Determinants of External Validity in CBC

Paper for publication in the 2003 Sawtooth Software Conference Proceedings

By Bjorn Arenoe, SKIM Analytical

Introduction

Ever since the early days of conjoint analysis, academic researchers have stressed the need for empirical evidence regarding its external validity (Green and Srinivasan, 1978; Wittink and Cattin, 1989; Green and Srinivasan, 1990). Even today, with traditional conjoint methods almost completely replaced by more advanced techniques (like CBC and ACA), the external validity issue remains largely unresolved. Because conjoint analysis is heavily used by managers and investigated by researchers, external validity is of capital interest (Natter, Feurstein and Kehl, 1999).

According to Natter, Feurstein and Kehl (1999), most studies on the validity and performance of conjoint approaches rely on internal validity measures like holdout samples or Monte Carlo Analysis. Also, a number of studies deal with holdout stimuli as a validity measure. Because these methods focus only on the internal validity of the choice tasks, they are unable to determine the success in predicting actual behaviour or market shares. Several papers have recently enriched the field. First of all, two empirical studies (Orme and Heft, 1999; Natter, Feurstein and Kehl, 1999) investigated the effects of using different estimation methods (i.e. Aggregate Logit, Latent Class and ICE) on market share predictions. Secondly, Golanty (1996) proposed a methodology to correct choice model results for unmet methodological assumptions. Finally, Wittink (2000) provided an extensive paper covering a range of factors that potentially influence the external validity of CBC studies. Although these papers contribute to our understanding of external validity, two blind spots remain. Firstly, the number of empirically investigated CBC studies is limited (three in Orme and Heft, 1999; one in Natter, Feurstein and Kehl, 1999). This lack of information makes generalisations of the findings to ‘a population of CBC studies’ very difficult. Secondly, no assessment was made of the performance of Hierarchical Bayes or techniques other than estimation methods (i.e. choice models and methodological corrections).

Objectives

CBC is often concerned with the prediction of market shares. In this context, the external validity of CBC can be defined as the accuracy with which a CBC market simulator predicts these real market shares. The objective of this study is to determine the effects of different CBC techniques on the external validity of CBC. The investigated techniques include three methods to estimate the utility values (Aggregate Logit, Individual Choice Estimation and Hierarchical Bayes), three models to aggregate utilities into predicted respondent choices (First Choice model, Randomised First Choice with only product variability, Randomised First Choice with both product and attribute variability) and two measures to correct for unmet methodological assumptions (weighting respondents by their purchase frequency and weighting estimated product shares by their distribution levels). A total of ten CBC studies were used to assess the effects of using the different techniques. All studies were conducted by Skim Analytical; a Dutch marketing research company specialised in CBC applications.

Measures of validity

Experimental research methods can be validated either internally or externally. Internal validity refers to the ability to attribute an observed effect to a specific variable of interest and not to other factors. In the context of CBC, internal validation often refers to the ability of an estimated model to predict other observations (i.e. holdouts) gathered in the same artificial environment (i.e. the interview)[1]. We see that many authors on CBC techniques use internal validation as the criterion for success for new techniques (Johnson, 1997; Huber, Orme and Miller, 1999; Sentis and Li, 2001).

External validity refers to the accuracy with which a research model makes inferences on the real world phenomenon for which it was designed. External validation assesses whether the research findings can be generalized beyond the research sample and interview situation. In the context of CBC, external validation of the SMRT – CBC market simulator provides an answer to the question whether the predicted choice shares of a set of products are in line with the actual market shares. External validity obviously is an important criterion as it can legitimise the use of CBC for marketing decision-making. Very few authors provide external validation of CBC techniques although many do acknowledge its importance. A proposed reason for this lack of evidence is that organisations have no real incentive to publish such results (Orme and Heft, 1999).

External validity of CBC can be assessed by a comparison of predicted market shares with real market shares. One way to do this, is to simulate a past market situation and compare the predicted shares with the real shares recorded during that time period. This approach is used in this study and in the two other important papers on external validity (Orme and Heft, 1999; Natter, Feurstein and Kehl, 1999). The degree of similarity in this study is recorded with two different measures: the Pearson correlation coefficient (R) between real and predicted shares and the Mean Absolute Error (MAE) between real and predicted shares.

Techniques

Three classes of CBC techniques are represented in this study. E*stimation methods* are the methodologies used for estimating utility values from respondent choices. Aggregate Logit estimates one set of utilities for the whole sample, hereby denying the existence of differences in preference structure between respondents. Individual Choice Estimation (ICE) tries to find a preference model for each individual respondent. The first step in ICE is to group respondents into segments (Latent Classes) that are more or less similar in their preference structure. During the second step, individual respondent utilities are calculated as a weighted sum of segment utilities. As ICE acknowledges heterogeneity in consumer preferences it is generally believed to outperform Aggregate Logit. In this study all ICE solutions are based on ten segments. Hierarchical Bayes (HB) is another

[1] Definitions by courtesy of Dick Wittink.

way to acknowledge heterogeneity in consumer preferences. This method tries to build individual preference models directly from respondent choices, replacing low quality individual information by group information if necessary. In general HB is believed to outperform ICE and Aggregate Logit, especially when the amount of choice information per respondent is limited.

Choice models are the methodologies used to transform utilities into predicted respondent choices. The First Choice model (FC) is the simplest way to predict respondent choices. According to this model, every consumer always chooses the product for which he has the highest predicted utility. In contrast, the Randomised First Choice model acknowledges that respondents sometimes switch to other preferred alternatives. It simulates this behaviour by adding random noise or 'variability' to the product or attribute utilities (Huber, Orme and Miller, 1999). RFC with product variability simulates consumers choosing different products on different occasions typically as a result of inconsistency in evaluating the alternatives. This RFC variant is mathematically equivalent to the Share of Preference (SOP) model. In other words: the Share of Preference model and the RFC model with product variability, although different in their model specifications, are interchangeable. RFC with product *and* attribute variability additionally simulates inconsistency in the relative weights that consumers apply to attributes. RFC with product and attribute variability is thought to generally outperform RFC with only product variability and FC. RFC with only product variability is thought to outperform FC. In order to find the optimal amounts of variability to add to the utilities, grid searches were used in this study (as suggested by Huber, Orme and Miller, 1999). This process took about five full working days to complete for all ten CBC studies.

Correctional measures are procedures that are applied to correct CBC results for unmet methodological assumptions. For instance, CBC assumes that all consumers buy with equal frequencies (every household buys an equal amount of product units during a given time period). Individual respondents' choices should therefore be duplicated proportionally to their purchase frequency. In this study, this is achieved by applying 'respondent weights' in Sawtooth's SMRT where every respondent's weight reflects the number of units that a respondent typically buys during a certain time period. These weights were calculated from a self-reported categorical variable added to the questionnaire. CBC assumes also that all the products in the base case have equal distribution levels. This assumption is obviously not met in the real world. In order to correct this problem predicted shares have to be weighted by their distribution levels and rescaled to unity. This can be achieved by applying 'external effects' in Sawtooth's SMRT. The distribution levels came from ACNielsen data and were defined as 'weighted distribution' levels: product's value sales generated by all resellers of that product as a percentage of the product category's value sales generated by all resellers of that product category. Finally, the assumption of CBC that respondents have equal awareness-levels for all products in a simulated market is typically not met. Although a correction for unequal awareness levels was initially included in the research design it turned out that awareness data was unavailable for most studies.

Hypotheses

In the previous section some brief comments were provided on the expected performance of the techniques relative to each other. This expected behaviour resulted in the following research hypothesis:

With respect to estimation methods:

H1: ICE provides higher external validity than Aggregate Logit.
(Denoted as: ICE > Aggregate Logit).

H2: HB provides higher external validity than Aggregate Logit.
(Denoted as: HB > Aggregate Logit).

H3: HB provides higher external validity than ICE.
(Denoted as: HB > ICE).

With respect to choice models:

H4: RFC with product variability provides higher external validity than FC.
(Denoted as: RFC + P > FC).

H5: RFC with product and attribute variability provides higher external validity than FC. (Denoted as: RFC + P + A > FC).

H6: RFC with product and attribute variability provides higher external validity than RFC with product variability. (RFC + P + A > RFC + P).

With respect to correctional measures:

H7: Using the purchase frequency correction provides higher external validity than not using the purchase frequency correction.
(Denoted as: PF > no PF).

H8: Using the distribution correction provides higher external validity than not using the distribution correction.
(Denoted as: DB > no DB).

Sample and validation data

The sample consists of ten commercially conducted CBC studies involving packaged goods. All the studied products are non-food items. All the interviews were administered by high quality fieldwork agencies using computer assisted personal interviewing (CAPI). Names of brands are disguised for reasons of confidentiality towards clients. All studies were intended to be representative for the consumer population under study. The same is true for the sample of products that makes up the base case in every study. All studies are designed to the best ability of the responsible project managers of SKIM. All studies were conducted in 2001 except for study J that was conducted in 2002. A study only qualified if all the information was available to estimate the effects for all techniques. This includes external information like distribution and purchase frequency measures in order to test propositions P7 and P8. Refer to table 1 for an overview of the design characteristics of each of the studies.

Table 1. Individual study characteristics

Study name	Product category	Country of study	Trade channel[a]	Attributes[b]	Sample size[c]	Base case size[d]	Market covered[e] (%)
A	Shampoo	Thailand	TT	Brand, price, SKU, anti-dandruff (y/n)	495	20	63
B	Shampoo	Thailand	MT	Brand, price, SKU, anti-dandruff (y/n)	909	30	53
C	Liquid surface cleaner	Mexico	Both	Brand, price, SKU, aroma, promotion	785	14	65
D	Fabric softener	Mexico	TT	Brand, price, SKU, promotion	243	12	78
E	Fabric softener	Mexico	MT	Brand, price, SKU, promotion	571	20	90
F	Shampoo	Germany	Both	Brand, price, SKU, anti-dandruff (y/n)	659	29	63
G	Dish washing detergent	Mexico	TT	Brand, price, SKU	302	14	92
H	Dish washing detergent	Mexico	MT	Brand, price, SKU	557	21	84
I	Female care	Brazil	both	Brand, price, SKU, wings (y/n)	962	15	59
J	Laundry detergent	United Kingdom	both	Brand, price, SKU, promotion, variant 1, variant 2, concentration	1566	30	51

[a] MT = Modern Trade; TT = Traditional Trade
[b] Attributes used in the CBC design
[c] Number of respondents
[d] Number of products in the base case
[e] Cumulative market share of the products in the base case

The interpretation of these characteristics is straightforward, except perhaps for the type of outlet channel studied. Each of the CBC studies is typically performed for either traditional trade, modern trade or for both trade types. *Traditional trade channels (TT)* is the term used for department stores, convenience stores, kiosks, etc. *Modern trade channels (MT)* consist of supermarkets and hypermarkets. Analysis of a separate trade channel is achieved by drawing an independent sample of consumers who *usually* buy the studied products through a certain trade channel.

The *real market share* of a product is defined as the unit sales of a product in a studied market as a fraction of the total unit sales of all the products in the studied market. The real market shares used for validation purposes were provided by the client and involve ACNielsen market share listings. These are typically measured through point of sale scanner data or through retail audits. Volume shares were converted to unit shares if necessary. Sales data is aggregated nationally over retailers, over two to three-monthly periods. The aggregation over such time periods is believed to neutralise any disturbing short-term promotional effects. Also the *real prices* during the studied time period were provided by the client.

Methodology

The ten CBC studies are analysed at the individual level. This means that a separate model is constructed for each CBC study, which describes the effects of using the techniques *within that particular CBC study*. An assessment of each hypothesis can now be made by *counting* the number of studies that support it. This limits the evaluation to a qualitative assessment, which is inevitable due to the small sample size (n=10).

The first step is to create a set of dummy variables to code the techniques. The first two columns in table 2 depict all the techniques described earlier. In order to transform all techniques into dummy variables, a base level for each class has to be determined. The base level of a dummy variable can be viewed as the 'default' technique of the class and the effects of the occurrence of the other techniques will be determined relative to the occurrence of the base level. For instance, in order to test hypothesis H1 (ICE > Aggregate Logit), Aggregate Logit has to be defined as the base level. The performance of ICE is now determined relative to that of Aggregate Logit. The last column assigns dummy variables to all techniques that are not base levels. Any dummy variable is assigned the value 0 if it attains the base level and the value 1 if it attains the corresponding technique. The coding used in table 2 is denoted as coding scheme 1.

The problem of coding scheme 1 is that hypothesis H3 (HB > ICE) and H6 (RFC+P+A > RFC+P) cannot be tested. This is because neither of the techniques considered in any one of these propositions is a base level in coding scheme 1. In order to test these two hypotheses we have to apply the alternative dummy variable coding depicted in table 3. This coding is denoted as coding scheme 2. The interpretation of table 3 is analogous to that of table 2.

Table 2. Coding scheme 1 (used for testing H1, H2, H4, H5, H7 and H8)

Class of techniques	Technique	Base level	Dummy variable	Hypothesis to be tested
Estimation method	Aggregate Logit	*		
	ICE		dICE	H1
	HB		dHB	H2
Choice model	FC	*		
	RFC + P		dRFCP	H4
	RFC + P + A		dRFCPA	H5
Purchase frequency weighting	Not applied (no PF)	*		
	Applied (PF)		dPF	H7
Distribution weighting	Not applied (no DB)	*		
	Applied (DB)		dDB	H8

Table 3. Coding scheme 2 (used for testing H3 and H6)

Class of techniques	Technique	Base level	Dummy variable	Hypothesis to be tested
Estimation method	Aggregate Logit		dLogit	
	ICE	*		
	HB		dHB2	H3
Choice model	FC		dFC	
	RFC + P	*		
	RFC + P + A		dRFCPA2	H6
Purchase frequency weighting	Not applied (no PF)	*		
	Applied (PF)		dPF2	
Distribution weighting	Not applied (no DB)	*		
	Applied (DB)		dDB2	

The approach to the analysis is to try and construct a full factorial experimental design with all the techniques. Three estimation methods, three choice models, and the application or absence of two different corrections thus result in 36 unique combinations of techniques (3*3*2*2). However, eight combinations are not possible because Aggregate Logit is not compatible with the First Choice model or with the purchase frequency correction. Therefore, the final design only consisted of 28 combinations of techniques. All 28 combinations were dummy variable coded according to coding scheme 1 and coding scheme 2. This double coding ensures the possibility of testing all hypotheses. Note that such a 'double' table was constructed for each of the ten CBC studies.

Each row in the resulting data matrix represents a unique design alternative. Each design alternative is fully described by either the first set of six dummies (coding scheme 1) or the second set of six dummies (coding scheme 2). The next step is to parameterise a market simulator according to the techniques within each row, thus 'feeding' the market simulator a specific design alternative. Although the real market shares of the products in a base case are fixed within each individual study, the way in which a market simulator *predicts* the corresponding choice shares is not. These choice shares are believed to vary with the use of the different techniques. Consequently, two unique external validity measures (MAE and R) can be calculated for each design alternative in the dataset.

The two measures of validity can each be regressed on the two sets of dummy variables. The resulting models describe the absolute effects on MAE and R when different techniques are applied. The estimation of all models was done by linear regression in SPSS. This assumes an additive relationship between the factors. Furthermore, no interaction effects between the techniques were assumed. Linear regression assumes a normally distributed dependent variable (Berenson and Levine, 1996). R and MAE have some properties that cause them to violate this assumption if they are used as a dependent variable. Because the distribution of the R-values is strongly left skewed, the R-values were transformed with Fisher's z' transformation before entering in the regression[2]. An attempt was made to transform MAE with a logistic transformation (ln [MAE]) but this did not yield satisfactory results. Therefore, no transformation was used for MAE.

As mentioned earlier, the First Choice model as well as the application of purchase frequency weighting is prohibited for the Aggregate Logit model. The interpretation of the estimated effects must therefore be limited to an overall determination of the magnitude of effects. The effects from the regression model are formally estimated as if all estimation methods could be freely combined with all choice models and purchase frequency correction schemes. Admittedly this is not completely methodologically correct. However, this approach was chosen for the strong desire to determine *independent* effects for estimation methods as well as for choice models. The omission of eight alternatives, all estimated with Aggregate Logit, resulted in a strong increase in collinearity between dummy variables dICE and dHB (correlation of R= -0.75; p=0.00; VIF for both dummies: 2.71). Similar, but weaker, effects occurred between the variables dRFCP and dRFCPA (correlation of R=-0.56; p=0.00; VIF for both dummies: 1.52). Between dummies from coding scheme 2, collinearity occurs to a lesser extent.

No correctional action was undertaken because the collinearity did not seem to affect the individual parameter estimates in either of the models (i.e. many models were able to estimate highly significant effects for both dummy variables within each pair of correlating dummy variables). Furthermore, in every model the bivariate correlations between the dummy variables of each correlating pair fell below the commonly used cut-off levels of 0.8 or 0.9 (Mason and Perreault, 1991). Finally, the VIF for neither variable in neither model fell above the absolute level of 10 which would signal harmful collinearity (Mason and Perreault, 1991).

[2] The Fisher z' transformation is defined as: $Z' = 0.5 \ln (1+R / 1-R)$. The final coefficients were converted back into R-values.

In summary, ten datasets were generated according to coding scheme 1 and another ten datasets were generated according to coding scheme 2 (each dataset describes one original study). Each dataset consists of 28 combinations of techniques, 28 corresponding values for the dependent variable MAE and 28 values for the dependent variable Z'. The regression models used for hypotheses H1, H2, H4, H5, H7 and H8 are thus defined for every individual study as:

$$MAE_i = \alpha + \beta_1 dICE_i + \beta_2 dHB_i + \beta_3 dRFCP_i + \beta_4 dRFCPA_i + \beta_5 dPF_i + \beta_6 dDB_i + \varepsilon_i$$

$$Z_i = \alpha + \beta_7 dICE_i + \beta_8 dHB_i + \beta_9 dRFCP_i + \beta_{10} dRFCPA_i + \beta_{11} dPF_i + \beta_{12} dDB_i + \varepsilon_i$$

Where:

i	=	Design alternative where i = {1..28}
MAE_i	=	External validity measured by MAE for study alternative i.
Z_i	=	External validity measured by Z' for study alternative i.
β_1 - β_{12}	=	Unstandardized regression coefficients for the dummy variables that were coded according to data matrix 1
α	=	Intercept
ε_i	=	Error term for study alternative i.

The regression models that were used for hypotheses H3 and H6 are defined for every individual study as:

$$MAE_i = \alpha + \beta_1 dLogit_i + \beta_2 dHB2_i + \beta_3 dFC_i + \beta_4 dRFCPA2_i + \beta_5 dPF2_i + \beta_6 dDB2_i + \varepsilon_i$$

$$Z_i = \alpha + \beta_7 dLogit_i + \beta_8 dHB2_i + \beta_9 dFC_i + \beta_{10} dRFCPA2_i + \beta_{11} dPF2_i + \beta_{12} dDB2_i + \varepsilon_i$$

Where:

i	=	Design alternative where i = {1..28}
MAE_i	=	External validity measured by MAE for study alternative i.
Z_i	=	External validity measured by Z' for study alternative i.
$\beta_1 - \beta_{12}$	=	Unstandardized regression coefficients for the dummy variables that were coded according to data matrix 2
α	=	Intercept
ε_i	=	Error term for study alternative i.

Note that variables dLogit, dFC, dPF2 and dDB2 from coding scheme 2 are discarded after the model has been estimated because they are not relevant to the hypotheses.

The values of the regression coefficients are interpreted as the amount with which the external validity measure increases when the dummy variable switches from the presence of the base level to the presence of the technique (assuming that the other dummies in the six-dummy model remain constant). The medians (m_i) and means (μ_i) of the regression coefficients are indicative for the magnitude of the effects in general. The standard deviations (σ_i) of the regression coefficients give an indication of the stability of these estimates across the studies.

Effects for the dummy variables are estimated for each study independently. The eight hypotheses can thus be accepted or rejected *for each individual study*. A hypothesis is supported by a study if there exists a significant positive (R models) or significant negative (MAE models) effect for the respective dummy variable (at or below the 0.05 significance level). The final assessment of a hypothesis is accomplished by counting the number of studies that show a significant positive (R) or negative (MAE) effect. No hard criteria are formulated for the final rejection or acceptance.

Refer to tables 5 and 6 for an overview of individual model statistics. All models were significant at the 0.01 level. As can be seen, the quality of the models is generally high. However, R^2 values are somewhat artificially inflated because the observations are not independent. The bottom rows in each table show the minimum, maximum, median and mean validity measures observed in all studies as well as standard deviations.

Table 5. Individual model statistics for MAE

	Individual study models									
Statistic	A	B	C	D	E	F	G	H	I	J
R^2	0.960	0.657	0.971	0.810	0.975	0.980	0.995	0.994	0.982	0.975
Std. Error	0.316	0.301	0.158	0.530	0.087	0.032	0.255	0.112	0.089	0.070
F	83.88	6.72	118.51	14.94	121.83	172.98	677.38	605.67	186.20	137.92
Observations: min	1.99	4.42	4.46	4.09	2.82	2.60	2.88	2.62	1.80	1.75
Observations: max	3.41	7.99	7.34	5.56	6.52	3.10	6.19	10.23	3.55	3.07
Observations: median	2.13	5.63	6.13	4.89	3.58	2.85	3.23	6.31	2.73	2.14
Observations: mean	2.33	5.85	5.87	4.73	4.01	2.84	3.93	6.05	2.58	2.18
Observations: std. dev	0.46	1.39	0.83	0.46	1.07	0.20	1.31	3.13	0.57	0.39

Table 6. Individual model statistics for R (R^2, Std. Error and F are based on z' values)

	Individual study models									
Statistic	A	B	C	D	E	F	G	H	I	J
R^2	0.821	0.725	0.947	0.878	0.936	0.958	0.916	0.781	0.972	0.952
Std. Error	0.158	0.132	0.057	0.082	0.043	0.016	0.241	0.166	0.049	0.080
F	16.07	9.24	62.71	25.18	51.22	79.839	38.172	12.49	119.42	70.055
Observations: min	0.27	-0.14	-0.07	0.11	0.49	0.51	-0.30	0.04	0.72	-0.26
Observations: max	0.75	0.60	0.51	0.53	0.84	0.64	0.49	0.98	0.93	0.66
Observations: median	0.59	0.08	0.22	0.31	0.73	0.59	-0.09	0.61	0.87	0.26
Observations: mean	0.56	0.21	0.27	0.35	0.73	0.58	0.08	0.59	0.86	0.27
Observations: std. dev	0.15	0.29	0.20	0.13	0.11	0.05	0.29	0.34	0.07	0.28

Results

Table 7 shows the unstandardized regression coefficients and their p-values needed for the evaluation of hypothesis H1 to H8 for validity measure MAE. All coefficients, as well as the median and mean values for the coefficients, indicate the absolute change of the Mean Absolute Error (in %-points) between real market shares and shares of choice, as a result of a switch from the base level technique to the technique described by the corresponding dummy variable. Note that positive coefficients denote a negative impact on validity, as MAE is a measure of error.

Table 8 shows the unstandardized regression coefficients and their p-values needed for the evaluation of hypothesis H1 to H8 for validity measure R. All coefficients, as well as the median and mean values for the coefficients, indicate the absolute change of the Pearson correlation coefficient between real market shares and shares of choice, as a result of a switch from the base level technique to the technique described by the corresponding dummy variable. Note that positive coefficients denote a positive impact on validity, as R is a measure of linear relationship.

Figure 9 shows the median and mean values of the regression coefficients for both the MAE models (top graph) and R models (bottom graph).

Table 7. Unstandardized regression coefficients and p-values for MAE

Study	dICE[a]		dHB[a]		dHB2[b]		dRFCP[c]		dRFCPA[c]		dRFCPA2[d]		dPF[e]		dDB[f]	
	b	p	b	p	b	p	b	p	b	p	b	p	b	p	b	p
A	0.00	0.99	0.05	0.79	0.06	0.67	-2.81	0.00	-2.81	0.00	0.00	1.00	0.01	0.95	-0.83	0.00
B	-0.55	0.01	-0.14	0.47	0.41	0.00	-0.66	0.00	-0.73	0.00	-0.07	0.60	0.05	0.66	0.03	0.83
C	0.45	0.00	0.32	0.00	-0.12	0.07	-1.01	0.00	-1.01	0.00	0.00	1.00	-0.01	0.84	-1.23	0.00
D	-1.04	0.01	0.35	0.30	1.39	0.00	-1.49	0.00	-1.51	0.00	-0.02	0.95	0.11	0.62	-0.44	0.04
E	-0.08	0.18	-0.20	0.00	-0.12	0.00	-0.67	0.00	-0.68	0.00	-0.01	0.76	-0.04	0.27	-0.65	0.00
F	-0.01	0.59	-0.09	0.00	-0.08	0.00	0.00	0.76	0.00	0.76	-0.01	0.50	0.02	0.14	-0.38	0.00
G	-0.37	0.03	0.07	0.66	0.44	0.00	-0.54	0.00	-0.63	0.00	-0.09	0.45	0.03	0.77	-6.10	0.00
H	0.04	0.55	0.19	0.01	0.15	0.00	-2.79	0.00	-2.81	0.00	-0.02	0.70	0.04	0.40	-0.12	0.01
I	-0.05	0.34	0.47	0.00	0.53	0.00	-0.09	0.05	-0.11	0.02	-0.02	0.61	0.15	0.00	-0.97	0.00
J	0.04	0.38	-0.05	0.32	-0.09	0.01	-0.71	0.00	-0.76	0.00	-0.06	0.09	0.02	0.60	-0.36	0.00
Median:	-0.03		0.06		0.11		-0.69		-0.75		-0.02		0.03		-0.55	
Mean:	-0.16		0.10		0.26		-1.08		-1.11		-0.03		0.04		-1.11	
Std. Dev:	0.41		0.23		0.47		1.00		0.99		0.03		0.06		1.80	

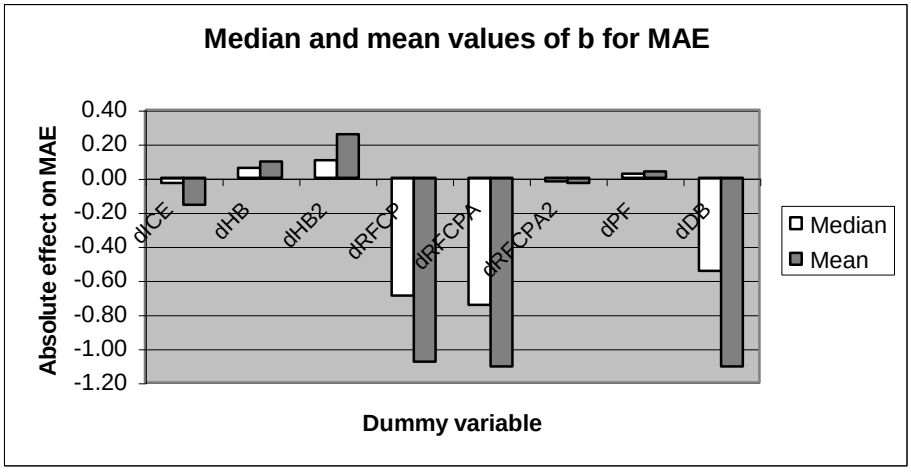
[a] base level: Aggregate Logit
[b] base level: Individual Choice Estimation
[c] base level: First Choice
[d] base level: RFC with product variability
[e] base level: No purchase frequency weighting
[f] base level: No distribution weighting

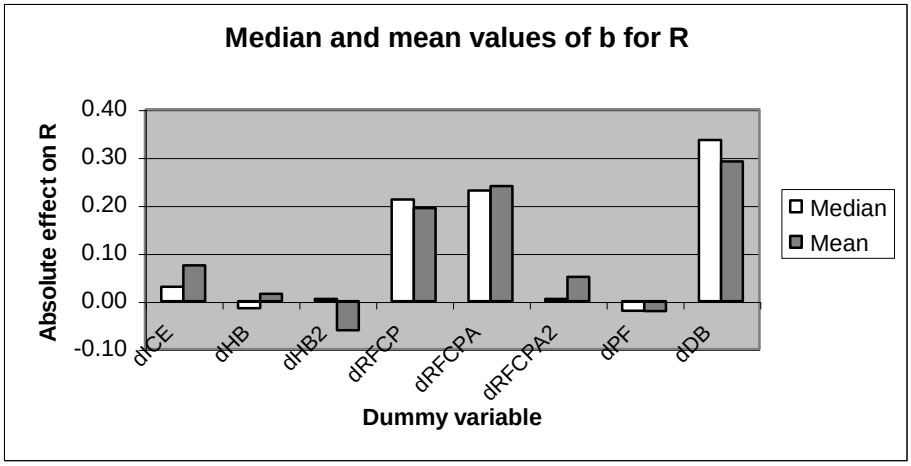
Table 8. Unstandardized regression coefficients and p-values for R

Study	dICE[a]		dHB[a]		dHB2[b]		dRFCP[c]		dRFCPA[c]		dRFCPA2[d]		dPF[e]		dDB[f]	
	b	p	b	p	b	p	b	p	b	p	b	p	b	p	b	p
A	-0.03	0.73	-0.01	0.90	0.02	0.74	0.24	0.00	0.24	0.00	0.00	1.00	-0.03	0.61	0.49	0.00
B	0.45	0.00	0.27	0.00	-0.20	0.00	0.09	0.19	0.21	0.00	0.12	0.05	-0.05	0.39	-0.16	0.00
C	0.03	0.44	0.03	0.37	0.00	0.86	0.30	0.00	0.30	0.00	0.00	1.00	0.03	0.23	0.31	0.00
D	0.30	0.00	-0.04	0.45	-0.34	0.00	0.22	0.00	0.22	0.00	0.00	0.99	-0.03	0.46	-0.03	0.27
E	0.03	0.24	0.08	0.01	0.05	0.01	0.21	0.00	0.21	0.00	0.00	0.92	0.02	0.39	0.21	0.00
F	-0.11	0.00	-0.03	0.01	0.08	0.00	0.00	0.98	0.00	0.98	0.00	0.96	-0.01	0.12	0.09	0.00
G	0.16	0.31	0.17	0.27	0.01	0.90	0.14	0.26	0.31	0.01	0.18	0.11	-0.01	0.90	0.87	0.00
H	-0.13	0.22	-0.16	0.13	-0.03	0.67	0.32	0.00	0.37	0.00	0.06	0.43	-0.06	0.42	0.36	0.00
I	0.11	0.00	-0.14	0.00	-0.24	0.00	0.01	0.59	0.02	0.41	0.01	0.75	-0.06	0.01	0.41	0.00
J	-0.06	0.22	-0.02	0.63	0.04	0.25	0.41	0.00	0.52	0.00	0.14	0.00	-0.01	0.81	0.36	0.00
Median:	0.03		-0.01		0.00		0.21		0.23		0.00		-0.02		0.34	
Mean:	0.07		0.02		-0.06		0.19		0.24		0.05		-0.02		0.29	
Std. Dev:	0.18		0.13		0.14		0.13		0.15		0.07		0.03		0.29	

[a] base level: Aggregate Logit
[b] base level: Individual Choice Estimation
[c] base level: First Choice
[d] base level: RFC with product variability
[e] base level: No purchase frequency weighting
[f] base level: No distribution weighting

Figure 9. Median and mean coefficient values for MAE (top) and R (bottom)





Estimation methods

Table 7 indicates that the use of ICE over Aggregate Logit results in an average decrease in MAE (over all ten studies) of 0.16 %-points. Table 8 indicates that the same change in estimation methods results in an average increase in R of 0.07. The effects of using ICE over Aggregate Logit can thus be regarded as very modest. In the same manner, the average effects of using HB over Aggregate Logit and HB over ICE are very small.

However, although the *average* effects of the estimation methods across the ten studies are modest, the relatively high standard deviations at the bottom rows of tables 7 and 8 indicate large variance *between* the coefficients. In other words: it seems that extreme positive and negative coefficients cancel each other out. If we look for instance at the effect on MAE of using ICE instead of Aggregate Logit, we see a set of coefficients ranging from a low of –1.04%-points to a high of 0.45%-points. This not only indicates that effect sizes vary heavily between studies but also that the direction of the effects (whether increasing or decreasing validity) varies between studies.

The findings with regard to the estimation methods can be considered surprising. Although in theory, ICE and HB are often believed to outperform Aggregate Logit, the empirical evidence suggests that this does not always hold in reality. Also the superiority of HB over ICE in the prediction of real market shares cannot be assumed. In general, there seems to be no clearly superior method that 'wins on all occasions'. The performance of each method instead seems to be different for different studies and is dependent on external factors. Possible factors might be the degree of heterogeneity in consumer preferences or the degree of similarity in product characteristics. It is also believed that the design of the CBC study (number of questions per respondent, number of concepts per task) has an effect on the relative performance of HB over ICE.

Choice models

The use of RFC with product variability over First Choice results in an average decrease in MAE of 1.08 %-points and an absolute increase in R of 0.19. These effects are much more pronounced than any of the average effects of the estimation methods. Furthermore, looking at the individual studies, we can see that the effects are much more stable. Randomised First Choice with product variability (RFC+P) as well as Randomised First Choice with both product and attribute variability (RFC+P+A) outperform First Choice (FC) on most occasions. However, RFC+P+A does not improve external validity much over RFC+P. Because RFC+P is equal to the SOP model, RFC+P+A seems to have limited added value over the much simpler and less time consuming SOP model. The process of determining the optimal amount of product and attribute variability in the RFC model is a tedious process, which does not really seem to pay off. Approximately 95% of the total data generation effort, being around fifty hours, went into the determination of these measures for all ten studies (although some optimising is required for SOP as well).

Note that it is no coincidence that all the effects for the choice models are zero or positive. RFC with only product variability is an extended form of FC where an optimal amount of random variability is determined. If adding variability results in a level of performance worse than FC, the amount of added variability can be set to zero and the RFC model would be equal to the FC model (this actually happens for study F). Hence, RFC can never perform worse than FC. The same holds for the performance of RFC with product and attribute variability over RFC with only product variability.

Purchase frequency correction

The use of purchase frequency weighting actually results in a (small) average decrease in validity (increase MAE of 0.04 %-points; decrease R of 0.02). A possible explanation for this finding is that people really buy different products with approximately equal frequency. However, this assumption seems implausible, as larger package sizes typically take longer to consume. It does also not explain the tendency towards *decreasing* validity. Therefore, a second explanation seems more plausible. Because purchase frequency was measured with a self-reported, categorical variable, it can easily be the case that this variable was not able to capture enough quantitative detail necessary for these purposes. It could thus add more noise than it explains, resulting in decreasing validity.

Distribution correction

The mean coefficients for the use of distribution weighting in the MAE model (-1.11%-points) as well as in the R model (0.29) indicate a strong average increase in external validity. At the level of individual studies, the distribution correction almost always results in an improvement in external validity. However, as with most techniques, the magnitude of the improvement can vary between studies and is dependent on external factors. The decision whether to apply distribution weighting or not can make or break a CBC study as it has the potential of turning an invalid study into an extremely valid one. A good example is study G where applying the distribution correction resulted in a reduction of MAE with more than 6%-points and an increase in R of almost 0.9 (although this is an extreme situation).

Assessment of the hypotheses

A qualitative assessment of the hypotheses can be made by counting the number of studies with a significant negative (MAE) or positive (R) effect for each of the corresponding dummy variables (see table 10). Studies with a significant negative MAE effect or a significant positive R effect indicate an improvement in external validity and hence are considered supportive to the respective hypothesis. Also the number of studies with a significant but opposite effect is reported for each hypothesis.

Table 10. Assessment of hypotheses (cells display number of studies from a total of 10)

Hypotheses	Description	Number of studies:			
		Supporting[a]		Supporting opposite[b]	
		MAE	R	MAE	R
H1	ICE > Logit	3	3	1	1
H2	HB > Logit	2	2	3	2
H3	HB > ICE	3	2	5	3
H4	RFC + p > FC	9	6	0	0
H5	RFC + p + a > FC	9	8	0	0
H6	RFC + p + a > RFC + p	0	2	0	0
H7	PF corr. > no PF corr.	0	0	1	1
H8	DB corr. > no DB corr.	9	8	0	1

[a] Number of studies that show a significant negative effect (MAE models) or positive effect (R models) for the dummy variable corresponding to the hypothesis at or below the 0.05 significance level.

[b] Number of studies that show a significant positive effect (MAE models) or negative effect (R models) for the dummy variable corresponding to the hypothesis at or below the 0.05 significance level.

I will not provide any hard criteria for the assessment of the hypotheses. I believe every reader has to decide for himself what to take away from the summary above. However, I believe it is fair to state that H1, H2, H3, H6 and H7 cannot be confirmed with respect to CBC studies for packaged goods *in general*. Accordingly, H4, H5 and H8 *can* be confirmed for these situations.

Recommendations

It seems that utilities from a CBC study should be estimated with all three methods (Aggregate Logit, ICE and HB) if possible. The market simulations resulting from all three methods should be compared on external validity and the best performing method should be chosen. It is advised to try and relate the performance of the methods to some specific external variables that are known in advance. Finding such a relationship (which makes it possible to exclude certain methods in advance) could save time as some of the methods typically take considerable time to estimate (i.e. HB). Candidates for such variables are measures for the heterogeneity between the respondents or the similarity of the attributes and levels of the products in the base case.

If there are no objections against the RFC model, than it can be used instead of the First Choice model. If there are objections against the RFC model, the Share of Preference model can be used as an alternative to the RFC model. Objections to the RFC model could exist because the model is difficult to understand and because the time needed to find the optimal amount of product and attribute variability is quite considerable.

Weighting respondents' choices by their purchase frequency as measured with categorical variables could actually make the results less valid. It is advisable however to experiment with other kinds of purchase frequency measures (e.g. quantitative measures extracted from panel data). Weighting products' shares of choice by their weighted distribution should always be tried as it almost always improves external validity.

Directions for future research

Future research in the area of external validation of CBC should focus on the following questions. Firstly, what are the determinants of the performance of Aggregate Logit, ICE and HB? Potential determinants include the amount of heterogeneity in consumer preferences, the degree of similarity in product characteristics and study design characteristics like number of choice tasks per respondent.

Secondly, what other factors (besides the techniques investigated in this study) determine external validity? Potential candidates are study design characteristics, sample design characteristics and characteristics of consumers and products in a particular market.

Thirdly, what is the effect of purchase frequency weighting if quantitative instead of qualitative variables are used for the determination of the weights? Consumer panel diaries or POS-level scanning data could perhaps be used to attain more precise purchase frequency measures.

And finally, what are the effects of the investigated techniques for products other than fast moving consumer goods? Because the structure of consumer preference typically differs between product categories, the performance of the techniques is probably different as well. For instance, a decision about the purchase of a car differs considerably from a decision about the purchase of a bottle of shampoo. Because consumers are expected to engage in less variety seeking when it comes to cars, the performance of RFC over FC will probably be less pronounced.

References

Berenson, Mark L. and Levine, David M. (1996), *Basic Business Statistics, Concepts and Applications,* Sixth edition, New Yersey: Prentice Hall, p.736

Golanty, John (1996), 'Using Discrete Choice Modelling to Estimate Market Share', *Marketing Research*, Vol. 7, p. 25

Green, Paul E. and Srinivasan V. (1978), 'Conjoint Analysis in consumer research: issues and outlook', *Journal of Consumer Research,* 5, p.338-357 and 371-376

Green, Paul E. and Srinivasan V. (1990), 'Conjoint Analysis in Marketing: New Developments with Implications for Research and Practice', *Journal of Marketing*, 4, p. 3-19

Huber, Joel, Orme, Bryan and Miller, Richard (1999), 'Dealing with Product Similarity in Conjoint Simulations', *Sawtooth Software Conference Proceedings,* p. 253-266

Johnson, Richard M. (1997), 'Individual Utilities from Choice Data: A New Method', *Sawtooth Software Conference Proceedings,* p. 191-208

Mason, Charlotte H. and Perreault, William D. Jr. (1991), 'Collinearity, Power, and Interpretation of Multiple Regression Analysis', *Journal of Marketing Research*, Volume XXVIII, August, p. 268-80

Natter, Martin, Feurstein, Markus and Leonhard Kehl (1999), 'Forecasting Scanner Data by Choice-Based Conjoint Models', *Sawtooth Software Conference Proceedings,* p. 169-181

Orme, Bryan K. and Mike Heft (1999),'Predicting Actual Sales with CBC: How Capturing Heterogeneity Improves Results', *Sawtooth Software Conference Proceedings,* p. 183-199

Sentis, Keith and Li, Lihua (2001), 'One Size Fits All or Custom Tailored: Which HB Fits Better?', *Sawtooth Software Conference Proceedings, p. 167-175*

Wittink, Dick R. (2000), Predictive Validation of Conjoint Analysis, *Sawtooth Software Conference Proceedings*, p.221-237

Wittink, Dick R., Cattin, Philippe (1989) 'Commercial Use of conjoint Analysis: An Update', *Journal of Marketing*, Vol.53 (July), p. 91-96