New Economic Approaches to Consumer Welfare and Nutrition
A Food & Agricultural Marketing Consortium Conference
Sponsored by the Economic Research Service
January 14-15, 1999

New Economic Approaches to Consumer Welfare and Nutrition

Conference Proceedings

Food and Agricultural Marketing Consortium
Alexandria, Virginia
January 14-15, 1999

Development of Scales to Measure Perceived Performance of the Food System and an Index of Food-Related Welfare.

Spencer Henson and Bruce Traill [1].

1. Introduction:

In talking about consumer welfare, economists employ a rather narrow definition based upon consumer and producer surplus. Changes in the level of consumer and/or producer surplus are brought about by changes in price, themselves often induced by government policy or the exploitation of market power. Knowledge of demand and supply elasticities permits welfare changes (so defined) to be easily measured, and agricultural economists have frequently demonstrated the welfare losses/gains to society from, for example, agricultural policies, oligopoly power in the food industries and food safety and environmental regulations (see for example Connor and Peterson, 1994; Buckwell *et al*, 1984; McNeil, 1980). Yet a casual observation of newspapers, television and the food industry press suggests that there are many non-price issues which excite and/or anger food consumers such as food irradiation, biotechnology, food safety, nutrition and health on the negative side, and choice, convenience, functional foods on the positive.

It is hard to imagine that the economics discipline can construct an analytical framework that is capable of adequately measuring the impact of such factors on consumer welfare. However, the marketing profession has had a long-standing interest in a similar multi-faceted issue, namely consumer satisfaction/dissatisfaction with products and/or marketing services in general. Woodruff and Gardial (1996) examine product-level customer satisfaction and explore a range of qualitative and quantitative methods to identify and assess customer satisfaction with the product itself, the purchase situation and after-sales service. Lundstrom and Lamont (1976) produce a single 'scale' (see below for a detailed discussion of scales) to measure customer discontent with

[1] The authors are respectively Lecturer and Professor in the Centre for Food Economics Research (CeFER), Department of Agricultural and Food Economics, The University of Reading,

the marketing-related practices of businesses. Account is taken of: 1) the product strategies of businesses; (2) business communications and information; (3) the impersonal nature of business and retail institutions; and (4) the broader socio-economic forces that are linked with the business system. These four categories of potential discontent are all presumed to be the manifest outcomes of general discontent and so are combined (i.e. weighted equally) in a single scale. By contrast Gaski and Etzel (1986) take the four basic elements of the marketing mix[2] and attempt to estimate consumer satisfaction with each of these and also their relative importance (i.e. different weights) in arriving at an overall index of consumer sentiment to marketing.

Some of the techniques used in these papers appear promising in addressing the issue of consumer welfare from food. The present article should be seen as a preliminary investigation into the usefulness of such methodologies, supported by a preliminary empirical exercise, but should not be seen as attempting to provide a definitive answer to the question of how much welfare consumers derive from food. Given the limited resources available, the empirical work is insufficiently representative to support any such claims.

2. Scales and indices:

The use of summated scales to represent 'constructs' or 'latent variables' has its origins in psychology, measuring intangible constructs such as 'stress' through individuals' responses to 'items' caused by stress (e.g. feeling depressed, not sleeping well). In fact, much of the earliest work on 'measurement theory' was associated with the testing of mental ability (De Vellis, 1991). However, the concept has been enthusiastically adopted by the market and consumer research professions to measure issues which might affect consumer buying behaviour such as consumer ethnocentrism (Shimp and Sharma, 1987), consumer discontent (Lundstrom and Lamont, 1976), consumer involvement

with products (Laurent and Kapferer, 1985); and exploratory tendencies in consumer behaviour (Raju, 1980)[3].

The fundamental notion is that the construct or latent variable[4] of interest manifests itself in a number of measurable items that may thus be considered to be indicators of the construct. The latent variable is thus regarded as a *cause* of the item score; that is the strength or quantity of the latent variable is presumed to cause a set of items to take on a certain value (DeVellis, 1991). In this way, a construct and its related items are quite different than an index. The latter may be built up from a number of items, but it is the items that cause the index rather than the reverse (Bollen, 1984). For example, an individual's general level of well-being may be caused by income, health and status, but well-being does not cause income, health and status. Thus an *index* of wellbeing could be constructed from (appropriately weighted) measures of income, health and status, but a wellbeing *scale* would be inappropriate.

The question arises, why measure a construct indirectly through summation of a number of item scores which have been caused by the construct rather than directly through the response to the score on a single question aimed directly at the construct in question? (for example, how satisfied are you with food nowadays?). The answer is that the sum (or average) of a number of items should be more accurate than the response to a single question[5] provided, of course, that all of the items used are genuinely caused by the construct in question. Further, the construct of interest is often more abstract than the manifest items it causes and consumers find it difficult to think about such abstract concepts. Much of scale development involves identifying potential items that could represent a construct of interest and excluding items found not to be appropriate.

[3] For a comprehensive listing, see Bearden, Netemeyer and Mobley (1993).
[4] The terms are interchangeable.
[5] For example, measuring IQ on the basis of a single question would be generally considered

In the present research, we begin with the notion that consumers' welfare from food is *caused by* their perceptions of the performance of the food system in relation to various aspects of food including safety, convenience, nutritional quality, taste etc. Furthermore, that these aspects of food may be viewed as psychological constructs which themselves 'cause' specific measurable attitudes. In turn, these attitudes can be measured by eliciting the level of agreement/disagreement with a series of statements[6], the scores from which can be summed to derive scales of the various psychological constructs. As we are also interested in overall welfare, we attempt to discover how important each of these constructs is to the consumer. This information is used to construct a welfare index that is an appropriately weighted sum of the various scale scores[7].

Scale development typically involves five major steps (Spector, 1992). Firstly, the construct(s) of interest must be clearly defined. Secondly, the scale is designed, meaning that a pool of items caused by the construct is identified, the response choices defined and the questionnaire constructed. Thirdly, the scale is piloted and then revised as necessary, usually with the elimination of some items. Fourthly, the survey is elicited and the scale's reliability and unidimensionality tested, followed by purification if necessary. Testing for reliability effectively involves a number of ways of examining correlations between items, which should all be positive given that they are caused by the same construct. Unidimensionality means that all of the items in a scale are caused by a single latent variable, which can be confirmed using factor analysis. Finally, the scale is validated, that is checked to ensure that it measures what it is intended to measure. Again there are a number of standard tests for validity and it is common to include validation questions in the questionnaire.

[6] For example, in relation to convenience, statements such as: 'modern supermarkets are very convenient places to shop' and 'affordable ready-to-eat food is widely available today'.

3. Scale development:

3.1. Construct definition, item pool generation and questionnaire design:

This section describes Stages one to three of the scale development process. As a first attempt to understand the main factors influencing consumer food-related welfare, we read extensively both the academic and industry literatures on consumer concerns about food, changing food consumption patterns and their causes (see for example Senauer *et al.*, 1991; Murcott, 1998; Askegaard and Madsen, 1995; Brunso, 1996; Wierenga *et al.*, 1997; Clancy, 1988; Food Marketing Institute, 1997; Food Marketing Institute, 1998). We were able to arrive at a list of major headings that seemed to us comprehensive and mutually exclusive in describing the main issues to consumers associated with the food system. We then administered an open-ended questionnaire to a convenience, but reasonably representative, sample (n=30) of consumers in the UK (Reading) and USA (Boston). The aim was both to check the validity of the major headings and to generate a pool of items. Respondents were asked to list factors that potentially affected the level of satisfaction obtained from food products under each of these headings. To ensure that our constructs were well defined we checked the items listed under each heading to discover whether any items appeared under more than one, either from the same or different respondents. This led to some refinement to the definition of the constructs[8]. Our final list was of nine constructs as detailed in Table 1.

Table 1. Food-related welfare constructs:

Construct	Examples of Issues
Safety	Standards of hygiene in food manufacture Safety of food additives Availability of information on food safety
Convenience	Convenience of modern supermarkets

[7] It should be highlighted that this part of the research is undertaken through questioning relating to single items (as described below), which we would expect to be less reliable than the scales, and a fruitful avenue for future research.

[8] For example, we originally distinguished between ethical issues such as animal welfare and environmental issues such as recycling, but discovered that consumers did not make this

	Time saved by use of ready-prepared foods Range of places available to eat out
Health and nutrition	Reliability of nutritional claims Availability of nutrition information Nutritional quality of processed food products
Cost	Competition between food suppliers and food prices Prices relative to peoples incomes Relationship between price and food quality
Taste	Taste of food today compared to the past Modern methods of food processing and the taste of food Emphasis of appearance of food over taste in food marketing
Ethical issues	Exploitation of workers in developing countries Animal welfare in food production Recycling of food packaging
Choice of food	Choice of alternative brands of food product Availability of traditional foods Choice of pack size
Behaviour of food companies	Money spent on food advertising Power of food companies over farmers Profits made by major food companies
Where food comes from	Proportion of food imported Labelling of country of origin Availability of food products produced locally

It is desirable to generate a sufficiently large pool of items that a reliable measure of the construct can be obtained after scale purification (which eliminates some items), but small enough not to cause respondent fatigue in the final questionnaire. In our case, we decided on seven items per construct. These were derived from the pool of items provided by respondents to the open-ended questionnaire and the existing literature on consumer-related scales (see for example Hofstede *et al* (1997); Holbrook, 1993; van Trijp *et al*., 1996; Gaski and Etzel, 1986; Oude Ophius, 1989; Mital and Lee, 1989; Steenkamp, 1989; Grunert and Juhl, 1995; Baumgartner and Steenkamp, 1996).

In the main questionnaire, items were grouped under the relevant construct heading. Items took the form of attitudinal statements[9] that respondents were

[9] [illegible]

asked to score on a seven-point Likert scales[10] from 'agree strongly' at one extreme to 'disagree strongly' at the other[11]. Care was taken to avoid unduly lengthy items, double negatives, double-barrelled items and jargon, since these are all potential sources of confusion (De Vellis, 1991). In accordance with recommended good practice, a number of the statements were intentionally 'inverted' to ensure that respondents thought about each question rather than hurriedly gave the same response to each[12]. Prior to data analysis, responses to such questions were reverse-scored to make them consistent with those that were positively-scored.

To assist in the assessment of the validity of the scales, a validation item was included for each construct: 'In general I am satisfied with the <*construct*> of food available today'. Each question was scored on a seven-point Likert scale from agree strongly' at one extreme to 'disagree strongly' at the other. The expectation was of a high correlation between the scales constructed from the individual items and the relevant validation items.

To obtain weights to attach to each of the nine constructs in calculating the overall food-related welfare index, respondents were asked to score each construct in terms of its influence on the total satisfaction obtained from food. For this purpose, respondents were presented with a ten-point interval scale from 'of utmost importance' at one extreme to 'not at all important' at the other. A ten-point rather than a seven-point scale was employed to reduce the tendency for bunching of responses at the top end of the scale.

Finally, the questionnaire was piloted in the UK and the US by a total of 15 consumers, resulting in minor modification and clarification and the replacement of a small number of items.

3.2. Data collection:

[10] For a discussion of Likert scales see Oppenheim (1992).
[11] A copy of the questionnaire is available from the authors on request.

The questionnaire was sent to a random national sample of 1,000 households each in the UK and USA[13]. The valid useable response rate (accounting for households that had gone away and returned incomplete questionnaires) was 33.6 per cent in the case of the UK and 9.7 per cent in the case of the USA. The reason for the relatively poor response in the case of the US sample is unknown, given that resource constraints prohibited a follow-up of non-respondents[14].

To assess the possibility of non-response bias, an extrapolation method was employed based on the notion that persons responding later are likely to be similar to non-respondents (Armstrong and Overton, 1977). On the basis of t-tests for independent samples, there was no significant difference (at the 5 per cent level) between late and early respondents in the UK sample, in terms of a range of socio-economic variables (most notably gender, age, income level and years of full-time education). This suggests that non-response bias was not a problem in the UK sample. It was not, however, possible to test for non-response bias in the case of the US sample.

3.3. Scale reliability and unidimensionality:

Having generated data employing seven items for each construct, the next task was to inspect the data to determine whether any items needed to be excluded from the scales. When developing a scale, the aim is to choose a set of items that most accurately reflect variation in the latent constructs. Not all items that were initially included in the scales will perform as expected and it is important that these items are identified in order to maximise reliability.

There are a number of attributes of the scales that should be examined to identify items that perform badly. Firstly, the sign and magnitude of

---

[12] For example 'I find the smaller number of local food shops today inconvenient'.

[13] Correct English spellings and words were used in the UK version, but were replaced by Americanisms for the US version.

[14] Perhaps a strategic error was made in the letter, proclaiming the value of such information to government in general and the interest of USDA in particular. Could this have run counter to

correlations between the individual items. Given that the scales consist of effect indicators, all items should be positively correlated (Bollen, 1984). Those items that correlate negatively or correlate weakly should be excluded. Secondly, the extent to which individual items correlate with the sum of all remaining items. In general, items with low item-total correlations should be excluded. Thirdly, items should ideally have a high variance, so that they better discriminate between individual respondents, although it is also desirable that the item mean is close to the centre of its range.

Having identified potential items for exclusion, the reliability of the scale should be examined. There are a variety of ways in which reliability can be assessed (see for example De Vellis, 1991; Carmines and Zeller, 1979; Sullivan and Feldman, 1994), the most commonly employed of which is Cronbach's alpha[15]. This measures the proportion of scale variance that is communal, resulting from covariation among the items. Since the construct is presumed to cause each of the item scores, 'good' items are positively correlated and α should be 'high'. If the items in the scale were completely orthogonal, scale variance would equal the sum of the individual item variances and α would take a value of zero (the lower bound). The upper bound for α approaches one, with values above 0.7 generally accepted as demonstrating that a scale is internally consistent (or reliable) (Nunnally, 1978). The aim of scale purification is to obtain a high α, which implies a reliable scale. However, whilst elimination of an item with a low item-total correlation raises α, fewer items in a scale reduces α.

Each construct was examined using the procedures detailed above in an attempt to eliminate superfluous items. For the majority of scales, all seven items performed well and the overall α coefficient was satisfactory. In certain cases, however, most notably the scales for safety and costs, items were excluded. Tables 2 and 3 reports reliability and related statistics for the US and

UK samples. Only one α-value is below 0.7, the safety construct for the UK sample, and this only very marginally at 0.69. Thus the battery of scales may be considered to be internally consistent and reliable for both the US and UK samples.

Table 2. Scale reliability and descriptive statistics for US sample:

Scale	Number of Items	Cronbach Alpha	Mean Scale Score	Standard Deviation	Average Item-Total Correlation
Safety	6	0.71	4.20	0.553	0.556
Convenience	5	0.72	5.22	0.505	0.620
Health & nutrition	7	0.74	3.81	0.678	0.570
Cost	5	0.71	3.56	0.449	0.632
Taste	7	0.77	3.79	0.524	0.626
Ethical issues	6	0.72	3.02	0.636	0.583
Choice	7	0.81	4.71	0.576	0.783
Behaviour of food companies	7	0.75	3.28	0.470	0.605
Where food comes from	6	0.70	3.72	0.574	0.598

Table 3. Scale reliability and descriptive statistics for UK sample:

Scale	Number of Items	Cronbach Alpha	Mean	Standard Deviation	Average Item-Total Correlation
Safety	6	0.69	4.10	0.761	0.488
Convenience	5	0.72	5.27	0.413	0.599
Health & nutrition	7	0.73	3.96	0.639	0.547
Cost	5	0.71	3.48	0.668	0.635
Taste	7	0.79	3.52	0.597	0.661
Ethical issues	6	0.74	2.73	0.638	0.573
Choice	7	0.75	4.72	0.684	0.578
Behaviour of food companies	7	0.74	3.07	0.732	0.564
Where food comes from	6	0.71	3.76	0.709	0.516

The unidimensionality of a scale is assessed by analysing the loading of items onto factors with a view to determining whether one broad or several specific constructs charactise the item set (De Vellis, 1991). This is generally assessed using principal axis factoring to examine whether all items in a scale load onto a single factor[16]. Whilst there are no rigorous criteria that can be applied to assess when factor loadings are significant, it is suggested that a minimum value of around 0.30 to 0.35 indicates that an item loads onto a factor (Spector, 1992).

Unidimensionality was tested for both the US and UK samples using principal axis factoring constrained to a single factor. Given that all items included in each of the purified scales loaded onto a single factor (they had coefficients greater than 0.35), there is evidence to suggest that the scales were indeed unidimensional.

3.4.Validation:

Whereas reliability concerns how much a variable influences a set of items, validity concerns whether the variable is the underlying cause of item covariation. Determining that a scale is reliable does not guarantee that the latent variable shared by the items is, in fact, the variable of interest to the scale developer. Validity is inferred from the manner in which the scale was constructed, its ability to predict specific events, or its relationship to measures of other constructs (DeVellis, 1991; Ghiselli *et al*., 1981). There are three basic types of validity:

- **Content Validity:** This concerns the manner in which a scale is constructed. In theory the items representing a construct should be randomly sampled from the universe of items caused by the construct, which is of course unknown.

- **Predictive validity** : To have predictive validity, a scale is required to be empirical associated with some pre-specified criterion. For example, it might be hypothesised that a high score on satisfaction of the food system in relation to food safety will be associated with lower consumption of organic food[17].

- **Construct Validity** : This is concerned with the theoretical relationship of a variable (in this case a score on a scale) to other variables. It is the extent to which a measures 'behaves' the way that the construct it is hypothesised

[17] An interesting extension of this research would be to analyse the relationship between the

to measure should behave with respect to established measures of other constructs. Thus, if we expect construct X to be positively related to construct A, negatively related to construct B, and unrelated to construct C, a scale designed to measure construct X should have a comparable relationship to measures of constructs, A, B and C (DeVellis, 1991).

Due to resource constraints, no explicit tests were undertaken to test the content and predictive validity of the food-related scales[18]. It is, however, suggested that greater emphasis should be placed on construct validity since this relates directly to the measurement of unobservable and abstract theoretical concepts (Carmines and Zeller, 1979). The construct validity of the scales was examined by assessing convergent, discriminant and nomological validity in turn.

Evidence of convergent validity is provided if a scale correlates highly with other methods designed to measure the same construct (Churchill, 1979). Discriminant validity is indicated when the measure in question has a low correlation with other measures that are hypothesised not to measure the same construct (Heeler and Ray, 1972). The most frequently applied method to assess convergent and discriminant validity is Campbell and Fiske's (1959) multitrait-multimethod matrix (MTMM) (Churchill, 1979; DeVellis 1991; Carmines and Zeller, 1979). The MTMM matrix reports the correlations between different constructs and their alternative measures. Given two or more constructs and two or more ways of measuring each, one would expect high correlations between alternative measures of the same construct, but low correlations between the measures of different constructs.

Although ideally, alternative multi-item scales for the constructs of interest are used to assess convergent and discriminant validity, in the current study single validation items for each construct were applied as alternative measurement

[18] Although evidence of content validity was provided by the comprehensive review of the

instruments in the MTMM matrix. This is in line with the procedure employed by Cardogan *et al* (199?) for example. Tables 5 and 6 report the MTMM matrix for the US and UK samples. Labels 1 to 9 refer to the validation items and 10 to 18 refer to the summated food-related welfare scales[19]. Matrix elements are correlation coefficients. Evidence of convergent and discriminant validity is provided if the correlations between alternative measures of the

[19] **Validity Items:** 1 =Safety; 2 = Convenience; 3= Health and nutrition; 4= Cost; 5= Taste; 6= Ethical issues; 7= Choice; 8 = Behaviour of food companies; 9 = Where food comes from.
Scale Items: 10 = Safety; 11 = Convenience; 12 = Health and nutrition; 13 = Cost; 14 = Taste; 15= Ethical; issues; 16 = Choice; 17 = Behaviour of food companies; 18 = Where food comes from.