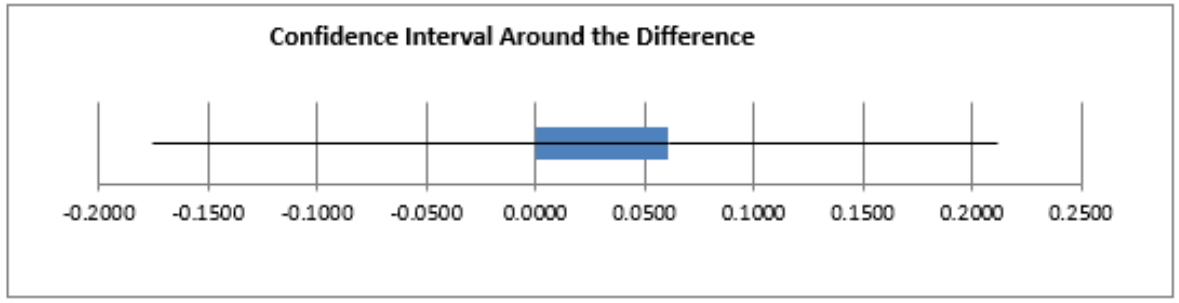
Population 1		Population 2	
Success*	Total*	Success*	Total*
50	58	12	13

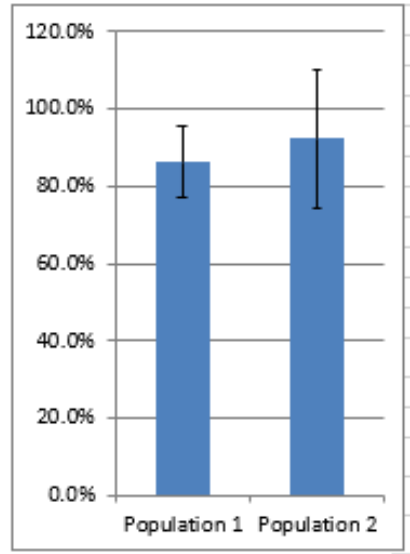
Confidence Level: 95%

Results		
	Population 1	Population 2
Completion Rate	86.21%	92.31%
Difference	**0.06101**	

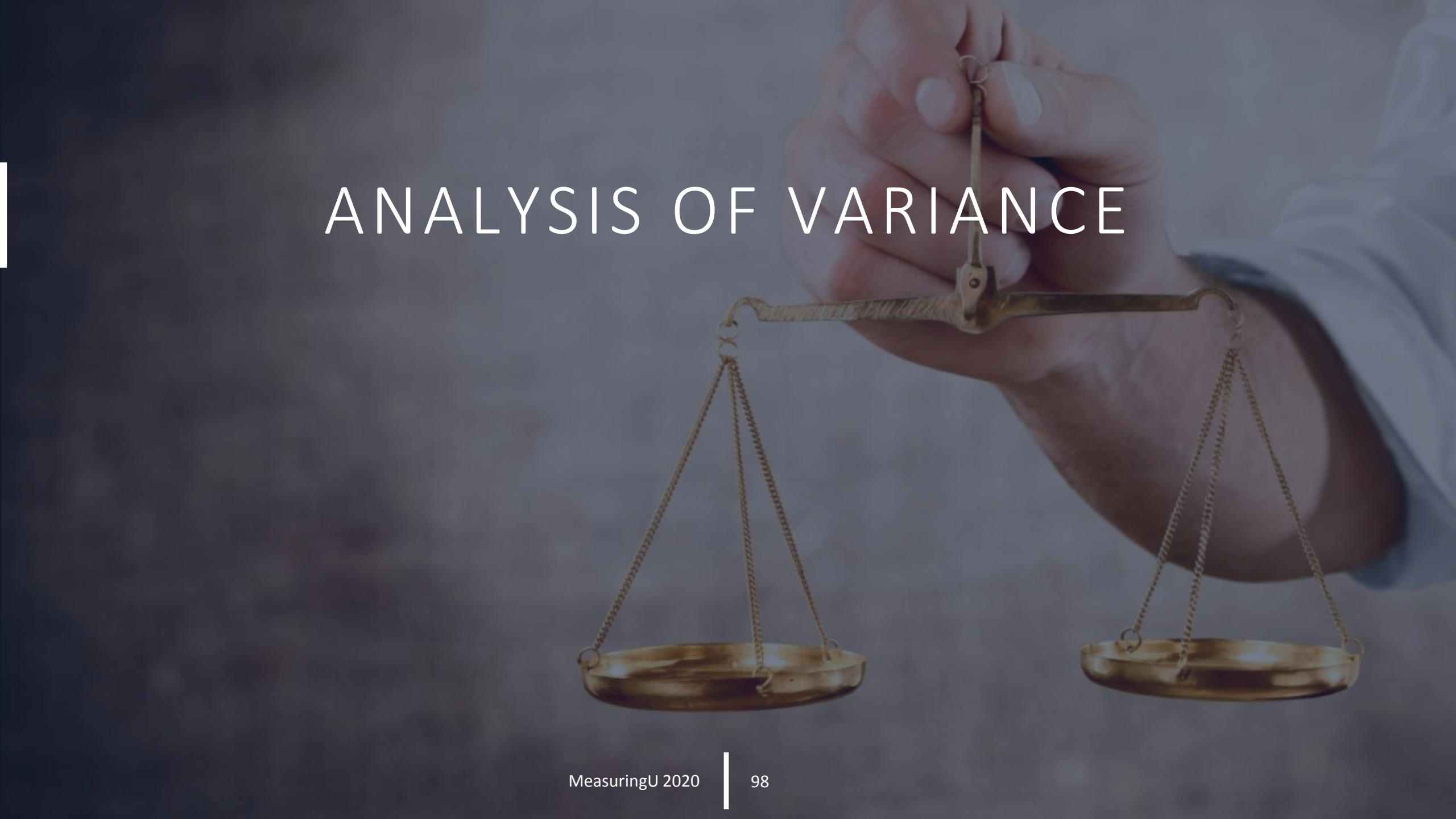
Confidence Interval Around Difference	Low	High
	-0.176	0.212

p-values	
P1 = P2	0.55295
P1 > P2	0.27648
P1 < P2	0.72352





ANALYSIS OF VARIANCE



ANOVA: Effect of fare on CSAT (between subjects)

Sort fares and copy CSAT values into statsUsabilityPak - ANOVA

What are the differences? Statistically significant? Practically?

1-Way ANOVA (Analysis of Variance)

Enter the raw data (e.g time, satisfaction data) in columns to compute an Analyis of Variance (Comparing 2+ Means) up to 1000 values

* *Required Fields*

Input

First Class	Main	Saver
5.0	2	1.0
5.0	1	5.0
4.0	2	4.0
5.0	4	5.0
5.0	5	4.0
5.0	5	5.0
5.0	5	5.0
5.0	4	4.0
	5	5.0
	4	5.0
	4	5
	5	5
	5	3
	5	
	5	
	5	
	5	
	2	
	5	
	3	
	4	
	5	
	5	
	5	

Results

Summary Table

Source	df	SS	MS	F	p
Factor (Between Group	2	2.179141	1.08957	0.96	**0.3908**
Error (Within Groups)	56	64	1		
Total	58	66.0339			

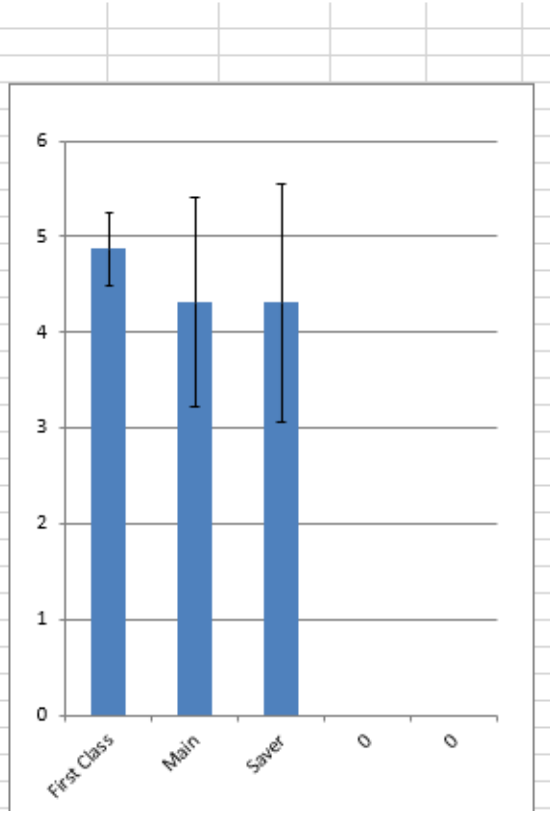
Fcritical Value
3.161861165

How to Report

There is a 60.922% chance at least 1 product has a different mean.

Calculations

	First Class	Main	Saver	0	0	Total
N	8	38	13	0	0	59
Mean	4.875	4.315789	4.30769			4.39
Standard Deviation	0.3535534	1.117557	1.18213			1.067
Variance	0.125	1.248933	1.39744			1.139
t-critical	2.3646243	2.026192	2.17881	####	#NUM!	
Margin	0.3786429	1.089984	1.24098	####	#####	



ANOVA: Compare three ratings (within subjects)

Copy Simple, Engaging, and VisAppeal into statsUsabilityPak - RM-ANOVA.

What are the differences? Statistically significant? Practically?

Repeated Measures ANOVA (Analysis of Variance)

Enter the raw data (e.g time, satisfaction data) in columns to compute an Analysis of Variance for Comparing up to 6 Within Subjects Conditions

* *Required Fields*

Input

Subject	Simple	Engaging	Visual Ap	4	5	6
3	5	4	5			
4	5	1	5			
5	4	2	3			
6	1	3	3			
7	5	3	5			
8	5	4	5			
10	5	3	5			
11	5	3	4			
13	5	4	3			
14	5	3	5			
16	2	1	3			
17	4	3	4			
18	5	3	5			
19	4	2	5			
20	5	3	5			
21	4	3	3			
22	5	3	4			
23	5	1	4			
24	3	2	5			
25	5	5	5			
26	5	4	4			
27	4	4	5			
28	5	3	4			
31	3	2	4			
32	5	4	5			
33	5	4	5			

Results

Summary Table

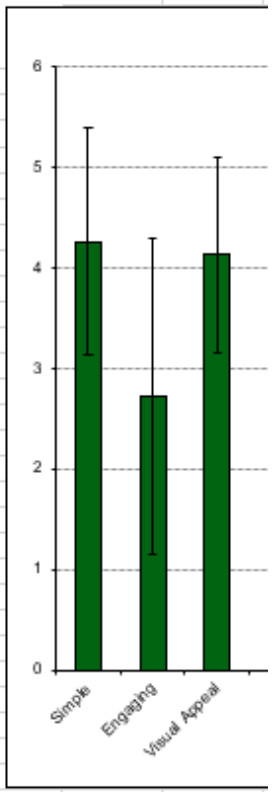
Source	df	SS	MS	F	p	
Between Subjects	96	267.6				
Factor (Within Subjects)	2	141.3	70.64	106.02	**0.000**	Sphericity Assumed
Greenhouse-Geisser	1.739	141.3	81.25		**0.000**	Greenhouse-Geisser
					0.000	Lower Boundary of P
Error (Subject x Factor)	192	127.94	0.67			
Greenhouse-Geisser	166.94	127.94	0.77			
Total	290	536.8				

How to Report

There is a 100.000% chance at least 1 mean is different.

Calculations

	Simple	Engaging	Visual Appeal	4	5	6	Total
N	97	97	97	0	0	0	291
Mean	4.257731959	2.72164948	4.134020619				3.704
SS Factor	29.69186712	93.6953036	17.89805269				141.3
Standard Deviation	1.175074027	1.3129346	0.996126519				1.358



MeasuringU

Usability, Customer Experience & Statistics

MeasuringU.com
@MeasuringU

MUIQ

Remote UX Testing Platform
(Desktop & Mobile)

UX Research

SUPR-Q

Measurement
& Statistical Analysis

Eye Tracking &
Lab Based Testing

MeasuringU is a research firm based in Denver, Colorado focusing on quantifying the user experience.

DenverUX BOOTCAMP

UX Measurement Boot Camp