Computer Simulated Shopping Experiments for Analyzing Dynamic Purchasing Patterns: Validation and Guidelines

Katia Campo[1], Els Gijsbrechts[2], and Fabienne Guerra[3]

[1] Senior Manager Modeling & Analytics Accuris Group Belgium, and UFSIA, Plantin & Moretuslei 220 b10, 2000 Antwerp, Belgium, Email: Katia.campo@accuris.com
[2] Professor, UFSIA, Prinsstraat 13, 200 Antwerp, and CREER, FUCAM, Chaussée de Binche 171, 7000 Mons, Belgium, Email: Els.gijsbrechts@ufsia.ac.be
[3] Professor, CREER, FUCAM, Chaussée de Binche 171, 7000 Mons, Belgium, Email:Guerra@message.fucam.ac.be

Acknowledgments

This research was possible thanks to a research grant made available by the Belgian Science Policy Department to the research center CREER (FUCAM) (PAI n° 03-53), and a special research grant from the University of Antwerp, UFSIA. The first author, Katia Campo, whishes to thank the Fund for Scientific Research of Flanders (Belgium) for allowing to participate in this research as Postdoctoral Fellow of the Fund for Scientific Research.
The authors thank Dr. Byron Sharp and two anonymous referees for their useful suggestions and comments.

Abstract

Computer simulated shopping experiments open up new opportunities for marketing research, allowing researchers to collect purchase data in a tightly controlled yet realistic environment, at relatively low cost and with a high degree of flexibility. While a number of authors have used computer simulations to generate purchase data, research on the validity of these outcomes is scarce. The present paper aims to improve insights into the validity of computer simulation experiments, by providing a replication and extension of the work by Burke et al. (1992). Experimental purchase data are generated and compared with real life scanner data to provide additional evidence for major biases examined by Burke et al., and to assess the effect of repetitive choice tasks and incomplete product assortments on the realism of purchase quantity and choice data. The results show that, overall, the experimental data correspond quite closely to actual purchase data. Consumers exhibit realistic purchase variation, and provide meaningful data throughout the experimental periods. Yet, some biases are found in category purchase quantities - which are overestimated - and in purchase shares of generic products - which are underestimated. While promotion quantity effects appear to be slightly overestimated, more research is needed on this issue.

Introduction

Computer simulated experiments have gained increasing attention from marketing researchers interested in studying dynamic buying behavior. Recent years have witnessed the development of software programs enabling consumers to engage in 'virtual shopping', that is, to repeatedly purchase items from a computer screen reproducing store shelves (the 'virtual store') under highly controlled conditions. Besides simulating the consumers' shopping behavior, these software packages typically allow researchers to collect additional information from consumers through computerized questionnaires. Computer simulated shopping programs have been developed in various degrees of sophistication, and are continuously evolving (Burke et al. 1992, Burke 1995, Burke 1996, Cohen and Gadd 1996). According to Burke (1996), they offer a viable alternative to other marketing research tools currently employed:

"...most marketing research techniques are, by and large, sadly outmoded. The tools most marketers employ are too expensive, vulnerable to observation and manipulation by competitors, contrived and unrealistic, or simply incapable of providing the information managers really need....Virtual shopping simulation can help managers make tactical decisions in areas such as new products and promotions, packaging and merchandising. Ultimately, the tool promises to change the way companies innovate and how they approach a variety of strategic issues that range from entering new markets to responding to a competitor's attack".

Computer simulated shopping experiments could offer 'the best of two worlds' for modeling consumer purchases over time. As far as other laboratory tests are concerned, Computer Simulated Shopping Experiments (CSSE) offer more realism (compared to rudimentary paper and pencil tests) or more flexibility at a much lower cost (compared to simulated test markets in laboratory stores). In comparison with real life test markets or even controlled field experiments, CSSE offer more control over extraneous variables and more flexibility in manipulating marketing variables and treatments, without exposure to competitive reactions or adverse consumer or distributor effects, and this at higher speed and lower cost (Burke 1996, Cohen and Gadd 1996, Brucks 1988).

Besides these advantages, CSSE have some inherent limitations. Being an articicial environment, CSSE cannot be expected to provide a completely realistic picture of marketplace phenomena. The artificial nature of the setting entails potential biases that should be carefully examined, not in order to remove them altogether, but to get a better grasp on their magnitude, implications, and underlying reasons. These insights serve two crucial purposes. First, by pinpointing areas where caution is needed, they prevent researchers and managers from making unjustified generalizations. Second, they foster evolvement in the development of CSSEs by indicating experimental features that may diminish certain biases.

Previous research on the issue is relatively scarce. An interesting study by Burke et al. in 1992, provides a first evaluation of the validity of CSSE for analyzing dynamic purchasing behavior. In a small scale study, the authors examine potential biases in CSSE results at different stages of the buying process, and hypothesize links with experimental features. A recent article by Cohen and Gadd (1996) provides further insights into characteristics that enhance the realism and the attractiveness of virtual shopping programs to respondents. The present paper intends to shed further light on the issues of CSSE validation and design. It builds upon the work by Burke et al., and examines some additional validation issues.

Previous Research and Purpose of the Paper

While a number of authors have used computer simulations to generate dynamic purchasing data, **research on the validity** of these outcomes is scarce. Burke et al. (1992) report on a first systematic evaluation of the quality and realism of computer simulated multi-period buying data. In their study, data obtained from a sample of 16 respondents participating in a rudimentary and a more realistic laboratory shopping experiment are compared to the respondents' actual purchases under identical marketing conditions. Burke et al. (1992) find that their more sophisticated computer experiment provides significantly better results than the rudimentary test. In a recent HBR article, Burke (1996) briefly refers to an additional validation study of dynamic virtual shopping data. In this study, 300 consumers were invited to make 6 shopping trips in a simulated computer store. These purchase data were compared to UPC scanner information obtained from a concurrently running conventional test market. Simple comparisons of consumers' reactions to package size and price manipulations in the virtual store with conventional test market data demonstrated consumer responses to be similar in the virtual and actual store environment. Also, simulated market shares for various brands of the tested product categories (cleaning products and health and beauty aids) closely matched scanner market shares. The author concludes that "the performance of the virtual store was particularly impressive given the number of ways in which the real and computer generated stores differed", yet emphasizes the need for "future benchmarking of virtual shopping results against existing methodologies".

Clearly, despite these appealing overall results, one should remain aware of the fact that CSSE are subject to external validity threats resulting from artificial experimental conditions. Artificial conditions may refer to the experimental task, stimuli, settings, manipulations or measurements (Lynch 1982). To the extent that unrealistic features in one of these areas interact with the treatment variables (e.g. recorded choice, purchase quantity or promotion responses), the generalizability of the experimental results may be threatened.

Table 1 provides an overview of potentially unrealistic CSSE features, and indicates how they may reduce external validity. First, the *task* of making fictitious purchases in a virtual store may be experienced as unrealistic for several reasons. The purchases bear no ‘real’ consequences and require less effort (respondents do not have to wander through the store to search for and/or compare alternatives placed at distant positions on the shelf). Respondents may therefore be willing to engage in more risky purchases (e.g. trying an unfamiliar brand), and /or in more active evaluation processes. As indicated by Burke et al. (1992), the extent to which the fictitious nature of the task affects purchase quantity and choice decisions depends on the type of products and buying behavior. For frequently purchased consumer goods, consumers tend to develop simple choice heuristics - based on cues such as brand, price and promotions - that can be easily reproduced in computer simulations (Burke et al. 1992, p.72).

Table 1: External validity: Realism of virtual shopping

UNREALISTIC ASPECTS OF CSSE	POTENTIAL VALIDITY THREATS
Experimental task	
Fictitious purchases	Exploratory choice behavior, higher PQ*
Absence of physical displacement	More extensive alternative evaluation
Pre-imposed shopping frequency	Bias in PQ/shopping occasion
Reduced number of product categories	More extensive alternative evaluation
Reduced interpurchase time	Stronger purchase event feedback
Absence of consumption experiences	
Stimuli	
In-store marketing stimuli	
- Products: presentation and assort.	
- Price	Bias in choice behavior
- Promotions	Bias in price sensitivity
Other factors of purchase decisions	Increased promotion sensitivity
- Time constraint	
- Budget constraint	More extensive alternative evaluation
- Space constraint	Higher PQ, reduced price sensitivity
- Inventory	Higher PQ
-	Bias in PQ
Settings	
Physical store environment	Rational, socially acceptable behavior
Store atmosphere	Purchases less influenced by others
People: shopping company	
Manipulations	
Promotion frequency and conditions	Bias in promotion effects
Other in-store stimuli	Bias in effect of other stimuli
Measures	
Choice	Less private label and generics purchases
Purchase quantity and timing	Higher purchase frequency and/or PQ

* PQ = purchase quantity

Another task-related issue is that in dynamic CSSE, data on subsequent shopping occasions are often collected in the course of a single experimental session. Such concentrated data collection is typically needed for practical reasons, or to reduce internal validity threats such as history, maturation and mortality effects. Yet, this 'time compression' may further reduce the realism of the purchasing task, and – following Burke et al. (1992) - strengthen purchase event feedback. Another reason why CSSE purchase simulation tasks may be experienced as unrealistic, is that some shopping decisions have to be 'pre-imposed' (e.g. a fixed shopping frequency, fixed and severely limited number of categories to make purchase decisions on). In cases where pre-imposed decision aspects deviate from the respondents' habitual buying behavior and affect other (measured) purchase decisions, the external validity of the results may be threatened (e.g. choice of one of two complementary products that are usually selected in combination).

Second, as demonstrated by Burke et al.'s (1992) comparison of the rudimentary and realistic laboratory results, the realism of in-store marketing *stimuli* and the way they are presented to the consumer (e.g. real versus fictitious brands, verbal versus visual announcements of promotions) may be crucial to the external validity of CSSE purchase data. Similar observations were made by Burke (1996) and Cohen and Gadd (1996), who emphasize the importance of store familiarity and availability of sufficient product and promotion information. Likewise, the extent to which real-life constraints and other factors influencing actual buying decisions can be reproduced in a realistic way, may affect biases in quantity and choice behavior. For instance, Burke et al. (1992) attribute purchase quantity biases to the absence of realistic time, budget and space constraints.

Third, the *settings* in which CSSE data are collected may have an impact on simulated purchase behavior. Laboratory environments may be unrealistic in terms of physical features (e.g. a 'sterile' environment may reinforce the impression of being observed), atmosphere (e.g. lack of a noisy and stimulating store atmosphere), or company (e.g. absence of family or friends who usually accompany the respondent on his/her shopping trips). From their pilot study of a 3-D modeled virtual store, Cohen and Gadd (1996) conclude, for example, that the stylized nature of their virtual store environment affects consumers' purchase behavior because the degree of sensory stimulation is much lower than that in a real store.

Fourth, *manipulations* of treatment variables should be in line with real life conditions in order to obtain realistic estimates of their effect. For instance, research on sales promotions effects has shown that the strength of the promotion response depends on the intrinsic attractiveness of the promotion advantage, but also on the way it is announced and on the frequency with which promotions have been offered in the past (see e.g. Blattberg and Neslin 1990). Hence, a lower/higher than usual promotion frequency could lead to an over-/underestimation of promotion effects.

Finally, *measures* and measurement procedures may influence experimental results. Knowing that their buying behavior is observed and recorded may encourage respondents to exhibit more rational or socially acceptable behavior, such as buying higher priced alternatives rather than cheap low quality brands (see below). It may also create a tendency to comply with the interviewer's expectations or objectives, e.g., the respondent may feel forced to purchase more or more frequently to accommodate the interviewer.

The **research objective** of this paper is twofold. *First*, we intend to gather additional evidence on the type and magnitude of biases in CSSE outcomes. The study of Burke et al. (1992) is a pilot study, comprising an in depth analysis of a small number of households, for which actual and simulated purchases are 'matched'. This paper opts for a between-subjects design (Carson et al. 1996), as suggested by Burke (1996). More specifically, we collect computer simulated shopping data from a large sample of households, and confront the results with information from an equally large scanner panel data set. The magnitude of our sample base allows us to shed further light on the external validity of CSSE outcomes, including method-specific household selection biases. *Second*, our study also offers some additional insights, by examining the impact of design variables not explicitly studied so far. To assess the impact of simulation length and the completeness of the on-screen assortment, we investigate the consistency in purchase behavior over subsequent simulation periods, as well as the choice behavior of consumers who do not find their preferred items on screen.

Research Hypotheses

As stated earlier, a *first* objective of our research is to shed more light on the *degree of realism of consumer purchase information obtained in CSSE*. Burke et al.’s (1992) validation study indicated that purchase quantities, choice shares, and choice dynamics more closely reflect actual buying behavior in realistic than in rudimentary computer experiments, but can still be biased. Building on Burke et al.’s theoretical reasoning and experimental results, hypotheses 1 to 4 specify expected biases in these 3 purchase behavior characteristics. All hypotheses focus on buying behavior for frequently purchased, non-impulse products, that are typically chosen on the basis of cues (Burke et al. 1992). Hypotheses 1 to 3 investigate the external validity of experimental shopping trips disregarding promotional effects, that is, when promotions are either absent or separately accounted for (by removing their effect on the test variables). In turn, hypotheses are formulated on the purchase quantity, brand choice, and purchase dynamics (degree of purchase variation) within a product category. Hypothesis 4 relates to consumers' differential reactions to in-store promotions in a CSSE setting, where the effect of promotions on purchase quantity as well as brand choice is considered.

Our *second* objective is to examine the *impact of two additional experimental features*. Hypotheses 5 and 6 concentrate on how the number of shopping trips, and the 'completeness' of the CSSE assortment, affect consumers' quantity and choice decisions. Below, we briefly present the hypotheses and associated comments.

Hypothesis 1*: *In the presence of inventory cues, the number of units purchased by consumers in a CSSE is in line with actual purchases.

Comments: Previous research suggests that laboratory shopping may lead to above normal purchase quantities, especially in the presence of promotions. Following Malhotra (1996) consumers “…*tend to increase the behavior being measured, such as food purchasing*”. As demonstrated by Burke et al.’s (1992) research findings, the bias in purchase quantity decisions can be substantially reduced by making the choice context more realistic. Although promotional purchases are still somewhat overestimated in their realistic experiment, Burke et al. argue that the bias could be further reduced by imposing additional constraints that restrict purchase quantities in real life, such as time, space and budget constraints. In line with these guidelines, Cohen and Gadd (1996) imposed a fictitious budget constraint based on the household's 'normal spending' in the product category, to reduce quantity biases. Feedback from respondents revealed, though, that the product-specific budget constraint was experienced as too restrictive and highly unrealistic. In other words, buyers do not appear to restrict purchase outlay for specific product categories. In line with the buying behavior modeling literature, we believe that purchase quantity decisions of frequently purchased, low cost products are rather restricted by handling and holding cost considerations ('storage constraints'). Empirical results confirm that consumers typically rely on inventories available at home to decide upon the purchase quantity and timing of frequently purchased consumer goods (see e.g. Sivakumar and Raj 1997; and Kumar, Karande and Reinartz 1998). Inventory cues may therefore act as 'realistic' purchase quantity constraint mechanisms.

Hypothesis 2: *In the context of CSSE, consumers tend to buy more national brands, and less private labels and generics, than in reality.*

Comments: This hypothesis is in line with previous research on the presence of response bias in interviews, and the social desirability to purchase national brands, rather than private labels or generics (Burke et al. 1992). In an experimental setting, consumers want to 'look good', or 'give the right answer' (Malhotra 1996). To the extent that distributor brands are perceived to be cheaper and of lower quality, respondents may be reluctant to admit buying these brands when they know they are being observed. Over-reporting of National Brand purchases - as opposed to private label purchases - is a well known problem in survey research, and has already been referred to by Wind and Lerner (1979) and Sudman (1964b).

Hypothesis 3: *In the context of CSSE, consumers seek less variation in choice behavior than in reality.*

Comments: Burke et al. (1992) hypothesize that the time compression factor increases the impact of past purchases on current purchases. They also argue that satiation is less likely to occur since respondents do not actually consume products (see e.g. McAlister and Pessemier 1982). Their conclusion is that time compression decreases switching, and reduces the number of different brands bought in the category, which was confirmed by the empirical results.

Hypothesis 4: *Consumers react more strongly to sales promotion offers in CSSE than in reality.*

Comments: Previous research suggests that consumers react more strongly to promotions in a laboratory setting because of higher promotion visibility (Gabor et al. 1970, Nevin 1974, Burke et al. 1992, Nagle and Holden 1995). Compared to the actual, large supermarket shelves, products and promotions are conveniently arranged on a small computer screen, placed in front of the respondent. In the same way as products at eye level attract more attention than products placed lower or higher on the shelf, CSSE promotions may be more clearly visible and may more easily attract respondents' attention than actual promotions (Corstjens and Corstjens 1995). CSSE promotions are therefore expected to exert a stronger influence on consumers' buying decisions than in the actual store environment. Burke et al.'s study pointed to overly strong reactions to promotions in the rudimentary as well as the realistic experiment

Hypothesis 5: *Consumers do not systematically change their purchase behavior (quantity and choice decisions) over subsequent simulation periods of a CSSE.*

Comments: A crucial question in analyzing dynamic behavior in a laboratory setting concerns the number of successive shopping trips that can be presented to the consumer. Using CSSE to study dynamic buying behavior is conditional upon consumers being able and willing to exhibit realistic buying patterns over subsequent periods. In a typical interview setting, potential threats to validity of information collected over time are testing and maturation effects (Cook and Campbell 1979, Malhotra 1996). Testing effects may occur if respondents first have to become accustomed with the interviewing procedure (in the case of CSSE, the purchase simulation procedure). In such cases, information on early purchases is unusable. This type of initialization effect can be circumvented, though, by providing clear instructions and familiarizing respondents with the procedure beforehand (Cohen and Gad 1996). Maturation effects, leading to deterioration of information quality towards the end of the interview, can be a result of boredom, fatigue or a lack of willingness to spend more time (Malhotra 1996, Carson et al. 1996). Like for other interview types, the risk of maturation effects in CSSE can be substantially reduced by keeping interview length – and thus the number of successive purchase occasions – within reasonable limits.

Hypothesis 6: *The absence of preferred items on the computer screen does not affect the quantity purchased in the product category, but leads to a choice bias in favor of major national brands.*

Comments: Hypotheses 1 to 5 are expected to hold for consumers who find their regularly purchased products on the computer screen, and are therefore able to realistically reproduce their real life purchasing patterns (Cohen and Gadd 1996). In many CSSE, it will be practically infeasible to include all the items in the category in the virtual store. Since consumers do not have the possibility to shop elsewhere in a CSSE, we do not expect assortment reductions to affect purchase quantity. In terms of choice, the consumer behavior literature indicates that buyers who are uncertain about their preferences for the (less known) available alternatives can be expected to make a choice on the basis of cues such as brand name and price (see e.g. Kapferer 1994, Kahn and McAlister 1997). In line with this research and with Hypothesis 2, we expect consumers who cannot find their regular product on the shelf to switch to a 'well known' and 'socially acceptable' alternative rather than to a (lower quality) distributor brand.

Description of the CSSE experiment

The CSSE used here is comparable, in degree of sophistication, to the 'realistic' experiment described in Burke et al (1992). The discussion that follows describes the major features of the

CSSE experiment and the actions taken to enhance the realism and external validity of the results.

Experimental task and stimuli

The experiment involved two product categories: jam – for which hypotheses will be tested – and a second category (paper towel or margarine depending on use rates) to make purchase simulations more realistic and reduce time compression effects (successive purchases in one category only, could strongly increase purchase feedback effects). The product categories selected contain a 'reasonable' number of items in their assortment, are regularly promoted, and are frequently purchased by a majority of consumers. Given the high purchase frequencies, consumers can be expected to be fairly familiar with items in the categories. Items in the product classes are typically chosen on the basis of cues that can be visualized, such as brand, pack, price, promotion, and product information.

Using scanned pictures of real products, actual shelf layouts were imitated on the screen. To make the task as realistic as possible, the actual choice context (shelf layout, prices) of a large supermarket chain was reproduced in the computer simulation. Of the actual store assortment, a limited assortment of 22 items was selected for the jam category, including the most popular product variants (tastes) and major national brands, in addition to distributor brands and generics. Product prices were displayed on a price bar below the shelf, but remained at the same level throughout the purchase simulation. Respondents could zoom in on an item, retrieve product information printed on the (actual) product package, or select one or several items for purchase from the fictitious shelf. Each respondent made purchase decisions over twelve successive weeks. In the task instructions preceding the purchase simulation, respondents were asked to imitate their real shopping behavior as accurately as possible. For instance, they were asked to mimic their true purchase frequency (they did not have to make a purchase every week) and choice pattern (they did not always have to buy the same item when they frequently switched among different items).

Additional cues displayed at the bottom of the screen were a time indicator (simulation week) and the household inventory level. The latter was updated weekly using information on the household's reported average consumption rate, and previous purchase decisions. As indicated above, information on the household inventory should make purchase quantity decisions more realistic. In addition, the fact that interviews were conducted in store (see below) implies that some time constraint is active. Indeed, if customers are intercepted in the course of a 'normal' shopping trip and are asked to perform their 'experimental' purchases on that occasion, these experimental purchases will take place under 'similar' time pressure as their normal shopping activities. More realistic time constraints may also help to reduce purchase quantity biases (Burke et al. 1992).

Settings

In the actual data collection stage, shoppers were intercepted and interviewed in five different outlets of the supermarket chain that had served as a model for the shelves presented in the CSSE. This guaranteed consumer familiarity with the virtual store shelves, and ensured that consumers found themselves in a noisy, stimulating, natural environment. As indicated earlier, previous research revealed that these interview conditions are particularly important in creating a realistic shopping atmosphere (Cohen and Gadd, 1996).

Manipulations

In the course of their shopping trips, consumers were confronted with in-store promotions of different types (direct price cuts, coupons, premiums, and quantity discounts). Promotions were offered for two test items in each category, and were randomly assigned to periods and consumers. No promotions occurred in the first two periods, and the promotion frequency was in line with the normal occurrence of in-store promotions in the category. Promotions were visualized by means of realistic on-pack labels in combination with colored shelf tags.

Measures

Information on the household's consumption rate and other characteristics – such as variety seeking tendency and socio-demographics – was collected by means of a computer questionnaire preceding the purchase simulations. At the same time, the questionnaire allowed respondents to become familiar with the virtual store's shelf layout and assortment, thereby reducing the risk of initialization effects (the CSSE shelf was displayed several times, for example, when consumers were asked to indicate their average use rate in the category). During the purchase simulations, item choice(s), purchase incidence and quantity were recorded. Unless the respondent explicitly requested help, the questionnaire was self-administered, so as to minimize interviewer bias. Yet, the continuous presence of two interviewers (see Methodology section), and the fact that respondents knew their purchase decisions were recorded, implies that measurement biases may not have been completely eliminated.

Methodology

Store managers of the interview stores agreed to make space available near the checkout counters to install tables with PC's. This was possible as interviews were conducted in spacious hypermarkets, in which wide aisles were available for special events or display activities. Respondents were intercepted either before or after they completed their shopping activities. Those willing to co-operate were lead to one of the PC's and given clear instructions on how to complete the interviews in a self-administered fashion. Two interviewers remained permanently available to supervise respondent activities and - if necessary - provide assistance. A special video-corner was installed near the tables to entertain the children accompanying respondents during the interview. In total, 506 respondents were interviewed, yielding some 3700 purchase occasions (i.e., simulation weeks in which at least one item was bought).

To validate our hypotheses, we confronted the CSSE outcomes with results obtained with actual scanner panel data. Note that these data are widely used in practice to establish purchase levels, market shares and promotional reactions. Data were available for 2 outlets of the considered retail chain, over 16 subsequent weeks, for all store customers possessing a chain loyalty card. In addition, information was collected on various types of in-store promotions during the period for which scanner panel data were available. In comparing the experimental with the scanner panel data, we corrected as much as possible for 'systematic' sample or 'selection' differences, and for differences in in-store conditions.

Matching Samples and Store Conditions

Inhabitants in the trading area of the outlets from which scanner panel data were available, have a socio-demographic profile that closely matches that of the outlets where CSSE interviews were conducted. In the scanner panel data set, we retain consumers who purchase at least one pack in the category per month: this minimum use rate is similar to the one used for screening CSSE respondents. For the jam category, this leaves us with information on 486 households in the experimental data set, and 283 households from the scanner panel data. While the scanner panel sample is based on consumers holding a chain loyalty card, the experimental sample is drawn from all consumers in the store. The implication of this difference in sampling frame is expected to be limited, as loyalty card penetration is considerable (chain surveys reveal that about 95% of store visitors are card holders), and does not imply exclusive patronage of that chain (many shoppers have loyalty cards for different chains). Yet, the risk that card holdership induces a bias in behavior for the scanner panel sample cannot be completely ruled out. Other potential differences between the experiment and real life setting concern the frequency with which consumers are confronted with the shelves (weekly visits in the experiment), and the width of the assortment presented (experiment comprises only subset of items). Information on scanner panel members' store visit frequency, and on whether CSSE consumers find their regularly purchased products on the computer shelves, was collected in the initial questionnaire. This information will be helpful in detecting potential biases in behavior arising from experimental store visit and assortment conditions (for instance, by examining differences in buying behavior between (i) consumers who visit the store on a weekly basis and those who have substantially lower or higher shopping frequencies, and (ii) between consumers who did or did not find their favorite product on the virtual store shelf).

As outlined above, the artificial shelves presented on screen comprise only a subset of the items sold in the chain's outlets (22 out of 57 items for jam). The shelf layout is a reproduction of real store planograms for the category. While regular price levels in our experiment match those in the scanner panel data set, promotional conditions are somewhat different between the two data sets (see Hypothesis Tests section).

Purchase Quantity and Choice Models

As indicated above, Hypotheses 1 to 3 investigate the validity of CSSE purchase decisions in the absence of promotions. Promotion effects on purchase quantity and choice decisions should therefore be removed in order to test these hypotheses. With this purpose, we estimate for each data base: (i) a purchase quantity model, linking the number of units bought by a consumer in a given category and week, to household and marketing variables, and (ii) a stochastic choice model, explaining the choice probability of any item in the assortment, given a category purchase, as a function of item characteristics, previous choice decisions, and promotion variables. Purchase quantity probabilities per consumer are modeled as a Poisson-based process (see e.g. Bucklin and Gupta 1992, Dillon and Gupta 1996), household choice probabilities as a Multinomial Logit model (see e.g. Guadagni and Little 1983, Gupta 1988, Bucklin and Gupta 1992). Model specifications can be found in Appendix 1 and Appendix 2 respectively.

Using the parameter estimates, purchase quantity and choice probabilities under a no-promotion condition can be computed by setting all promotion variables equal to zero (see Hypotheses Tests section). Comparison of these values for the scanner panel and CSSE data sets allows to test Hypotheses 1 to 3. In addition, estimated promotion parameters can provide insight in to the realism of CSSE promotion effects on purchase quantity and choice decisions (Hypothesis 4).

Tests on the stability of purchase quantity and choice behavior in the CSSE (Hypothesis 5) are carried out by introducing weekly parameters into the choice and quantity models for the experimental data set (i.e., for each simulation week, a dummy variable is incorporated into the model specification; see Appendix 1 and Appendix 2). When respondents systematically change their buying behavior over subsequent simulation periods, at least some of the week parameters should be significantly different from zero. Positive and significant parameters for the last weeks in the purchase quantity model would, for instance, indicate a tendency to increase purchase quantity towards the end of the purchase simulation.

Hypothesis 6, finally, is tested by comparing weekly purchase quantities and choice shares of national and store brands between (1) the subsample of respondents whose favorite item is available in the CSSE assortment, and (2) the subsample of respondents who did not find their most preferred item on the virtual store shelf. Since not all consumers face the same CSSE promotions, comparisons are carried out for computed quantities and choice shares under the no-promotion condition.

Estimation Results

The purchase quantity model is estimated using maximum likelihood procedures. Estimation for the experimental data set is based on 11 weeks, week 1 being an initialization week for the lagged promotion effects. For the scanner panel data, weeks 1 to 4 serve as initialization weeks over which the household's use rate in the category is calculated, and model estimation is based on purchases from weeks 5 to 16. Both models provide a satisfactory fit, with Mean Absolute Percentage Errors (MAPE) for weekly purchase quantities of .059 in the experimental data set, and of .099 in the scanner data set. The estimation results - which can be obtained from the authors on request - reveal a highly significant positive effect for use rate, and negative impact for inventory level. Estimated promotion coefficients are either insignificant, or have the expected sign: a more detailed discussion of promotion effects is provided when discussing Hypothesis 4.

Choice models are estimated by maximum likelihood procedures, using information from weeks 5 to 12 for the experimental data set, and from weeks 5 to 16 for the scanner data set. Weeks 1 to 4 again serve as initialization weeks, to compute the household's intrinsic preferences (loyalty variables) and variety seeking tendency. Goodness of fit appears to be satisfactory for both data sets, with MAPE in brand shares of .080 and .091 for the experimental and scanning data set, respectively. Estimation results further reveal that most coefficients are significant and have the expected sign. To test whether the MNL model's IIA assumption is justified, a procedure suggested by Fader and Hardie (1996) was applied. Results demonstrate that both data sets comply with the IIA assumption.

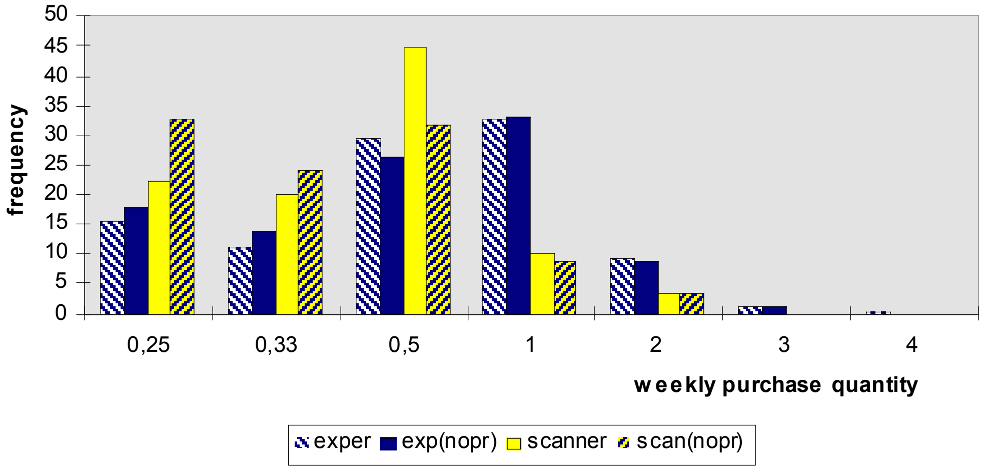
Hypothesis Tests

Hypothesis 1

In the presence of inventory cues, the number of units purchased by consumers in a CSSE is in line with actual purchases.

To check whether experimental purchases match what consumers buy in real life settings, we compare the weekly experimental purchase rates with those obtained from the scanner panel data set. Week 1 is left out of this analysis as it is considered exceptional (starting inventory is zero for all consumers), and since the purchase quantity model is estimated for weeks 2 to 12. Next to observed experimental purchases, which may be affected by promotional activities in the experiment, we calculate expected weekly purchase quantities per consumer in the absence of promotions. These expected quantities are obtained from the purchase quantity model through 'dynamic' simulation, taking into account that consumers who buy less in a given week when the promotion is absent, will end up with a lower inventory level at the beginning of the following week, and hence are more likely to buy larger quantities then. For purposes of comparison, we also include expected quantities from the Poisson model when promotions are present.

Figure 1.
Frequency Distribution of Weekly Purchase Quantities in Experiment and Scanner Data Set



Exper : observed weekly purchase quantities, experiment
Exp(nopr) : simulated weekly purchase quantities, experiment, no promotion
Scanner : observed weekly purchase quantities, scanner data
Scanner(nopr): simulated weekly purchase quantities, scanner data, no promotion

Figure 1 shows a histogram of weekly purchases per consumer as they are (i) observed in the experiment, (ii) simulated in the experiment under the no-promotion condition, (ii) observed from the scanner panel, and (iv) simulated for the scanner panel in the absence of promotions. Note that in the scanner panel data, eliminating the impact of promotions leads to a more pronounced shift towards lower quantities[1] than for the experimental data. This follows from the frequency as well as from the nature of the promotional actions: the scanner panel data are characterized by somewhat more frequent promotions that, in contrast with those in the experimental setting, may encompass various tastes of the promoted brand simultaneously and are featured in retailer

[1] This shift is especially apparent at the lower end of the graph. A tentative explanation for this phenomenon is that promotional purchase quantity shifts are largely due to purchase acceleration (see e.g. Gupta, 1988), implying short term peaks and post-promotion dips in category purchases. For heavy users, who consume at least one pack per week and are likely to buy the product on a weekly basis, almost all post promotion dips are likely to occur within the sixteen-week period, and to compensate for the immediate sales peaks. For light users (weekly consumption below one pack), post promotion dips will to a considerable extent be observed after the observation period, and not yet intervene in the simulated quantities for the sixteen-week period. The necessary downward correction within the observation period is therefore larger for light users, at the low end of the graph.

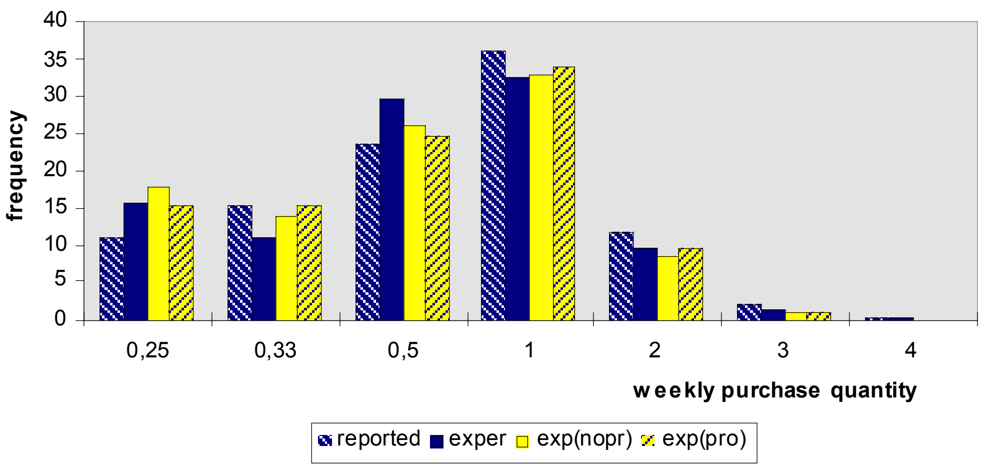
leaflets. From Figure 1, we observe that the experiment has a higher frequency for larger weekly quantities per consumer (1 or 2 units per week) compared to the scanning data set. The proportion of buyers adopting only 1 unit every 3 weeks (use rate .33) or every 4 weeks (use rate .25) is more important in the scanner panel than among CSSE consumers. A non-parametric test on the difference between the quantity distributions indicates that the deviations are significant (Man-Whitney z=-8.765; p=0.00). Table 2, which summarizes means and standard deviations for the various weekly purchasing rates, provides further support for this finding. We therefore have to reject our hypothesis that experimental purchase rates coincide with real life purchase frequencies. Despite the presence of the inventory cue (which acts as an implicit space constraint), we find that experimental purchase quantities systematically exceed scanner purchases.

Table 2.
Means (Standard Deviations) of weekly purchases in the experiment and scanner data set

	Experiment	Scanning
Observed	.82 (.59)	.53 (.35)
Simulated, with promotions	.79 (.55)	.56 (.40)
Simulated, without promotions	.76 (.54)	.48 (.34)

These differences are not likely to result from card holdership bias: ceteris paribus, one would expect shoppers possessing a loyalty card (and from which the scanner sample was drawn) to purchase at least as much in the chain as non-card holders. Further analysis revealed that purchase quantity deviations cannot be attributed to differences in shopping frequency between experimental and scanner panel consumers. An interesting finding, though, is that within the experimental setting, consumers' purchases coincide with self-reported use rates. Figure 2 provides a histogram of consumers' self-reported use rates, and observed as well as simulated experimental purchase quantities. A Wilcoxon Matched-Pairs Signed-Ranks test confirms that reported and observed experimental use rates have the same distribution (the test-value is equal to -.575, p=.565). Having specified a use rate, consumers tend to engage in experimental purchases consistent with this estimate. They are encouraged to do so since their inventory level throughout the experiment is based on the use rate estimate. This finding offers a tentative explanation for why Hypothesis 1 fails. If consumers overestimate their use rate, this will be translated into exaggerated experimental purchase quantities. Earlier studies have indeed shown that in an interview setting, respondents have a tendency to overstate purchases (Sudman 1964, Wind and Lerner 1979). Under these assumptions, a meaningful approach for future CSSE would be to estimate respondents' use rates on the basis of their past observed purchase behavior, and to rely upon this information for experimental inventory updates instead of upon their self-reported use rates.

Figure 2.
Frequency Distribution of Reported and Observed Purchase Quantities (experiment)



Reported : reported purchase quantity per week
Exper : observed purchase quantity per week (experiment)
Exp(nopr) : simulated purchase quantity, no promotion

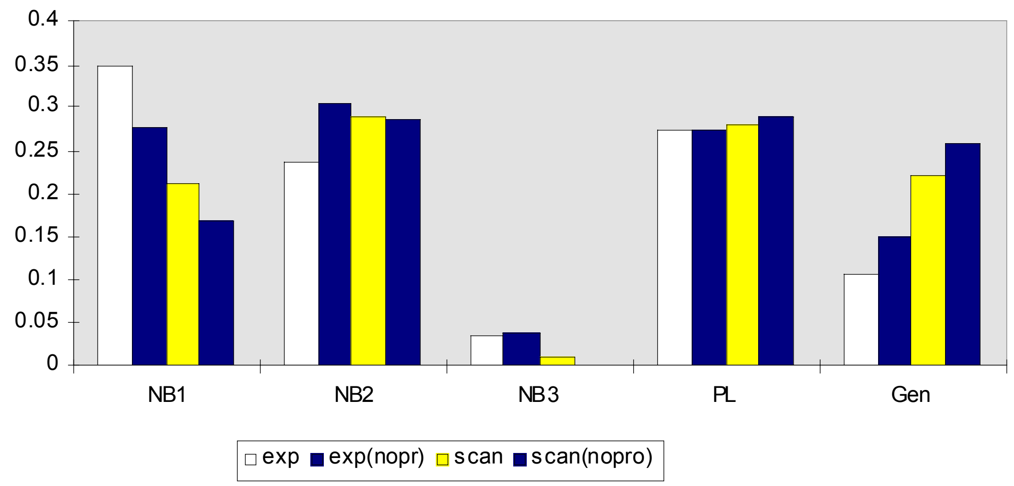
Hypothesis 2

In the context of CSSE, consumers tend to buy more national brands, and less private labels and generics, than in reality.

In comparing market share levels across data sets, we have to account for differences in promotional conditions and assortments: two minor national brands of the scanner panel store's assortment are not included in the experiment, while for other national brands and the private label some taste variants are missing. To make the results comparable, we therefore predict market shares for both data sets in the absence of promotions, and rescale the (predicted) scanner data market shares of items included in the experiment so that they sum to one. Figure 3 provides a summary of these calculations. It reports, for the brands and items included in the experiment: (i) the observed market shares in the experimental data set, (ii) predicted shares in the experiment under the no-promotion condition, (iii) observed scanner data shares of items available in the experiment (rescaled so that they sum to one), and (iv) predicted scanner data shares in the absence of promotions, for the items available in the experiment, rescaled to sum to one. Note that correcting for promotions leads to a marked decline in share for NB1: this brand is regularly promoted in both the experimental and the scanning data set. NB2 is never on promotion in the CSSE, and subject to occasional promotional activities in the scanning data set. This explains why it benefits from promotional corrections in the CSSE data (by capturing some of NB1's share), while those corrections do not make a difference in the scanning data set (own promotional share lost, but compensated by absence of promotions for NB1). A similar finding holds for the private label, which is promoted in both data sets, but much less heavily or effectively so than NB1. The Generic, finally, is never on deal, and benefits from the elimination of promotion effects by capturing part of NB1's share.

Turning to Hypothesis 2, some interesting patterns emerge from Figure 3. For two of the national brands (NB1 and NB3), experimental shares significantly and systematically exceed true shares (based on a t-test of the significance between two estimated proportions). This finding is in line with our hypothesis. For the generic product, observed experimental share is significantly below observed scanning share, and predicted experimental share without promotions is substantially below predicted scanner share. This observation, too, is consistent with our expectations. Yet, the findings for the Private Label do not coincide with the hypothesized pattern. After correcting for promotional differences, Private Label share in the experiment is comparable to the market portion it captures in the scanner panel data set. Finally, experimental share for one National Brand NB2 is in line with the rescaled scanning data share.

Figure 3.
Experimental versus Scanning Market Shares for Items Available in the Experiment



Exp : observed market share, experiment
Exp(nopr) : simulated market share, experiment, no promotion
Scan : observed market share (rescaled), scanner data
Scan(nopr) : simulated market share (rescaled), scanner data, no promotion

We conclude from this discussion that the results provide partial support for Hypothesis 2. Like Burke et al. (1992), we find that choice shares tend to be overestimated for national brands, and that the experimental data set systematically underestimates the true share of generic products. Yet, no such underestimation is found for the private label. A plausible explanation is, that customers of the analyzed supermarket chain have come to perceive private labels as valuable and socially acceptable alternatives, in contrast to generics which are still attributed a negative sign value. As suggested in a recent article by Bronnenberg and Wathieu (1996), the quality gap between national brands and private labels appears to be closing. Dhar and Hoch's (1997) study further demonstrates that the performance of store brands varies widely among retailers and product categories, depending on factors such as the quality and breadth of private label offerings, the retailer's overall pricing strategy, and degree of promotional support for private labels. The realism of private labels' choice shares in CSSE's may therefore depend on these factors and associated buyer perceptions, yielding biased estimates for some retail chains and/or product categories but not for others.

Hypothesis 3

In the context of CSSE, consumers seek less variation in choice behavior than in reality.

To test for differences in purchase variation between experimental and real life settings, we compute the level of entropy in the purchase history of CSSE and scanner panel households. Entropy is a widely accepted measure of purchase variation, and indicates whether a household tends to concentrate purchases in a single item (entropy equal to 0) or to spread purchases over several items (entropy>0; the measure equals 1 at maximum level of variation) (see e.g. Van Trijp and Steenkamp, 1990). Since jam items differ in brand and taste, both attributes are accounted for

Table 3.
Mean (Standard Deviation) of Entropy in the experiment and scanner data set

	Experiment	Scanning
absolute entropy	1.18 (.45)	1.00 (.39)
relative entropy	.57 (23)	.54 (.21)

in our entropy computations.

Table 3 reports the average entropy measure, and its standard deviation, across households in both data sets. As assortment depth and purchase quantities differ across data sets, 'absolute' entropy levels are not really comparable. We therefore calculate household's relative entropy levels (see Appendix 2), the mean and standard deviations of which are also reported in Table 3. Clearly, entropy levels are highly comparable across data sets. A non-parametric test reveals that while absolute entropies differ (Man-Whitney z=-5.14, p=0), there is no significant deviation between relative entropy distributions in both data sets (z=-1.14, p=.254): we therefore reject Hypothesis 3. A possible reason why dynamic buying behavior turns out to be realistic in our experiment, in contrast to Burke et al.'s (1992) findings, is that out interview instructions made variety seeking behavior more 'rationally acceptable' to consumers (see Malhotra 1996). As indicated above, the task to imitate real purchase behavior was illustrated by a number of examples, including the tendency to switch among different items or to stay with the same item(s). Burke et al.'s respondents may have restrained from switching, fearing that such behavior would be considered to be irrational and inconsistent

Hypothesis 4

Consumers react more strongly to sales promotion offers in CSSE than in reality.

Table 4 summarizes the results for various promotion types in the two data sets: it reports the promotion parameters (β_{pro}) and their standard deviations. As the scanner data set did not include coupons or gift promotions, the comparison is limited to price cuts and quantity discounts. Table 4 reports on promotional effects for national brands and private labels separately, which have been found to generate different promotional reactions (see e.g. Blattberg and Neslin 1990, Sethuraman 1995 and 1996). A distinction is also maintained between featured and non-featured price cuts for the scanner data set. In the computer experiment, all promotions are automatically in-store and thus non-featured.

Table 4.

Promotion parameters β_{pro} (and their standard deviations) for various brand types and instruments, in the experiment and scanner data set.

Table 4a: PURCHASE QUANTITY

Promotion instrument	Brand type	Scanning Data	Experimental Data
Featured Price Cut:	National Brand	-	-
	Private Label	.13 (.01)	-
Non-featured Price Cut:	National Brand	.02 (.02)	.16 (.09)
	Private Label	-	.04 (.10)
Quantity Discount:	National Brand	.05 (.02)	.15 (.10)
	Private Label	-	.06 (.10)

Table 4b: CHOICE

Promotion instrument	Brand type	Scanning Data	Experimental Data
Featured Price Cut:	National Brand	-	-
	Private Label	1.74 (.64)	-
Non-featured Price Cut:	National Brand	.48 (.64)	2.77 (.21)
	Private Label	-	2.03 (.26)
Quantity Discount:	National Brand	1.00 (.56)	2.78 (.24)
		1.70 (.99)	
	Private Label	-	1.76 (.27)

[a] Featured promotion in the scanner panel data set, non-featured in the experimental data set.

[b] Bold = significant at the 10% level

Even though the difference in promotion conditions (national brand versus private label, featured or non-featured) complicates the comparison between data sets, some interesting patterns emerge. For the *purchase quantity model*, the exponent of the promotion parameter ($\exp(\beta_{pro})$) can be easily interpreted: it quantifies the ratio of expected weekly purchase quantity with and without a(n) (additional) promotion action. From the purchase quantity model estimates in Table 4a, it seems that comparable promotions in the scanner data set are somewhat less effective than experimental promotions: both the non-featured price cut and the quantity discount for the national brand have a significantly larger impact in the experimental than in the scanning setting (based on a t-test on parameter differences). One reason might be that scanning promotions on an item more often coincide with promotions of other items of the same brand, which reduces their impact. On the other hand, it seems like the on-shelf promotions in the experiment act more like ‘displayed’ promotions, which typically entail more pronounced consumer reactions (see e.g. Blattberg and Neslin 1990, Narasimhan et al. 1996). Though no specific display materials are used to support them, the fact that CSSE consumers are placed in a front view of the (reduced) shelf, instead of walking through the aisle in a noisy store environment, may lift promotions up to a ‘display’ status. Comparing the relative effectiveness of promotions within the experimental setting, we find National Brand promotions to be significantly more effective than Private Label deals of the same type. This observation is in line with findings in the literature on asymmetric promotion effects between brand types (see e.g. Tellis and Zufryden 1995, Sethuraman 1995 and 1996), and indicates that –though the absolute promotion effect is overrated – comparisons across brands still seem realistic.

Table 4b reports promotion parameters and their standard deviations for the choice models in the scanner and the experimental data set. Being multinomial logit model parameters obtained from different samples, their absolute values cannot be readily compared to yield insights into 'true' differences in promotional effectiveness. Parameter differences might reflect deviations in the variance or uncertainty in consumer decisions in both settings (the scale parameter in the multinomial logit model) rather than differences in promotion reactions as such. Informal inspection of model estimates reveals that most parameter values - including those associated with explanatory variables other than promotions like loyalty, variety seeking and last purchase variables - are larger in magnitude in the experimental model compared to the scanning model. This suggests that the difference in promotion parameters reported in Table 4b might well be due to experimental conditions affecting the variability in consumer decisions overall (increase in the scale parameter) rather than to a particularly stronger promotion response[2]. Clearly, further research is needed on this issue. The promotion parameters from the choice models do, however, allow to compare the relative impact of various promotions within the experimental setting. Like for the purchase quantity model, National Brand promotions are found to generate a significantly stronger response than their Private Label counterparts.

We conclude that the preliminary results roughly point in the direction of Hypothesis 4 for purchase quantity effects of promotions. In addition, the estimates reveal that, in the experimental setting, National Brand Promotions generate stronger response than similar Private Label activities, a finding in line with expectations from the literature. This suggests that, even

[2] Since the explanatory variables in both choice models differ, the formal test suggested by Swait and Louvière

though some doubt remains on the generalizability of absolute promotion estimates, CSSE do yield meaningful insights into the relative effectiveness of promotions for different brands.

Hypothesis 5

Consumers do not systematically change their purchase behavior (quantity and choice decisions) over subsequent simulation periods of a CSSE.

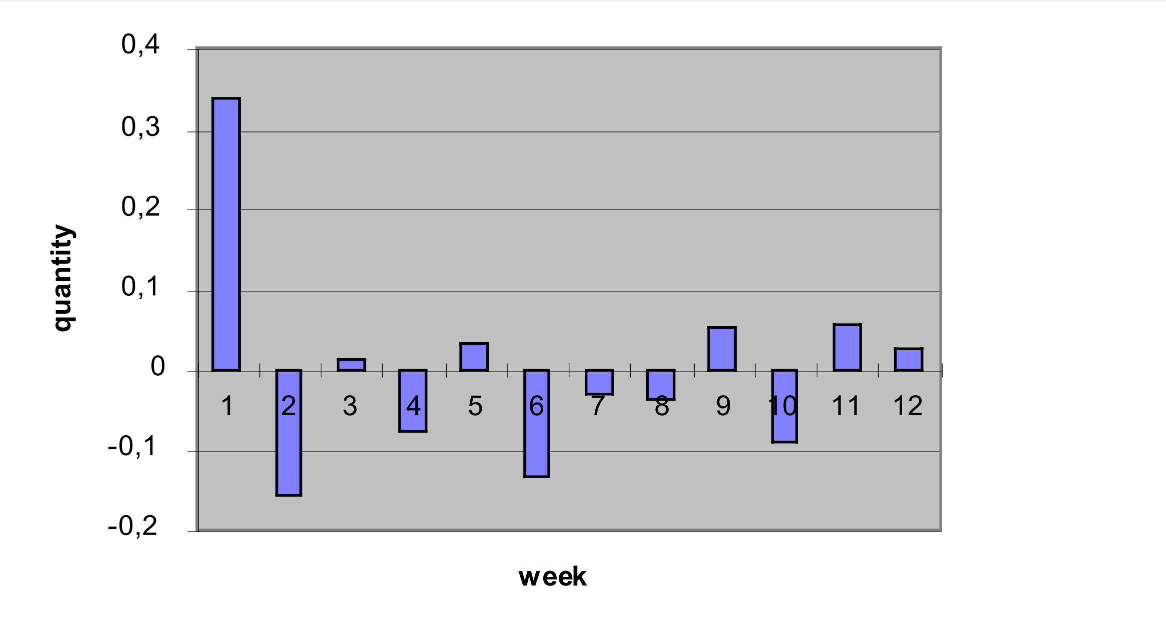
To assess whether systematic differences in quantity and choice appear over the 12 subsequent weeks, we re-estimate the purchase quantity and choice models for the experimental data set including weekly dummy variables. In the purchase quantity model, only weeks 1 and 2 lead to statistically significant deviations[3]. These deviations could be interpreted as 'initialization' effects. As respondents were given the opportunity to become familiarized with the simulation procedure and virtual store shelves beforehand, these higher initial purchase quantities are probably due to the fact that inventories are set to zero in the beginning of the experiment. To avoid this initialization effect, future studies could start with a non-zero inventory (for instance, an average inventory or self-reported inventory level). The pattern of weekly parameters over the subsequent weeks displayed in Figure 4 'confirms' this interpretation of an initialization effect that dies out: there is no evidence of a trend in the quantities bought throughout the experimental history, nor is there a systematic end-of-period effect.

In the choice model, the majority of weekly parameters are not significantly different from zero. From the 154 parameters estimated, only 5 are significant at the 10% significance level. From this we can conclude that there is no systematic tendency for items to be more or less often bought in early versus late weeks of the experimental period.

(1993) to estimate relative scale values and derive true differences in promotion parameters cannot be applied.
[3] This result was obtained irrespective of which week was chosen as the reference week in the estimation stage.

From the above discussion, we conclude that Hypothesis 5 can be accepted. Apart from exceptionally high inventory build-up purchases in the first week, consumers behave in a realistic manner throughout the observation period: their purchase quantities and choices do not

Figure 4.
Purchase Quantity Pattern over Subsequent Weeks
(weekly parameters are 'rescaled' against the average over all periods)



systematically change as the experiment proceeds.

Hypothesis 6

The absence of preferred items on the computer screen does not affect the quantity purchased in the product category, but leads to a choice bias in favor of major national brands

To check for biases induced by the reduced experimental assortment, we compare the weekly purchase rate of respondents who do find their preferred products on the computer screen, with that of the remaining respondents.

Closer analysis reveals that of the 486 respondents for which experimental information is obtained, 438 report finding their preferred item on the simulated shelf. Only a small group of respondents (48) fall under the ‘do not find their preferred item’ condition. Both respondent groups exhibit similar socio-demographic and behavioral profiles: there is no sigificant difference in distributions of gender, age, family composition, income or occupation distributions, and both groups have comparable variety seeking tendencies. Of those respondents who could not find their favorite item in the virtual store assortment, 26% indicated to prefer another variant of one of the experimental brands (13% of NB2 and 13% of the store’s private label), while 22% prefer another high quality brand available in the store but not presented on screen, and 44% have a preference for brands not available in the intercept store. Hence, the majority of this group prefers an alternative that is quite similar to the ones incorporated in the experimental assortment; only 5% of the sample appears to prefer a non-available brand.

With an average of .83 (standard deviation .60), the weekly purchase quantity of those who can locate their preferred item is not substantially higher than that of the remaining respondents (for which the average amounts to .70, with a standard deviation of .50). Since not all respondents are confronted with the same promotion scenarios, we also compare the simulated weekly quantities in the absence of promotions: these amount to .79 (SD: .56) for the first group of 438 consumers, and to .66 (SD: .47) for the remaining 48 respondents. A non-parametric test on the comparability of the purchase rate distributions in both subgroups suggests that the difference is at best marginally significant, for observed as well as for simulated average weekly quantities (for observed quantities, Mann-Whitney $z=-1.378$, $p=.168$; the Mann-Whitney test statistic for simulated no-promotion quantities $=-1.621$, $p=.105$). Whether purchase rate differences are a result of assortment reductions, can be further explored by analyzing deviations between actual and self-reported purchase rates for the sub-sample of respondents that did not find their favorite item on the CSSE shelves (cf. Hypothesis 1). If respondents who cannot find their preferred item on screen have a tendency to buy less, this should be reflected in observed experimental quantities below self-reported use rates for this group. A Wilcoxon Matched-Pairs Signed Ranks test on reported versus observed weekly purchase quantities indicates that the distribution is not significantly different (z-statistic: -1.103, p=.27). We conclude that not finding their preferred items on screen does not induce respondents to buy substantially less, yet our results suggest that some caution is in order, especially in view of the small number of respondents who do not find their preferred item.

In a similar vein, we compare the market share of the various brands among respondents who do find their preferred item, and others. Table 5 reports the observed market shares for the experimental brands in both respondent groups. Based on a test of equality of proportions, we find that for 2 of the national brands (NB1 and NB3), a significant difference is found. For the private label, the difference is 'marginally' significant. Yet, the difference does not go in the hypothesized direction: contrary to expectations, we observe from table 5 that the smaller national brand NB3 enjoys an increase, and the larger NB1 a decrease. Hence, we conclude that the second part of hypothesis 6 cannot be accepted. An alternative hypothesis, supported by recent developments in the promotion and stock-out literature, is that consumers - instead of shifting to the major national brands - tend to choose a substitute item within the same price/quality tier (Tellis and Zyfryden 1995, Sethuraman 1995 and 1996, Bronnenberg and Wathieu 1996, Campo et al. 1999). Our data seem to provide indications in that direction. As indicated above, most of the respondents who cannot find their favorite item, prefer either a high quality national brand or a private label item. Our result in Table 5 that the Private Label and high quality tier (NB2 and NB3) - in contrast with the more popular NB1 - enjoy a higher share, is therefore in line with the premise of within-tier substitution. Yet, like before, given the small number of respondents whose preferred item is not displayed, the results are to be treated with caution.

Table 5.
Observed Experimental Market Shares
for respondents who find their preferred items, and others

	Respondents who find their preferred items % market share	Others % market share
NB1	36	29
NB2	23	26
NB3	3	7
Private Label	27	29
Generic	11	10

Implications

As suggested by a number of recent articles, computer simulated shopping experiments open up a whole new and exciting range of research possibilities. The present research confirms most of Burke et al.'s (1992) findings, and sheds further light on what uses can be made of the outcomes of these types of experiments, and where the pitfalls are.

First, like Burke et al. (1992) we find that respondents tend to buy larger quantities in the experiment than in real life situations, despite the introduction of inventory cues which served as a storage constraint. Yet, the close correspondence between reported and actual use rates in the CSSE suggests that such inventory cues have the potential to increase the realism of purchase quantity decisions, on condition that a more accurate measure of real use rates can be applied. Second, in line with Burke et al.'s findings, choices in the CSSE context are biased in the direction of less generic brand purchases. For private labels, though, no such bias is observed. A potential explanation, which is confirmed by the chain's managers, is that consumers have come to judge these private label items as 'valuable' alternatives for choice that are accepted by their peers. This shift in private label positioning is not unique to the analyzed supermarket chain, but appears to be a recent trend observed for many retailer brands (see e.g. Corstjens and Corstjens

1995). Our findings further reveal that, in contrast to Burke et al. (1992), consumers' purchase variation is not systematically lower in our computer experiment. This might be attributed to a reduced 'rationality bias' as a result of appropriate purchase task instructions. In terms of promotional impact, our study confirms Burke et al.'s findings and provides some interesting additional indications. Like Burke et al., we find that experimental promotions tend to be more effective than (non-displayed) in-store promotions. Our study suggests, even though no clear display material is involved, the on-screen presentation makes experimental promotions act like displayed promotions. For the choice decision, however, this observation cannot be confirmed. Yet, while the matter of absolute promotion effects remains largely unresolved, the study suggests that CSSE can be used to infer relative effects of brand promotions. Our findings also suggest that CSSE yield valid information on purchase behavior over successive periods. Throughout the 12-week horizon covered by the experiment, consumers kept on buying in a similar (realistic) manner. Yet, it should be noted that some starting up effects in the first week or weeks are inevitable, rendering these weeks less useful for analysis. Also, from consumer reactions, we feel that prolonging the experiment over more than 12 weeks is not advisable. Not only would this reduce response rates due to excessive time requirements, it would render the exercise too boring for consumers and therefore jeopardize response quality in later periods. We therefore posit that while relevant dynamic information can be obtained, the length of the observation period must remain limited. Finally, the absence of favored items on the screen does not reduce purchase quantities. While consumers who cannot locate their preferred items on the screen exhibit different choice probabilities for the remaining items than other respondents, there is no clear shift towards popular national brands. Rather, these consumers seem to switch to experimental items belonging to the same 'price/quality tier' as their original preferences, but more research is needed to confirm this result.

In summary, our research suggests that CSSE can yield meaningful insights into consumers' dynamic purchasing strategies and promotion reactions. These results are obtained for a 'realistic' computer experiment that, while being clearly more sophisticated than paper and pencil tests, is far from being a 'futuristic' gadget. To the extent that results from this type of experiment are valid and useful this will certainly also be true for more advanced virtual shopping programs bound to be developed in the future. Several guidelines can be derived from the present and previously obtained results for the development of these programs. For one, like Burke et al. (1992) we find that a realistic representation of products and other in-store marketing stimuli is crucial for the external validity of the results. Future research should investigate whether on–screen presentation leads to an increased visibility and over-estimated effect of these stimuli, as the reported sales promotion results seem to suggest. Next, an effort should be made to build in purchase constraints that restrict actual buying decisions. Depending on the category considered, constraints can be budget, time or storage oriented. For frequently purchased consumer goods, our results lead to the conjecture that inventory cues can improve the realism of purchase quantity decisions, provided that they are based on accurate estimates of consumption rates. Another important aspect – though it clearly needs further research – appears to be the composition of the experimental assortment. Our results suggest that virtual store assortments need not be complete to yield useful purchase quantity data, at least as long as they incorporate (the major) alternatives of each price/quality tier present in actual stores. And finally, although valid information can be collected for several subsequent purchase occasions in a single computer session, one should account for the possibility of initialization biases in the first purchase decisions made. Clearly, our analysis is only of limited scope: it is confined to only one product category, and based on a comparison of two consumer panels. Even though both samples were carefully matched, sample sizes are important and the statistical tests robust, confirmation from additional studies might be useful.

References

Blattberg, R.C. and S.A.Neslin (1990), *Sales Promotions: Concepts, Methods and Strategies*, Prentice Hall: Englewood Cliffs, N.J.

Bronnenberg, B.J. and L. Wathieu (1996), "Asymmetric Promotion Effects and Brand Positioning", *Marketing Science*, 15(4), 379-394.

Brucks, M. (1988), "Search Monitor: An Approach for Computer-Controlled Experiments Involving Consumer

Information Search", *Journal of Consumer Research*, p.117-121.
Bucklin, R. and S. Gupta (1992), "Brand Choice, Purchase Incidence and Segmentation: an integrated approach", *Journal of Marketing Research*, 29 (May), p.201-215.
Burke, R. (1995),"Virtual Shopping", *OR/MS today*, August, p.28-34.
____ (1996), "Virtual Shopping: Breakthrough in Marketing Research", *Harvard Business Review*, March-April, p.120-131.
____, B. Harlam, B. Kahn and L. Lodish (1992), "Comparing Dynamic consumer Choice in Real and Computer-simulated Environments", *Journal of Consumer Research*, June, p.71-82.
Campo, K. (1997), *Variety Seeking and the Sensitivity to In-Store Promotions*, unpublished doctoral dissertation, University of Antwerp-UFSIA: Antwerp.
Campo, K., E.Gijsbrechts and P.Nisol (1999), *Towards a Conceptual Framework of Out-of-Stock Behavior: The Impact of Product, Consumer and Situation Characteristics on Out-of-Stock Reactions*, Working paper, University of Antwerp-UFSIA, Department of Business Economics.
Carson, R.; J. Louviere, D. Anderson, P. Arabie, D. Bunch, D. Hensher, R. Johnson, W. Kuhfeld, D. Steinberg, J. Swait, H. Timmermans and J. Wiley (1994), "Experimental Analysis of Choice", *Marketing Letters*, 5 (4), p.351-368.
Cohen S. and M. Gadd (1996), "Virtual Reality Shopping Simulation for the Modern Marketer", *Marketing and Research Today*, February, p.18-26.
Cook, Th.D. and D.T.Campbell (1979), *Quasi-Experimentation: Design and Analysis Issues for Field Settings*, Houghton Mifflin Company: Boston.
Corstjens, J. and M.Corstjens (1995), *Store Wars: The Battle for Mindspace and Shelfspace*, John Wiley and Sons: Chichester-New York.
Dhar, S.K. and S.J.Hoch (1997), "Why Store Brand Penetration Varies by Retailer", *Marketing Science*, 16(3), p.208-227.
Dillon, W. and S. Gupta (1996), "A Segment Level Model for Category Volume and Choice", *Marketing Science*, 15 (1), p.38-59.
Fader, P.S. and B.G.S.Hardie (1996), "Modeling Consumer Choice among SKU's", *Journal of Marketing Research*, 33(November), p.442-452.
Gabor, A., C.W.J.Granger and A.P.Sowter (1970), "Real and Hypothetical Shop Situations in Market Research, *Journal of Marketing Research*, 7(August), 355-359.
Guadagni, P.M. and J.D.C.Little(1983), "A Logit Model of Brand Choice Calibrated on Scanner Data", *Marketing Science*, 2(3), 203-238.
Gupta, S (1988), "Impact of Sales Promotions on When, What, and How Much to Buy", *Journal of Marketing Research*, 25(November), 342-355.
Kahn, B.E. and L.McAlister(1997), *Grocery Revolution: The New Focus on the Consumer*, Addison Wasley: Reading, Massachusetts.
Kapferer, J.-N.(1994), *Strategic Brand Management*, McMillan: New York.
Kumar, V., K.Karande and W.J.Reinartz (1998), "The Impact of Internal and External Reference Prices on Brand Choice: The Moderating Role of Contextual Variables", *Journal of Retailing*, 74(3), 401-426.
Lynch, J.G.JR. (1982), "On the External Validity of Experiments in Consumer Research", *Journal of consumer Research*, 9(December), 225-239.
Malhotra, N.K.(1996), *Marketing Research: An Applied Approach*, Prentice Hall: Englewood Cliffs, N.J.
McAlister, L. and E.Pessemier (1982), "Variety Seeking Behavior: An Interdisciplinary Review", *Journal of Consumer Research,* 9(December), 311-322.
Nagle, T.T. and R.K.Holden (1995), *The Strategy and Tactics of Pricing: A Guide to Profitable Decision Making*, Englewood Clifs, NJ: Prentice Hall.
Narasimhan, C.; S. Neslin and S. Sen (1996), "Promotional Elasticities and Category Characteristics", *Journal of Marketing*, 60, April, p.17-30.
Neslin, S. and L. Schneider(1996), "Consumer Inventory Sensitivity and the Postpromotion Dip", *Marketing Letters*, 7 (1), p.77-94.
Nevin, J.R.(1974), "Laboratory Experiments for Estimating Consumer Demand: A Validation Study", *Journal of Marketing Research*, 9(August), 261-268.
Papatla, P.and L.Krishnamurthi (1992), "A Probit Model of Choice Dynamics", *Marketing Science* , 11(2), p.189-206.
Sethuraman, R. (1995), "A Meta-Analysis of National Brand and Store Brand Cross-Price Elasticities", *Marketing Letters*, 6(4), 275-280.
____ (1996), "A Model of How Discounting High-Priced Brands Affects the Sales of Low-Priced Brands", *Journal of Marketing Research*, 13(3), 263-269.

Sivakumar, K. and S.P.Raj (1997), "Quality Tier Competition: How Price Change Influences Brand Choice and Category Choice", *Journal of Marketing*, 61(July), 71-84.

Sudman S. (1964a), "On the Accuracy of Recording of Consumer Panels: I", *Journal of Marketing Research,* 1(May), 14-20.

____ (1964b) "On the Accuracy of Recording of Consumer Panels: II", *Journal of Marketing Research,* 1(August), 69-83.

Swait, J. and J. Louviere (1993), "The Role of the Scale Parameter in the Estimation and Comparison of Multinomial Logit Models", *Journal of Marketing Research*, 30(3), 305-314.

Tellis, G.J. and F.Zufryden (1995), "Tackling the Retailer Decision Maze: Which Brands to Discount, How Much, when and Why?", *Marketing Science*, 14(3, part 2), 271-299.

Van Trijp, H. and J.B. Steenkamp (1990), "An Investigation into the Validity of Measures for Variation in Consumption Used in Economics and Marketing", *European Review of Agricultural Economics*, 17(1), p.19-41.

Van Trijp, H.; W. Hoyer and J. Inman (1996), "Why switch? Product Category level Explanations for True Variety-Seeking Behavior", *Journal of Marketing Research*, 33, August, p.281-292.

Wind, Y. and D.Lerner (1979), "On the Measurement of Purchase Data: Surveys Versus Purchase Diaries", *Journal of Marketing Research*, 16(February), 39-47.

Appendix 1: Purchase Quantity Model

Purchase quantity probabilities per consumer are modeled as a Poisson-based process with the following structure:

$$P(y_{ht} = n_h) = \frac{exp(n_h . \lambda_{ht}).exp(exp(\lambda_{ht}))}{n_h!} \qquad \textbf{(1)}$$

where $P(y_{ht}=n_h)$ represents the probability that household h purchases n_h units in week t. In accordance with Dillon and Gupta (1996), we use an exponential function for the expected purchase rate to ensure positive values. Similar to Bucklin and Gupta (1992) and Gupta (1988), we specify the expected weekly quantity, $exp(\lambda_{ht})$, as a function of household use rate, inventory level, and promotions in the category (see equation 2). While the 'use rate estimate' allows to capture differences between households, the inventory variable mainly accounts for intertemporal variations in ♉$_{ht}$.

$$\lambda_{h,t} = \alpha_{s,h} + \beta_1 * US_h + \beta_2 * Inv_{h,t} + \sum_{k} \beta_{pro,k} * Prom_{k,h,t} \qquad \textbf{(2)}$$

where

$$Inv_{h,t} = Inv_{h,t-1} + N^h_{t-1} - US_h$$
$$N^h_{t-1} = number\ of\ units\ bought\ by\ h\ in\ t-1 \qquad \textbf{(3)}$$

$\alpha_{s,h}$ = a segment specific constant.

The Poisson model is estimated using a latent class approach, where segment-specific -s account for unexplained heterogeneity between households.
US_h= household use rate. US_h is either measured as self-reported use rate (experimental data set), or estimated from the household's purchases during a four week initialization period (scanner panel data set).
$Inv_{h,t}$ = a dynamically updated inventory level estimate, representing the number of units the household has in stock at the beginning of week t. Inventory levels are weekly updated, based on information on inventory and purchases in the previous week, and data on weekly use rates of the household:
$Prom^k_{ht}$ = the level of the k'th promotion variable, observed by household h in period t :
$Price(j)_t$: number of brands of type j for which a price cut was offered in period t,
j = PL (private label), NB (national brand).
$Coup(j)_t$: number of brands of type j for which a coupon was offered in period t,
j = PL (private label), NB (national brand).
$Quan(j)_t$: number of brands of type j for which a quantity discount was offered in period t,
j = PL (private label), NB (national brand).
$Gift(j)_t$: number of brands of type j for which a gift or premium was offered in period t,
j = PL (private label), NB (national brand).
$Shelf(j)_t$: number of brands of type j that received extra shelf facings in period t,
j = PL (private label), NB (national brand).

To test hypothesis 6, additional terms are added to equation 2, consisting of the product between week-parameters ♉ $_{w,\ I}$ (i=week number) and weekly dummy variables $D_{w,\ I}$ (equal to 1 for observations corresponding to the i'th simulation week, and equal to 0 elsewhere). Week 5 serves as the reference week, for which no dummy variable is included.

Appendix 2: Item choice model

Household *item choices* conditional on purchase are estimated using a simple MNL type model. Our specification is based on the structure suggested by Guadagni and Little (1983), subject to some minor modifications. Following a suggestion by Bucklin and Gupta (1992), we replace the original dynamic 'loyalty' variables in the G&L model by static measures of household preferences, and supplement them with 'last purchase variables' (as specified in equation 5). This better allows to distinguish between intrinsic preference heterogeneity, and state dependence. Relevant item characteristics refer to brand and taste, no size differences are encountered. We further supplement Guadagni and Little's model with an interaction effect between the household's tendency to seek variation in behavior (see equation 5) and the past purchase variable, thereby allowing some households to exhibit reinforcement behavior, and others to show

$$p^{h}_{j,t} = \frac{exp(v^{h}_{j,t} + u^{h}_{j,t})}{\sum_{l} (exp(v^{h}_{l,t} + u^{h}_{l,t})} \qquad \textbf{(4)}$$

a preference for switching. The specification for the choice model is given in equation 4.

where
$P^{h}_{j,t}$ = probability that household h chooses item j on occasion t
$u^{h}_{j,t}$ = random component of utility of item j for household h on occasion t,
$v^{h}_{j,t}$ = systematic component of utility of item j for household h on occasion t,

The systematic utility is further given by:

$$v^{h}_{j,t} = \sum_{b} D_{j,b} * (\alpha_{1,b} + \alpha_{2,b} \cdot Loyb^{h}_{j,t}) + \sum_{v} D_{j,v} * (\alpha_{1,v} + \alpha_{2,v} \cdot Loyv^{h}_{j,t}) +$$
$$\gamma_{1} \cdot Lastp^{h}_{j,t} + \gamma_{2} \cdot Lastp^{h}_{j,t} \cdot VST^{h} + \sum_{k} \cdot \beta^{k}_{pro} \cdot Prom^{k}_{j,t} \qquad \textbf{(5)}$$

where
$D_{j,b}$ ($D_{j,v}$) are brand(variant) dummy variables, equal to 1 if item j belongs to brand b (variant v), and equal to 0 otherwise.
$\alpha_{1,b}$ ($\alpha_{1,v}$) are brand-(variant-) specific intercepts (see Fader and Hardie 1996).
$Loyb^{h}$ ($Loyv^{h}$) are static loyalty variables indicating the fraction of consumer purchases in the initialization period accounted for by brand b (variant v),
$Lastp^{h}_{j,t}$ indicates the difference between item j and the item(s) purchased on the previous occasion, the difference is equal to zero if the items have the same brand and variant (taste), .5 if either the brand or the variant is different, and 1 if both differ.

$Prom_{kht}$ = the level of the k'th promotion variable encountered by household h on purchase occasion t.
VST^h= household h's tendency to seek variation in buying behavior, approximated by the entropy measure computed over a 4 week initialization period. The entropy measure was rescaled to account for differences in assortment size and purchase quantity (determining the opportunities for variation in buying behavior; see Campo 1997). After rescaling, the entropy lies between 0 (no variation in behavior) and 1 (maximum variation in behavior).

$$VST^h = \frac{ent^h}{entmax^h} \qquad \textbf{(6)}$$

$$ent^h = {}_{j} - share_j^h * ln(share_j^h)$$

$$entmax^h = ln(M^h)$$

$$M^h = min(J, Q^h)$$

$$J = \text{number of items in the assortment}$$

$$Q^h = \text{overall purchase quantity of } h \text{ during the initialization period}$$

$$share_j^h = \text{share of item } j \text{ in household } h\text{'s purchases during the initialization period}$$

While this model remains fairly simple, it provides a good description of household behavior, and serves our purpose of removing major immediate promotion impacts.

In order to test for systematic changes in choice behavior over subsequent simulation periods (hypothesis 6), the following terms are added to the systematic utility function (equation 5):

$${}_{w}\beta_{w,j} * D_w$$

Where w is a week indicator, and D_w a week-dummy variable equal to 1 for observations corresponding to week w, and equal to 0 elsewhere.