Journal of Agricultural and Applied Economics, 44,1(February 2012):49–62

A Comparison of Choice Experiments and Actual Grocery Store Behavior: An Empirical Application to Seafood Products

Darren Hudson, R. Karina Gallardo, and Terrill R. Hanson

In this paper we compare results from an in-store field experiment and a mail survey choice experiment (CE) to investigate CE's capacity in predicting grocery store market share. For the comparison, we used three seafood products: freshwater prawns, marine shrimp, and lobster. CE estimates were obtained via four econometric models: the conditional logit, the random parameter logit, the heteroskedastic extreme value, and the multinomial probit. We found that the level of control in the grocery store experiment and the choice of econometric model influenced the capacity of CE to predict grocery store market shares.

Key Words: choice experiment, grocery store, hypothetical bias

JEL Classifications: C35, Q13

Discrete choice experiments have become a popular method of estimating willingness to pay (WTP) and market share predictions for products and services. The method's popularity is understandable given its consistency with Lancaster's (1966) demand theory (Louviere, Hensher, and Swait, 2000), the ability to handle a number of attributes simultaneously in a controlled, orthogonal experimental design, and the ability to generate a large number of observations on choice from a relatively small number of respondents. Despite its popularity in applied analysis, a number of critical questions remain open as to the validity of choice experiments in predicting actual behavior.

The predictive capacity of choice experiments (CEs) has been investigated from a number of perspectives. Aggregate predictions of market share (Page and Rosenbaum, 1987; Srinivisan et al., 1981), as well as individual level predictions of behavior (Srinivisan, 1988; Srinivisan and Park, 1997) have been examined. A central critique of CEs is that by relying on hypothetical choices respondents give biased results, causing a systematic difference between elicited and actual statistics (i.e., WTP and market share). Hypothetical bias, as this systematic difference is known, has been widely documented in the literature (Fox et al., 1998; List and Gallet, 2001; List and Shogren, 1998; Little and Berrens, 2004; Murphy et al., 2005).

Darren Hudson is Larry Combest Chair of Agricultural Competitiveness, Department of Agricultural and Applied Economics, Texas Tech University, Lubbock, TX. R. Karina Gallardo is assistant professor extension specialist, School of Economic Sciences – Tree Fruit Research and Extension Center, Washington State University, Wenatchee, WA. Terrill R. Hanson is associate professor extension specialist, Department of Fisheries and Allied Aquaculture, Auburn University, Auburn, AL.

The authors acknowledge the funding by the Mississippi Agricultural and Forestry Experiment Station, the product support of Dolores Fratesi and the U.S. Freshwater Prawn Growers Association, and the Kroger grocery store in Germantown, TN for allowing us to test the product. The authors also acknowledge the useful comments and suggestions of Jayson Lusk, Robert Rosenman, and numerous seminar participants.

However, unlike other hypothetical methods such as contingent valuation, CE questions are typically posed in a manner more like true choice environments, leading to a maintained hypothesis that CEs are less prone to hypothetical bias (Adamowicz et al., 1998).

Recent research has employed experimental economics principles related to nonhypothetical choices to test this maintained hypothesis. Here nonhypothetical refers to incentive compatible mechanisms, carried out in a laboratory[1] setting (Alfnes et al., 2006; Carlsson et al., 2001; Chang, Lusk, and Norwood, 2009; Ding, Grewal, and Liechty, 2005; Lusk and Schroeder, 2004; Miller et al., 2011; Sattler and Volckner, 2002). These studies generally find hypothetical bias in predicted CE WTP and market shares relative to values derived from incentive compatible experiments. But it is unclear if hypothetical bias also extends to differences between predicted WTP values. Lusk and Schroeder (2004) and Carlsson and Martinsson (2001) both found evidence that marginal WTP values were not different between hypothetical and nonhypothetical settings, but Carlsson et al. (2001) and Miller et al. (2011) found evidence of differences.

While these studies provide insight in the external validity of choice experiments, they (except Chang, Lusk, and Norwood, 2009) are confined to a laboratory setting. Laboratory experiments do offer a high degree of control over decision variables of interest (Hudson, 2003), but their sterile nature makes generalization to more complex realistic situations difficult. That is, laboratory experiments are conducted "out of context," which may lead respondents to focus all attention on the decision task. Conversely, real shoppers are attempting to make a myriad of choices in a confusing environment, which increases cognitive effort and may lead to a different set of decisions as compared with decisions made in isolation in the laboratory. Thus, while a laboratory experiment may provide a refined test, its conclusions may not extend to actual consumer behavior in a shopping context. Moreover, differences between laboratory settings and actual field behavior could be contextual. For example, several studies have focused on two behavioral issues, one the Hawthorne effect or individuals' awareness that their behavior is being studied and second, sample selection of individuals participating in laboratory experiments compared with real market shoppers (Harrison and List, 2004; Levitt and List, 2007; and List, 2006). To this particular, Chang, Lusk, and Norwood (2009) argue that one should not always expect identical behavior in the laboratory and in the field and that different economic models might explain differences in the environments in question.

In this context, we find it useful to compare results of CEs with actual purchasing behavior to more fully explore the external validity of CEs. Previous studies have focused on similar comparisons (Brookshire, Coursey, and Schulze, 1987; Chang, Lusk, and Norwood, 2009; Lusk, Pruitt, and Norwood, 2006; Shogren et al., 1999). Brookshire, Coursey, and Schulze (1987) compared demand protocols obtained via experimental auctions and door-to-door sales. They did not find significant differences in demand behavior across the two settings. Shogren et al. (1999) compared consumer behavior under experimental auctions, mail survey CE, and a grocery store experiment. They found that the hypothetical CE yielded a higher WTP and market share estimate than the grocery store. However, no formal measurement of hypothetical bias was made in this study. Lusk, Pruitt, and Norwood (2006) compared market shares from an incentive compatible field experiment at a grocery store with market shares from actual sales data. They found that the field experiment overestimated market shares results but that bootstrapped confidence intervals overlapped suggesting that experiment results were reasonably accurate predictors of consumer behavior. Chang, Lusk, and Norwood (2009) compared three elicitation formats (hypothetical CE, incentive compatible rankings, and grocery store sales) for three different product categories (ground beef, wheat flour, and dishwashing

[1] Laboratory, here, refers to conducting experiments in a confined environment that is controlled and isolated from other external environments. Computer labs, classrooms, etc., are typical examples of economic laboratories, but laboratory refers to any experiment not conducted in the field or natural environment.

liquid). They found that the incentive compatible method outperformed the hypothetical CE in predicting actual market shares.

Overall, past studies suggest that lack of control over the store setting (e.g., prices for substitutes, information given to shoppers in the grocery store) makes it difficult to compare actual purchase behavior with hypothetical settings, resulting in confusion over whether observed differences are a result of the lack of control or actual hypothetical bias. We argue that controlling sacrifices realism, leading to grocery store settings that do not necessarily correspond to actual behavior. Hence, measuring the predictive ability of hypothetical CEs by implying that the grocery store setting reflects true choice behavior might not be accurate and comparisons should be made with caution (Chang, Lusk, and Norwood, 2009).

This paper presents an analysis that compares a mail survey CE with a grocery store experiment, with the objective of measuring the predictive capacity of CE in relation with observed market outcomes in a specific context. Rather than conducting in-store auctions or "taste tests," this study places the product in the store where the price of the product of interest is controlled, but the shopper is unaware of the experimental design making his/her experience identical to an ordinary shopping experience. This approach necessarily means sacrificing some control over the external environment, but improves the realism of the experiment.

The product used in the analysis was the freshwater prawn (FP), which is similar in appearance to marine shrimp. Prawns make an interesting subject for analysis because it is a relatively new product in the U.S. market but is similar to existing products (marine shrimp and lobster), which adds evidence from a different perspective to the predictive power of CE in a setting where a new product is introduced into a market with existing substitutes.

Methods

The methodological approach centers around two related experiments conducted concurrently during January to March 2004 in Germantown, TN (a suburb of Memphis, TN). This site was chosen for two important reasons. First, it represents an affluent, suburban community that is the most likely target market for the FPs. Second, and most importantly, it was the location where the grocery chain agreed to test the product. First, we describe the in-store experiment. Second, we describe the mail survey CE. Finally, the procedures used in the analysis are presented.

Grocery Store Experiment

The grocery store experiment was conducted in a major grocery chain in Germantown, which is the largest of four grocery supermarkets in this community of approximately 40,000 people. The store management agreed to stock the FPs in their fresh seafood counter. They also allowed the researchers to set the price for the FPs on a weekly basis and collect data on sales of FPs, as well as the competing products, marine shrimp and lobster. Researchers provided the store personnel with brochures containing FP's nutritional information and pictures, so shoppers could get familiar with the relatively new seafood. These brochures were displayed at the fresh seafood counter, and were visible for customers. Prices for marine shrimp and lobster were set by the store. The FPs for this experiment were obtained from members of the U.S. Freshwater Prawn Growers Association and an agricultural experiment station. The store had the incentive to keep the revenues generated by the FP sales.

Given that marine shrimp and FPs are close in composition, look, taste, and texture, it was assumed that FPs would be priced similarly to marine shrimp. Regional grocery stores were contacted to determine a reasonable range of prices for marine shrimp. Data collected over a 3-week period at three regional grocery stores showed that marine shrimp prices typically ranged from $7 to $13 per pound, depending on variety and size. This study focuses on "large-size" shrimp and prawn; this size is equivalent to 23–45 units per pound.

Five price levels ($5.99–$13.99/lb for large prawns in $2/lb increments) for the FPs were randomly assigned to different weeks as shown in Table 1. The price range was established to encompass the normal range of prices observed

in the stores for shrimp. Also listed in Table 1 are the average weekly prices for marine shrimp and lobster during the same period. Note that while researchers had control over prawn prices, there was no control over shrimp and lobster prices as they were set by the grocery store. Data were collected for all three products every other week for 5 weeks.[2] FPs were offered every other week to give time for shoppers to "forget" the prices during the previous period to reduce attempts to predict the pricing pattern.

Daily transactions data were collected by the grocery store on FP sales during each week of the study as well as sales and prices of both fresh shrimp and lobster. These weekly data were used to calculate the market share and its standard deviation for each product. A sample of 1,000 random draws from an assumed normal distribution centered on the market share with the sample standard deviation calculated from the data was taken.[3] This process generated a distribution of shares for each product from which comparisons could be made to CE results.

CE Survey

A mail survey was conducted in Germantown, TN during the same period as the in-store experiment. A random sample of 2,000 names from Germantown (the same zip code as the store) was purchased from a commercial marketing firm. A Dillman three-wave design was used—survey, then reminder card, then survey—to mitigate nonresponse bias (Dillman, 1978; Hudson et al., 2004; Pennings, Irwin, and Good, 2002). Researchers enclosed, in the survey envelope, the same informational brochure displayed at the grocery store. The survey collected basic data on consumption patterns and attitudes toward seafood, demographic variables, as well as the CE.

The CE was constructed in a manner similar to Lusk and Schroeder (2004), whereby respondents faced a series of choices on product type—in this case, prawns, shrimp, and lobster—where only the price of the product was allowed to vary. Each category was one pound of product, with shrimp and prawns being the same count size (23–45 count). Thus, the stated price in the CE experiment was on a per pound basis. This procedure was designed to match the count sizes in the store experiment for the "large" category. An example of a CE scenario is shown in Figure 1. A similar set of prices was used for the CE as for the grocery store experiment to ensure comparability. More specifically, the prawn prices were the same as used in the store (a set of five price levels ranging from $5.99 to $13.99 in $2 increments). The price ranges for shrimp and lobster were consistent with the store, but not all prices used in the survey were observed in the store for shrimp and lobster over the test period. For example, in the survey we used a set of five prices for marine shrimp ranging from $5.99 to $13.99 in $2 increments, and prices in the grocery store ranged from $6.99 to $8.99 (see Table 1). As for lobsters in the survey we used a price range from $6.99 to $14.99 in $2 increments, and prices in the grocery store ranged from $12.00 to $12.99 (see Table 1). This shows that although pricing points were not identical across settings, store prices were within the bounds of the prices used in the survey. Note that shrimp prices were comparable, but lobster prices in the store were concentrated at the high end of the price range used in the survey. This concentration for lobster prices ultimately had some impact on the result for lobsters as will be discussed later in the paper.

There were five price levels for each product. Because of the large number of potential choice sets ($5^3=125$), a fractional factorial design was used. The fractional factorial is a subset of the full factorial. In this case, we chose a fractional factorial that minimizes correlation

[2] This study was part of a larger study that used other FPs' forms and sizes in other weeks of the experiment. We did not test for covariance across random prices for shrimp and lobster generated by the grocery store, because there were not enough observations and we have no reason to suspect covariance across prices for the three products being studied.

[3] The market shares were assumed to follow a normal distribution. There is no *a priori* reason to suspect non-normality. Given that these are averages of random variables, the Central Limit Theorem suggests that an assumption of normality is justified.

Table 1. Randomly Assigned Prices and Quantities Sold for Freshwater Prawns and Average Marine Shrimp and Lobster Prices, Grocery Store Experiment, Germantown, TN, 2004

Product		Week 1[a]	Week 3	Week 5	Week 7	Week 9	Weighted Average Price[b]	Total Quantity Sold (lbs)
Prawns	Weekly price ($/lb)	9.99	13.99	11.99	5.99	7.99	9.48	35.7
	Quantity sold (lbs)	0.00	13.17	0.50	14.00	8.00	[7.73–11.25][c]	
Shrimp	Weekly price ($/lb)	8.49	8.99	8.99	7.16	6.99	7.87	122.5
	Quantity sold (lbs)	12.50	23.50	18.00	38.00	30.50	[8.37–9.07]	
Lobster	Weekly price ($/lb)	12.49	12.99	12.99	12.00	12.50	12.70	132.0
	Quantity sold (lbs)	13.00	24.00	50.00	20.00	25.00	[11.46–13.93]	

[a] Products were offered every other week.
[b] Lobster was only sold live, but the price was quoted in $/lb.
[c] Numbers in brackets are 95% confidence intervals

among the attributes subject to identification of the main effects, or the resulting fractional factorial is D-efficient (see Kuhfeld, Tobias, and Garratt, 1994).[4] The result was 25 choice sets, but with this number of sets, respondent fatigue may still be a problem (Bradley and Daly, 1994). Thus, the 25 choice sets were randomly blocked into two different groups—one with 12 and one with 13 choice sets. These two different versions were randomly assigned to individuals, resulting in 1,000 people initially receiving version 1 and 1,000 people initially receiving version 2. An example of the choice set is presented in Figure 1.

The parallel data from the grocery store and the CE are interesting in a number of respects. First, the researchers had direct control of FP prices in both the grocery store and the CE. As Lusk and Schroeder (2004) point out, this is a preferred method of testing external validity, but it is often difficult to get retailers to agree to participate due to the proprietary nature of the data. In this case, we had full cooperation of the grocery store, leading to a direct test of external validity. One can certainly argue that there are other grocery stores in the area offering consumers a choice not captured in the CE. However, no other grocery stores in the area were offering freshwater prawns. A second important feature is that data were collected at the same time and in the same location, as the mail survey CE. This prevents confounding potential seasonal or location effects.

Yet, there is some difficulty in using this procedure as well. Unless demographic data of the grocery store shoppers is collected, it is difficult to know whether differences arising between CE and grocery store results arise from differences in the sample or hypothetical bias.[5] We feel that collecting demographic data from each store shopper is impractical. Also, it may make shoppers aware that they are being studied and influence their behavior. However, having demographic data from the survey sample, (which is confined to a specific zip code within the community where the store was located) with given demographic characteristics, allows one

[4] Strictly speaking, the D-efficient criteria generated orthogonal designs that are level balanced. But, Huber and Zwerina (1996) also suggest that designs must meet the additional criteria of utility balance and minimal overlap to be "optimal." These issues were not addressed here. Carlsson and Martinsson (2003) discuss alternative choice set formation techniques that can be used to induce utility balance and minimal overlap as well as the D-efficient criteria used here. Also note that experimental design is evolving and studies such as Street and Burgess (2007) and Rose and Scarpa (2008) show evidence of highly statistically efficient designs with new evaluation criteria and generation algorithms able to provide more design choices.

[5] For future research it might be useful to collect demographic data from a "shopper's card" or some other device. However, in this study, the seafood department only inserts a "seafood" UPC code on purchases from the fresh counter, so it is impossible to trace what "seafood" products were being purchased and match them directly to demographic data. Data for this analysis were collected directly by the seafood department, and linking to shopper information was not possible.

Attribute	**Farm-Raised, Freshwater Prawns 1 lb. Count: 23-45**	**Wild-Caught Marine Shrimp 1 lb. Count 23-45**	**Wild-Caught Marine Lobster 1 lb.**	**None**
Price ($/lb)	$5.99	$5.99	$6.99	
I would choose… (Please Check)				

Figure 1. Example of the Choice Set Used in the Mail Survey, Germantown, TN, 2004

to assume that patrons of this grocery store represent a random sample from the surveyed zip code community at large.

Finally, the method presented here assumes minimal uncertainty about the product in question, since shrimp and lobster are "familiar" products to most U.S. grocery shoppers. In relation to prawns, the relatively new seafood, we provided in both settings (survey and grocery store) an informational brochure as an attempt to reduce uncertainty due to unfamiliarity to the maximum extent possible.[6] Adamowicz et al. (1998) discuss alternative approaches when consumers may have some uncertainty about the product under question.

Data Analysis

Responses to CE questions were analyzed according to random utility theory, which holds that utility is given by:

$$U_{ij} = V_{ij} + \varepsilon_{ij}, \tag{1}$$

where U_{ij} is utility for the i^{th} individual choosing the j^{th} product (j = prawns, shrimp, lobster, and none), V_{ij} is the deterministic portion of the utility for individual i and product or alternative j, and ϵ_{ij} is the random component of the utility. If we assume that consumers wish to maximize subjective utility: $U_i = \max[U_1, U_2, \ldots, U_j]$, consumers will only choose product j if $U_{ij} \geq U_{ik}$. The probability that consumer i chooses alternative j from a set of k alternatives is given by:

$$\Pr(j \text{ is chosen}) = \Pr\left\{V_{ij} + \varepsilon_{ij} \geq V_{ik} + \varepsilon_{ik}; \forall \mathbf{k} \in \mathbf{C_i}\right\}, \tag{2}$$

where C_i is the set of all consumer choice alternatives {C = prawns, shrimp, lobster, and none}.

Four estimation methods were employed in this study—conditional logit (CL), random parameters logit (RPL), heteroskedastic extreme value (HEV), and multinomial probit (MNP). The reason for the different models is that all these model forms are common in the literature, but each has relative strengths and weaknesses. Model selection is driven by a number of issues ranging from econometric concerns about error structure to issues related to preference heterogeneity across respondents. We report estimates of each of these specifications so as to explore how robust our conclusions on the presence of hypothetical bias are to the underlying assumptions embedded in each specification.

The most common method of estimating parameters for this model is the multinomial/CL approach, which assumes that the error terms on utility are independent and identically distributed with a Type I extreme value distribution. Given these assumptions, the probability that consumer i chooses alternative j is modeled as:

$$\Pr(j \text{ is chosen}) = \frac{e^{V_{ij}}}{\sum e^{V_{ik}}}. \tag{3}$$

The CL approach suffers from the assumption of independence of irrelevant alternatives (IIA), or that model errors are independently

[6] Perhaps enclosing an informational brochure in every mailing, making the recipients a "captive" audience, could skew the mail survey results. However, we felt that this possibility was less of a potential problem compared with the mail respondents not having a picture and access to information that may be gathered in the store. There is no evidence to either support or refute a hypothesis of information-induced bias, but the reader should be aware of that possibility.

and identically distributed across alternatives. Several other approaches relax the IIA assumption, although in different ways. The HEV model assumes that errors are independently but not identically distributed across the alternatives (prawns, shrimp, and lobster) (Bhat, 1995). From a slightly different perspective, the MNP relaxes the IIA assumption by assuming that the errors across alternatives are normally distributed. To operationalize the MNP model, we assume that all off-diagonal covariances are zero, but we allow for free estimation of the variance of alternatives. This produces a model very similar in structure to the HEV model except that errors are distributed normally rather than as extreme value. Yet, another method of relaxing the IIA assumption is through the RPL model (Revelt and Train, 1998). Here, taste parameters are assumed to be random within the population with a given distribution (in this case, normal). We allow the alternative specific constants to vary randomly within the population and hold the price invariant across individuals or fixed. All models were estimated using SAS® (SAS Institute, Inc., Cary, NC).

Estimating the market share from the grocery store experiment is straightforward. We divided the quantity sold of each product (i.e., prawn, shrimp, or lobster) by the total quantity sold of all three products during the 5 weeks the experiment took place. Forecasted market shares from the CE models were estimated by substituting the prices of each product into Equation (3) for each specification approach used (i.e., CL, HEV, RPL, and MP).

To examine the difference between CE and grocery store market share distributions, we follow the combinatorial procedure introduced by Poe, Giraud, and Loomis (2005). The combinatorial approach takes the difference between the i^{th} element of one distribution (for example, 1,000 bootstrapped values from grocery store prawn market share) and every element of the second distribution (for example, 1,000 bootstrapped values from prawns' market share from the CE). In this manner, the procedure constructs every possible difference between the two distributions (1,000*1,000 = 1 million differences). Within this distribution, the percentage of observations greater than zero is the unbiased, nonparametric p value, which indicates that the mean of the first grocery store market share distribution is statistically greater than the mean of the second market share CE distribution (Poe, Giraud, and Loomis, 2005).

Results

Sample Characteristics

Of the 2,000 original surveys mailed, 91 were returned with incorrect addresses, leaving an effective sample of 1,909. Of these, 550 were returned (response rate = 28.8%), but only 523 were usable (usable response rate = 27%). While somewhat lower than desired, the response rate was still within the acceptable norm for mail surveys (Dillman, 1978). The demographic characteristics were compared with the U.S. Census for Germantown (Table 2). As can be seen, income and ethnicity for the sample were not significantly different from the census using a chi-square test ($p > 0.05$). Age is not included in the table because the sample was restricted to individuals with mailing addresses, which necessarily precludes children whose numbers are reflected in the census. While income and ethnicity are well represented, education and gender are different at the statistically significant level of 0.10. Education can be somewhat misleading as the mail sample uses categories to approximate years of education. Males appear to be overrepresented in the sample compared with the general population. Considering this disparity in male representation in the mail sample and that, typically female heads of households are more likely to do grocery shopping; we used weighted data in the econometric analysis. That is, data from the survey were weighted by the proportion of males in the sample to the proportion of males in the population (ratio = 1.27). In other words, all observations from male respondents were divided by 1.27 to correct for overrepresentation following the weighting procedure used by Lusk, Roosen, and Fox (2003).

Grocery Store Results

Overall, 36 pounds of large size prawns were sold in the fresh seafood counter over the

Table 2. Comparison of Response and Census Demographic Characteristics, Germantown, TN, 2004

Demographics	Survey (N = 523)%	U.S. Census (N = 37,348)%
Gender[a]		
Percent male	61.71	48.70
Household income[b]		
Less than $25,000	3.83	6.10
$25,000–$50,000	10.81	12.80
$50,000–$75,000	15.54	18.10
$75,000–$100,000	19.14	16.30
$100,000 or more	50.68	46.80
Education[a]		
Less than high school	0.40	2.00
High school	4.18	11.00
Some college	12.75	22.40
Completed college	42.63	42.30
Beyond Bachelors degree	40.04	22.40
Ethnicity[b]		
Caucasian	95.09	92.90
African American	1.02	2.30
Native American	0.61	0.20
Asian	2.04	3.50
Hispanic	0.61	1.10
Other	0.61	–

[a] Sample and census significantly different using a χ^2 test ($p < 0.10$).
[b] Sample and census not significantly different using a χ^2 test ($p > 0.05$).

5-week period (compared with 122.5 pounds of marine shrimp in the fresh counter and 132 pounds of lobster). The weighted average prices of prawns, shrimp, and lobster with their associated 95% confidence intervals are shown in Table 1. As can be seen, prawn weighted average prices are higher, but lie within the overlap of the 95% confidence intervals, than marine shrimp average prices. This suggests that prawns are viewed as close substitutes for marine shrimp.

CE Results

Table 3 shows the results of the CL, RPL, HEV, and MNP models. Alternative specific constants (ASCs) for all three products are significantly different from zero, indicating that all products were preferred to "none." Additionally, all price coefficients are negative and statistically significantly different from zero, indicating that increases in price lead to a decreased probability of choice. About the RPL model, none of the standard deviations are statistically significantly different from zero, suggesting preference homogeneity across respondents. A Hausman test to verify the IIA assumption was conducted. Results show that one fails to reject the IIA assumption ($\chi^2 = 0$, $p = 1$). Another CL assumption is that error variances across options are constant. We verified this assumption by conducting a likelihood ratio test to check for error variance variability. Test results imply that one can reject the hypothesis of constant variances ($\chi^2 = 4412$, $p = 0$), implying that HEV would yield more robust results than CL. In sum, from the test results, one can conclude that the HEV and MNP models yield more robust estimates compared with CL and RPL models. An additional likelihood ratio test comparing HEV and MNP likelihood functions showed that MNP is superior to HEV ($\chi^2 = 130$, $p = 0$).

We present in Table 4 the grocery store and CE market shares for each product and its corresponding bootstrapped confidence interval. One can observe that market share estimates vary significantly across products and models. The HEV and MNP models yield market share estimates closer to the grocery store market shares for shrimp, but not for prawns or lobster. For prawns, CE seems to overestimate the store market share; whereas for lobster, the CE underestimates this store share. Prawns' grocery store and CE market share are depicted in Figure 2.

Comparisons

Table 5 shows the comparison between the market share estimates from the grocery store and the mail survey CE using the Poe, Giraud, and Loomis (2005) combinatorial approach. One can observe that grocery store estimates are statistically significantly lower than CE estimates for prawns and marine shrimp for all models. A plausible reason to explain such differences is that shoppers were somewhat unfamiliar with prawns within the first weeks of the study, implying that potential acquaintance

Table 3. Conditional Logit, Random Parameters Logit, Heteroskedastic Extreme Value, and Multinomial Probit Model Results for CE Responses, Germantown, TN, 2004

	Conditional Logit		Random Parameters Logit		Heteroskedastic Extreme Value		Multinomial Probit	
Parameter	Estimate	Standard error	Estimate	Standard error	Estimate	Standard error	Estimate	Standard error
Prawns[a]	4.701*	0.160	4.803*	0.163	2.795*	0.144	1.902*	0.091
Shrimp	4.949*	0.153	5.013*	0.154	3.146*	0.165	2.124*	0.124
Lobster	4.285*	0.167	4.794*	0.299	3.014*	0.183	2.458*	0.145
Price prawns	−0.670*	0.023	−0.681*	0.022	−0.326*	0.024	−0.221*	0.016
Price shrimp	−0.646*	0.021	−0.653*	0.020	−0.357*	0.026	−0.240*	0.019
Price lobster	−0.544*	0.021	−0.621*	0.043	−0.354*	0.029	−0.299*	0.021
Scale Parameters/Standard Deviations								
Prawns	–	–	0.003[b]	2.928	2.128*[c]	0.203	0.102[d]	0.140
Shrimp	–	–	−0.010	1.954	1.831*	0.158	0.419*	0.150
Lobster	–	–	0.920	0.258	1.307*	0.119	1.000	
None					1.000		1.000	
Number of observations	5650.000		5650.000		5650.000		5650.000	
Log-likelihood	−5448.000		−5445.000		−5466.000		−5391.000	
Likelihood ratio index	0.304		−0.695				0.290	

[a] Refers to the alternative specific constant.
[b] Refers to the standard deviation of the parameter distribution for the random parameters logit model.
[c] Refers to the scale parameter for the heteroskedastic extreme value model.
[d] Refers to the standard deviation of the parameter distribution for the multinomial probit model.
* Statistically significant at the $p < 0.001$ level.

Table 4. CE Model Market Share and 95% Confidence Intervals and Grocery Store Market Share and 95% Confidence Intervals

					Grocery Store Market Share	
Product	CL	RPL	HEV	MNP	With Valentine's Day Weekend Sales	Without Valentine's Day Weekend Sales
Prawns	16.90% [15.13–18.65][a]	17.08% [14.35–20.17]	26.09% [23.05–29.23]	27.42% [24.03–30.84]	12.29% [0.00–15.27]	14.85% [0.00–15.57]
Shrimp	76.71% [74.65–78.65]	78.85% [75.13–81.83]	62.41% [57.71–66.75]	48.79% [42.53–54.32]	42.22% [19.00–63.64]	51.01% [28.00–72.65]
Lobster	6.39% [5.36–7.53]	4.07% [2.65–6.24]	11.48% [8.75–14.95]	23.21% [17.08–32.10]	45.49% [23.10–68.63]	34.14% [12.10–57.63]

[a] Numbers in brackets are 95% confidence intervals derived from the bootstrapping of 1,000 observations on the market share from the model estimates using the Krinsky-Robb procedure.

effects might exist. As for lobster, results show that market share grocery stores estimates are statistically significantly higher than CE estimates. Note that the CE encompassed a wide range of prices, but prices in the store were clustered at the high end of that distribution, generating a relatively tight simulated distribution for store prices. The other products experienced a wider range of prices in the store—prawns by design, and shrimp by virtue of the natural change in prices over the time period of the experiment. Other factors that could explain differences were the out-of-ordinary lobster sales that happened during Valentine's Day weekend. We estimated comparisons across CE and grocery store market shares without sales that happened during this weekend. Results are somewhat different. For FPs, there is not a statistically significant difference across grocery store and CE market shares, for all four models. This shows that when controlling for the out-of-ordinary sales on this weekend, the CE correctly predicted market shares for prawns under the four econometric models. However, for shrimp, CE market share is statistically significantly higher than grocery store market share under the RPL model; under the CL, HEV, and MNP there are no statistically significant differences. For lobster, grocery store market share is statistically significantly higher than

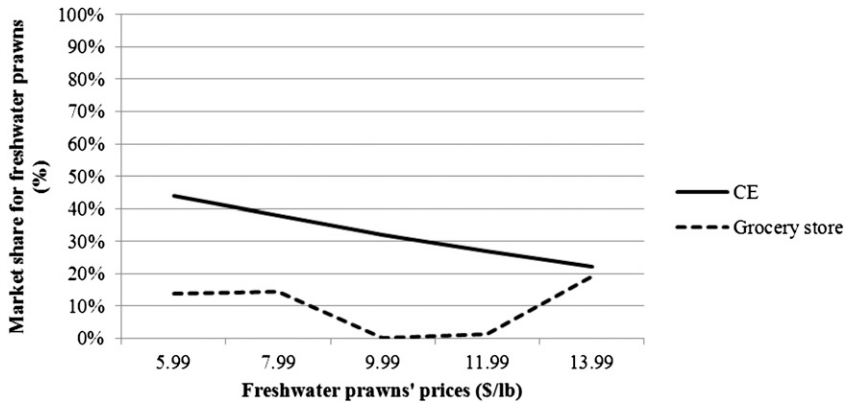

Figure 2. Predicted Market Shares from the CE and Actual Market Shares from the Grocery Store, Freshwater Prawns, Germantown, TN, 2004 (Note: Prices for marine shrimp and lobster are held at the weighted average price observed in the grocery store for the CE market share calculation. Also, weekly sales of prawns at each price level are compared with the weekly average marine shrimp and lobster sales for the grocery store market share)

Table 5. One-Sided *p*-Values from the Combinatorial Method Comparison of Market Share Distributions from the Grocery Store and Mail Survey CE

	Market Share			
Product	CL	RPL	HEV	MNP
Including Valentine's Day Weekend Sales				
Prawns	1.00	0.99	1.00	1.00
Shrimp	0.96	1.00	1.00	0.96
Lobster	0.00	0.00	0.00	0.00
Not Including Valentine's Day Weekend Sales				
Prawns	0.59	0.54	0.59	0.61
Shrimp	0.43	0.98	0.81	0.30
Lobster	0.05	0.02	0.25	0.23

[a] *p*-values represent the *p*-value of a one-sided test of grocery store > mail survey CE market share. The one-sided *p*-value of mail survey CE market share > grocery store is simply $1 - p$-value reported in the table. A two-sided test for statistical differences is simply 2 * *p*-value in the table (Poe, Giraud, and Loomis, 2005).

CE market share under the CL and RPL models while, however under the HEV and MNP models, there are no statistically significant differences. These findings show that lack of control, that is, the holiday behavior, at the grocery store experiment and the econometric model impact the ability of CEs to correctly predict grocery store market shares.

Discussion

Results in this paper somewhat agree with Chang, Lusk, and Norwood (2009) in that CEs poorly predict grocery store market shares. In such a study, albeit all prices were under control, CEs did not perform as well as incentive compatible formats in predicting grocery store market shares; even though such incentive compatible exercises were conducted at a laboratory setting. These findings underscore the importance of incentive compatible mechanisms along with the environment where the elicitation experiment takes place. For example, Lusk, Pruitt, and Norwood (2006) show that frame field experiments, that is, incentive compatible experiments conducted at the grocery store, yield reasonably accurate market share predictions.

Clearly, the level of control and the choice of estimation method seem to influence the likelihood of hypothetical bias. When not controlling for the lobster sales on Valentine's Day weekend, CEs poorly predicted grocery store market share for all three seafood products under the four econometric specifications. However, when controlling for the sales on this weekend, CE accurately predicted the grocery store market share for FPs. However, this was not consistent through all the three products under analyses. Chang, Lusk, and Norwood (2009) noted that most literature on this topic show that relaxing the assumptions of the CL improves in-sample and out-of-sample predictions, but that there is no conclusive evidence on this issue. In our case, test statistics show evidence of heteroskedastic error variances across alternatives and that the IIA assumption held, leading one to conclude that HEV and MNP models are superior to RPL and CL. This is somewhat validated by the out-of-sample validation showing that HEV and MNP yielded market share estimators closer to actual market shares, for shrimp and lobster.

This leads to the question: what extent are results affected by the experimental design, which assumes different price distributions between the two settings. Previous research noted that reference prices could introduce some effects on value elicitation (Drichoutis et al., 2008). Our study suffers control limitations in price setting for shrimp, lobster, and potential substitutes in the grocery store. While the primary investigation centers on prawns, the lack of control on other prices impacted the results. This lack of control obviously had a more pronounced impact on lobster results, and perfect control of all goods would have greatly enhanced overall findings. Although there is no reason to expect different behavior for other types of goods, tests with other products would also enhance generalization of results. Nonetheless, the case we present extends the debate about potential bias into the natural shopping environment and reaches a somewhat similar conclusion as Chang, Lusk, and Norwood (2009).

Another potential shortcoming of this analysis is the relatively small sample size in the grocery store. The study was conducted over a

10-week period (with 5 weeks of observations on the product in question). Although this represents a substantial period of time for an in-store experiment and prawn purchases over this period were comparable to other competing seafood products in the fresh counter, it remains an open question whether longer periods of time would have resulted in different weighted average prices and premiums. There is likely to be some acquaintance effects, as shoppers were unfamiliar to prawns during the first weeks of the study. One can also wonder about seasonality effects; on this, note that the mail survey was conducted at the same time as the grocery store experiment to control for any "seasonal bias" that might exist in consumers' minds.

Conclusions

This paper presents a case study that compares results from two elicitation formats, an in-store field experiment and a mail survey choice experiment (CE). While previous studies have examined the issue of hypothetical bias, this case adds the feature of having a real-world experiment conducted concurrently with the mail survey CE at the same geographical location. Our findings show that CE market share estimates were statistically significantly different from the grocery store market shares. Also, we found that results are sensitive to the choice of estimation method. For the specific case studied, the heteroskedastic extreme value (HEV) and multinominal probit (MNP) models seem to yield more robust results than the conditional logit (CL) and random parameters logit (RPL). In general, these methods assist in functional form choice, but proper choice is contingent on the underlying problem/product being addressed, and, thus, we cannot offer a general definitive conclusion as to the most appropriate model.

This paper underscores the need for additional work in this area. To improve upon this approach, it would be desirable to obtain shoppers' demographic data. Due to the proprietary nature of such data, it may be difficult to obtain, but would certainly allow for a richer analysis of preferences in comparison with hypothetical surveys. Moreover, the long-term design of the experiment poses limitations as it includes learning effects. These effects are observed in the market share for prawns during the first two weeks of the experiment and could be attributed to the fact that little was known about prawns by grocery store patrons. In this sense, CEs appear to predict more accurately market shares after the introduction period. Further research should address this question by separating acquaintance effects at constant prices.[7] In addition, it would enhance the robustness of the study if the grocery store allowed controlling prices of all relevant products. Here, we controlled prawn prices directly, but could only observe prices for other products with no control. This posed problems for shrimp and lobster, leading to decreased confidence in being able to analyze cross-price effects. Given the relatively small market for these seafood products, a similar examination in more widely consumed/lower priced products should be conducted to determine sensitivity to product price and familiarity.

[Received September 2010; Accepted September 2011.]

References

Adamowicz, W., P. Boxall, M. Williams, and J.J. Louviere. "Stated Preference Approaches for Measuring Passive Unit Values: Choice Experiments and Contingent Valuation." *American Journal of Agricultural Economics* 80(1998): 64–75.

Alfnes, F., A.G. Guttormsen, G. Steine, and K. Kolstad. "Consumers' Willingness to Pay for the Color of Salmon: A Choice Experiment with Real Economic Incentives." *American Journal of Agricultural Economics* 88(2006):1050–61.

Bhat, C.R. "A Heteroscedastic Extreme Value Model of Intercity Travel Model Choice." *Transportation Research B* 29(1995):471–83.

Bradley, M., and A. Daly. "Use of the Logit Scaling Approach to Test for Rank-Order and Fatigue Effects in Stated Preference Data." *Transportation* 21(1994):167–84.

Brookshire, D.S., D.L. Coursey, and W.D. Schulze. "The External Validity of Experimental Economics Techniques: Analysis of Demand Behavior." *Economic Inquiry* 25(1987):239–50.

[7] We acknowledge an anonymous reviewer for providing the comment on acquaintance effects.

Carlsson, F., and P. Martinsson. "Do Hypothetical and Actual Marginal Willingness to Pay Differ in Choice Experiments." *Journal of Environmental Economics and Management*, 41(2001):179–92.

———. "Design Techniques for Stated Preference Methods in Health Economics." *Health Economics* 12(2003):281–94.

Chang, J.B., Lusk, J.L., and F.B. Norwood. "How Closely Do Hypothetical Surveys and Laboratory Experiments Predict Field Behavior?" *American Journal of Agricultural Economics* 91(2009):518–34.

Dillman, D.A. *Mail and Telephone Surveys: The Total Design Method.* New York: John Wiley & Sons, 1978.

Ding, M., R. Grewal, and J. Liechty. "Incentive-Aligned Conjoint Analysis." *Journal of Marketing Research* 42(2005):67–93.

Drichoutis, A.C., P. Lazaridis, R.M. Nayga. "The Role of Reference Prices in Experimental Auctions." *Economic Letters* 99(2008):446–48.

Fox, J.A., J.F. Shogren, D.J. Hayes, and J.B. Kliebenstein. "CVM-X: Calibrating Contingent Values with Experimental Auction Markets." *American Journal of Agricultural Economics* 80(1998):455–65.

Harrison, G., and J.A. List. "Field Experiments." *Journal of Economic Literatures* 42(2004): 1009–55.

Huber, J., and K. Zwerina. "The Importance of Utility Balance in Efficient Choice Designs." *Journal of Marketing Research* 33(1996):307–17.

Hudson, D. "Problem Solving and Hypothesis Testing Using Economic Experiments." *Journal of Agricultural and Applied Economics* 35(2003): 337–47.

Hudson, D., D. Hite, L. Seah, and T. Haab. "Telephone Presurveys, Sample Selection, and Non-Response Bias to Mail and Internet Surveys in Economic Research." *Applied Economics Letters* 11(2004):237–40.

Kuhfeld, W.F., R.D. Tobias, and M. Garratt. "Efficient Experimental Design with Marketing Research Applications." *Journal of Marketing Research* 31(1994):545–57.

Lancaster, K.J. "A New Approach to Consumer Theory." *Journal of Political Economy* 74(1966): 132–57.

Levitt, S.D., and J.A. List. "What do Laboratory Experiments Measuring Social Preferences Reveal About the Real World." *Journal of Economic Perspectives* 21(2007):153–74.

List, J.A. "The Behavioralist Meets the Market: Measuring Social Preferences and Reputation Effect in Actual Transactions." *Journal of Political Economy* 114(2006):1–37.

List, J., and C.A. Gallet. "What Experimental Protocol Influence Disparities between Actual and Hypothetical Stated Values." *Environmental and Resource Economics* 20(2001):241–54.

List, J., and J.F. Shogren. "Calibrating the Differences between Actual and Hypothetical Valuations in a Field Experiment." *Journal of Economic Behavior and Organization* 37(1998):193–205.

Little, J., and R. Berrens. "Explaining Disparities between Actual and Hypothetical States Values: Further Investigation Using Meta-Analysis." *Economics Bulletin*, 3(2004):1–13.

Louviere, J.J., D.A. Hensher, and J.D. Swait. *Stated Choice Methods: Analysis and Application*, United Kingdom: Cambridge University Press, 2000.

Lusk, J.L., J.R. Pruitt, and B. Norwood. "External Validity of a Framed Field Experiment." *Economic Letters* 93(2006):285–90.

Lusk, J.L., J. Roosen, and J.A. Fox. "Demand for Beef from Cattle Administered Growth Hormones or Fed Genetically Modified Corn: A Comparison of Consumers in France, Germany, the United Kingdom, and the United States." *American Journal of Agricultural Economics* 85(2003):16–29.

Lusk, J.L., and T.C. Schroeder. "Are Choice Experiments Incentive Compatible: A Test with Quality Differentiated Beef Steaks." *American Journal of Agricultural Economics* 86(2004): 467–82.

Miller, K.M., R. Hofstetter, H. Krohmer, and Z.J. Zhang. "How Should Consumers' Willingness to Pay Be Measured? An Empirical Comparison of State-of-the-Art Approaches." *Journal of Marketing Research* 48(2011):172–84.

Murphy, J.J., P.G. Allen, T.H. Stevens, and D. Weatherhead. "A Meta-Analysis of Hypothetical Bias in Stated Preference Valuation." *Environmental and Resource Economics* 30(2005): 313–25.

Page, A.L., and H.F. Rosenbaum. "Redesigning Product Lines with Conjoint Analysis: How Sunbeam Does It." *Journal of Product Innovation Management* 4(1987):120–37.

Pennings, J., S.H. Irwin, and D.L. Good. "Surveying Farmers: A Case Study." *Review of Agricultural Economics* 24(2002):266–72.

Poe, G.L., K. Giraud, and J.B. Loomis. "Computations Methods for Measuring the Difference of Empirical Distributions." *American Journal of Agricultural Economics* 87(2005): 353–65.

Revelt, D., and K.E. Train. "Mixed Logit with Repeated Choices: Households' Choices of Appliance Efficiency Level." *Review of Economics and Statistics* 80(1998):647–57.

Rose, J., and R. Scarpa. "Experimental Designs for Environmental Valuation with Choice Experiments: A Monte Carlo Investigation." *Australian Journal of Agricultural and Resource Economics* 52(2008):253–82.

Sattler, H., and F. Volckner. "Methods of Measuring Consumers' Willingness to Pay." Research Papers in Marketing and Retailing, Institute of Marketing, Retailing and Management Science, University of Hamburg, Germany, 2002.

Shogren, J.F., J.A. Fox, D.J. Hayes, J. Roosen. "Observed Choices for Food Safety in Retail, Survey, and Auction Markets." *American Journal of Agricultural Economics* 81(1999): 1192–99.

Srinivisan, V. "A Conjunctive-Compensatory Approach to the Self-Explication of Multiattributed Preferences." *Decision Sciences* 19(1988):295–305.

Srinivisan, V., P.G. Flaschsbart, J.S. Dajani, and R.G. Hartley. "Forecasting the Effectiveness of Work-Trip Gasoline Conservation Policies through Conjoint Analysis." *Journal of Marketing* 45(1981):157–72.

Srinivisan, V., and C. Park. "Surprising Robustness of Self-Explicated Approach to Customer Preference Structure Measurement." *Journal of Marketing Research* 34(1997): 286–91.

Street, D.J., and L. Burgess. *The Construction of Optimal Stated Choice Experiments: Theory and Methods*. New York: Wiley, 2007.