Alternative Approaches to MaxDiff with Large Sets of Disparate Items

- Presented at the Sawtooth Software Conference

October 2007

Dr. Phil Hendrix
immr

1 (770) 612-1488
phil.hendrix@immr.org

Stuart Drucker
Drucker Analytics

1 (312) 642-1202
sdrucker@pipeline.net

Contents

- Applications of MaxDiff to Large Item Sets
- Complications that Arise and Consequences
- Mitigating the Consequences
- Augmented MaxDiff and Tailored MaxDiff
- Study Design
- Results
- Discussion
- Q&A

Acknowledgements

Programming/Hosting:
Ordinary and Tailored MaxDiff

Programming/Hosting:
Q-Sort

Sample

Applications of MaxDiff to Large Item Sets

Why MaxDiff?

- Ø Versatility
- Ø Flexibility
- Ø Ease-of-use for respondents
- Ø Discrimination
- Ø Robustness

Large Item Sets

Theoretically, no limit	In practice, 40+ items

Examples

Moriarity – MSI 2007 ü Examined 50+ items ü Brands, celebrities, and icons	Hendrix and Drucker – 2007 ü Examined 40 items ü Benefits of new wireless services

Illustration – Benefits of Wireless Services

Alerts	§ Weather § Other Time Sensitive
Banking	§ Check accounts § Send/receive money § Pay for items
Calling	§ Reliable/problem-free calls in buildings, public places § Free domestic calls § Low-cost internat'l calling
E-mail	§ Read/reply to e-mail
Entertainment	§ FM/Satellite radio § Television (live) § Television + DVR § Music
Functions	§ Electronic Ticketing § Remote – home § Remote – TV § Security Monitor
Games	§ Against others § Individual

Health	§ Track exercise § Medical profile § Monitor condition
Imaging	§ Take/send photos § Record/send video § Video phone
Information	§ Learn/receive instructions
Internet	§ Access/use Internet
Messaging	§ Send/receive text message
Navigation	§ Directions § Real-time Routing § Tour Guide
Services	§ Concierge
Shopping	§ Business Finder § Comparison shopper § Customized offers § Digital Profile
Social	§ Social Networking § Find others' locations

Common Ingredients

- Items to scale in terms of some property
  - Desirability
  - Importance
  - Value
  - Similarity/proximity
  - etc.
- Respondents able to compare items

MaxDiff particularly well-suited

immr

Complications and Consequences

As the number of items increases, complications arise when applying Ordinary MaxDiff

	For Example:	
	40 items requires...	**60 items requires...**
Ü **Design gets large**	24 Choice Sets	36 Choice Sets
Ü **Task gets burdensome**	8 - 10 mins.	10-15 mins.

Assumptions:

- Five items per task yields recommended average of 3x/item exposure to capture HB estimates with “reasonable precision” (Orme, 2005)
- Four to five items is optimal for MaxDiff experiments (Chrzan and Patterson, 2006)

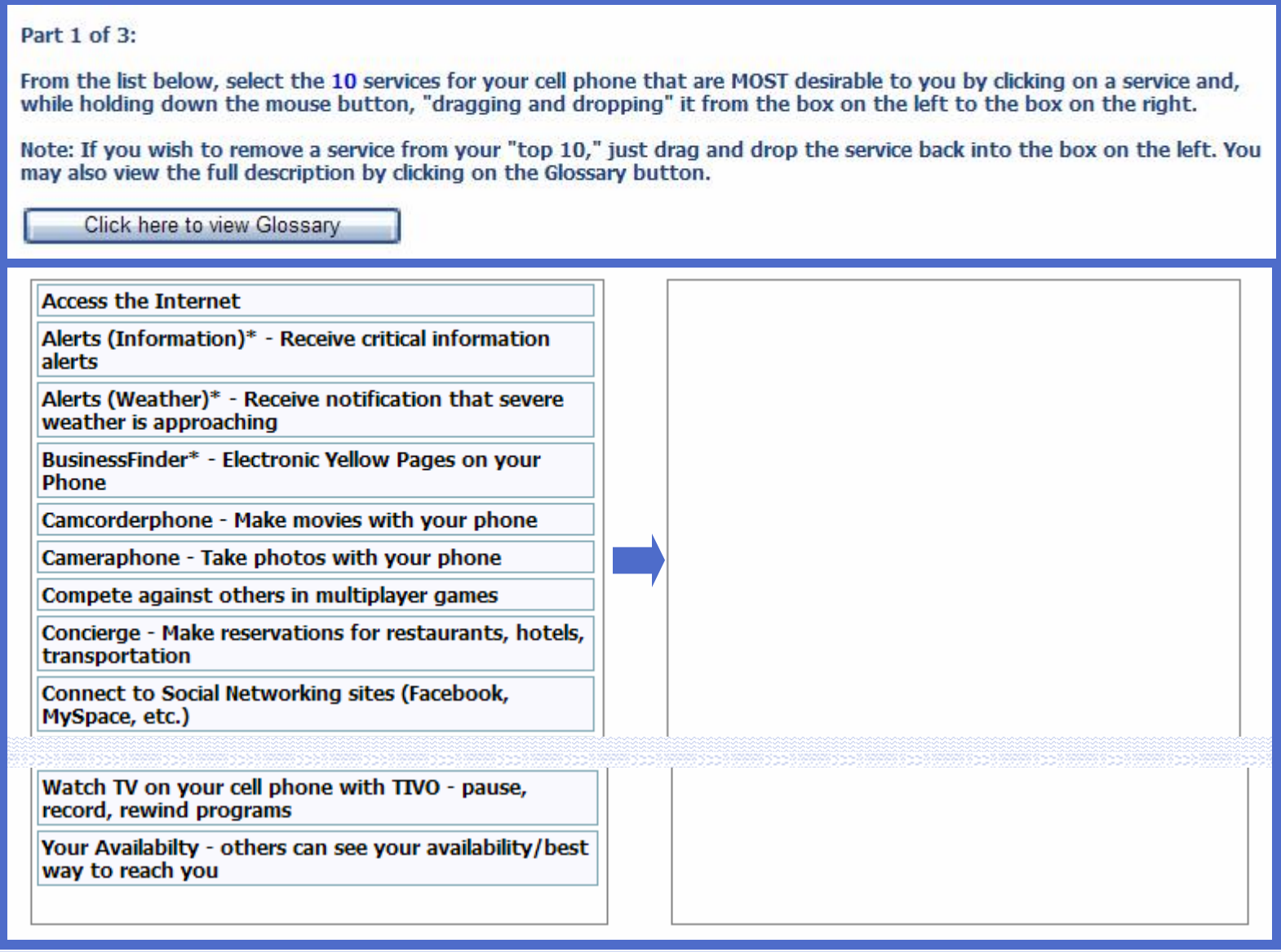
Handling Many Items – Prior Approaches

Prior Approaches	Description
Ü **Ordinary MaxDiff**	§ Prior to generating design, items equally likely to be presented across respondents
Ü **Adaptive MaxDiff (Orme 2006)**	§ Choice sets dynamically constructed for each respondent, based on responses within own set of MaxDiff tasks

Handling Many Items – Alternative Approaches

New Approach	Description
Ü **Augmented MaxDiff**	§ Ordinary MaxDiff, with data from Q-Sort augmenting MaxDiff exercise § From comparisons of items across categories (e.g., items in Category 1 preferred over items in Categories 2, 3, and 4; etc.), we construct sets of judgments that reflect preference structure of each item § Concatenated data set utilized in model estimation
Ü **Tailored MaxDiff**	§ Customized Maxdiff + Q-Sort § For each respondent, MaxDiff choice sets constructed from a subset of items, chosen by disproportionate sampling that corresponds to respondent's preferences revealed from a priori Q-Sort § Data from Q-Sort augments information from MaxDiff exercise as per Augmented MaxDiff § Concatenated data set utilized in model estimation

Q-Sort Task, using Drag and Drop UI



immr

Pros, Cons of Alternative Approaches

Approach	Pros	Cons
Ü Ordinary MaxDiff	§ Straightforward to design, implement, analyze	§ As # of items grows, insufficient data for individual-level utility estimates without increasing number of tasks § Potential for respondent fatigue
Ü Adaptive MaxDiff	§ Robust estimates § More meaningful tasks § Better estimation of "best" preferences than Ordinary MaxDiff	§ Requires custom programming § Sensitivity to decision rules TBD § Potential loss of estimation precision on "worst" preferences

Approach	Potential Pros	Cons
Ü Augmented MaxDiff	§ Respondent engagement § Reduce # of MaxDiff tasks § All items shown in MaxDiff § No loss in precision of estimates for average item	§ R must complete Q-Sort § Data from Q-Sort must be formatted appropriately to estimate HB models in practical amount of time (e.g., LT two days) § Sensitivity to degree of categorization in Q-Sort TBD
Ü Tailored MaxDiff	§ Respondent engagement § Reduce # of MaxDiff tasks § Subset of items shown in MaxDiff § No loss in precision of estimates for most preferred (best) items	§ R must complete Q-Sort in advance of MaxDiff section of study § Requires custom programming § Subset of items shown in MaxDiff § Sensitivity to item sampling TBD

Overview and Description of Process

Q-Sort

- R sorts all p items into m groups, based on criterion
- *For example, sort 40 items into 4 quartiles, based on desirability*

Rank Items in Selected Categories (optional)

- *For example, R rank orders items in top quartile*

Augmented MaxDiff

- Create couplets from Q-Sort data, showing ordered relationships for each item in Q-Sort groups
- *For example, all items in the first quartile are preferred to those in the second, third, and fourth quartiles*

Tailored MaxDiff

- Select subset of items shown to R using disproportionate stratified sampling
- *For example, select 10 items from R's first quartile, 8 from the second, etc.*

Models Examined

- Ordinary MaxDiff
- Augmented MaxDiff, using data from Q-Sort
  a. Four quartiles
  b. Top quartile, middle half, bottom quartile
  c. Top quartile vs. all other
  d. All 4 quartiles, but smaller number of MaxDiff tasks
- Tailored MaxDiff (four Q-sort quartiles)

Augmented MaxDiff

Augmented MaxDiff utilizes data provided by a respondent from an auxillary task - in this case, a Q-Sort - to:

- Create additional judgments that can be unfolded into respondent-level preferences. Underlying both MaxDiff and Q-Sort is Thurstone's Law of Comparative Judgment
  - MaxDiff - from most and least preferred items in a given choice set, we can infer that the most preferred choice beats all other items in that set, and the least preferred choice is beaten by all remaining items in the set
  - Q-Sort - after respondents assign the items to a pre-specified number of categories, we infer that items in:

    Category 1 > Category 2 > Category 3 > Category 4

    Category 2 > Category 3 > Category 4

    Category 3 > Category 4

Tailored MaxDiff

Tailored MaxDiff utilizes data provided by a respondent from an initial task - in this case, a Q-Sort - to:

- Select a subset of items, using a disproportionate sampling rule, to individually tailor the MaxDiff sets of items shown to each respondent
- Allow a particular respondent to focus more of his/her comparisons on items most important to that respondent
- Approach is consistent in spirit to Tailored Testing (also known as Adaptive Testing, based on Item Response Theory[1])

[1]Handbook of Modern Item Response Theory. W. J. van der Linden and R. K. Hambleton, eds. Springer, 1997

Example of Tailored MaxDiff for Respondent

Quartile 1	
Item	Sampled
1	x
2	x
4	x
12	x
15	x
16	x
25	x
27	x
29	x
39	x

Quartile 2	
Item	Sampled
3	x
10	x
18	x
21	x
22	x
24	x
28	x
37	x
34	
40	

Quartile 3	
Item	Sampled
5	x
14	x
30	x
38	x
8	
11	
13	
33	
35	
36	

Quartile 4	
Item	Sampled
20	x
29	x
6	
7	
9	
17	
23	
26	
31	
32	

- Sampled 100% items from Quartile 1 (100%), 80% from Quartile 2, 40% from Quartile 3, and 20% from Quartile 4 – accounts for 60% of all items (24 / 40)
- This subset was then matched to an experimentally generated design for 24 items such that each item was seen an average of 3x (theoretically), assuming four items per choice set

Study Design – Sample and MaxDiffs

Population	U.S. cell phone owners age 18-64
Sample size	n = 619
Sample Strata	Gender x Age (M/F x Age 18-29; 30-39; 40-49; 50-64)
Items	40 Wireless Services
Assignment	Random assignment within Strata to one of three cells

MaxDiff Models	# of Items (Total/Subset)	Maxdiff Tasks per Respondent	# of Items per Set	N (Cell size)
1. Ordinary MaxDiff	40	24	5	152
2. Augmented MaxDiff	40	16	5	245
3. Tailored MaxDiff	40/24	18	4	222

MaxDiff Task, Item Explanations

Which one of the following services for your cell phone would be most desirable, and which one would be least desirable?

(Assume that each service is available, reliable and easy to use, and free of charge)

Most Desirable	Table 2 of 24	Least Desirable
[◉]	BusinessFinder* - Electronic Yellow Pages on your Phone	[○]
[○]	Watch TV with TIVO - pause, record, rewind TV programs	[○]
[○]	Text Messaging - send/receive text messages	[◉]
[○]	Camcorderphone - Make movies with your phone	[○]
[○]	Store, share personal profile*	[○]

*Automatically detects your location and provides information specific to your location.

Glossary

1	**Access the Internet**	Access the internet, visit websites, upload/download files
2	**Alerts (Information)* - Receive critical information**	Receive alerts containing time-sensitive information (e.g., change in stock price,
9	**Connect to Social Networking sites (Facebook, MySpace, etc.)**	One-click access to Facebook, Myspace, Twango, other sites where you can share photos, stay connected with friends and family
10	**Control appliances, lighting in your home**	Control lights, appliances, temperature in your home.
11	**Coupons/offers* - receive coupons and offers**	Receive special offers and discount coupons from selected businesses for products and services that you use. Digital coupons/offers transmitted to your phoneto redeem, simply scan digital image at checkout. Offers can be "location-based," e.g., made available when you enter a store.
12	**Determine others' availability, best way to reach**	Indicates if the person you wish to contact is available, and best way to reach them

MaxDiff + Q-Sort – Data Structure

Illustration of Data Structure for Augmented and Tailored MaxDiff

MaxDiff Task

ID	Item 1	Item 2	Item 3	Item 4	Preferred Most	Preferred Least
1000016	1	3	5	20	1	5

Results of Q-Sort (excerpted for items in task above)

ID	Item	Meets Threshold for Quartile 1	Meets Threshold for Quartile 2	Meets Threshold for Quartile 3	Unique Judgments	
1000016	1	Yes			1	
1000016	3	No	Yes		2	
1000016	5	No	No	Yes	3	
1000016	20	No	No	No	3	

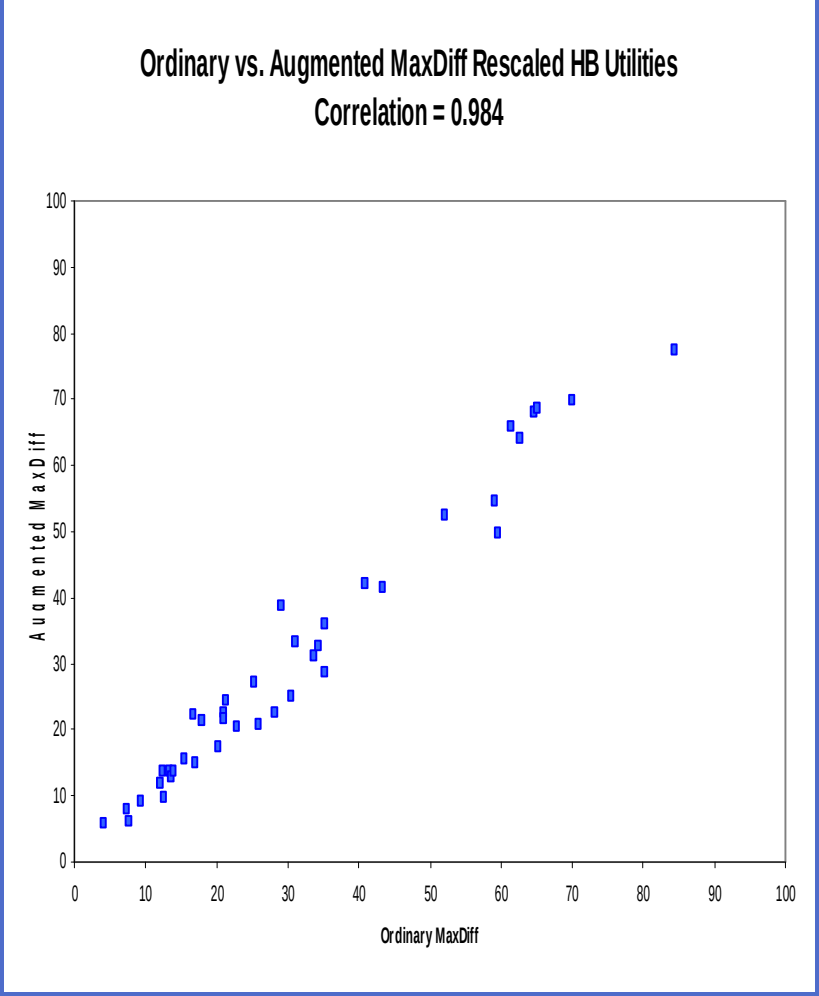
MaxDiff Ü Two multinomial logit choices (most and least preferred)

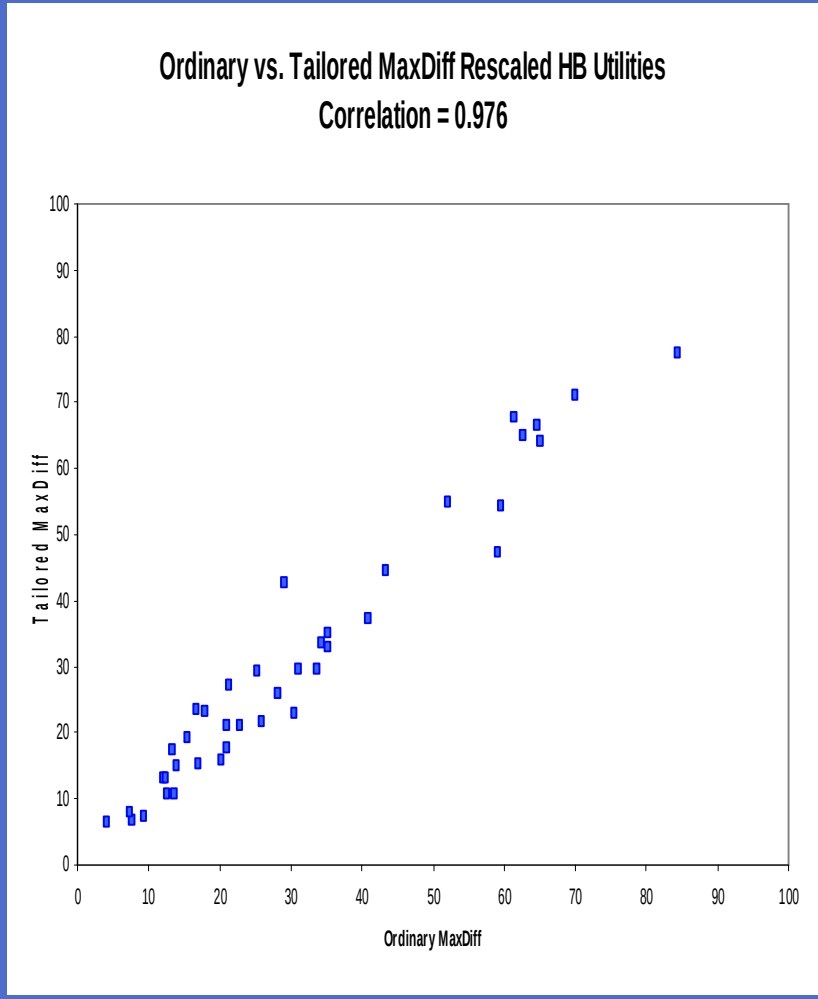
Q-Sort Ü Sets of binary logit choices (meet or doesn't meet threshold)

Model Estimation and Fit Assessment

- Hierarchical Bayesian estimation using CBC/HB
- 30,000 burn-in iterations,10,000 draws saved
- Respondents ranked five randomly selected items from most desired to least desired
- 3 separate holdout tasks per respondent

Convergence of Estimates across Methods





Average Utilities for Services by MaxDiff Method

Item	Statement	Ordinary MaxDiff	Augmented MaxDiff	Tailored MaxDiff	Rank: Ordinary MaxDiff	Rank: Augmented MaxDiff	Rank: Tailored MaxDiff
20	Make free domestic calls	84.4	77.6	77.3	1	1	1
13	Directions	70.0	69.8	71.1	2	2	2
33	Text Messaging	65.0	68.6	64.1	3	3	6
14	E-mail	64.7	68.0	66.5	4	4	4
6	Cameraphone	62.7	64.1	65.0	5	6	5
1	Access the Internet	61.4	65.9	67.5	6	5	3
29	Real-time routing, based on conditions	59.5	49.6	54.3	7	9	8
18	Landline reception at home/public places	59.2	54.4	47.2	8	7	9
3	Alerts (Weather)	52.1	52.5	54.8	9	8	7
2	Alerts (Information)	43.3	41.4	44.5	10	11	10
23	Mobile Tickets - tickets stored on your phone	12.1	11.9	13.2	36	35	33
9	Connect to Social Networking sites	9.4	9.2	7.2	37	37	38
27	Receive lessons, instructions	7.6	6.1	6.7	38	39	39
32	Store, share personal profile	7.4	7.8	8.0	39	38	37
7	Compete against others in multiplayer games	4.1	5.7	6.5	40	40	40

Results - Prediction of Holdout Tasks

- Both Augmented and Tailored MaxDiff outperformed Ordinary MaxDiff in predicting the average number of items correctly
- Only Augmented MaxDiff had highly statistically different results for both highest and lowest rank – Tailored MaxDiff had a directional difference for predicting the highest (best) rank
- Spearman's Rho correlation of predicted vs. actual ranking is highest for Augmented MaxDiff, followed by Tailored MaxDiff

MaxDiff Models	Model Fit			Spearman's rho
	Bests %	Worsts %	Avg # Items Correct	
Œ Ordinary MaxDiff	59.4%	55.7%	2.21	.635
• a Augmented MaxDiff	62.4%††	62.2%††	2.50††	.707
Ž Tailored MaxDiff	64.7%†	56.6%	2.40††	.676

††Significantly different from Model 1 at p=0.05 †Significantly different from Model 1 at p=0.10

Augmented MaxDiff: Reducing Information

- Neither reducing the number of Q-Sort Categories by collapsing the number of judgments nor running a separate model where we removed ½ of the MaxDiff tasks had an adverse impact on predicting the highest (best) rank %
- However, reducing the number of Q-Sort Categoryies analytically resulted in a statistical difference relative to predicting the lowest rank %, and a simple top 10 vs. all other approach yielded a statistically different number of average items predicted correctly

Augmented MaxDiff Models	Model Fit			Spearman's rho
	Bests %	Worsts %	Avg # Items Correct	
• a **All four quartiles**	62.4%	62.2%	2.50	.707
• b **Three Categorys: 10/20/10**	62.0%	59.2%††	2.45	.693
• c **Two Categorys: 10/30**	62.6%	56.3%††	2.35††	.682
• d **Four quartiles/8 tasks**	61.1%	61.0%	2.43	.696

††Significantly different from Model 2a at p=0.05 †Significantly different from Model 2a at p=0.10

Respondents' Ratings of Tasks

	Respondents' Ratings on a 1-7 (Strongly Disagree to Strongly Agree) Scale				
	Enjoyable	*Confusing*	*Easy*	*Made me feel like Clicking*	*Allowed you to express your opinion*
Œ **Ordinary MaxDiff**	4.88	2.62	5.47	3.13	5.35
• **Augmented MaxDiff**	5.20††	2.32†	5.65	2.70††	5.72††
Ž **Tailored MaxDiff**	5.19††	2.57	5.61	2.77†	5.68††

††Significantly different from Model 1 at p=0.05 †Significantly different from Model 1 at p=0.10

Conclusions

1. Augmented MaxDiff showed a clear ability to predict a higher best rank % and worst rank % correctly vs. Ordinary MaxDiff, as well as a higher average number of items
2. Tailored MaxDiff predicted a higher number of items correctly than Ordinary MaxDiff, but only directionally the best rank %
3. Collapsing four quartiles into three or even two Categories to augment an Ordinary MaxDiff exercise can be considered if one is willing to accept a decreased hit rate for the worst %
4. With a four quartile Q-Sort, including as few as eight MaxDiff tasks with a Q-Sort exercise will predict key fit measures as well as with the same approach with 16 MaxDiff tasks
5. Both Augmented and Tailored MaxDiff methods result in an exercise that respondents find to be enjoyable and allow them to express their opinions
6. There is no clear winner between Augmented and Tailored Maxdiff, but if one had to choose, Augmented MaxDiff with a full Q-Sort shows overall better prediction ability than Tailored MaxDiff in recovering rankings compared to the Ordinary MaxDiff method

Limitations

1. This study was conducted with a moderately large number of items (40) – results may vary with smaller or larger sets of items

2. Disproportionate sampling weights for the Tailored MaxDiff were determined through judgment; a comprehensive investigation of differential sampling schemes needs to be considered before drawing definitive conclusions

3. The decision to use equally sized quartiles for executing the Q-Sort exercise was driven by previous work; other schemes should be tested as well

4. The assumption is that individual level utilities are necessary

5. Given the sample size, we did not investigate differences between methods on potential subgroups

6. Programming the Tailored MaxDiff as well as formatting the Q-sort results into a form to concatenate with MaxDiff data requires hands-on skills

Recommendations for Future Research

1. Replicating the study for a larger number of items, other product categories, and populations, as well as with attitudinal research (e.g. lifestyle characteristics, psychographics, and the like)

2. Conducting Augmented Maxdiff with a larger number of items, with a variable number of Categories in the Q-Sort, and evaluating the impact on holdout task prediction

3. Testing Tailored MaxDiff with differential disproportionate weighting schemes and number of items selected to retain for the MaxDiff portion of the exercise

4. Evaluating both Augmented as well as Tailored MaxDiff results with respect to distinguishing segments, line optimization, etc.