Sawtooth Software

The survey software of choice

Applied MaxDiff

Webinar

Applied MaxDiff Webinar

Keith Chrzan, Sawtooth Software

February 27, 2020

Sawtooth Software
The survey software of choice

Topics

- Introduction to MaxDiff
- Designing a MaxDiff Experiment
- Estimating Utilities
- Rescaling Utilities
- Sample Size
- Anchored MaxDiff
- Large Numbers of Items
- Subsequent Analyses
- Profile Case MaxDiff

Introduction

- Researchers often need to compare items on lists
  - Brands
  - Attributes
  - Message elements
  - Advertising executions
  - Product concepts
  - Product improvements
  - Ubiquitous need
- Rating scales often perform poorly
  - Scale use bias
  - Lack of discrimination
  - Lack of predictive validity

Enter MaxDiff

- First proposed by Finn and Louviere (1992) MaxDiff (or Best-Worst Scaling) is a multiple choice extension of the classic method of paired comparisons
- We know that people rank the top and bottom items on lists more reliably than they rank the things in the middle, so MaxDiff capitalizes on this by having respondents choose the best and worst thing on a list

Example Question

Which one thing would you **most** like to do on your next vacation and which one thing would you **least** like to do?

Most	Activity	Least
☐	Visit an art museum	☐
☐	Suntan on a sandy beach	☐
☐	Shop in a trendy area	☐
☐	Ride a horse	☐
☐	Zipline through the treetops	☐

- If we had to 20 such items to test, we might ask 12 questions like this one, each with a different set of 5 items

MaxDiff is Respondent-Friendly

- The task is easy, capitalizing on what we humans do well
- It’s so easy, in fact, that MaxDiff is often used in interviews by healthcare researchers surveying elderly, sick or otherwise impaired populations

MaxDiff is Information-Dense

- With just 2 mouse clicks we can learn that a respondent (for example)
  - Prefers a sandy beach to
    - Horseback riding
    - Ziplining
    - Going to a museum
    - Trendy shopping
  - And that the respondent likes ziplining less than
    - Horseback riding
    - Trendy shopping
    - Going to a museum

- In other words, we learn about preferences among 7 pairs of items from just 2 mouse clicks
- MaxDiff is a very efficient way to collect preference information

Typical Applications

- Measuring the relative appeal of
  - New products
  - Concepts
  - Varieties
  - Flavors
  - Menu items
  - Colors
- Measuring attribute importance
- Measuring the strength of different advertising message elements or executions

Liking

Now imagine that you've had an evening meal at the casual dining restaurant and that you've decided to buy a dessert. Considering only these 4 desserts, which would you like the Most and which would you like the Least?

(1 of 12)

Most		Least
○	Coconut cream pie	○
○	Mango lassi	○
○	Crème brûlée	○
○	Vanilla milk shake	○

Importance

Please indicate how important each of these aspects of a casual dining restaurant is to you.
Considering only these 4 features, which is the Most Important and which is the Least Important?

(1 of 10)

Most Important		Least Important
○	Comfortable environment	○
○	Reasonable prices	○
○	Pace of meal	○
○	Prompt greeting	○

Designing a MaxDiff Experiment

- Sometimes we can create simple designs with standard balanced incomplete block designs (BIBD) pulled from an experimental design catalog (Cochran and Cox 1957)
- More often, however, we use computer search algorithms to create good designs ("near BIBDs") for any number of items, which they do by balancing . . .
  - How often each item appears
  - How often each pair of items appear together
  - How often each item inhabits the each (top, middle, bottom) position in the question

Some Design Decisions

- Number of items - in commercial applications we've seen pressure to increase the number of items so that studies with scores of items have become common
- Number of items per question – empirical testing suggests using 4-5 items per question for most applications
- Number of questions per respondent
  - We like to have enough questions for each respondent to see each item 3-4 times
  - MaxDiff questions feel repetitive, however so for large numbers of items we often opt to keep the number of questions below 20 or 30, even at the cost of showing each item fewer times

- Once we've made these decisions our computer search will produce a requested number of equivalent blocks (or versions) of the design so that different respondents might see different (high quality) blocks of questions

Prohibitions

- Conjoint analysis can be sensitive to prohibitions, which quickly degrade the quality of the design as they increase in number
- MaxDiff is much less sensitive to the presence of prohibitions

Estimating Utilities

- Analysts have several options for creating their MaxDiff utilities
  - Simple count-based methods
  - Model-based methods
    - Aggregate MNL
    - LC-MNL
    - HB-MNL/mixed logit
    - On-the-fly estimation

Count-Based Methods

- One can simply subtract the number of times an item is selected as worst from the number of times it is selected as best (Louviere, Flynn & Marley 2015)
- Alternatively you can take the natural log of the best count divided by the worst count, ln(B/W), as described in Louviere, Flynn & Marley (2015)
- Lipovetsky and Conklin (2014) propose a more complex ratio of counts that they call the analytical best worst score

- At the sample level, all three of these methods produce utilities that are highly correlated with the utilities we get from statistical modeling

Aggregate MNL

- Multinomial logit (MNL) is a statistical model that identifies a set of utilities that best predict some observed set of choices that respondents make
- We can use a single (aggregate) MNL to calculate a set of MaxDiff utilities for an entire group of respondents
- You can run separate MNL models for separate subgroups of respondents
  - For males and females
  - For the different regions of the world or of a country
  - For high, medium and low volume customers
  - For loyal customers, switchers and defectors
  - Etc.
- Of course this gets repetitive if you want to get utilities for lots of subgroups

Latent Subgroups

- Usually pre-identified subgroups of respondents don't align perfectly with differences in preferences
- As a result, sometimes we want to identify subgroups of respondents who have similar utilities; because the variables that define group membership don't exist in our data set, they are called hidden or "latent"
- Latent Class MNL allows us to find these groups

Latent Class MNL

- Latent Class MNL simultaneously identifies
  - Segments of respondents
  - The sizes of the segments and
  - The utilities for each of the segments
- As such, Latent Class MNL is perfect fit for conducting segmentation with MaxDiff utility data
- Latent Class MNL even has goodness-of-fit statistics to help us determine how many subgroups our population contains

Hierarchical Bayesian (HB) MNL

- Usually, however, we want to have utilities for each respondent in our survey
  - Maybe we want to be able to slice and dice our utilities using crosstab software, as doing so will be easier than rerunning an aggregate logit over and over
  - Or maybe we want to run some analyses, like simulations or TURF analysis (described below) which require respondent-level utilities

On-The-Fly Utility Estimation

- If we've shown each item 3-4 times per respondent, we can get a quick and dirty read of the utilities while the respondent is still in the survey
- Having these utilities in real time enables us to do some interesting things
  - Using the utilities to focus subsequent open end questions:
    - "It looks like you liked the 60 day spa membership as the perk you most want to receive. Can you tell me why you liked that one the best?"
    - "It looks like you're willing to experience some pretty severe side effects of your treatment – can you tell me a little more about that?"
  - Respondent-friendly anchoring questions (more later)

Rescaling Utilities

- The multinomial logit model scales utilities to predict respondents' observed choices
- As a result logit-scaled utilities allow us to run simulations
- But for some other needs, different utility scaling options may be more appropriate

Rescaling for Comparison

- For some technical reasons we needn't go into here, logit-scaled utilities are larger for respondents who make more consistent choices and smaller for respondents who make less consistent choices
- More and less consistent choices may owe to
  - Respondents paying more or less attention to the survey (due to fatigue, interest in the topic)
  - Respondents differing in cognitive ability
  - Respondents differing in knowledge or experience with the topic of the study
  - Respondents learning their preferences in the course of the MaxDiff experiment
- Sometimes we rescale the utilities so that all respondents have the same range in utilities (i.e. the same magnitude)

Probability Rescaling

- Many audiences find ratio scaled numbers more meaningful, something we can do with probability rescaling
- Utilities rescaled as probabilities
  - Sum to 100%
  - Allow you to interpret a score of 60 as being 4 times as valuable as a score of 15, or 6 times as valuable as a score of 10

Sample Size

- MaxDiff was originally devised to allow respondent-level utility estimates
  - In fact, before the advent of HB analysis, MaxDiff was one of our best ways to get respondent-level utilities (more later)
  - So it scales down very nicely for small samples
  - In studies ranging from as low as 20 to as many as thousands respondents, we can get good data on individual respondents' preferences
- Generalizing sample results to populations requires the same kind of thinking about precision and power that you would do for any other research study
  - General rules of thumb like "a minimum of 300 or 200 per separately reportable subgroup, whichever is greater" usually hold for MaxDiff as well
  - In the book we also cover specific calculations for power analysis, for that handful of clients who need them (typically academics or grant-funded researchers interested in justifying their research designs to journal editors or grant committees)

MaxDiff Utilities are Relative

- We don’t know if all items are good, all are bad, or some are good and some are bad.

Considering only these four diseases, which is the Most Preferred and which is the Least Preferred?

(1 of 20)

Most Preferred		Least Preferred
○	Gum disease	○
○	Broken leg	○
○	Heart attack	○
○	Brain aneurysm	○

Click the 'Next' button to continue...

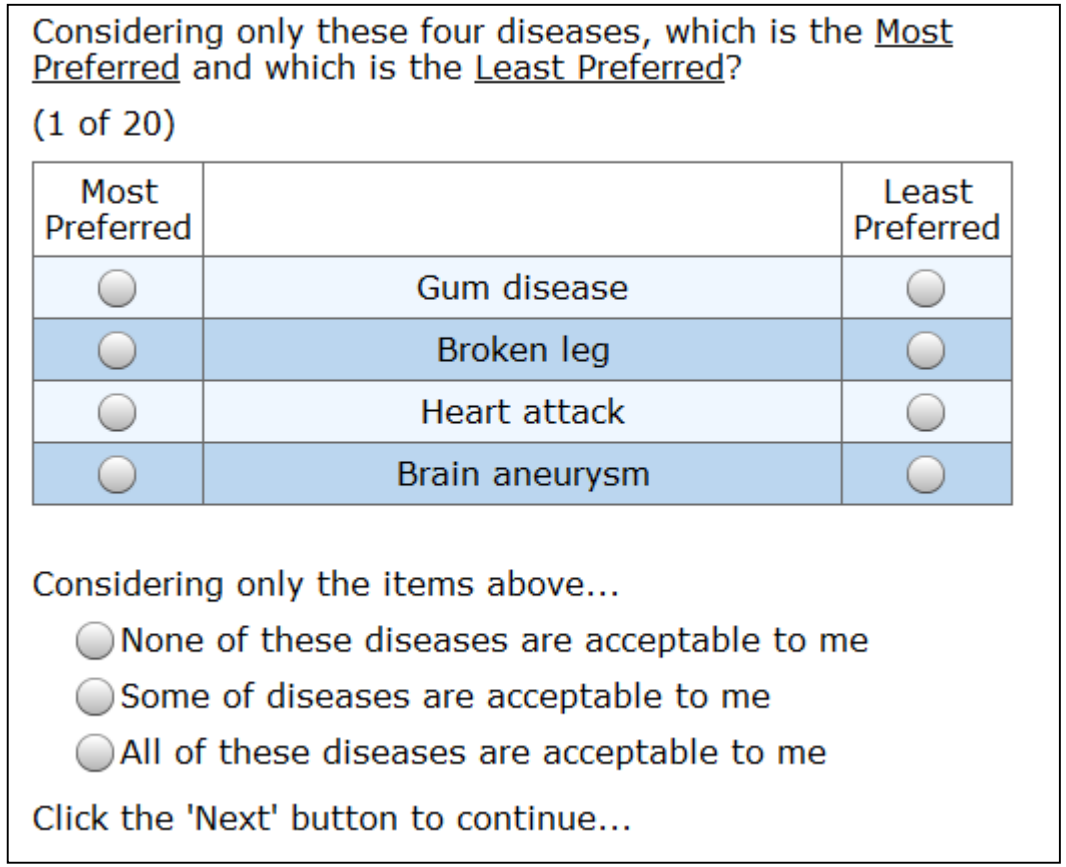
Anchoring MaxDiff Utilities

- Have respondents tell you which of the items from the entire list are acceptable and which are not
  - Or important/not
  - Most preferred/less preferred
  - Would increase chance of buying or not, etc.
- This is the “direct approach”
- There’s also an indirect approach using dual response question

Direct Approach

- Before or after the MaxDiff exercise, simply add a select question to your survey, like a “Yes/No” for each item on the list
- Or ask just a subset of the items
  - Use on-the-fly utilities to identify two items a given respondent likes more than the rest, two toward the bottom of the list and two in the middle
  - Better still, use the on-the-fly utilities to select items evenly spaced throughout a respondents’ rank ordered preferences, like the 1st, 6th, 11th, 16th and 21st ranked items from a list of 21

Dual Response Approach



Considering only these four diseases, which is the Most Preferred and which is the Least Preferred?

(1 of 20)

Most Preferred		Least Preferred
○	Gum disease	○
○	Broken leg	○
○	Heart attack	○
○	Brain aneurysm	○

Considering only the items above...

- ○ None of these diseases are acceptable to me
- ○ Some of diseases are acceptable to me
- ○ All of these diseases are acceptable to me

Click the 'Next' button to continue...

- Note that respondents don't actually indicate which items are preferred/acceptable
- We infer the anchor indirectly

Indirect Approach Analysis

- Simply check the box to include indirect data in analysis
  - “All are good” coded so utilities are higher than anchor value
  - “Some are good” coded so Best utility is higher and Worst utility is lower than anchor value
  - “None are good” coded so all item utilities are lower than anchor value

Large Numbers of Items

- MaxDiff has been a hit with researchers
- This has led to its use in more kinds of research
- It has also led end users to want to push its limits: "If we can do 20 items, how about 30? What about 50? Or 100? 200?"

Handling Many Items

- Several ways presented at Sawtooth Software Conferences
  - Augmented MaxDiff
  - Tailored MaxDiff
  - Express MaxDiff
  - Sparse MaxDiff
- One won out – Sparse MaxDiff

Sparse MaxDiff

- Sparse = show each item fewer than 3 times per respondent
  - 60 items shown in 30 sets of quads (each item shown twice)
  - 100 items shown in 20 sets of quints (each item shown once)
  - 36 items shown in 9 sets of quads (each item shown once)

Bandit MaxDiff

- Sometimes we only want to identify winners and we don’t need respondent level utilities
- In cases like this we can use Bandit MaxDiff, an organized way to have the survey adapt across respondents, so that earlier respondents identify more and less liked items and later respondents sample liked items more heavily
- The result is great precision about the winning items, with less precision about the large number of losers

Subsequent Analyses

- We talked about
  - Sub-group analyses
  - Tailoring subsequent questions with on-the-fly utilities
  - Segmentation
- Also
  - Simulations: we can predict the share of respondents preferring one item from a set of other items
  - TURF analysis: we can identify the "reach" of a bundle of items (e.g. "what's the set of 6 ice cream flavors that gives the most respondents a flavor they like")

Profile Case MaxDiff

- Sometimes we want to create a hybrid of MaxDiff and conjoint analysis
- This has been called
  - “Best-worst conjoint” or
  - “Best-worst case 2 scaling” or
  - “The profile case of best-worst scaling” (as opposed to the “item case” covered so far)
- Unlike other forms of conjoint analysis, this allows cross-attribute level comparisons
- This may be an easier way to administer conjoint analysis questions to some audiences (e.g. it’s commonly used in healthcare research)

Profile Case MaxDiff

Which of these features of a subscription TV service would most make you want to subscribe and which would least make you want to subscribe?

(1 of 11)

Most		Least
◯	Premium movie channels: HBO, Showtime	◯
◯	200 cable channels	◯
◯	2 free pay-per-view events per year	◯
◯	One time set-up cost: $199	◯
◯	Monthly cost: $24.99	◯

Combining BW-Case 2 with CBC

- We can augment the best-worst profile question with a follow-up choice comparing the entire product profile to a none alternative

Which of these features of a subscription TV service would most make you want to subscribe and which would least make you want to subscribe?

(2 of 11)

Most		Least
◯	One time set-up cost: $149	◯
◯	Premium movie channels: HBO, Showtime, STARZ, Epix	◯
◯	150 cable channels	◯
◯	Monthly cost: $19.99	◯
◯	2 free pay-per-view events per year	◯

Given what you know about the cost of TV services, would you subscribe to this one or not?

- ◯ No
- ◯ Yes

Or Add a Purchase Intent Question Instead

Which of these features of a subscription TV service would most make you want to subscribe and which would least make you want to subscribe?

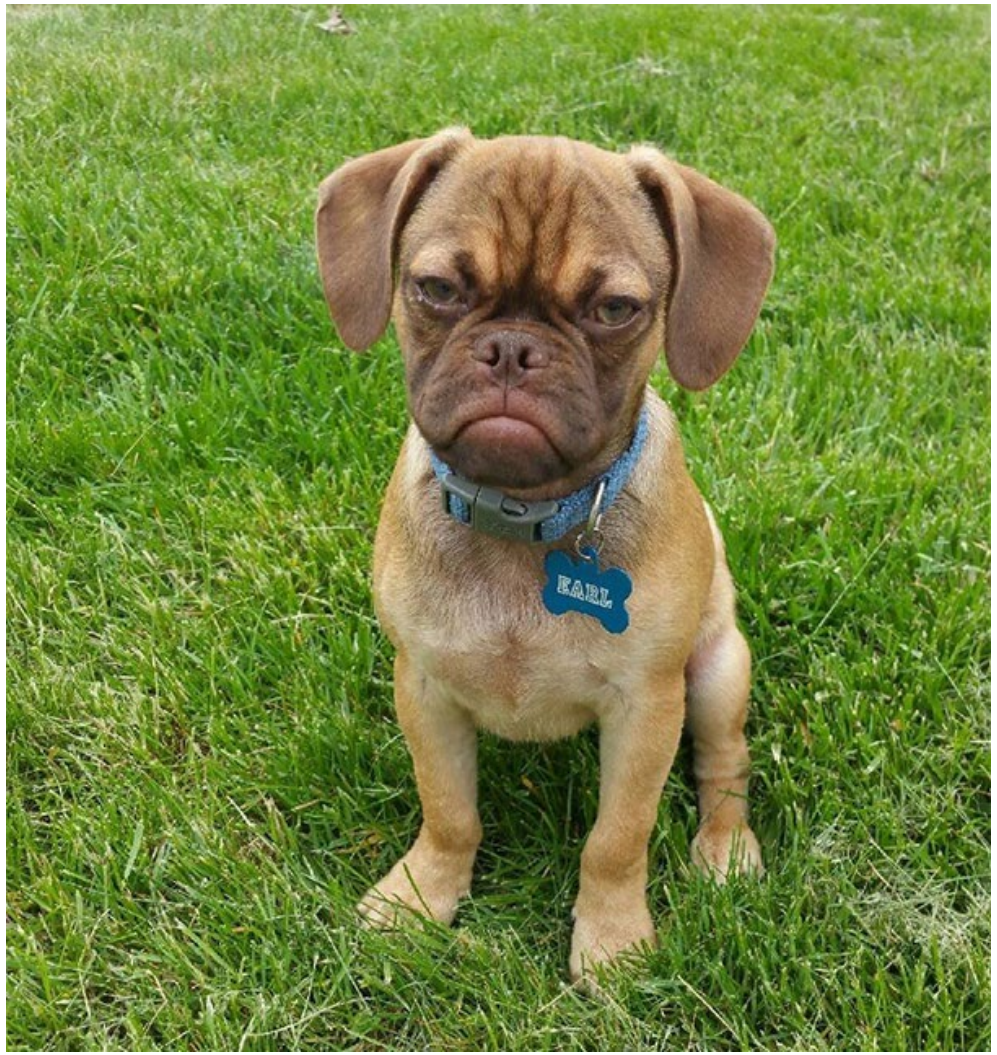
(3 of 11)

Most		Least
◯	One time set-up cost: $199	◯
◯	200 cable channels	◯
◯	Monthly cost: $19.99	◯
◯	Premium movie channels: HBO, Showtime, STARZ, Epix	◯
◯	6 free pay-per-view events per year	◯

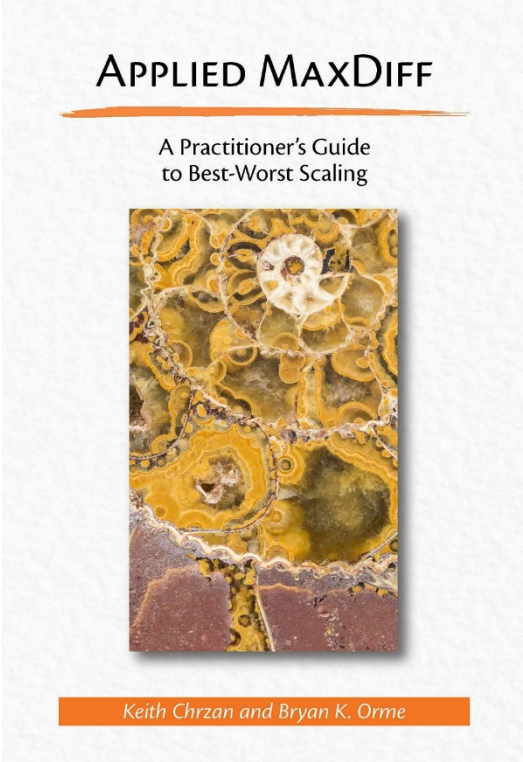
How likely would you be to subscribe to this TV service?

- ◯ I definitely would NOT subscribe
- ◯ I probably would NOT subscribe
- ◯ I might or might not subscribe
- ◯ I probably would subscribe
- ◯ I definitely would subscribe

Summary



For more information . . .



Questions?

keith@sawtoothsoftware.com