PROCEEDINGS OF THE SAWTOOTH SOFTWARE CONFERENCE

March 2012

"THE INDIVIDUAL CHOICE TASK THRESHOLD" NEED FOR VARIABLE NUMBER OF CHOICE TASKS

PETER KURZ
TNS INFRATEST FORSCHUNG GMBH
STEFAN BINNER
BMS - MARKETING RESEARCH + STRATEGY

SUMMARY

This paper reflects on the fact that CBC is the most widely used conjoint methodology in our industry, and that the authors' experience shows that fewer tasks are often all that is needed (fewer than typical norms in the industry). Furthermore, with the widespread use of online panel sample, panel providers and respondents are pushing for fewer tasks in CBC questionnaires. By observing individual respondents completing CBC surveys, we show that many respondents reach a point in the CBC survey where they disengage and begin to give data of questionable quality—a point we call the "Individual Choice Task Threshold (ICT)." We re-analyzed twelve commercial CBC data sets to see if quantitative measures could be used to somehow detect the threshold point at which respondents disengaged. The idea is that if somehow this could be detected in real-time, then the survey could not ask any more tasks, respondent burden could be reduced, and fewer overall choice tasks would need to be asked for the sample, while maintaining equal or better results. Even if this were never possible, we might benefit from finding the ICT after the fact and at least dropping the later tasks from the analysis.

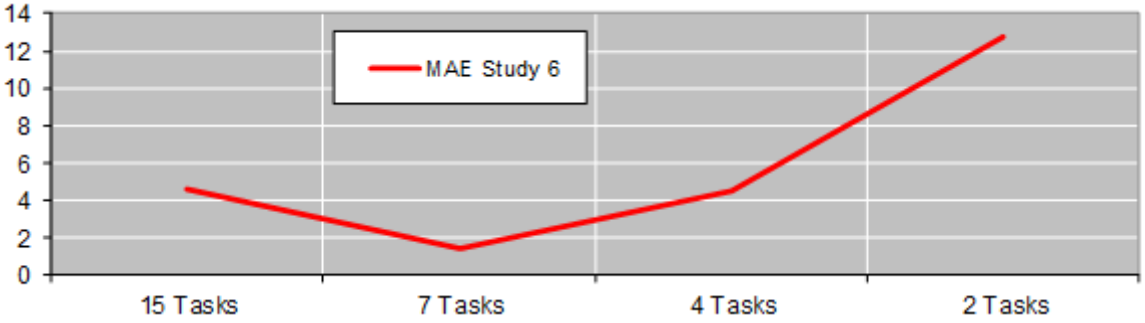
The first part of the analysis focuses on measures of internal fit to the tasks used in estimation (RLH), which is used also in an indexed form to take the decrease of the RLH measure into account when more information (higher number of choice tasks) is available. Using the existing data sets, it could be demonstrated that strategically throwing away 38% of the choice tasks would not lead to a very large decrease in the predictive quality of the CBC models. The results show clearly that most respondents use simplification strategies in later compared to earlier tasks. Based on these findings, the last part of the paper presents some ideas as to how future development could take the ICT in online surveys into account to shorten the interview. But, up to now there are no solution available to do this in real-time and apply it during data collection. Further research and more computational power are needed to solve this problem.

Introduction

Since its introduction, CBC (choice based conjoint) has become the most established method for conjoint analysis. Thousands of conjoint studies are conducted by the market research industry every year. Especially with the application of HB, alternative specific designs, disproportional sampling and long experience, CBC usually leads to high-quality results, successful projects and satisfied clients.

Despite the success researchers have with CBC, there are some observations which may be troubling:

- **Are fewer tasks enough?**
  Prior research (Kurz, Binner; 2010) showed some surprising results: Not using the last choice tasks did not automatically decrease the quality of simulations! As following example shows, the MAE[8] was unchanged even when we reduced the number of tasks from 15 to only four.



- **Do we bedevil respondents?**
  Answering conjoint choice tasks is a complex and monotonous exercise. More and more often one can hear or read comments from respondents complaining about conjoint surveys. Many find the interviews too long and complex. Especially in view of the consolidation of online panels, one more often finds experienced respondents who express their dissatisfaction. As a consequence, higher incentives are necessary to keep them answering.

- **Is there a choice task threshold?**
  Observations during personal conjoint interviews regularly show that some respondents mentally opt-out very early and some apply simplification in the choice tasks, while others remain engaged until the end of the exercise. This indicates that every respondent has her own "Choice Task Threshold" after which she is no longer contributing useful answers.

- **How can they be so fast?**
  As is widely known (Johnson, Orme; 1996) the answer times for the last choice tasks tend to decrease, which could be an indication of simplification or less concentration.

[8] MAE (Mean Absolute Error) calculations in this paper are done at the individual-level. For the holdout, the selected alternative gets 100% share, and the other alternatives 0%. For predictions, the share of preference (logit) rule is used. The average of the absolute differences between predicted and holdout shares is computed as MAE for each respondent, and the results averaged across respondents.

Hypotheses for This Paper

Based on the above observations we formulated the following hypotheses:

H1: The observed task simplification of some respondents leads to less accurate prediction of market shares.

H2: If respondents lose their interest in the conjoint exercise, this at least results in more noisy data.

H3: Our results can therefore be improved when we use only those "good" tasks in which respondents were concentrating and paid attention.

H4: The shift in price sensitivity of later tasks (Johnson, Orme; 1996) could also be avoided by using "good" tasks only.

H5: Finally, interview time could be shortened, if we are able to avoid "bad tasks" during the interview, leading to reduced cost or a larger sample for the same cost.

Basis for the Analysis

For the purposes of this paper 12 commercial studies with a total of 4,952 respondents and 67,413 choice tasks covering nearly all industries and topics were analyzed. The 12 studies included "brand + price" CBC designs as well as designs with many attributes (between 10-15 choice tasks and 14 to 105 estimated parameters). The selected studies cover all main market fields such as industrials, durables, and FMCG, in both B2B and B2C markets. They were conducted in all parts of the world using state-of the-art computer aided interview delivery and mainly recruited from online panels. Finally all 12 studies were developed in Sawtooth Software's CBC software, with all analysis done with HB using default settings.

	Project 1	Project 2	Project 3	Project 4	Project 5	Project 6
Industry:	Airline	Automotive	Media	FMCG	Finance	Technology
Target Group:	B2C	B2C	B2C	B2C	B2C	B2C
Sample Size:	N=434	N=240	N=840	N=460	N=513	N=802
Interview Delivery:	CAWI	CAWI	CAWI	CAWI	CAWI	CAWI
# Choice Tasks:	14	10	15	15	15	15
Holdout (#):	Random/Fix 6	Random	Random/Fix11	Random/Fix12	Random	Random
# Est. Parameter:	39	25	42	105	19	84
# Concepts / Task:	5	4	3	9	3	3
Conjoint Method:	STD CBC	STD CBC	CBC ASD	CBC ASD	STD CBC	CBC ASD
Number of Choices	6076	2040	12600	6900	7695	12835
	Project 7	Project 8	Project 9	Project 10	Project 11	Project 12
Industry:	FMCG	Construction	Construction	Durable	Food	Durable
Target Group:	B2C	B2C	B2C	B2B	B2C	B2B
Sample Size:	N=300	N=212	N=212	N=260	N=300	N=379
Interview Delivery:	CAWI	CAWI	CAWI	CAWI	CAPI	CAWI
# Choice Tasks:	13	10	10	10	12	10
Holdout (#):	Fix 5	Random	Random	Random	Random	Random
# Est. Parameter:	39	24	24	20	14	15
# Concepts / Task:	3	12	13	7	4	13
Conjoint Method:	CBC ASD	CBC ASD	CBC ASD	CBC ASD	CBC ASD	CBC ASD
Number of choices	3900	2120	2120	2600	3600	4927

Figure 1: Overview of studies analyzed

All 12 studies were analyzed, task by task, in terms of

- Individual time needed to answer the tasks
- Individual RLH reached per task (absolute and indexed)
- Individual utilities derived with different numbers of tasks
- Hit Rates and MAE for fixed and random tasks with different numbers of tasks (for consistency, second-to-last random task was used for fit measurement).

The computations were monitored by four standard measures, which are based on the calculation of the probability of each respondent choosing as she did on each task, by applying a logit model using current estimates of each respondent's part worths. The likelihood is the product of those probabilities over all respondents and tasks. Usually the logarithm of this measure is used, the “log likelihood”. Measures were:

Percent Certainty - indicates how much better the solution is than chance, as compared to a “perfect" solution.
Root Likelihood (RLH) - nth root of the likelihood, where n is the total number of choices made by all respondents in all tasks. RLH is therefore the geometric mean of the predicted probabilities.
Variance - the average of the current estimate of the variances of part worths, across respondents.
RMS - the root mean square of all part worth estimates, across all part worths and over all respondents.

Answer Times and RLH

As expected, the average answer times decline to a low and stable level within a few choice tasks:

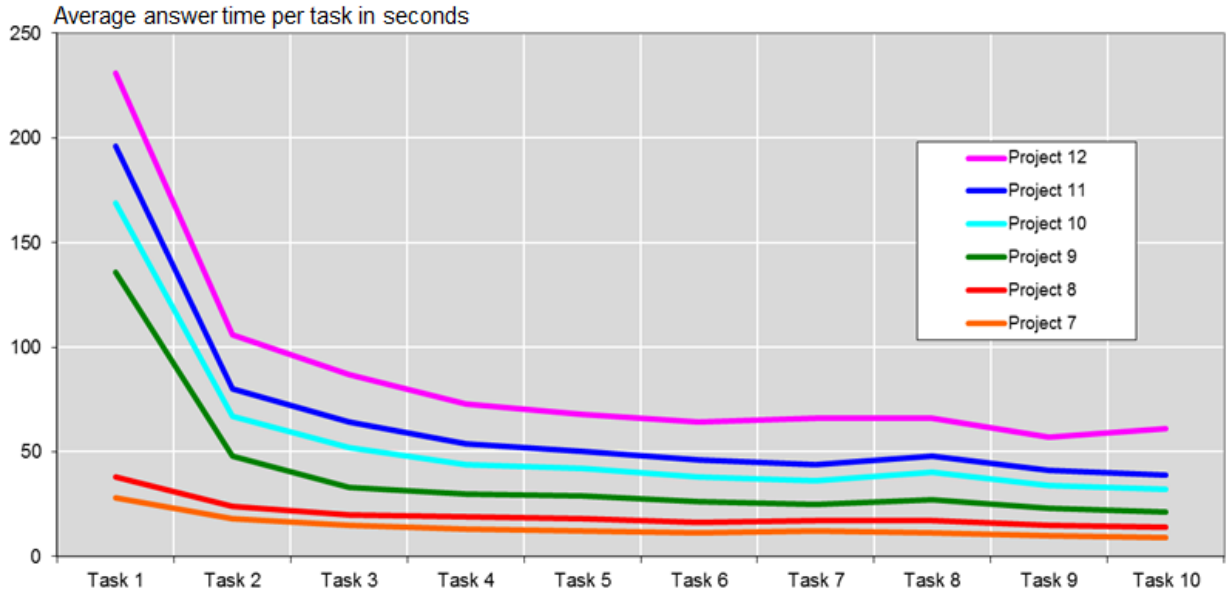


Figure 2: Average answer times throughout conjoint questionnaires

However there was no significant correlation between the average answer time and the RLH level. In addition, changes in individual answering time between choice tasks had no significant influence on the RLH of the answer. Slower or faster answer times for individual choice tasks are not good indicators for less attention, shortcutting or reaching the choice task threshold, as the times can be influenced by respondents' environment or the utility balance of the tasks. Furthermore the aggregated RLH was stable or even declining from choice task to choice task:

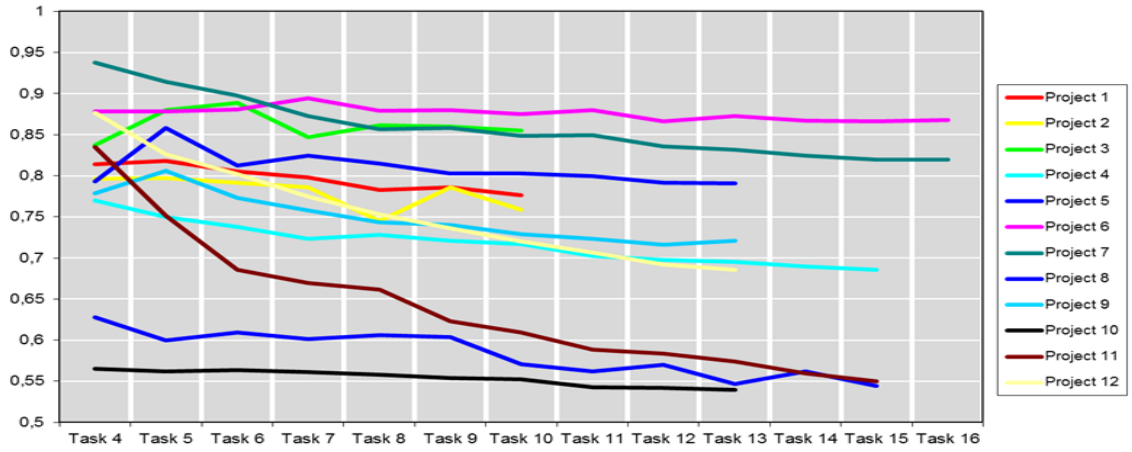


Figure 3: Aggregated RLH throughout conjoint questionnaires

This result was confirmed by all four measures for monitoring the HB process, as shown in the next four graphs. In these figures, and in most later results, the data plotted for "Task 4" is the overall figure from an analysis using only the first 4 tasks for each respondent, while that for

"Task 5" is from a re-analysis using only the first 5 tasks for each respondent, etc. In effect, "Task n" reflects a possible level of the ICT (individual choice threshold) that is being evaluated. We did not analyze using three or fewer tasks.

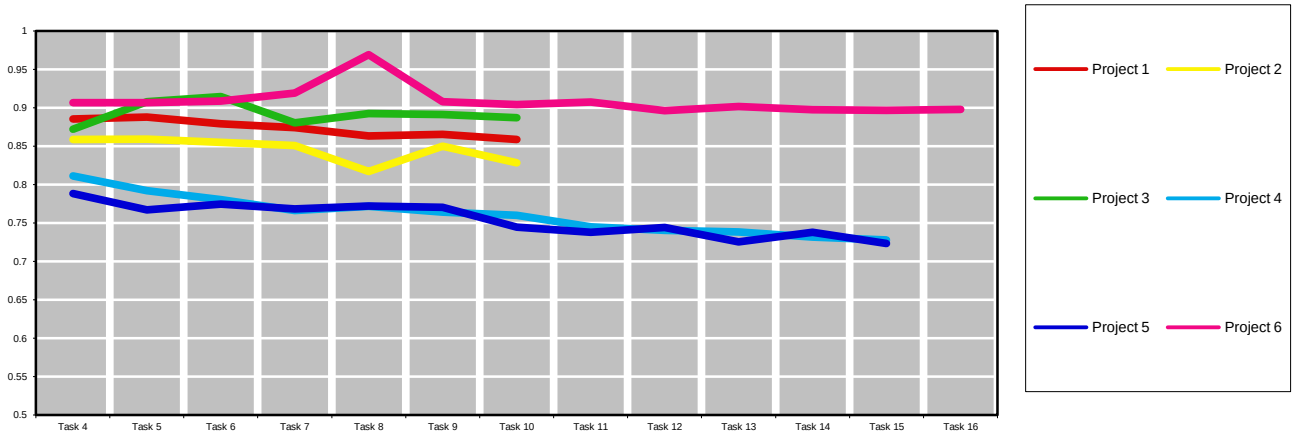


Figure 4a: Different measures throughout conjoint questionnaires – Percent Certainty

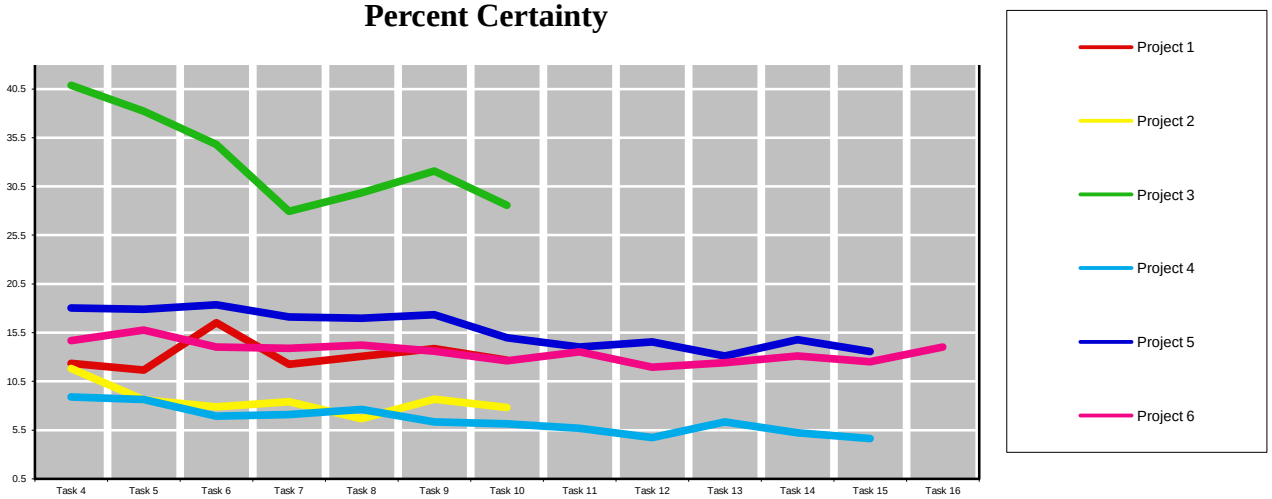


Figure 4b: Different measures throughout conjoint questionnaires – Root Likelihood (RLH)

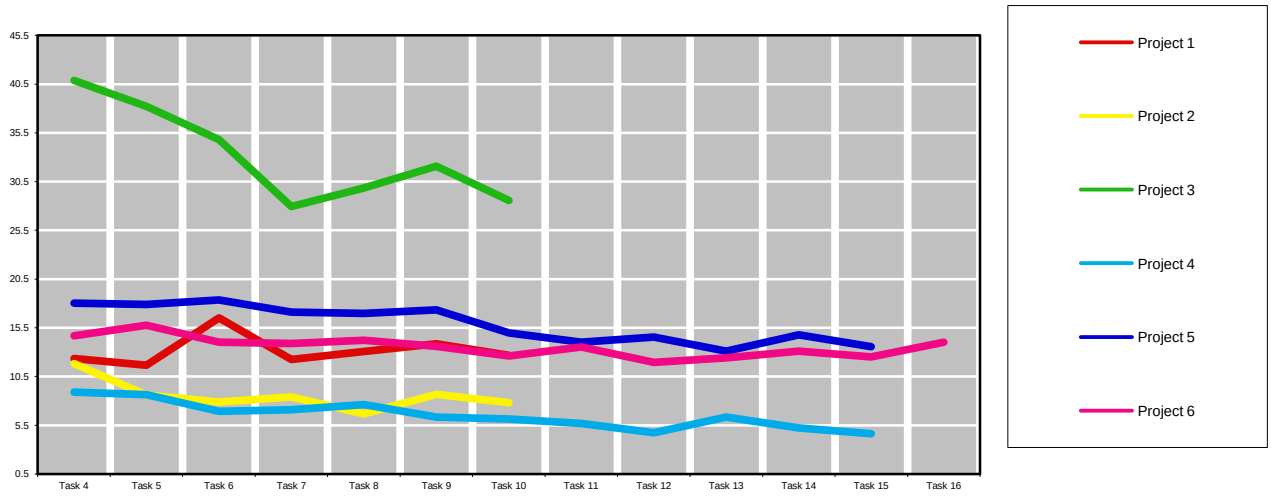


Figure 4c: Different measures throughout conjoint questionnaires – Variance

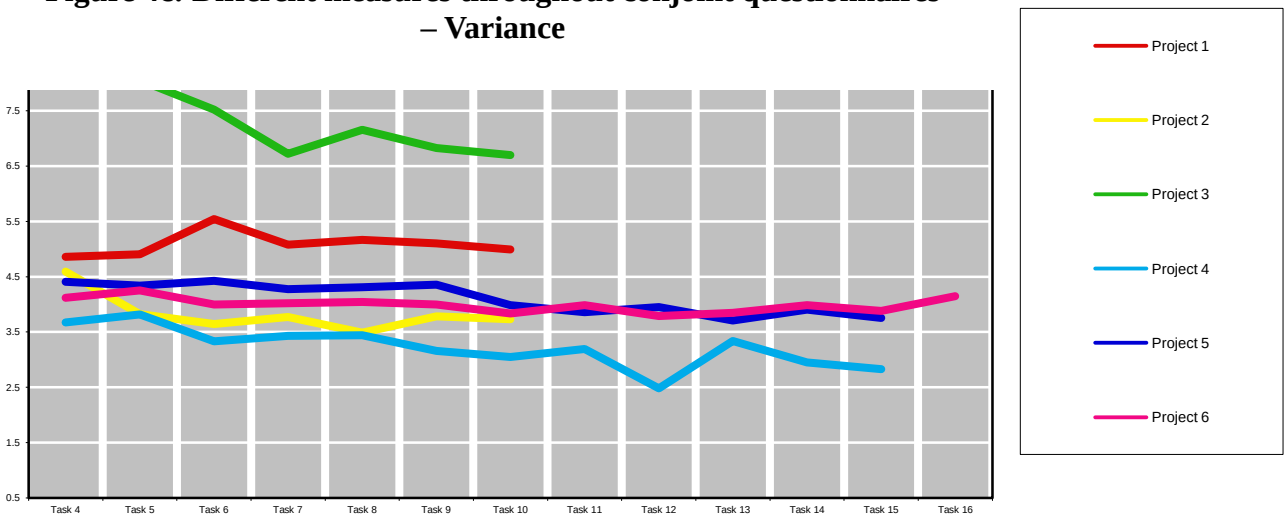


Figure 4d: Different measures throughout conjoint questionnaires – Parameter RMS

On an aggregated level we see that from 4th to last choice task all four measures tend to the same values.

Would that mean that four choice tasks are enough? This is certainly not the case. These measures are only sample averages and don't provide any information about the heterogeneity of respondents.

The individual RLH results showed a different picture, consistently in all the 12 studies. Here as an example is a cutout of individual data from one of our studies:

Example (Individual data Study 7):

Task 4	Task 5	Task 6	Task 7	Task 8	Task 9	Task 10	Task 11	Task 12
523	435	460	408	445	450	504	515	529
690	676	579	586	557	576	572	509	448
634	681	688	716	741	732	762	765	757
506	563	467	431	482	382	404	437	439
746	764	697	719	748	714	663	646	639
611	601	618	649	681	702	718	727	731
709	778	763	793	808	780	794	772	785
539	522	579	379	374	393	417	412	414
857	828	806	831	843	861	865	858	852
718	685	496	513	428	460	397	441	394
798	811	798	782	653	678	686	483	440
726	729	728	738	679	691	685	675	696
752	653	511	430	465	443	475	515	521

Figure 5: Typical individual RLH results

The highest individual RLHs (marked in red and underlined) are reached at a different number of choice tasks for each respondent. Also in a closer look at the data, where we counted how often the maximum individual RLH was reached at each number of tasks, we found no clear pattern.

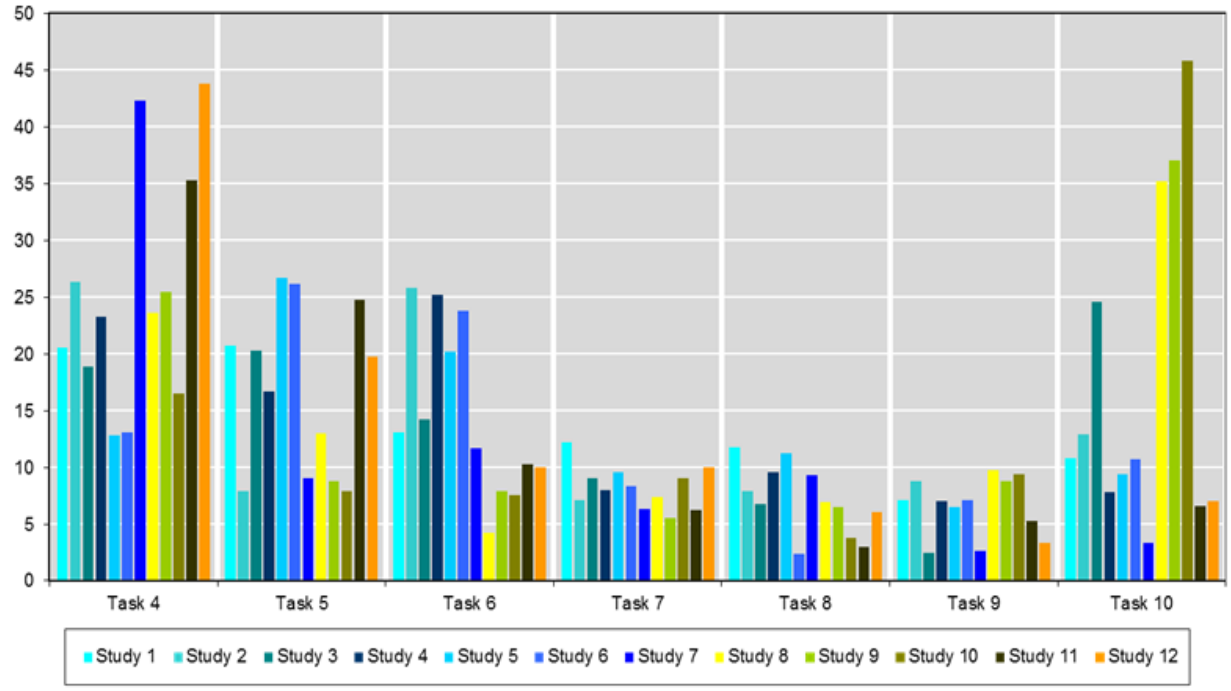


Figure 6: Highest individual RLH per task in %

Knowing that the RLH measure is dependent on the amount of information we use for the HB estimation, RLH might not be the best measure for comparisons across different numbers of tasks. RLH inherently decreases with the number of choice tasks used for deriving the part worth utilities (simply because having more tasks makes overfitting less possible). Therefore we repeated our analysis by creating an index taking the aggregated difference in RLH from one choice task to the next into account when identifying individual decreases in RLH. However, the results were the same: much individual variation in when the highest RLH was achieved.

Impact on Share Prediction

Our next question is whether there is a difference in share prediction with different numbers of choice tasks. Therefore we created datasets with increasing number of choice tasks and ran simulations against hold-out tasks. The MAE (mean absolute error) tends to decline with an increasing number of choice tasks. However, as Figure 7 shows, in some studies there is a stagnation after 6 or 7 choice tasks; in one study the MAE was even increasing after that. Overall there we can say that not all projects benefit from additional choice tasks.

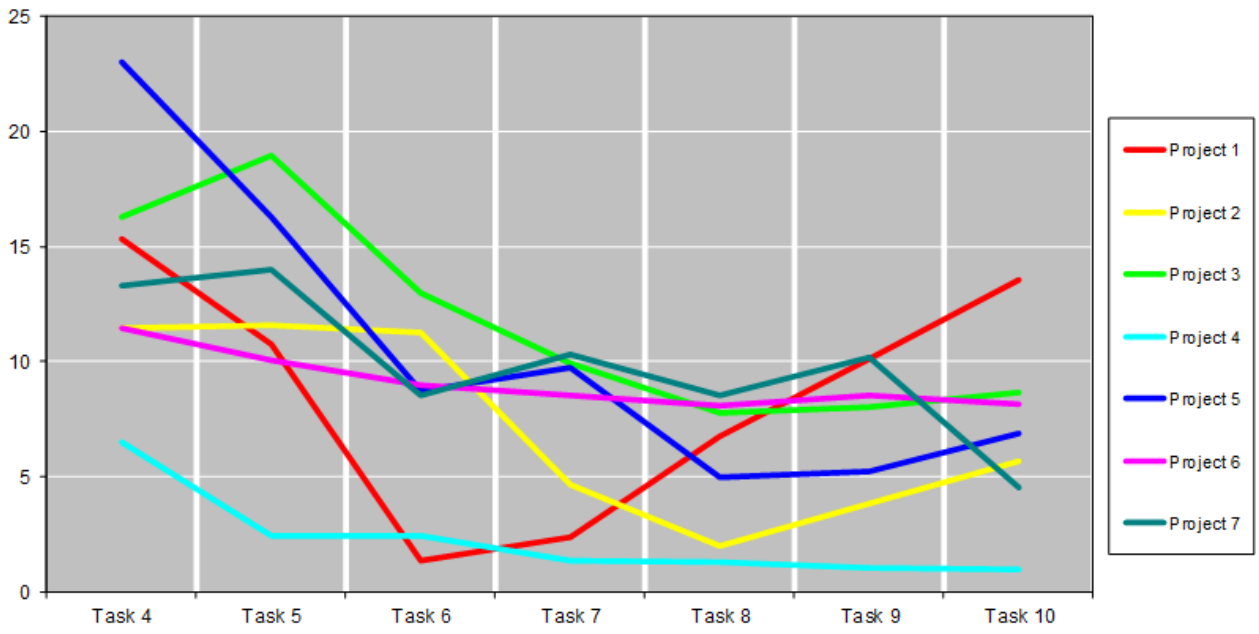


Figure 7: MAE on hold-out tasks by task

Also the position of the choice task after which a significant individual RLH decrease could be measured was widely distributed. However, it seemed that beginning with task 7 this effect can be measured more and more often:

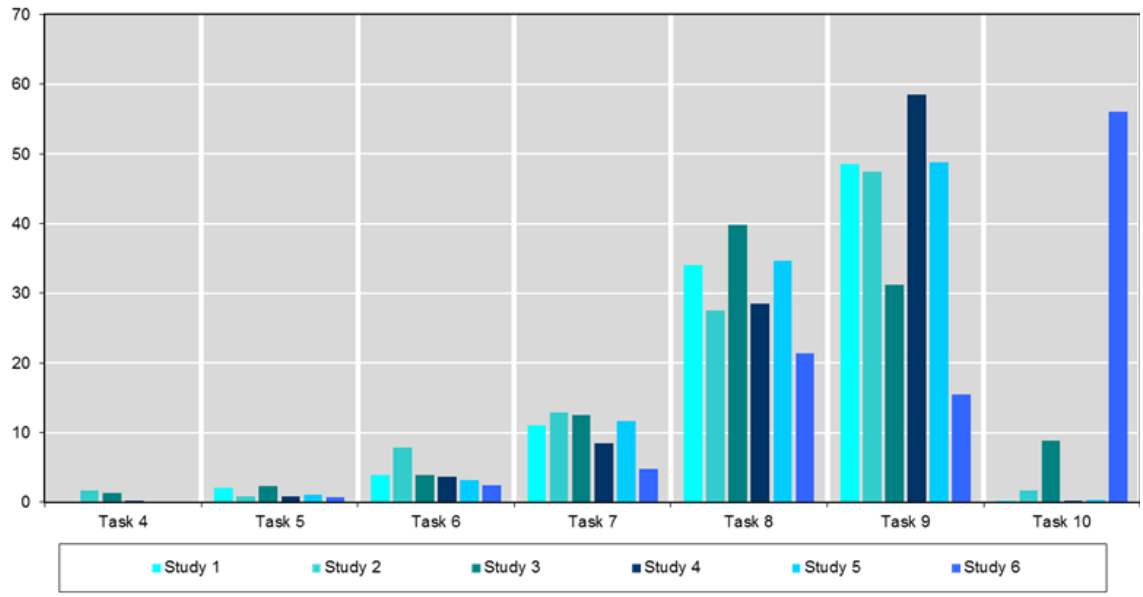


Figure 8: significant individual RLH decrease by choice task

Again using datasets with increasing number of choice tasks we calculated individual hit rates. These are also distributed all over the tasks without a clear indication of an overall choice task threshold.

Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Task 7	Task 8	Task 9	Task 10	Task 11	Task 12	Task 13
0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1	1
0	1	1	1	1	1	1	1	1	1	1	0	0
1	1	1	1	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	1	0	1	1	1	1	0	0	0	0
1	1	0	1	1	0	1	1	1	1	1	1	1
0	1	1	1	1	1	0	0	0	1	1	1	1
1	1	1	1	1	1	1	0	0	0	0	0	0
1	0	1	1	1	0	0	0	0	0	0	1	1
0	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1
0	0	1	1	1	0	0	1	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	1	1	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	1	0
1	1	0	0	0	0	0	0	0	0	0	0	0

Figure 9: Example data: individual hit rates by choice task

If we accumulate the respondents with at least one hit and don't take into account the later tasks of the respondent, we reach a high level of hits after only a few choice tasks:

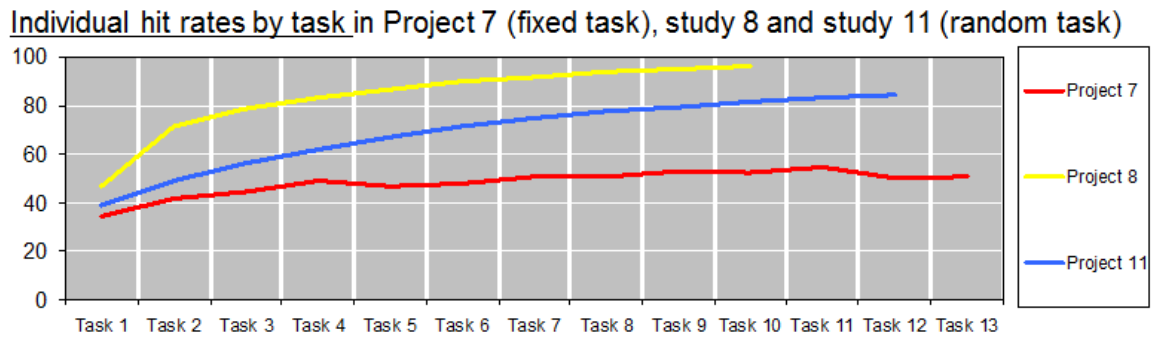


Figure 10: Accumulated individual hit rates by task (fixed holdout task used for study 7; for all other projects the first task which is not used for estimation (n+1) is used as holdout)

Looking at the three studies for the task in which an individual hit has first been reached we see after only a few choice tasks there is only very slow improvement.

Individual Choice Task Optimization

Based on our conclusions so far we optimized our data sets: We used only those choice tasks up to the RLH maximum on an individual level (meaning different individual interview lengths within the different studies). Applying this rule we could eliminate about 38% of all choice tasks. Were these choice tasks necessary at all?

Looking at total answer time and aggregated hit rates we found quite similar results for the ICT (individual choice task threshold) optimized data sets. In reality this would have led to a time saving potential of about 35.4%.

		Study1	Study2	Study3	Study4	Study5	Study6	Study7	Study8
All Choice Tasks	Choice Tasks	6.076	2.040	12.600	6.900	7.695	12.835	3.900	2.120
	Answer Time (hours)	47,3	9,6	66,5	26,8	27,8	71,3	17,5	4,3
	Hit Rate (percent)	**69,2**	**73,6**	**50,7**	**96,3**	**78,2**	**85,4**	**50,7**	**96,3**
ICT Optimization	Choice Tasks	3.957	1.499	7.823	3.879	4.957	7.953	1.957	1.579
	Answer Time (hours)	30,8	7,1	41,3	15,8	17,9	44,2	10,6	3,0
	Hit Rate (percent)	**68,0**	**68,4**	**48,3**	**85,6**	**64,2**	**78,1**	**48,3**	**85,6**
	Time Saving Potential	**35,9%**	**26,0%**	**37,9%**	**41,1%**	**35,6%**	**38,0%**	**39,4%**	**30,2%**

Figure 11: Effect of ICT optimization

The above results lead to the conclusion that Individual Choice Task Thresholds (ICT) exist. Unfortunately, since they are not correlated with the answer times, they can't simply be detected by time control. Using an individual number of choice tasks according to the individual choice task threshold does not end up with large differences in the hit rates. The small decrease we can see in all of our eight studies does not seem to change the interpretation of the results in any way

and therefore we are focusing more on the costs and interview length. An individual number of choice tasks shortens the interview significantly which would have several positive effects such as:

- lower field costs (time is money)
- less frustration of respondents who are either challenged or bored
- keeping the panels open for conjoint exercises

However, there is another interesting aspect: while the hit rates are quite similar between the full and ICT optimized data sets, we observed important differences in the results:

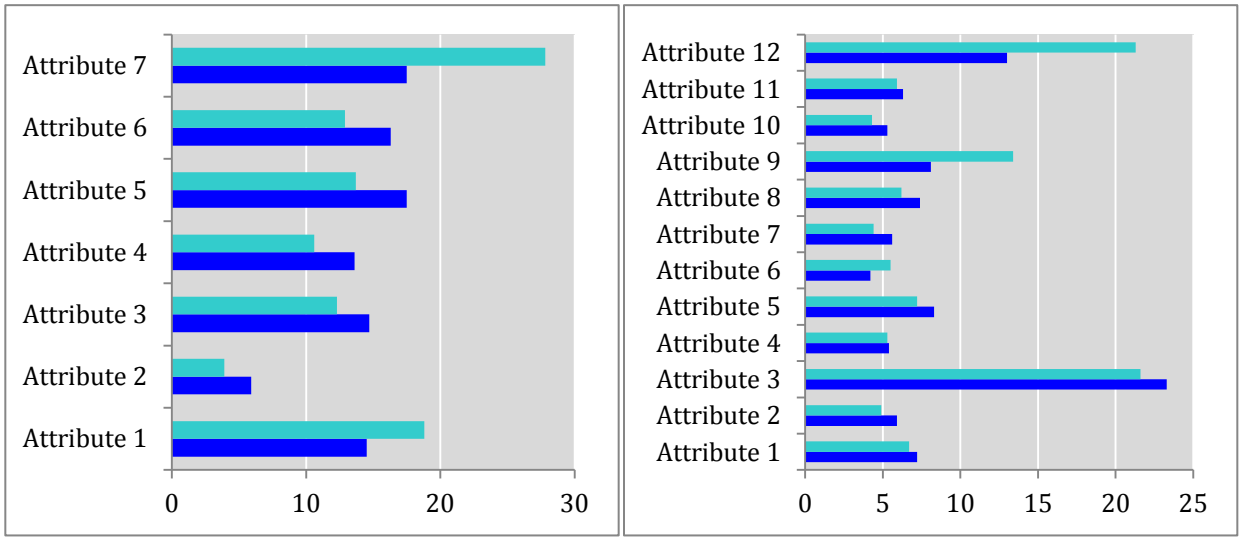


Figure 12: Differences in Average Importances between all tasks and ICT optimized datasets

It seems that some attributes become more important during those choice tasks which do not contribute to better individual results and were therefore excluded in the ICT optimization. This definitely leads to wrong interpretation of the impact of certain attributes.

Comparison of the Posterior Distribution

As shown in the previous section the individual hit rate sometimes increases, while share prediction results show a higher MAE. Simply looking at the aggregated results, it seems counter- intuitive that if one gets better hit rates with more choice tasks the MAE is getting worse.

A closer look into the data structure is necessary to see how this effect happens. The posterior distributions of the HB estimates and their standard deviations can provide more information. If the ICT really exists, a decreasing number of attributes with part worth utilities significantly different from zero must be found when increasing the number of choice tasks in the estimation. The reason for this effect is a simplification strategy, namely focusing on a lower

number of attributes the more choice tasks we ask. If it's possible to show this decrease in the number of part worth utilities significantly different from zero, we could conclude that respondents concentrate on only a small number of attributes when answering a larger number of choice tasks.

To show this effect we calculated[9] the number of attributes with significantly non-zero part worth utilities for each respondent, after analyzing with each additional choice task. Comparing the results of the twelve conjoint studies we found that the more choice tasks we use in the estimation, the more respondents take into account a smaller number of attributes when answering.

The more choice tasks we show, the more respondents consider only one or two attributes before making their choice. In contrast, at the beginning of the choice exercise, respondents show a more complex answering behavior and take into account a higher number of attributes when making their choice.

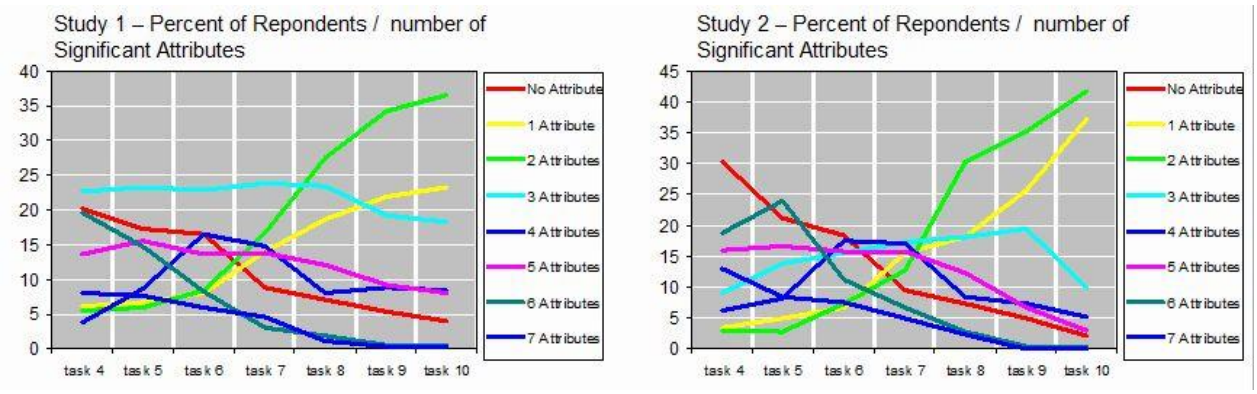


Figure 13: Number of attributes taken into account / % of respondents

The tendency toward taking fewer attributes into account before selecting a concept in the tasks could be observed in all twelve analyzed studies. The number of attributes in the choice experiment seems to have a much smaller influence on using the simplification strategy than does the number of choice tasks the respondent has to answer. By far the largest respondent group takes only two attributes, brand and price, into account in the later tasks. Only very small numbers of respondents focus on more than four attributes in later choices. Especially in studies with many different feature attributes, respondents tend to simplify very early. In Brand-Price only CBC we can show a very fast convergence into two different simplification strategies, one group taking only brand, and the other only price, into account before making their choice.

[9] We used the individual respondent beta draws from HB and report attributes that included levels where 95% or more draws have the same sign. This shows how many attributes the respondents takes into account when making their choices.

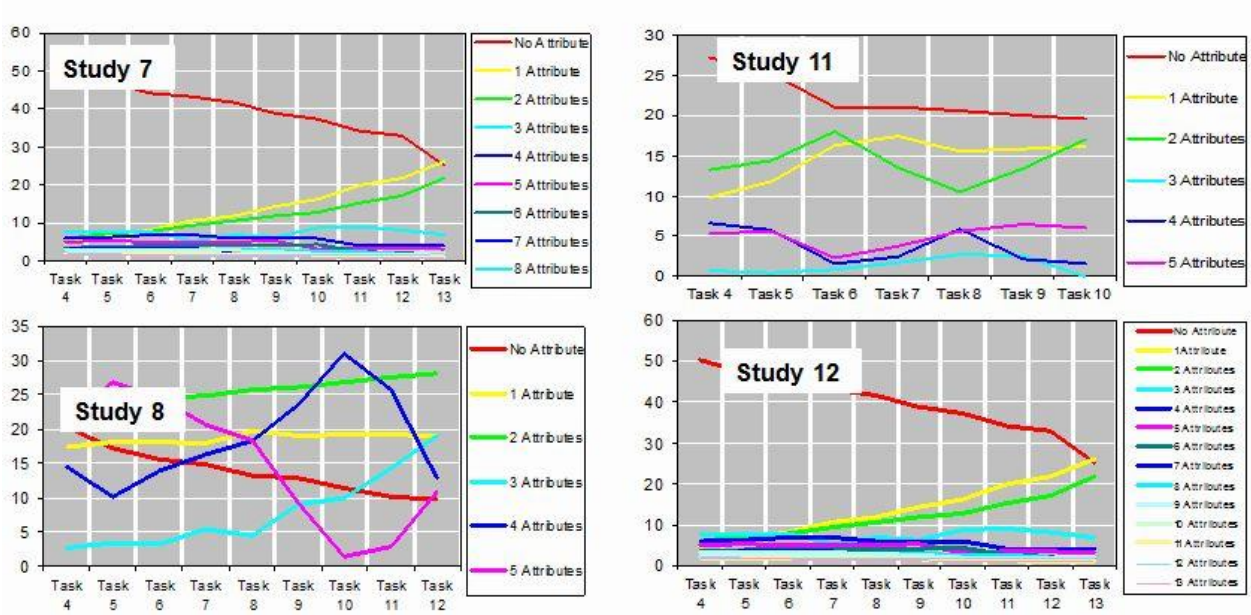


Figure 14: Number of attributes taken into account / in % of respondents

The decreasing number of attributes that are taken into account seems to be at least one of the reasons for increasing hit rates when asking more tasks, but increasing MAE in our share predictions.

A slight tendency towards higher MAE for high involvement products underlines the finding, because it's plausible, that for high involvement goods the simplification doesn't take place with the same intensity that it does for low involvement goods. In the case of low involvement, one can argue that the simplification makes a lot more sense because it reflects reality. People don't look at all the tiny differences between products if they are not really interested in this category.

The findings from our twelve studies show that—in at least ten of them—simplification confounds the real decision process and leads to incorrect share predictions. A simple example illustrates these findings. If respondents focus mainly on brand and price after a certain number of choice tasks, then their hit rate on a holdout or random task in a late position is improved because the choice behavior is much easier to predict. However, if in reality their real purchase decision is based on more attributes than brand and price, then the prediction of market shares based on their part worth utilities is inferior. That is the reason why better hit rates don't always result in better share prediction.

Is This a New Problem?

Maybe it is. Today we conduct most interviews in online panels and this environment has changed during the last decade dramatically.

We observe a concentration of panels. Many mergers and acquisitions take place in this field. Often, high quality panels are taken over by ones with lower quality, resulting in one lower quality panel.

More and more conjoint exercises are conducted around the world and therefore panel respondents are more often used to answering conjoint studies. In the past, most of the respondents were seeing a conjoint exercise for the first time when answering and didn't really know what the next screen would bring up. Now, panelists more and more know the flow, especially of the widely used classical CBC exercises, and know how they can speed up when answering without being identified as bad interviewees. This comes along with the fact that most of the panelists only respond for points, miles or other incentives. So from their motivation, it's best to finish the exercises as fast and efficiently as possible to earn incentives as efficiently as possible. In many of the feedback mails we received from panelists, after participating in a conjoint interview in the last year, we could read that the more engaging conjoint exercises are seen as a higher burden in earning points than simple scaled questions are. So panel providers are asked to offer higher incentives for these burdensome exercises in order to motivate panelists to take part in future conjoint studies. Many of the respondents say they will stop future interviews with conjoint exercises if the incentives are not higher compared to studies with conventional questions only.

Furthermore, we see in feedback mails that many of the panelists are less motivated and not willing to answer boring choice tasks any longer. The monotonous repetition of a large number of very similar-looking questions is often their main complaint. So from this feedback we could, conclude that we are forced to program more interesting choice exercises in future and make our studies more appealing to our panelists.

Nevertheless neither higher incentives nor fancy programming of the survey can protect us from the "speed up" behavior and the simplification strategies of the professionals in the panel, because more points can be earned in an efficient way by again speeding up the exercise through simplification.

To Put it in a Nutshell

The analysis conducted for this paper clearly shows that Individual Choice Task Thresholds (ICT) really exist. In our twelve analyzed studies we could see that respondents have a very diverse answering behavior and individually different choice task thresholds.

Unfortunately there is no simple rule for detecting the ICT during the online interview. Neither the absolute time a respondent needs to answer a question, nor the change in time needed for a single answer, provides useful information about the answering quality and behavior of a respondent. A speed-up on the one hand could be a sign of a simplification or, on the other hand, a learning process enabling the respondent to make decisions much faster. For this reason, analyzing the time stamps from questions alone doesn't really help. Only a more complex analysis taking more advanced measures into account could provide us with the information we need to decide if a respondent has reached her ICT.

The key findings of our work are:

Less is more: For a large number of respondents we gain better or equal hit-rates and share predictions when we use only a smaller number of choice tasks in order to avoid simplification.

More is dangerous: The analysis of the individual posterior distributions shows that a large number of respondents tend to simplify their answers in later choice tasks. In earlier choice tasks, one observes higher number of attributes with significantly non-zero utility values than in later ones.

More is expensive: Optimizing the number of choice tasks could reduce the total number of tasks without a significant quality decrease, thus saving about a third of interview time (for the conjoint part) and avoiding interviewee annoyance.

CONCLUSION

A better monitoring of individual answering behavior during the interview in order to avoid exceeding the ICT will be essential in the future.

A combination of monitoring the answering time and a maximum likelihood estimation predicting the next choice of a concept, based on respondents' previous answers could be an idea for future research. Another new approach could be a surveying system with individual number of choice tasks i.e. derived by measuring individual RLH against the aggregated RLH (Indexed LH) after each choice task or by comparing the posterior distributions from one task to the next. The system should stop the interview if there is no further improvement in RLH or simplification strategy or consistently lower measures for interview quality are detected. As computational power is increasing from year to year it seems possible to realize such online computations during online interviews in the near future (Kurz, Sikorski; 2011).

Another interesting approach that could be developed would be the combination of higher incentives for panelists and a more appealing conjoint programming in combination with an adaptive design algorithm. In the work of Toubia, Hauser and Garcia (2007) and Yu, Goos and Vandebroek (2011), we find new ideas for adaptive design algorithms with the aim to produce choice tasks with higher utility balance for each individual respondent. Perhaps combining higher utility balance with criteria that stop the interview when reaching a certain ICT would provide a solution. Calculating the maximum likelihood after each choice task and forecasting the answer to the next choice task would allow stopping the interview after this forecast produces a hit two or three times. One could then assume that enough information from this respondent is collected. This technique could avoid respondents simply clicking through conjoint questionnaires in order to earn the points. To be sure, this exercise is more burdensome, because choice tasks with higher utility balance are much harder to answer for respondents. But such adaptive interviews can be rewarded with using fewer choice tasks when answer quality is high and respondents can earn the same amount of incentives with an individually different length of the interview. Such a combination of adaptive algorithms and higher incentives could be most beneficial, and avoid only paying the higher incentives without getting improved quality for the conjoint interviews. In cases when no alternative-specific designs are needed[10] a solution can also be ACBC, because ACBC's more appealing interview leads to more respondent

[10] At the time of writing, Sawtooth Software's current version of ACBC software cannot do alternative-specific designs. Sawtooth Software says that soon this will be available.

engagement. ACBC uses different stages with differently formatted choice questions, and in the minds of most respondents is not as boring as CBC tasks. Maybe interspersing non-conjoint questions and transitional text-only pages throughout the CBC Tasks could help a little to reduce the monotony and avoid burning out respondents during the CBC exercise.

REFERENCES

Johnson, R., Orme, B., (1996) "How Many Questions Should You Ask in Choice-Based Conjoint Studies?" in Sawtooth Software, Inc. Technical Papers.

Kurz, P., Binner, S., (2010) "Added Value through Covariates in HB," in Proceedings of the 15th Sawtooth Software Conference, pp.269-282.

Kurz, P., Sikorski, A.,(2011) "Application Specific Communication Stack for Computationally Intensive Market Research Internet Information System" in: Proceedings of the 14th International Conference on Business Information Systems, pp. 207-217.

Toubia, O., Hauser, J. R., Garcia, R., (2007) "Probabilistic polyhedral methods for adaptive choice based conjoint analysis: Theory and application" in Marketing Science, 26(5), pp. 596–610.

Yu, J., Goos, P., Vandebroek, M., (2011) "Individually adapted sequential Bayesian conjoint-choice designs in the presence of consumer heterogeneity" in International Journal of Research in Marketing, vol. 28, pp. 378-388.