How many attributes should be used in a paired comparison task?

An empirical examination using a new validation approach

Heinz Holling[+]

Torsten Melles

Wolfram Reiners

February 1998

[+] Heinz Holling is a Professor for Organizational Psychology at the Westfälische Wilhelms-Universität Münster, Germany. E-mail: holling@psy.uni-muenster.de

Abstract

Conjoint analysis is one of the mostly used methods in the analysis of preferences and the prediction of choices. Today, most applications of conjoint analysis use paired comparisons. Paired comparisons on a graded scale are a substantive element of the adaptive conjoint analysis (ACA), a computer supported procedure that is predominantly used in marketing research. Contrasting this heavy application, little is known about how many attributes to present in a pair and how many judgments to be made.

We present an empirical study testing the validity of paired comparison tasks using two, three, and four attributes. This examination has been made by using a new approach of validation. Our results show that paired comparisons with two attributes lead to more accurate estimations than with three or four attributes. Furthermore, pairs with two attributes demand less time and are subjectively considered more beneficial in some sense.

Background

Talking about conjoint analysis may be somewhat confusing because this term names a lot of heterogenous approaches and procedures. They all have the following basic idea in common: There are real preferences for options as well as their attributes and levels. This prefences can be measured by utilities. Utilities that describe the value of an attribute level are called part-worths. Assuming a specific computation rule of judgments, these part-worths can be derived from multiattributive judgments.

The differences between the methods that share this idea arise from the way the judgments are collected and the procedure that is used to estimate the part-worths. In most cases the estimation is done by an ordinary least square (OLS) regression. The judgment task can require ratings of single concepts, a rank order of a set of concepts or paired comparisons. The concepts can be full profiles (consisting of all possible attributes) or partial profiles that consist of a subset of attributes. The presentation of these profiles and the collection of judgments can be done by computeraided personal interviews, computeraided non-personal interviews, paper-and-pencil or by telephone.

This paper examines the question of how many attributes and how many pairs should be used in computeraided personal interviews using graded paired comparisons. This is an important question taking the high number of applications into account that make use of this method.

The Paired Comparison Task in Adaptive Conjoint Analysis

As we already mentioned, using paired comparisons requires the researcher to decide how many pairs to be shown as well as how many attributes to be displayed in the pairs. Which effects have to be expected from these two decisions?

The *number of pairs* in a paired comparison task is comparable to the number of judgments which must be done by the subject. According to statistics, the number of pairs should be as high as possible to obtain the maximum amount of information. On the other hand, with an increasing amount of judgments an increasing amount of time has to be spent on the interview. Therefore, the motivation of the respondents to provide further information is expected to decrease. Moreover, a high number of judgments can lead to a task overload. This term names the fact that people are not able and willing to respond to a high number of tasks. Like in the case of information overload the accuracy of responses decreases at some point of the task.

The *number of attributes* displayed in the pairs defines its dimensional complexity (Huber & Hansen, 1986). In ACA, the most popular computerized conjoint procedure that makes use of paired comparisons, it is technically fixed to a range from 2 to 5. According to statistics, one achieves a higher amount of information through a higher dimensional complexity. A high dimensional complexity, however, makes the task more difficult for the subject. This can lead to an information overload and to an increase in response errors as well as to a decrease in motivation. Psychological findings indicate the possibility of an information overload through a high dimensional complexity.

In sum, with an increasing number of pairs and growing dimensional complexity, demand on the subject increases. This increase may be dependent on the subject's motivation and cognitive abilities. However, most studies try to draw *general* guidelines for constructing the paired comparison task from empirical results.

Results of Former Studies

Monte Carlo simulations that assume a low or no response error results confirm statistical expectations: The larger the number of pairs and/or the higher the dimensional complexity, the higher the accuracy of part-worth estimations. On the other hand empirical studies show that statistics can not sufficiently explain the quality of part-worth estimations that are derived from real judgments.

Huber and Hansen (1986) asked students about their preferences regarding apartments. They varied the number of attributes displayed in the pairs using 2, 3 and 4 attributes. The accuracy of their estimations was measured by the goodness of fit of the regression model and predictive validity using a holdout task. A holdout task takes a rating or ranking of a few full profiles or a choice between them. These judgments are usually not used for part-worth estimations. In the case of rankings the ranks are compared to the predicted ranks of the profiles. Huber and Hansen's results showed no differences between the three tasks regarding predictive validity. Nevertheless, using 2 attributes led to a higher goodness of fit. For 16 pairs the task was more time consuming with a dimensional complexity of 3 and 4 attributes (12.91 or 13.5 minutes) than of 2 attributes (11.36 minutes). Furthermore, the subjects judged the paired comparison task with 2 attributes to be more enjoyable, stimulating and relaxing. Huber and Hansen therefore recommend to use profiles with a dimensional complexity of 2 or 3.

Tests for the optimal number of pairs do not show any conclusive results. Finkbeiner and Platz (1986, quoted from Agarwal, 1989) noticed that ACA using 16 pairs showed a higher predictive validity (choice validation) than in the case of 8 pairs. The aggregate estimates of the part-worths were nearly indifferent.

Agarwal and Green (1989, quoted from Green, Krieger & Agarwal 1991) also tested apartment preferences of students using ACA. They defined six attributes characterizing the apartments. Each subject judged 15 pairs with a dimensional complexity of 2. The authors remarked only a slightly higher predictive validity regarding the holdouts by increasing the number of pairs. This finding matches the results from Green, Schaffer and Patterson (1991). These authors analyzed the preferences of students regarding cars (8 attributes, each with 4 levels). They asked their subjects to compare up to 20 pairs. The results show that additional pairs only lead to a slight improvement of predictive validity.

Only few studies analyzed the *joined* effects of the number of pairs and dimensional complexity. Agarwal (1988a, quoted from Agarwal, 1988b) tested the validity of parameter estimations by varying dimensional complexity (2 vs. 3 attributes) and the number of pairs (max. 15 or max. 30) on an individual level. He used a houldout task to measure predictive validity. Agarwal concluded that the number of pairs had a weak yet significant positive effect on predictive validity if the profiles were composed of two attributes. This effect did not occur using 3 attributes. Moreover, the difference between partial profiles with 2 and 3 attributes was not significant. Nevertheless, the tasks with 2 attributes were judged as simpler and needed less time serving the same level of predictive validity.

Agarwal (1989) examined a possible interaction between dimensional complexity and the number of pairs in an additional study. The number of attributes was 2 versus 4; the number of pairs was 0, 9, 18 or 36. With 0 pairs, the part-worths were solely derived from ACA's self-explicated judgments. Again, predictive validity was measured by a holdout task (ratings of 4 full profiles). Agarwal could not prove a difference between profiles with 2 and 3 attributes on an individual level. On the other hand, on an aggregate level (estimation of market share), 2 attributes produced the better estimations. The paired comparison task provided no increase in accuracy on an individual level. This means that the holdout ratings could be best predicted through ACA without any pairs. On an aggregate level, 18 paired comparisons were found to be optimal. With a dimensional complexity of 2 attributes, the mean error of prediction decreased up to 18 pairs and after 36 pairs it was higher than with 0 pairs. Using 4 attributes, the average prediction error was the lowest with 18 pairs. A higher number of paired comparisons provided no difference regarding validity.

Nearly all of the former studies tested the validity of estimations on an individual level using a holdout task. Only Huber and Hansen (1986) considered predictive validity and the goodness of fit of their regression model. Due to this fact it is important to view the main characteristics of a holdout task. As we mentioned earlier a holdout task is composed of a limited number of full profiles. Finkbeiner and Platz (1986) asked their subjects to choose between those profiles, Agarwal (1989) let his subjects rate the profiles. More often subjects are asked to rank the holdouts. Performing a holdout task is a conflicting job for the researcher. On the one hand, a high number of profiles complicates the task for the respondent and leads to an increasing amount of mistakes. Due to this a low number of profiles would be preferable. On the other hand, if the holdout sample is composed of only a few profiles, it is not possible to clearly draw back from a measured preference structure to the holdout rank order. There might be more than one estimation of the parameters that is matching the rank order. In consequence a high coefficient of predictive validity may appear which can be traced back to the ambiguity of the criterion. Moreover, the holdout judgments are themselves not fully reliable. Taking this into account a higher number of holdout profiles would be preferable. Up to now there is no conclusive empirical evidence on how to perform a holdout task.

An additional problem is the obvious similarity between the classic full profile analysis task type and the holdout task type. The higher the number of attributes that are included into the profiles of the conjoint task (e.g. one paired comparison) the more similar it is to the holdout task. For instance, measuring preferences regarding apartments defined by five attributes using all of these attributes in each pair means that these profiles are equivalent to full profiles just like the holdouts. Due to this, the validity of paired comparisons with higher dimensional complexity is expected to be overestimated compared to paired comparisons with a lower dimensional complexity if we use a holdout task.

A further problem regarding the holdout task is that it can produce different levels of difficulty for different people. The difficulty is indicated by difference of ones utilities between the profiles. For some subjects the difference may be relatively large and, therefore, the task is easy. For others, the same task can be more difficult taking the smaller difference of the same profiles into account. The more difficult the holdout task, the higher the expected amount of response errors. Due to this the part-worth estimations are less valid and reliable. This effect takes place on an individual level. Therefore, the accuracy of the estimations can only be compared between respondents who judge holdouts with a similar level of difficulty.

Problems with testing the quality of conjoint analysis using a holdout tasks have rarely been discussed. Only Huber, Wittink, Fiedler and Miller (1993) as well as Orme, Alpert and Christensen (1997) deal with this issue and offer guidelines for the construction of holdout tasks. Moreover, they mention that tests that indicated a low level of reliability can be caused by unreliable responses in the holdout task.

Empirical Study

The goal of this study is to analyze the effects of dimensional complexity and the number of pairs on the quality of the parameter estimations and the subjective assessment of the task. We defined the following hypothesis to be tested:

> Hypothesis 1: The validity of part-worth estimations depends on the dimensional complexity of the profiles used in a conjoint task.
>
> Hypothesis 2: The validity of part-worth estimations depends on the number of pairs that are judged in a paired comparison task.
>
> Hypothesis 3: The validity of part-worth estimations depends on the joined effects of dimensional complexity and number of pairs.
>
> Hypothesis 4: There are differences in subjective assessments in tasks displaying different numbers of attributes.

Method

Three levels regarding the number of attributes were defined to examine the effects of dimensional complexity. The number of attributes was 2, 3 and 4. The number of pairs extended from 1 to 40. For inferential statistical tests four levels were defined (10, 20, 30 and 40 pairs).

Effects on validity of these experimental variations were examined by different methods. A desirable measure of validity would be one which reproduces individual preferences exactly as possible so that the estimated part-worths can be directly compared to the true utilities. Since such a measure does not exist, a new method to evaluate different procedures has been chosen. The true utilities were defined by the researcher and have been introduced to the subjects within the framework an extended learning task. After finishing the learning phase, the estimation of the part-worths via the conjoint tasks followed. These estimations have been tested through a comparison with the true utilities. The correspondence of the true and the estimated part-worths from the paired comparison tasks has been measured by the root of their mean squared difference. This measure indicates the root mean square error (RMSE) (fomula 1).

formula 1 $$RMSE = \sqrt{\frac{\sum_{j=1}^{J}\sum_{i=1}^{I}(u_{ij} - w_{ij})^2}{h}}$$

with

u_{ij} estimated part-worth of level i from the attribute j

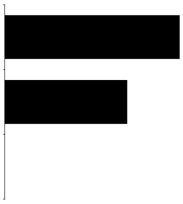
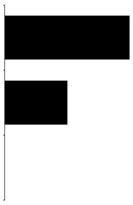
w_{ij} true part-worth of level i from the attribute j

h number of all levels used in the study

Moreover, a traditional full profile conjoint analysis was used as a second method to test for validity. Each respondent provided a ranking of 15 cards that were drawn from a fractional factorial design at two times of the interview. 10 of the 15 stimuli of both holdout samples differed between the sets and five were identical. These identical holdout cards provide a possibility to measure the retest-reliability of the rankings. Part-worths were estimated for each of the rankings. The comparisons of the estimations from the full-profile analyses with those that were derived from the paired comparison tasks represent a cross-validity test. Seen as holdouts the rank order of the profiles serves as a criterion for testing the predictive validity in a traditional manner. It is also possible to compare the full profile estimations and the paired comparison estimations against the true utilities. This represents a direct measure of the validity for two popular conjoint methods.

Student apartments were chosen as preferential objects due to the fact that choices between apartments is familiar and sometimes meaningful to students. Due to this, the learning task should be not too difficult and promote their motivation to participate. Many studies that deal with conjoint analysis ask respondents to indicate their preferences regarding apartments (e.g. Agarwal and Green, 1989; Corstjens and Gautschi, 1983; Green, Krieger and Schaffer, 1993). Five attributes with three levels each were defined. The attributes and their levels as well as the predefined true utilities are shown in Table 1.

Table 1: Attributes, levels and utilities introduced in the learning task

Attributes	Levels	Utilities
1. Distance to Institute	1 km	100
	3 km	50
	5 km	0
2. Size of the Apartment	35 m²	90
	25 m²	45
	15 m²	0
3. Type of House	Student Dorm	80
	One Family House	8
	Apartment House	0
4. Condition	new	70 
	renovated	49
	old	0
5. Rent	15 DM/m²	50 
	20 DM/m²	25
	25 DM/m²	0

Experimental Procedure

Subjects were tested during one-day group sessions. The experiment lasted about 4.5 hours. The learning task required the subjects to put themselves into the place of a fictitious person. The apartment preferences of this person covered the predefined utilities. In the experimental tasks, the subjects should have acted and indicated preferences in the way the fictitious person would have done. The subjects did not retain the numerical part-worths. Instead, graphical representations and verbal descriptions were used to introduce utilities. Furthermore, comparisons between partial and full profiles were carried out. Subjects evaluated these profiles in groups and received feedback in a discussion of the right evaluation. The learning task was composed of four parts. After each part a learning test was carried out. The whole learning task took about three hours.

The learning phase was followed by the experimental conjoint tasks. The first and the last task were the full profile tasks. Between the two full profile tasks, each participant worked on three paired comparison tasks with a dimensional complexity of 2, 3 and 4 each consisting of 40 pairs. The sequence of the three tasks was completely balanced between subjects. The paired comparisons were carried out with the help of the software 'ALASCA' (Holling, Jütting & Großmann, 1998) using an adaptive design. The subjects judged the pairs on a graded scale. Part-worths were derived using an OLS-regression.

After each of the three paired comparison task the subjects were asked to judge the task on different dimensions. They rated their interest in the task, the general difficulty of the task as well as the difficulty of concentrating and finding an answer.

Results

24 students with different majors from the ages of 19 to 30 (M=21.6) participated in the study. They were randomly assigned to one of six groups. Each group consisted of 5 to 7 subjects. This should have been a manageable size to conduct the learning tasks and to give each subject a feedback. An analysis of the learning results showed that the utilities had been sufficiently 'internalized'. We do not offer a detailed presentation of these results in this paper. This is done by Melles (1996).

Root Mean Square Error

After 40 pairs there was no difference between part-worths derived from the tasks with different numbers of attributes. This result holds for both validity criteria. The deviation from the true part-worths was very low in all three cases, ranging between 6 and 8%. This indicates a high level of accuracy.

Difference between the true and the estimated parameters measured by RMSE occur over the course of the tasks. A two-factor (4 x 3)-ANOVA with the within-subjects factors *number of pairs* and *dimensional complexity* showed significant effects on both factors confirming hypothesis 1 and 2 (number of pairs: $F_{3,69}$=214.03, p<.01; dimensional complexity: $F_{2,46}$=3.93, p<.05). The effect of the first factor offered significant differences between each of the levels 10, 20, 30 and 40 pairs[1] ($F_{1,23}$>15 p<.009). The best estimation is achieved only after a total number of 40 paired comparisons. The higher the number of paired comparisons the lower the root mean square error.

[1] The RMSE of each level was determined by averaging RSME of five paired comparisons in order to avert variations due to chance.

The effect of the second factor *dimensional complexity* is due to the differences between 2 versus 3 and 2 versus 4 attributes. The estimations using 3 and 4 attributes do not differ significantly from each other regarding RMSE ($F_{1,23}$=0.01, p>.9). Paired comparisons with only two attributes provided the lowest root mean square error.

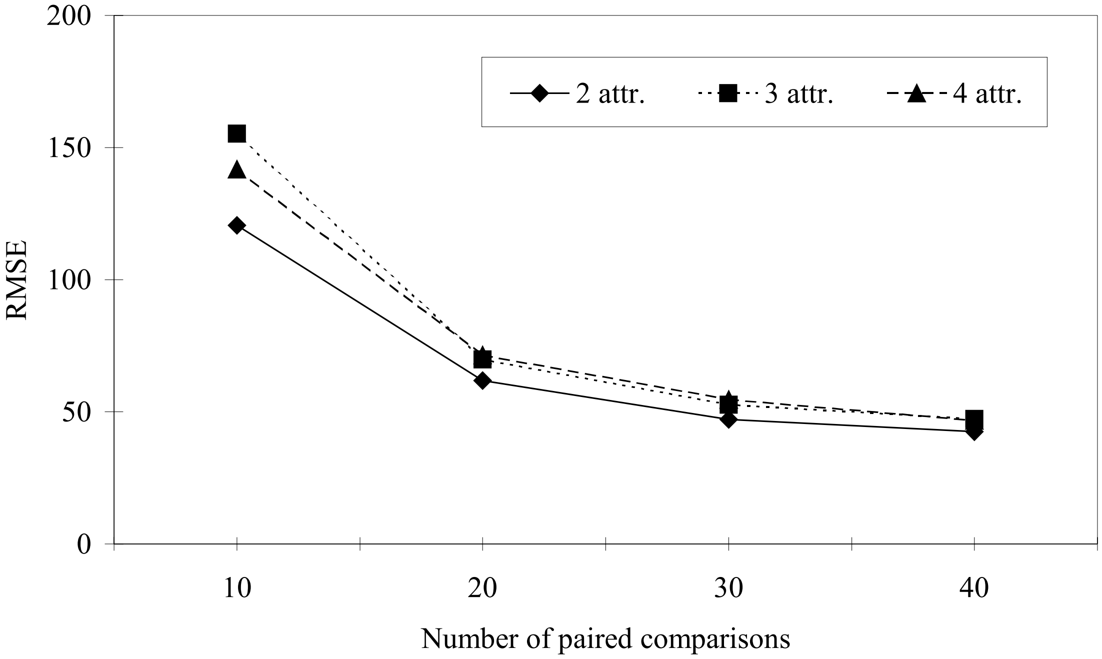


Figure 1: Mean error after 10, 20, 30 and 40 paired comparisons.

We also observed an interaction effect between the number of pairs and dimensional complexity ($F_{6,138}$=4.03, p<.01). This means that the RMSE over the course of paired comparisons is not independent of the dimensional complexity confirming hypothesis 3. This interaction effect is attributable to the levels 10 and 20 pairs. It can be explained by the fact that the RMSE with profiles of 2 attributes after 10 pairs is already greatly reduced. The improvement through the following 10 pairs is smaller than in the paired comparison tasks with profiles of 3 and 4 attributes where the error after 10 pairs is much worse.

Predictive Validity

The estimated part-worths of the paired comparisons serve to predict the rank orders of the cards in the two holdout tasks. In this case we consider the holdout judgments to be reliable. The mean correlation (averaging Fisher-Z-transformed correlations) between the identical cards in the two holdout tasks was .95 (ranging from .6 to 1.0) indicating a high level of reliability.

The predicted and the estimated ranks were compared via correlation (Kendalls Tau b). This correlation coefficient is a measure of predicted validity for the paired comparison tasks. Table 2 shows the mean correlation after 10, 15, 20, 25, 30, 35 and 40 pairs in the paired comparison task with holdout task 1 serving as criterion. Table 3 shows the correlation with holdout task 2 serving as criterion.

In both cases course of the correlations is similar. Up to a number of around 10 to 15 pairs, the accuracy of the estimations increases dramatically. Afterwards, the additional benefit decreases. However, the most exact estimations are obtained after 40 pairs. Considering the accuracy of part-worth estimations, the three tasks differ only slightly from each other. A two-factorial ANOVA with the within-subjects-factor *number of pairs* and the between-subjects-factor *dimensional complexity* showed no significant effect on the factor *dimensional complexity* with holdout task 1 ($F_{2,68}$=0.248, p>.7) or holdout task 2 ($F_{2,68}$=0.587, p>.5) as criterion. This means that paired comparisons of different dimensional complexity do not differ from each other according to predictive validity. This finding does not support hypothesis 2.

Table 2: Rank correlation of the predicted and empirical ranks of the holdout profiles in holdout task 1

	Number of pairs						
	10	**15**	**20**	**25**	**30**	**35**	**40**
2 attributes	.51	.63	.66	.68	.70	.71	.72
3 attributes	.47	.60	.66	.68	.69	.72	.72
4 attributes	.55	.60	.63	.66	.68	.69	.71

Contrasting the similarity between the paired comparison tasks regarding predictive validity there was an evident difference between the two holdout tasks. The correlations between the predicted ranks and the empirical ranks was higher in the case of the first holdout task. This may be due to a position effect in the experimental design as well as to a different level of difficulty of the two tasks.

Table 3: Rank correlation of the predicted and empirical ranks of the holdout profiles in holdout task 2

	Number of pairs						
	10	**15**	**20**	**25**	**30**	**35**	**40**
2 attributes	.48	.55	.58	.60	.60	.62	.62
3 attributes	.47	.54	.59	.59	.59	.61	.61
4 attributes	.52	.57	.58	.61	.61	.60	.61

Time Spent on the Task

Differences between the experimental tasks arise if the time the respondents spent on the task is taken into consideration. The higher the dimensional complexity the more time is required to form a response. If 2 attributes were displayed the task using 40 paired comparisons lasted 7.6 minutes on an average. The task with 3 attributes needed 10.4 minutes and with 4 attributes 11.9 minutes. Moreover, the tasks using 15 full profiles took much more time than the paired comparison tasks. Ranking these cards needed 22.3 minutes on an average.

In each of the paired comparison tasks the time spent on a pair decreased over the course of the task. With 2 attributes a decrease in time was evident, especially from the first third of the task to the second third. With profiles of 3 attributes, the reduction in time occurs going from the second to the last third. These results support the hypothesis that task effects occur in the paired comparison tasks due to different dimensional complexity. This effect may be due to a training effect that is later noticeable on account of the higher task complexity. Another explanation may be a changing state of motivation that forces to spend a lower amount of time on the task. In this sense, the faster decrease with 2 attributes can be an indicator of an earlier effect of boredom. However, the further increase of part-worth accuracy and the judgments of the respondents regarding the experienced effects on themselves contradict this explanation. The analysis of subjective ratings shows that profiles with 2, 3 and 4 attributes were judged as equally interesting. Moreover, paired comparisons with 2 attributes were judged as more positive on the other dimensions. Ratings and comparative assessments unanimously show that paired comparisons with 2 attributes were generally judged easier. Also, it is easier to concentrate on the subject here. These results confirm to hypothesis 4.

If we consider the accuracy of part-worth estimations dependent on the time spent on the three tasks the differences between them become more evident (figure 2). A two-factorial (3 x 3)-ANOVA with the within-subjects factors *time spent on the task* (with the levels 2, 4 and 6 minutes) and *dimensional complexity* (with 2, 3 and 4 attributes) offered significant effects on both factors (time spent on the task: $F_{2,46}$=233.78, p<.01; dimensional complexity: $F_{2,46}$=15.68, p<.01). Detailed analysis showed equally significant differences between the levels of the factor *time spent on the task* ($F_{1,23}$>4.3, p<.05). Therefore, the more time spent on the task the better the part-worth estimations.

All comparisons of the three levels of *dimensional complexity* embody significant differences ($F_{1,23}$>6.8, p<.05). Furthermore, an interaction effect exists between the factors *time spent on the task* and *dimensional complexity* ($F_{4,92}$=10.22, p<.01). With 2 attributes, a relatively low RMSE is achieved after two minutes and a further improvement is smaller than with paired comparisons of 3 and 4 attributes.

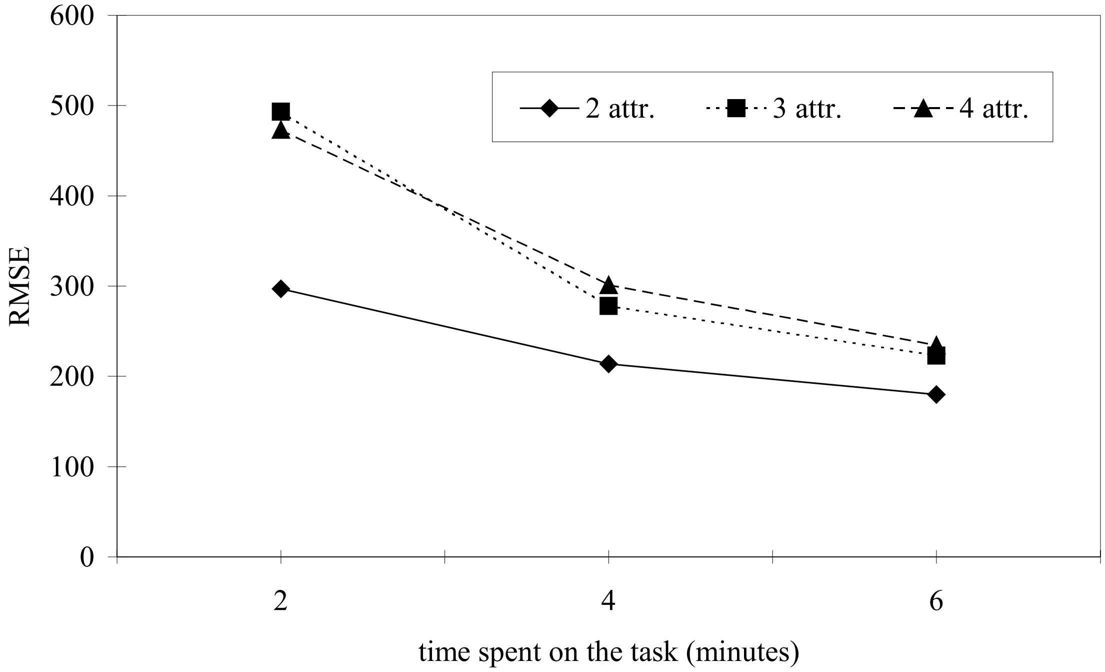


Figure 2: Accuracy of part-worth estimations depending on the time spent on the three tasks.

Discussion

The results show that the most effective parameter estimations are achieved using pairs with 2 attributes. The quality of the estimated part-worths with a dimensional complexity of 2 is the highest taking RMSE into account. Moreover, paired comparisons with 2 attributes need the lowest amount of time. However, the advantage 2 attributes was not supported by the holdout criterion. Regarding RMSE, the advantage is most evident with a small number of paired comparisons. The mathematical disadvantage of fewer data is more than compensated through effective information processing. The pairs with 2 attributes need less cognitive resources and are judged as the easiest tasks as well as the most suitable tasks to achieve a high level of concentration. Further experiments are needed to analyze the role of motivational and cognitive factors while working on the task in detail. Information boards, eye tracing techniques, the thinking aloud technique (verbal protocol) and physiological measurements could give further insight into judgmental processes and decision making in experimental tasks. Insights gained from these techniques may be a further contribution for an ideal formation of the paired comparison task.

Comparing our results with former studies that deal with the same issue is somewhat difficult. We did not analyze the *additional* increase of accuracy through the paired comparison task following a self-explicated phase. Instead of, we analyzed the part-worth estimations exclusively derived from paired comparisons. If the part-worth estimation through the paired comparison task is built on estimations from the self-explicated phase the improvement through the pairs depends on the quality of the self-explicated estimations. In the submitted analysis the self-explicated phase and these estimations were not carried out. Instead, all parameters started with the same initial start value, corresponding to a worse initial parameter estimation. The better the initial estimation the smaller the improvement due to the paired comparison task. With very good initial estimations it is possible that no improvement through the pairs can be achieved. Especially with a dimensional complexity of 3 and 4 attributes, one can be sure that an improvement in accuracy of part-worths won't be evident with a low number of pairs. Moreover, it will not reach the same quality as estimations with a complexity of 2 attributes. This assumption is supported by Agarwal's findings. Agarwal (1988a) carried out an ACA including the self-explicated estimations. Using 2 attributes predictive validity was higher if the number of pairs was high. This was not the case with a dimensional complexity of 3. The pairs with 2 attributes led obviously to an improvement in part-worth estimations and the pairs with 3 attributes, in comparison, did not. In a study from the year 1989 Agarwal found that profiles of 4 could offer an increase of information through 18 pair comparisons on an aggregated level, but not with 9 or 36 pairs. However, paired comparisons with a dimensional complexity of 2 produced an improvement already after 9 pairs. These results are thoroughly consistent with our findings.

In the submitted analysis some problems that arise from holdout tasks were averted by using true utilities. Due to identical part-worths, the holdout tasks offered for all subjects the same level of difficulty. Moreover, the new method offers a second measure of validity that may avoid several difficulties that the holdout task cannot. Its main advantage lies in the fact that it is a *direct* measurement of validity. No criterion task is needed. The main disadvantage of this method arises from the unproven external validity of learned preferences. A second disadvantage lies beyond the efforts that must be spent on the learning task. Regardless of the appropriateness of this method it seems necessary to use more than one method to test for validity. This demand is supported by findings of Acito and Jain (1980). These authors tested the agreement between different evaluation procedures. They compared Kruskals stress value (Kruskal, 1965) that measures the goodness of fit with violations of *a priori* sign expectations, as well as with predictive validity regarding a holdout task. Acito and Jain discovered that the results of the different experiments correlated, but that the correlation was not very strong. The decision for one or more criteria should be made dependent of specific applications. Nevertheless, we can conclude that with the introduction of true utilities using a learning task, a new, promising alternative to standard methods lies ahead.

References

Acito, F. & Jain, A.K. (1980). Evaluation of conjoint analysis results: A comparison of methods. *Journal of Marketing Research, 17*, 106-112.

Agarwal, K.M. & Green, P.E. (1991). Adaptive conjoint analysis versus self-explicated models: Some empirical results. *International Journal of Research in Marketing, 8*, 141-146.

Agarwal, M.K. (1988a). *An empirical comparison of traditional conjoint and adaptive conjoint analysis* (Working Paper No. 88140). New York: School of Management, State University of New York at Binghamton.

Agarwal, M.K. (1988b). Comparison of conjoint methods. In R.M. Johnson (Ed.), *Proceedings of the Sawtooth Software Conference of Perceptual Mapping, Conjoint Analysis, and Computer Interviewing* (No. 2, pp. 51-57). Ketchum, ID: Sawtooth Software.

Agarwal, M.K. (1989). How many pairs should we use in adaptive conjoint analysis? An empirical analysis. In American Marketing Association (Ed.), *AMA Winter Educators' Conference Proceedings* (pp. 7-11). Chicago: American Marketing Association.

Corstjens, M.L. & Gautschi, D.A. (1983). Conjoint analysis: A comparative analysis of specification tests for the utility function. *Management Science, 29* (12), 1393-1413.

Finkbeiner, C.T. & Platz, P.J. (1986, October). *Computerized versus paper and pencil methods: A comparison study*. Paper presented at the Association for Consumer Research Conference, Toronto.

Green, P.E., Krieger, A.M. & Agarwal, M.K. (1991). Adaptive conjoint analysis: Some caveats and suggestions. *Journal of Marketing Research, 28*, 215-222.

Green, P.E., Krieger, A.M. & Schaffer, C.M. (1993a). An empirical test of optimal respondent weighting in conjoint analysis. *Journal of the Academy of Marketing Science, 21*, 345-351.

Green, P.E., Schaffer, C.M. & Patterson, K.M. (1991). A validation study of Sawtooth Software's adaptive conjoint analysis. In M. Metegrano (Ed.), *1991 Sawtooth Software Conference Proceedings* (pp. 303-314). Ketchum, ID: Sawtooth Software.

Holling, H., Jütting, A. & Großmann, H. (1998). *ALASCA. Unpublished software for collecting and analyzing conjoint data*. Westfälische Wilhelms-Universität Münster, Psychologisches Intitut IV.

Huber, J. & Hansen, D. (1986). Testing the impact of dimensional complexity and affective differences of paired concepts in adaptive conjoint analysis. In M. Wallendorf & P. Anderson (Eds.), *Advances in consumer research* (No. 14, pp. 159-163). Provo, UT: Association for Consumer Research.

Huber, J., Wittink, D.R., Fiedler, J.A. & Miller, R. (1993). The effectiveness of alternative preference elicitation procedures in predicting choice. *Journal of Marketing Research, 30*, 105-114.

Kruskal, J.B. (1965). Analysis of factorial experiments by estimating a monotone transformation of the data. *Journal of the Royal Statistical Society, Series B, 27*, 251-263.

Melles, T. (1996). *Optimierung der Paarvergleichsaufgabe im Rahmen der adaptiven Conjoint Analyse*. Unveröffentlichte Diplomarbeit, Westfälische Wilhelms-Universität Münster.

Orme, B.K., Alpert, M.I. & Christensen, E. (1997). Assessing the validity of conjoint analysis - continued. In *Proceedings of the Sixth Sawtooth Software Conference* (pp. 209-225). Sequim, WA: Sawtooth Software.