OTTO-VON-GUERICKE-UNIVERSITY MAGDEBURG
FACULTY OF ECONOMICS AND MANAGEMENT

Do Consumers' Stated Preferences in Choice Models Depend on Differences in Stimulus Presentation: 2D versus 3D Presentation?

Alma Berneburg

FEMM Working Paper No. 23, October 2007

F E M M

Faculty of Economics and Management Magdeburg

Working Paper Series

Otto-von-Guericke-University Magdeburg
Faculty of Economics and Management
P.O. Box 4120
39016 Magdeburg, Germany
http://www.ww.uni-magdeburg.de/

"Do Consumers' Stated Preferences in Choice Models Depend on Differences in Stimulus Presentation: 2D versus 3D Presentation?"

Alma Berneburg[1]

University of Applied Sciences Merseburg, Geusaer Straße, 06217 Merseburg, Germany

Abstract

This study tries to contribute to the branch of research that is engaged in the analysis of different stimulus presentation formats and their influence on quality and validity of the test results in a Conjoint Analysis. This topic has gained special attention as new techniques became available that enable the inclusion of holographic three-dimensional stimuli in the research of consumers' preferences. Especially for examining design-related questions this proves very interesting.

The study compares the results of two Choice Based Conjoint analyses with one presenting the test object via computer-based 2D-pictures and the other using a holographic 3D-simulation. For the attributes at hand no differences between the results of the 2D- and 3D-test can be isolated on an aggregate level.

Keywords: Stimulus presentation formats, Dimension effects, Consumer decision making, Choice experiments

JEL codes

[1] Corresponding author. Tel.: +49 3461 46 24 32; fax: +49 3461 46 24 22.
E-mail address: Alma.Berneburg@hs-merseburg.de

When dealing with the measurement of consumer preferences Conjoint Analysis (CA) is one of the most commonly used techniques in market research. It can be applied in varying operational areas ranging from specifying price-response-functions to measuring product preferences in an innovation process (e.g. Wittink & Cattin, 1989 or Wittink, Vriens, & Burhenne, 1994). One of the factors that may crucially affect the validity of the analysis is the stimulus presentation format. Most of the research so far has been focused on the comparison of verbal product descriptions to two-dimensional product image stimuli (e.g. Louviere, Schroeder, Louviere, & Woodworth, 1987, de Bont, 1992, Huisman, 1997 or Vriens, Loosschilder, Rosbergen, & Wittink, 1998) or to physical stimuli such as dummies or actual products (e.g. Anderson, 1987 or Sattler, 1994). Only very few studies have dealt with the inclusion of multi-medial stimuli in the CA (Ernst, & Sattler, 2000).

In this study, we introduce an alternative virtual stimulus, the holographic three-dimensional product image (3D-stimulus), and experimentally assess its influence on the outcomes of a conjoint analysis to reveal potential differences between alternative visual stimuli (e.g. presenting an attribute either two- or three-dimensionally). Two Choice Based Conjoint (CBC) analyses are performed, one presenting the test object via computer-based 2D-pictures and the other using a holographic 3D-simulation. To be able to evaluate the test results against a benchmark, market shares are estimated on the basis of the respondents' part-worth utilities. These market share estimations are first compared internally with respect to convergent and predictive validity. Additionally, the external validity is assessed using actual market share data provided by a major market research company.

The paper is structured as follows: After the literature review, a further section describes the current study, introduces the three fundamental hypotheses and gives details on the used methods

and the general test design of the study. In a next step the findings are presented and discussed and the implications of these results for the hypotheses are being displayed. Finally a summary section concludes the paper.

1. Literature review

In general, one would assume that respondents react highly different to a product whether it is realistically and spatially represented or whether it is described verbally. Furthermore, it can be assumed that these differences have consequences for the quality of the test results. It hence is not surprising that in the last two decades some studies have tried to research the above assumptions and to gain further insight into the importance of stimulus presentation formats.

In the following an overview over the existing studies will be given. While not all of the considered comparative studies are based on Conjoint Analyses, the author assumes the results to be nevertheless adjuvant for the research question to be answered.

Albeit the intuitiveness of the assumption that, depending on the used stimulus presentation format, non-converging test results are likely, not all studies support this hypothesis. Anderson (1987) performed a comparative study, in which an actual product (a product innovation in industrial clothing), a written description and a combination of the two were analysed. When looking at the estimated part-worth utilities, Anderson found that the different types of representation gave a convergent behaviour of the results and therefore written descriptions can be used “as a proximate representation of actual products” for industrial research studies (p. 43). The study of Louviere et al. (1987) – dealing with test persons’ preferences for state parks and laying a special focus on the comparability of the used stimuli – also showed only little statistical evidence for differences in the models developed from either verbal or visual stimuli. Ernst and Sattler (2000)

involved pictures and sounds in a conjoint study dealing with alarm clocks and compared their findings to results of a traditional CA only using verbal descriptions. Surprisingly, they found only little difference between the two forms of stimulus presentation, as well.

But there also are studies that come to contrary conclusions: In their comparative study from 1981, Holbrook and Moore measured consumers' preferences by presenting the stimuli (sweaters, a mainly aesthetic product) both visually (sketched) and verbally. They hypothesized that pictorial presentations generate a more global focus and many interaction effects while verbal presentations should generate a more restricted focus and fewer interactions. Their study's results generally supported these hypotheses.

Domzal and Unger (1985) replicated the analysis from Holbrook and Moore but applied wrist watches as a more functional product. Their findings, however, did not support all of the statements of Holbrook and Moore, but in general the results of verbal versus visual presentation format were non-convergent as well. Similar results can be found in MacKay, Ellis and Zinnes (1986), who explored differences between configurations derived from graphic and verbal stimuli in a MDS analysis and measured a non-convergent behaviour of the results. Weisenfeld (1989) researched the convergent validity for three different product categories (cigarettes, meat salad and backpacks) by comparing written descriptions on product cards to real products. As the majority of test objects under study showed non-converging results, one can assume this to be a hint for a different processing of varied stimulus presentation formats.

Louviere's (1987) assumption that, when the stimuli's correspondence is assured, there will be no serious differences in the models developed from either verbal or visual stimuli gets actually disproved by the findings of de Bont (1992). In his comparative study dealing with filter coffeemakers, de Bont followed Edell and Staelin (1983) and Unnava and Burnkrant (1991) and used a special procedure to assure the comparability of the stimuli. He discovered that even with corre-

sponding forms of stimulus presentation there were significant differences in the evaluation of the attribute "form". Vriens (1995) traced back the differing results he found in his study mainly to the discrepancies between stimuli that include a "design-attribute" and stimuli that don't. He concluded that for design-attributes visual stimuli are essential to gain unbiased conjoint results if marketplace decisions are based on pictures or physical products as well. Huisman (1997) in fact found non-converging results in his comparative study of verbal versus multi-media stimuli (which just means visualized stimuli in his study), but also found that the effective interview time in both samples was surprisingly identical. There was no hint towards the conclusion that visualization could lead to shorter interviews. Vriens et al. (1998) added the statement that pictorial stimuli improve a participant's understanding of the attributes (especially of design attributes), while verbal stimuli seem to facilitate a respondent's judgement.

Summing up, one can say that the existing studies mostly find differences when comparing different forms of stimulus presentation. There are hints for different respondent reactions depending on the stimulus presentations – when looking at verbal vs. alternative stimuli. No large-scale representative CBC-study up to now has taken into consideration that there may be crucial differences in the processing of visual stimuli with varying dimensions[2].

Holographic 3D-stimuli have the potential to contain many advantages: The sometimes still quite artificial CA-study could be designed in a more realistic manner and the major disadvantages of simple 2D-stimuli, the missing ability to create complex objects that provide detailed impressions of depth or distance (Loosschilder, Rosbergen, & Wittink, 1995), could be counterbalanced (for

[2] Berneburg (2007) presents a small-scale study comparing two- and three-dimensional stimuli to physical dummies and finds the survey results with the two-dimensional stimuli to be different from the other two types of stimuli. That study, however, differs from this investigation in numerous crucial ways. First, a non-representative student sample was used. Second, only a single stimulus was used. Third, the external validity was not tested.

details see the technical appendix). This is especially important in the case of products which respondents would like to experience from different angles (e.g. automobiles, mobile phones, furniture, or package designs) and in a realistic manner. The more detailed the information of the product or concept (the stimulus in general), the higher the degree of realism and the more certain a participant can respond to the stimulus – a fact that might result in a higher external validity (Vriens, 1995). According to de Bont (1992), consumers at an intermediate degree of realism have to put much more effort in interpreting the information and may hence come to different interpretations of the same information. Loosschilder et al. (1995) supportingly state that especially for product choices that are strongly guided by styling components simple stimuli are not realistic enough and, hence, are not adequate. Loosschilder et al. define realism similarly to de Bont (1992) as "the degree to which the representation resembles the actual product" (p. 19). Johnson, Meyer, and Ghose (1986) supportingly state that a model should be calibrated on stimuli to represent the real world as closely as possible to gather Pareto optimality. Green, Helsen, and Shandler (1988), in a slightly different context (they researched conjoint internal validity under alternative profile presentations), repeat this claim. Also, as "the realism of stimuli is a determinant of the validity of consumer evaluations" (Loosschilder et al., 1995, p. 21), 3D-stimuli should help to enhance the validity of test results especially for design and packaging matters while still using the given resources more efficiently than when using actual physical stimuli (with verbal- or 2D-stimuli, of course, being less expensive, but at the same time less realistic). 3D-stimuli furthermore allow the evaluation of product innovations against a realistic background at an earlier stage – especially when dealing with products with completely new visual attributes this could be a drastic advantage.

But while the degree of realism of a test environment is enhanced by complex stimuli, they also make it more difficult for a test person to isolate single determining attributes out of the whole stimulus presentation (Smead, Wilcox, & Wilkes, 1981 and Moore & Holbrook, 1982) in the more demanding task for the test persons (Anderson, 1987). As a result, test persons could be seduced into using simplification strategies. They could be led to a decision based on only very few isolated product attributes, that are relevant to them (Boecker & Schweikl, 1988), rather than taking all characteristics of the product into account.

This threat to evaluative tasks with a high degree of realism seems to diminish, however, when considering, that in real life consumers also have to deal with integrated concepts and whole product profiles when making a buying decision. In the context of early product-concepts Finn (1985) states that "for the prediction to have value, the stimulus presented at the time of the concept test must convey to the subjects the same meaning that they would extract from a marketplace exposure to the product at a later time of launch" (p. 37). Following this statement, one should be aiming for a most realistic stimulus presentation at all times, in order to reduce the threat of biased results. Vriens et al. (1998) additionally point out that if marketplace choices are based on an inspection of the actual product or its visualization, then visual stimuli should enhance the degree of realism in a choice task and therewith the quality of the evaluation process by providing greater external validity. Making a task more simple to the respondent might result in missing the actual target: surveying the consumers' real behaviour. And the threat of simplification strategies can be annulled by referring to the fact, that presumably the simplification strategies a participant uses in a complex evaluation task with a high degree of realism is the same strategy that would be used in real life at the marketplace (Loosschilder et al., 1995).

2. The current study

Almost no study up to now has taken into consideration, that there might be crucial differences between varying *visual* stimuli, whereas in 1992, de Bont already required further research into the importance of the attributes "form" and "design" when considering pictorial presentation formats with differing degrees of realism. With the 3D-visualization technology now available, testing objects with mainly visual-based attribute specifications may result in interesting discrepancies between different stimulus presentations in terms of the test results' quality, the external validity or the consequences for possible application areas. These differences are the main factors under investigation in the research at hand.

2.1 Hypotheses

Three-dimensional spatial and vivid simulations may provide more information and, thus, form a better foundation for a CBC-preference statement. To test for this effect, the following hypotheses were formulated:

H1: 3D-stimuli lead to test results that differ from those gained by using 2D-stimuli in terms of convergent validity.

H2: 3D-stimuli lead to a higher predictive validity than 2D-stimuli.

H3: 3D-stimuli lead to a higher external validity than 2D-stimuli.

In order to look into these hypotheses the following test design was used.

2.2 Methodology

In marketing research processes, Conjoint Analyses is an often used tool for measuring consumers' product preferences. Especially the Choice Based Conjoint Analysis has gained major attention in the last couple of years (Hartmann & Sattler, 2002). A comparative survey of Sawtooth

Software users over the last 4 years illustrated, that CBC nowadays is the most used version (75%) among the three main variants (ACA, CBC and CVA) of Conjoint Analyses (Sawtooth Software, Inc. 2006).

CBC goes back to Louviere and Woodworth (1983) and, similarly to the traditional Conjoint Analysis, its intention is to determine consumers' product preferences and to express these preferences in terms of part worth utilities. CBC-analyses however add some major advantages to the traditional Conjoint Analysis by enhancing the degree of realism in the survey and increasing the external validity of the results. CBC-surveys consist of consumers expressing their preferences by simply choosing their preferred single product concept from a variety of concepts, rather than rating or ranking them. Therefore, the task is closer to a real buying decision at the point of sale in the consumers´ everyday life: choosing a preferred concept is similar to what consumers actually do in the market day by day.

This study was conducted in cooperation with the GfK, an international market research institute. The general configurations and processes therefore have been adjusted to the GfK standard procedures, in this special case to the GfK Price Challenger (Wildner, 1998 and 2003). The Price Challenger (PC) is a special form of CBC in which the whole test product is evaluated and only prices are allowed to vary. It therefore is suitable for problems especially resulting from price research. Its structure in general is quite similar to a Choice Based Conjoint study: test persons repeatedly select their most preferred product out of a choice of several concepts. The selection however only takes place within the test person's relevant set that has been selected in a short questioning prior to the actual Conjoint study. Moreover, in contrast to a classical CBC-study in the actual choice tasks the test person is allowed to pick more than just one test product if desired. When evaluating the test persons' responses, aggregate market shares are estimated with a spe-

cific simulation tool based on the respondents' individual utilities and buying probabilities (Wildner, 2003).

2.3 Test design

In August 2006 two comparative CBC-studies using identical test designs were performed. One used 2D- the other 3D-stimuli to present the test object: shampoo.

The studies were constructed as follows:

1. In order to control for potentially relevant background variables, a preliminary questionnaire was included in all interviews. This questionnaire covered socio-demographic topics (e.g. age, sex, etc.) and issues related to the main topic of the survey, namely the usage of hair care products (e.g. frequency of usage and purchase of shampoo, location of purchase, etc.). In this first stage, subjects that were not in the target group (e.g. non-buyers) were excluded.
2. A relevant set was individually selected, in order to ensure the familiarity of the respondents with the used attributes.
3. In both studies the actual CBC-survey then specified 10 randomized choice tasks for each respondent and additionally one holdout task (choice task 6)[3] in order to provide a proximal indication of predictive validity (11 choice tasks overall). Every choice task consisted of up to 5 test objects (stimuli) derived from each respondent's individual relevant set with randomized price assignments and the option not to buy any of the products at the given price (None-Option).

[3] Typically the holdout task is located about midway in the survey. At this point the respondents got used to the task but fatiguing effects have not yet occurred. In contrast to the constant holdout tasks in a classical CBC-survey, due to the specific relevant sets and the individualized choice tasks, the composition of the holdout task (choice task 6) differed from respondent to respondent. The hit rate was calculated on an aggregate level due to the overall percentage of matches between the most preferred product (according to the individual part-worth utilities) and the actual purchase in the holdout task.

As the inclusion of a relevant set is not a standard procedure in a classical CBC-study, this step shall be explained in a little more detail. The selection process was performed in sequencing steps with the following questions (Wildner, 2003):

TABLE 1

QUESTIONS OF THE RELEVANT SET SELECTION PROCESS

step	question
1	Which one of the following products do you know?
2	Which one of the following (known) products have you bought in the last 6 months? (filter: known products from step 1)
3	Which one of the following (already bought) products you would not buy again? (filter: bought products from step 2)
4	Which one of the following (known but not bought) products you would additionally consider for buying? (filter: un-bought products from step 2)

The individual relevant set of a respondent therefore consists of the products the test person

- knows, had already bought and was satisfied enough to buy it again or
- knows, had not bought so far, but would realistically consider the purchase.

Normally, a relevant set extracted with this selection process consists of 3 to 6 products (Wildner, 2003).

In the case when a test person selected more than 5 products in this selection process for the relevant set, he was requested to name the 5 products he would buy most frequently. When in contrast a test person selected less than 5 products, no further adjustments to the relevant set were necessary. As the result, every choice task consisted of up to 5 test objects and the additional None-Option.

In table 2 the respective means and standard deviations of each of the steps of the relevant set selection process for the 2D- and the 3D-study are displayed:

TABLE 2

MEANS AND STANDARD DEVIATIONS OF THE AMOUNTS OF CHOSEN PRODUCTS IN THE RESPECTIVE STEPS OF THE RELEVANT SET SELECTION PROCESS

	2D		3D	
	Mean	Standard Deviation	Mean	Standard Deviation
known products	13.53	2.52	13.08	2.77
bought products	3.18	2.22	3.05	1.91
not bought again[a]	.52	.91	.41	.75
perhaps bought[b]	2.57	1.85	1.99	1.59
relevant set	4.04	1.25	4.09	1.20

[a] subset of known and bought products

[b] subset of known but un-bought products

There are no crucial differences between the results of this selection process for the two different study specifications. The dimension of the stimulus presentation format does not seem to influence the size of a test person's relevant set.

The question that arises is whether the composition of the relevant sets varied systematically across the two treatments. To check for differences, the shares of the products in the respective relevant sets are considered on an aggregate level and then are compared between the 2D- and the 3D-sample. To account for differing sample sizes, the absolute amount of entries into the relevant set gets weighted in proportion to the respective sample size ($N_2 = 259$ and $N_3 = 205$) for every single product *i*:

$$\frac{rs_{i2}}{N_2} - \frac{rs_{i3}}{N_3} = \Delta_i\% \qquad (1)$$

with rs_{i2} being the absolute amount of entries for a product i (i = 1,…,19) in the relevant sets of all test persons in the 2D-sample and rs_{i3} being the absolute amount of entries for a product i in the relevant sets of all test persons in the 3D-sample.

In table 3 the respective shares of the 19 products in the relevant sets and the differences Δ_i are displayed:

TABLE 3

DIFFERENCES Δ_i BETWEEN THE SHARES OF PRODUCTS IN THE RESPECTIVE RELEVANT SETS (in %)

Products i	Shares in 2D-sample	Shares in 3D-sample	Δ_i
i1	39.080	42.439	-3.359
i2	9.962	15.122	-5.160
i3	42.146	44.390	-2.245
i4	28.736	43.902	-15.167
i5	9.962	9.268	0.693
i6	33.716	33.659	0.058
i7	16.475	21.951	-5.476
i8	7.280	5.854	1.426
i9	28.352	25.366	2.987
i10	36.782	34.146	2.635
i11	40.613	36.098	4.515
i12	40.996	35.610	5.386
i13	39.847	35.122	4.725
i14	4.598	2.439	2.159
i15	2.682	0.976	1.706
i16	2.682	1.463	1.219
i17	1.149	0.488	0.662
i18	15.326	18.537	-3.211
i19	3.448	2.439	1.009

The only substantial difference we observe concerns product i4 that is chosen significantly more often into their relevant set by the subjects in the 3D-sample than in the 2D-sample. But as the differences Δ_i overall, however, show no systematically biased distribution ($\sum_{i=1}^{19} \Delta_i = -5.44$),

the dimension of the stimulus presentation format seems to make no difference to the composition of a respondent's relevant set.

2.4 Sample

An overall sample of 464 test persons was drawn emanating from a homogeneous survey population of customers from a big department store. A between-subject-design was engaged to minimize distorting learning effects (Agarwal & Green, 1991) and in order to prevent an overtaxing of the test persons´ readiness and patience (Huber, Wittink, Fiedler, & Miller, 1993). The test persons were randomly assigned to one of the two samples and results of 259 persons in the 2D-study were compared to those of 205 test persons in the 3D-study. The smaller sample size in the 3D-case resulted from the fact, that the setup of the 3D-technique was somewhat more time-consuming than in the case of the 2D-technique. Therefore, the 2D-survey started slightly earlier and more test persons could be recruited to this sample.

As the two samples were randomly drawn from the same basic population, a comparison of the results for the two alternative stimulus presentations should be possible. In order to assure that there were no major differences between the two groups in terms of relevant background variables, the data gained from the preliminary questionnaire was analysed and resulted in the conclusion that the respective samples were homogenous and the results are hence comparable.[4]

2.5 Attributes and levels

The attribute "product" consisted of 19 different shampoos, representing 80% of the German shampoo-market:

[4] The detailed evaluation of the test persons' characteristics will not be presented in this paper, as it would unnecessarily extend the analysis. The results can, nonetheless, be requested from the author.

FIGURE 1

IMAGES OF THE 19 DIFFERENT LEVELS OF THE ATTRIBUTE "PRODUCT"

The picture quality of the 3D-simulations was comparable to the photorealistic pictures above but supplemented with an additional spatial impression of depth and perspective. Furthermore it was possible to pick the product from the virtual shelve and turn it in front of the test person.

The attribute "price" for the 19 different shampoos ranged between 11 levels specific to each product. The price ranges were defined according to actual marketplace data to assure realistic impressions for the test persons:

TABLE 4

19 DIFFERENT PRICE RANGES AS LEVELS OF THE ATTRIBUTE "PRICE"

product	price ranges[a]										
	1	2	3	4	5	6	7	8	9	10	11
i1	1.69	1.79	1.89	1.99	2.09	2.19	2.29	2.39	2.49	2.59	2.69
i2	1.59	1.69	1.79	1.89	1.99	2.09	2.19	2.29	2.39	2.49	2.59
i3	1.99	2.09	2.19	2.29	2.39	2.49	2.59	2.69	2.79	2.89	2.99
i4	1.89	1.99	2.09	2.19	2.29	2.39	2.49	2.59	2.69	2.79	2.89
i5	0.99	1.09	1.19	1.29	1.39	1.49	1.59	1.69	1.79	1.89	1.99
i6	1.39	1.49	1.59	1.69	1.79	1.89	1.99	2.09	2.19	2.29	2.39
i7	3.89	3.99	4.09	4.19	4.29	4.39	4.49	4.59	4.69	4.79	4.89
i8	0.89	0.99	1.09	1.19	1.29	1.39	1.49	1.59	1.69	1.79	1.89
i9	2.59	2.69	2.79	2.89	2.99	3.09	3.19	3.29	3.39	3.49	3.59
i10	1.49	1.59	1.69	1.79	1.89	1.99	2.09	2.19	2.29	2.39	2.49
i11	1.49	1.59	1.69	1.79	1.89	1.99	2.09	2.19	2.29	2.39	2.49
i12	1.99	2.09	2.19	2.29	2.39	2.49	2.59	2.69	2.79	2.89	2.99
i13	0.99	1.09	1.19	1.29	1.39	1.49	1.59	1.69	1.79	1.89	1.99
i14	0.39	0.49	0.59	0.69	0.69	0.79	0.79	0.89	0.99	1.09	1.19
i15	0.39	0.49	0.59	0.69	0.69	0.79	0.79	0.89	0.99	1.09	1.19
i16	0.39	0.49	0.59	0.69	0.69	0.79	0.79	0.89	0.99	1.09	1.19
i17	0.39	0.49	0.59	0.69	0.69	0.79	0.79	0.89	0.99	1.09	1.19
i18	0.49	0.59	0.69	0.69	0.79	0.79	0.89	0.99	1.09	1.19	1.29
i19	0.49	0.59	0.69	0.69	0.79	0.79	0.89	0.99	1.09	1.19	1.29

[a] most common price = current market price

As a result, one received the test object "shampoo" with 19 levels of the attribute "product" and with 11 levels of the attribute "price".

3. Findings and General Discussion

3.1 Part-worth Utilities

In a first step the part-worth utilities for the attribute "product" of the 2D- and the 3D-sample were compared (table 5). The utilities have been standardized (with 0 as the least preferred and 1 as the most preferred products from the purchase acts) to make them comparable on an individual level as well:

TABLE 5

PART-WORTH UTILITIES FOR THE ATTRIBUTE ”PRODUCT” ON AN AGGREGATE LEVEL

products	dimension	N	Mean[a]	Standard Deviation	Mean Standard Error
i1	2D	70	.560	.413	.049
	3D	72	.585	.400	.047
i2	2D	20	.487	.462	.103
	3D	19	.392	.448	.103
i3	2D	66	.619	.436	.054
	3D	61	.581	.432	.055
i4	2D	51	.547	.424	.059
	3D	71	.588	.421	.050
i5	2D	17	.302	.385	.093
	3D	18	.327	.420	.099
i6	2D	60	.552	.421	.054
	3D	54	.531	.420	.057
i7	2D	15	.989	.044	.011
	3D	21	1.000	.000	.000
i8	2D	14	.398	.485	.130
	3D	9	.456	.409	.136
i9	2D	45	.799	.355	.053
	3D	28	.830	.290	.055
i10	2D	78	.546	.446	.050
	3D	59	.564	.416	.054
i11	2D	69	.571	.434	.052
	3D	57	.531	.424	.056
i12	2D	65	.549	.421	.052
	3D	48	.530	.386	.056
i13	2D	76	.575	.441	.051
	3D	64	.323	.415	.052
i14	2D	10	.394	.398	.126
	3D	5	.200	.447	.200
i15	2D	6	.167	.408	.167
	3D	2	.000	.000	.000
i16	2D	5	.217	.381	.170
	3D	3	.342	.570	.329
i17	2D	3	.000	.000	.000
	3D	1	.132	.	.
i18	2D	30	.432	.462	.084
	3D	34	.190	.349	.060
i19	2D	6	.305	.475	.194
	3D	5	.432	.522	.234

[a] The means represent the individual utilities on an aggregate level.

As one can observe only very few relevant sets contained the products i14, i15, i16, i17 and i19 and thus they only entered in a small amount of choice tasks ($N \leq 10$). As these products were private labels, their small *N* might result from the fact that the survey took place in a big German department store (Kaufhof) and the test persons recruited there probably had small exposure to the specific retailers of the private labels. Since sample sizes are too small to allow for reliable statistics, these products are not incorporated in the following analyses.

When now looking for the stimulus dimension's influence on the respondents' evaluative answers, first of all a t-test for the equality of the respective means was conducted. All in all 12 test products show no differences between the two stimulus presentation formats. Table 6 displays the remaining two products with significant differences in the means:

TABLE 6

T-TEST FOR EQUALITY OF MEANS

	t	df	sig.	Mean Difference	Standard Error Difference	95% Confidence Intervall of the Mean	
						Lower	Upper
Utility i13	3.463	138	.001	.252	.073	.108	.396
Utility i18	2.346	53.661	.023	.243	.103	.035	.450

As one can see, with the utilities of product i13 ($t = 3.463, p = .001$) and product i18 ($t = 2.346, p = 0.023$) there are significant differences in the respective means to be identified. Astonishingly, in both cases the utilities for the 3D-case are lower than for the 2D-case: i13 with $M_{2D} = .575$ vs. $M_{3D} = .323$ and i18 with $M_{2D} = .432$ vs. $M_{3D} = .190$. As a possible explanation it should be pointed out that both products belong to the low-price segment and it is possible

that the 3D-stimulus presentation created a high-quality impression that clashed with the low-price image these products hold in the market.

Overall, however, the dimension of the stimulus presentation format seems to make little difference to the utilities of the test products on an aggregate level.

3.2 Validation Process

As the special setup of the PC is aligned on the estimation of market shares, the following analyses will base on these estimations as well. Especially, when considering that the external validity of the two competing stimulus presentation formats shall be evaluated against the background of real market data (real market shares), this approach seems consequential.

3.2.1 Convergent Validity

According to Campbell and Fiske (1959), convergent validity describes the phenomenon that the scale items in a given construct move in the same direction and, thus, highly correlate. Convergent validity therefore differs from reliability to the effect that tests of reliability only include the scale items for a single construct without comparing them to other constructs.

In this specific study, convergent validity is measured by comparing the estimated market shares of the 2D-analysis to the ones of the 3D-study. Table 7 displays the respective correlations of the estimated market shares based on the respective part-worth utilities for the 14 products which have not been excluded earlier ($N_{MS} = 14$):

TABLE 7

CORRELATIONS OF THE ESTIMATED MARKET SHARES OF 2D- VS. 3D-STUDY (%)

	Aggregate Level	Male	Female	15-19 y.	20-29 y.	30-39 y.	40-49 y.	50-59 y.	≥ 60 y.
Spearman's rho	.895(**)	.864(**)	.833(**)	.746(**)	.908(**)	.530(*)	.763(**)	.462	.123
N_{2D}		49	201	28	98	42	48	22	21
N_{3D}		50	155	25	88	39	23	15	15

** significant at 1%-level
* significant at 5%-level

According to Spearman's rank correlation coefficient rho as a nonparametric rank statistic, which seems appropriate due to the small N_{MS}, the strength of the association between the two variables is .895 on the aggregate level. This means the two variables are highly correlated which according to Campbell & Fiske (1959) can be taken as a sign of high convergent validity.

The comparison of the estimated market shares was not only performed on the aggregate level (2D vs. 3D) but for a deeper understanding of the dimension effects, sub-samples have been used to compare the two stimulus presentation formats in a more detailed manner as well.

Specifications of the sub-samples are:

- respondent's sex: male and female
- respondent's age: 15-19 years, 20-29 years, 30-39 years, 40-49 years, 50-59 years, ≥ 60 years

These specifications are in line with the assumption that rather than the frequency of shampoo purchase or other hair care related issues most likely personal characteristics like age or sex are the reason why respondent's reactions might change according to the dimension of the stimulus presentation.

Analyzing the data on the level of the sub-samples, a similar picture as on the aggregate level can be observed: When looking for gender related effects, there are no significant differences. Male and female respondents both show a highly correlated answering behaviour between the respective 2D- and the 3D-samples with $\rho_m = .864$ and $\rho_f = .833$.

In the case of the test person's age, in general little difference between 2D and 3D can be observed either. Except for the sub-samples "50-59 years" ($\rho = .462$) and "≥ 60 years" ($\rho = .123$), every other comparison of the corresponding sub-samples shows significant correlations of the market shares between the 2D- and the 3D-study. Worth mentioning is the fact that the market share estimations seem to diverge as age increases, as indicated by a Spearman's rho of $-.714$, which in this case indicates a substantial negative correlation between the correlations of the market share estimates of the two studies and the age group of the respondents. Hence, it seems the younger the respondent the better the 2D- and 3D-estimates match.[5][6]

3.2.2 Predictive Validity

The predictive validity is obtained through hit rates. The hit rate is the percentage of matches achieved when comparing the actual respondent's decision in the holdout task (purchase/non-purchase) and the estimation based on the calculated part-worth utilities (most preferred product). As the GfK's procedure includes the option to choose more than one product in each choice task, the standard hit rates needed to be adapted to this option by additionally calculating weighted hit rates (table 8):

[5] This result has to be taken with some caution. Classifying subjects into age groups reduces the number of observations to six. Furthermore the negative correlation seems to be driven by the two eldest age groups.
[6] The question which stimulus presentation format performs better will be covered when looking at the external validity.

TABLE 8

HIT RATES FOR PREDICTIVE VALIDITY (%)

	unweighted		weighted		random choice	
	Mean	Std.dev.	Mean	Std.dev.	Mean	Std.dev.
PC: 2D *(N = 259)*	.734	.443	.565	.421	.217	.081
PC: 3D *(N = 205)*	.742	.439	.580	.418	.213	.073

Since the chance of obtaining a hit obviously increases as the amount of chosen products in the holdout task rises, simply measuring the predictive validity as the percentage of achieved hits (unweighted hit rate) overestimates the goodness of fit. This problem is solved by weighting the hit rate with the number of chosen products of each holdout task. Each hit now was "punished" with a weighting factor and the weighted hit rate poses as follows:

$$whr_j = \frac{1}{pc_j} \cdot hr_j \tag{2}$$

where whr_j is the weighted hit rate, hr_j is the unweighted hit rate and pc_j is the sum of product choices for each respondent j in the respective holdout task.

The higher the number of chosen products, the greater the "punishment" and thus the smaller the value of the respective hit that is considered in the overall hit rate.

Assuming the respondent's first choice in the holdout task will in most cases also provide the highest utility, the weighted hit rate will thus display the minimum level of the predictive validity. If the choice task had only allowed for one pick, these cases would have generated a full hit. The punishment with the weighting factor, however, does not consider the order of choices but fully amerces the fact of multi-choice holdout tasks. One therefore can assume the "real hit rate"

to lay somewhere in between the weighted hit rate as a lower and the unweighted hit rate as an upper bound with a likely tendency towards the upper limit.

The random choice hit rate, displayed in the last column of table 8, indicates the hit rate that would have resulted if test persons were to have unstable preferences and were to pick products at random. This is a lower benchmark for the other two hit rates.

The predictive validity is convincing in both cases (2D and 3D) with no observable differences between the two treatments due to the stimulus presentation format. It is to mention, though, that the goodness and consistency of the answering behaviour – as measured by the standard deviations – is quite heterogeneous within both samples.

3.2.3 External Validity

Very few studies use external validity measures (Natter & Feurstein, 2002). A possible reason might be the fact that testing for external validity involves comparing study-based market share estimations to real market share data, which often are difficult to obtain. In the mid 1990s, however, some prominent researchers already expressed the need for a detailed analysis of external validity measures for conjoint analyses (e.g. Carson et al., 1994 or Neslin et al., 1994).

In the study at hand real market data was used that is provided by the GfK to measure the external validity. As the data is to be treated confidentially, the real market shares are not displayed in this paper and only the following measures shall be presented:

TABLE 9

CORRELATIONS OF THE ESTIMATED MARKET SHARES VS. THE REAL MARKET SHARES (%) AND ROOT MEAN SQUARE ERROR (RMSE)

	Spearman's rho							RMSE
	aggregate	15-19 y.	20-29 y.	30-39 y.	40-49 y.	50-59 y.	≥ 60 y.	
2D vs. real	.648(*)	.840(**)	.499	.389	.570(*)	.695(**)	.524	4.96
3D vs. real	.560(*)	.670(**)	.578(*)	.354	.538(*)	.130	.310	6.14
rank order		1	2	3	4	5	6	

** significant at 1%-level
* significant at 5%-level

With .648 for Spearman's rho the 2D-estimates on the aggregate level show a higher correlation with the panel data than the 3D-estimates with .560 – with both correlations being significant on a 5%-level. The same picture shows when looking at the root mean square errors. The Root Mean Square Error of the 2D-estimates with 4.96 percentage points is smaller than the one of the 3D-study with 6.14 percentage points. But since the Diebold-Mariano test for predictive accuracy (Diebold & Mariano, 1995) with $t = 1.338$ and a p-value of .204 indicates that the null hypothesis of similar forecast errors cannot be rejected, these differences do not seem to be significant. One hence is inclined to conclude that both stimuli encompass comparable external validity on the aggregate level.

Since only the age-based sub-samples resulted in non-convergent results for the 2D- and the 3D-treatment, the gender sub-samples will not be considered here and only the respective initial correlations of the age-based sub-samples will be observed with regard to their rank order. The results paint an interesting picture: With Spearman's $\rho = -.086$ there seems to be no significant correlation between the rank order and the initial correlation of 2D-estimations and real data. This indicates that the correlations between the 2D market shares and the real data do not systematically in- or decrease with the age of the respondent. When looking at the 3D-estimations this pic-

ture is quite different. With Spearman's $\rho = -.886$ the rank order and the initial correlations between 3D-estimates and real panel data are correlated at a 5%-level of significance. This implies that there is a systematic decline of the correlations between 3D market shares and real data as the age of the respondent rises. In other words, the elder the test persons the bigger the divergence between the market share estimations gained with the 3D-stimuli and the real data.

4. Summary and Outlook

4.1 Discussion of the Hypotheses

The primary goal of this study was to compare a three-dimensional stimulus presentation format to a two-dimensional one. Special focus was placed upon the different forms of validity. The three underlying hypotheses therefore dealt with the quality of the respective survey results (convergent validity), the goodness of the estimations based on these results (internal predictive validity) and the comparison of the study results to real market data (external validity).

The results of the 2D- and the 3D-study reveal an overall convergent behaviour. The dimension within the visual stimulus presentation does not seem to make a difference to the results of the choice tasks (at least in this study with the attributes and levels at hand) and H1 has to be rejected. The predictive validity paints a similar picture: Albeit assuming the degree of realism to be higher when including spatial 3D-stimuli, the predictive validity does not significantly differ for the two forms of stimulus presentation. H2 has to be rejected as well. Finally, external validity, as the ultimate criterion of interest when trying to capture consumers' actual behaviour, between 2D- and 3D-stimulus presentations does not differ substantially either. H3 therefore also has to be rejected. All in all, no significant differences between the aggregate market share estimations based on two-dimensional and three-dimensional stimuli appear.

4.2 Possible Explanations

One possible explanation for this study's results could be that three-dimensional stimuli simply do not influence the amount of information transported to the test person and therefore are not able to generate a higher degree of realism and quality-wise enhanced test results in a Conjoint Analysis. This assumption is rather unlikely as first investigations of the effects of three-dimensional stimuli come to contrary conclusions (Berneburg 2007).

Another possible explanation for the astonishingly homogeneous results of the two studies could be delivered by Söderman (2005). In his study on VR-applications in product evaluations with direct involvement of potential customers he states that "the degree of realism of a product representation is subordinate to the product knowledge required that the participants are familiar (i.e. high product knowledge) with the product or the type of product." This means that stimulus presentation formats with a low degree of realism might still be very efficient if the respondents already have high product knowledge: information which was not delivered by the stimulus could be substituted by the respondent's prior knowledge. Arguments in the same direction come from Schoormans et al. (1995) or Alba and Hutchinson (1987). Adopting this idea to the study at hand means that if products are well known to the respondents, existing differences in the information content of the stimulus could be nullified. This study operated with test objects that were very well known to the respondents (shampoo as a well-known product from the consumers' everyday life). This fact was additionally strengthened by the relevant set selection process: test persons mainly evaluated products on which they had great expertise. Therefore it possibly made no difference if the 3D stimulus presentation delivered more information as the respondents simply "added" the missing information in the 2D case from their memory.

Leaving the aggregate level and looking at the sub-samples, one could assume some influencing effects in the 3D-case: Looking at the convergent and the external validity, there are hints that the test persons' age has some effect on the test results in the 3D-treatment. The results showed that the elder a respondent, the more the market share estimations seem to diverge from the benchmark of real data. Elder test persons potentially are distracted by the new technique and the 3D-impressions. Possible explanations therefore might be manifold:

- ⇨ physical restrictions (physical factors)
- ⇨ stress of the experimental situation (emotional psychic factors)
- ⇨ recessive memory (cognitive psychic factors)

As there are only first hints at these effects, deeper analyses have to follow to come to a well-founded conclusion on 3D-stimuli in a Conjoint Analysis on individual level.

4.3 Theoretical and Managerial Implications

Based on the findings above, it can be hypothesize that if a test person already knows an attribute quite well and ideally is able to form a mental picture without any additional medial aid, the advantages of the visual presentation and especially the dimension of the stimulus are superposed by the product knowledge. But the less the product is known to the test person, the more important a simulation at the highest possible degree of realism will become, in order to form a solid base for the respondent's preference statement. An important question for theory and practice in market research will therefore not only be how much information and degree of realism a test person needs to come to a well-founded preference statement, but in fact how much he already holds, as this information could be a precondition, when trying to decide for the best possible stimulus presentation format and dimension in terms of quality and cost issues.

4.4 Future Research

A next study should concentrate on the comparison of two- and three-dimensional stimulus presentation formats with respect to different attribute categories: On the one hand there should be attributes and levels that are well known to the respondent and on the other hand there should simultaneously be attributes and levels included, that the test persons are unfamiliar with and that need an elaborate inspection before the respondent is able to give a well-founded statement of his preferences. A special attention in this context should be paid to the 3D-effects on a test persons answering behaviour against the background of the test person's age (direct 3D-impact on the test person) and against the background of the test product's price levels (indirect 3D-effect on the test product).

REFERENCES

Agarwal, M. K., & Green, P. E. (1991). Adaptive Conjoint Analysis versus Self Explicated Models: Some Empirical Results. *International Journal of Research in Marketing*, Vol. 8, 141-146.

Alba, J. W., & Hutchinson, J. W. (1987). Dimensions of consumer expertise. *Journal of Consumer Research,* 13(4), 411-454.

Anderson, J. C. (1987). The Effect of Type of Representation on Judgements of New Product Acceptance. *Industrial Marketing & Purchasing,* Vol. 2, 1987, 29-46.

Berneburg, A. B. (2007). Interactive 3D Simulations in Measuring Consumer Preferences: Friend or Foe to Test Results?, *Journal of Interactive Advertising*, Vol. 8 (1) (forthcoming issue)

Boecker, F., & Schweikl, H. (1988). Better preference prediction with individualized sets of relevant attributes. *International Journal of Research in Marketing,* 5(1), 15-24.

de Bont, C. (1992). *Consumer Evaluations of Early Product-Concepts,* Delft: Delft University Press.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix. *Psychological Bulletin,* 56:2, 81-105.

Carson, R. T. et al. (1994). Experimental Analysis of Choice. *Marketing Letters,* 5:4, 351-368.

Diebold, F. X., & Mariano, R. S. (1995), "Comparing Predictive Accuracy", *Journal of Business and Economics Statistics*, Vol. 13, 253-63.

Domzal, T. J., & Unger, L. S. (1985). Judgments of Verbal versus Pictorial Presentations of a Product with Functional and Aesthetic Features. *Advances in Consumer Research,* Vol. 12, 268-272.

Edell, J. A., & Staelin, R. (1983). The information processing of pictures in print advertisements. *Journal of Consumer Research,* Vol. 1. 45-61.

Ernst, O., & Sattler, H. (2000). Validität multimedialer Conjoint-Analysen. Ein empirischer Vergleich alternativer Produktpräsentationsformen [Validity of Multi-Medial Conjoint-Analyses: an empirical Comparison of alternative Forms of Product Presentation]. *Marketing ZFP*, Vol. 22 (2), 161-172 (in German).

Finn, A. (1985). A theory of the consumer evaluation process for new product concepts. In J. N. Sheth (Ed.), *Research in consumer behaviour* (pp. 35-65).

Green, P. E., Helsen, K., & Shandler, B. (1988). Conjoint Internal Validity Under Alternative Profile Presentations. *Journal of Consumer Research,* Vol. 15, 392-397.

Hartmann, A., & Sattler, H. (2002). Commercial Use of Conjoint Analysis in Germany, Austria and Switzerland?. Working Paper, University of Hamburg.

Holbrook, M. B., & Moore, W. L. (1981). Feature Interactions in Consumer Judgements of Verbal versus Pictorial Presentations. *Journal of Consumer Research,* 8 (June), 103-113.

Huber, J. C., Wittink, D. R., Fiedler J. A., & Miller, R. L. (1993). The Effectiveness of Alternative Preference Elicitation Procedures in Predicting Choice. *Journal of Marketing Research,* Vol. 3, 105-114.

Huisman, D. (1997). Creating End User Value with Multi-Media Interviewing Systems. *Proceedings of the Sawtooth Software Conference,* Seattle, 49-55.

Johnson, E. J., Meyer, R. J., & Ghose, S. (1986). When Choice Models Fail: Compensatory Representations in Efficient Sets. Working Paper, Carnegie-Mellon University, Pittsburgh, PA 15213.

Loosschilder, G. H., Rosbergen, E., & Wittink, D. R. (1995). Pictorial Stimuli in Conjoint Analysis - to Support Product Styling Decisions. *Journal of the Market Research Society,* Vol. 37, 17-34.

Louviere, J. J., Schroeder, H., Louviere, C. H., & Woodworth, G. C. (1987). Do the Parameters of Choice Models Depend on Differences in Stimulus Presentation: Visual versus Verbal Presentation?. *Advances in Consumer Research,* Vol. 14, 79-82.

Louviere, J. J., & Woodworth, G. G. (1983). Design and Analysis of Simulated Consumer Choice or Allocation Experiments: An Approach Based on Aggregate Data. *Journal of Marketing Research,* Vol. 2, 350-367.

MacKay, D. B., Ellis, M., & Zinnes, J. L. (1986). Graphic and Verbal Presentation of Stimuli: A Probabilistic MDS Analysis. *Advances in Consumer Research,* Vol. 13, 529-533.

Moore, W. L., & Holbrook, M. B. (1982). On the Predictive Validity of Joint-Space Models in Consumer Evaluations of New Concepts. *Journal of Consumer Research,* Vol. 9, 206-210.

Natter, M., & Feurstein, M. (2002). Real world performance of choice-based conjoint models. *European Journal of Operational Research,* 137, 448-458

Neslin, S., Allenby, G., Ehrenberg, A., Hoch, S., Laurent, G., Leone, R., Little, J., Lodish, L., Shoemaker, R., & Wittink, D. R. (1994). A Research Agenda for Making Scanner Data More Useful to Managers. *Marketing Letters*, 5:4, 395-412.

Sattler, H. (1994). Die Validität von Produkttests: Ein empirischer Vergleich zwischen hypothetischer und realer Produktpräsentation [Validity of Product Tests: an empirical Comparison of hypothetical and real Forms of Product Presentation]. *Marketing ZFP,* 16(1), 31-41 (in German).

Sawtooth Software, Inc. (2006). 2006 Customer Survey: CBC Most Used. *Sawtooth Solutions*, Spring 2006, 1-2.

Schoormans, J. P. L., Orrt, R. J., & de Bont, C. J. P. M. (1995). Enhancing concept test validity by using expert consumers. *Journal of Product Innovation Management,* Vol. 12, 153-162.

Smead, R. J., Wilcox, J. B., & Wilkes, R. E. (1981). How Valid are Product Descriptions and Protocols in Choice Experiments. *Journal of Consumer Research,* Vol. 8, 37-42.

Söderman, M. (2005). Virtual reality in product evaluations with potential customers: An exploratory study comparing virtual reality with conventional product representations. *Journal of Engineering Design,* Vol. 16, No. 3, 311-328.

Unnava, H. R., & Burnkrant, R. E. (1991). An imagery-processing view of the role of pictures in print advertising. *Journal of Marketing Research,* Vol. 28, 226-231.

Vriens, M. (1995), *Conjoint-Analysis in Marketing – Developments in Stimulus Representation and Segmentation Methods,* Capelle a/d Ijsssel: Labyrint Publication.

Vriens, M., Loosschilder, G. H., Rosbergen, E., & Wittink, D. R. (1998). Verbal versus Realistic Pictorial Representations in Conjoint-Analysis with Design Attributes. *Journal of Product Innovation Management,* Vol. 15, 455-467.

Weisenfeld, U. (1989). *Die Einflüsse von Verfahrensvariationen und der Art des Kaufentscheidungsprozesses auf die Reliabilität der Conjoint-Analyse* [Influences of variations in process and form of buying decision on the reliability of a Conjoint-Analysis], Berlin: Duncker & Humblot (in German).

Wildner, R. (1998). The Introduction of the Euro - the Importance of Understanding Consumer Reactions. *marketing and research today*, 11/1998, 141-147.

Wildner, R. (2003). Marktforschung für den Preis [Market research for the price]. *Jahrbuch der Absatz- und Verbrauchsforschung*, Bd. 49, 1, 4-26 (in German).

Wittink, D. R., & Cattin, P. (1989). Commercial Use of Conjoint Analysis: an Update. *Journal of Marketing,* Vol. 53, 91-96.

Wittink, D. R., Vriens, M., & Burhenne, W. (1994). Commercial Use of Conjoint Analysis in Europe: Results and Critical Reflections. *International Journal of Research in Marketing,* Vol. 11, (1), 41-52.

TECHNICAL APPENDIX

To give a detailed explanation of the new technique, the following important components will be introduced:

The Fraunhofer HHI **3D Kiosk system** offers simple intuitive handling as it presents any kind of object in a virtual manner on a large screen in photorealistic 3D quality (1600 by 1200 pixels; 21.3" screen). 3D-objects seem to float in front of the display for which no further technical additive like 3D-glasses or -helmet is necessary.

An **eye tracking system** enables the permanent detection of the test persons´ position so that the three-dimensional projection can be perpetuated even with the test person moving in front of the screen. No glasses or helmets are needed to achieve the three-dimensional impression as the monitor projects a single picture in each eye separately to produce a stereovision effect.

A camera based **hand tracker** detects hand gestures. It recognizes the position of the finger tip, which can be used for pointing at or for moving virtual objects represented with the stereoscopic display. The test person in front of the screen is able to touch, pick and turn the three-dimensional objects as if they were real without any further support of a stylus, glove or the like (virtual 3D touch screen).