



AgEcon SEARCH
RESEARCH IN AGRICULTURAL & APPLIED ECONOMICS

The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search
<http://ageconsearch.umn.edu>
aesearch@umn.edu

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

Principles of Principal Component Analysis

Catherine A. Durham and Robert P. King

With increasing frequency consumer studies are supplementing demographic and price variables with responses to an extended set of Likert-scale questions to elicit information on consumer motivations and attitudes. Principal component analysis (PCA) is a statistical tool that reduces a large number of variables to a smaller set of "components" that describe as much as possible of the variation in the original variables. Attitudinal responses can then be represented by component scores in statistical models. This paper reviews fundamental principles of PCA and concludes with a proposal for collaborative efforts to standardize attitudinal questions and PCA of responses across studies.

Researchers interested in factors affecting consumer purchase decisions or willingness to pay for new product attributes often use extended sets of questions designed to elicit information on environmental, ethical, or health attitudes and motivations. Typically, these questions have Likert-scale responses, and it is not unusual for researchers to include from 20 to 30 such questions in a survey instrument. Respondents find them easy to answer, and they can be a valuable source of nuanced information on motivations that, while not easily observed, can have a profound effect on consumer decisions.

Including responses to such questions individually in a statistical analysis can pose difficult challenges, however. Even when the number of respondents is large enough to overcome limits on degrees of freedom, responses are often highly correlated across questions, and interpretation of parameter estimates associated with each question can be problematic. One way to address this problem is to cluster questions and create arbitrary "scores" that are simple sums or averages of Likert-scale responses for each cluster. A more formal—and we believe superior—alternative is to use principal component analysis (PCA) to analyze relationships among responses and estimate component weights that can be used to construct component scores.

PCA is a statistical tool for data reduction (Garson 2009). Standard PCA methods establish a

procedure for reducing a large number of variables to a smaller set of "components" that describe a known portion of the total variation in the original variables. Each component includes a cluster of variables, not chosen arbitrarily but identified as a group based on statistical association. Component weights estimated in the PCA process can be used to construct component scores that are likely to have more information content than arbitrary scores that are sums or averages of responses for a cluster of questions.

"Principal component analysis" and "factor analysis" are terms that are often used interchangeably, but this can lead to confusion. "Principal component analysis" and "principal factor analysis" are alternative methods for extracting components or factors from a set of data. Though similar, the method of extracting factors produces different representations of the amount of variation in the data set that they explain, based on intended use. As Garson notes (2009, p. 5), principal factor analysis "seeks the least number of factors that can account for the common variance (correlation) of a set of variables, whereas the more common principal components analysis (PCA) in its full form seeks the set of factors which can account for all the common and unique (specific plus error) variance in a set of variables." Garson goes on to note that PCA is generally preferred for data reduction, and it is the method used in this report.

PCA has been used in several recent studies. For example, Johnston et al. (2001) used this approach to evaluate the impact of environmental interests on ecolabel preferences, and Durham (2007) used it to examine the impact of health concerns and environmental attitudes on organic preferences. These studies draw upon Roberts (1996), who segmented consumers for their environmental orientations, and

Durham is Associate Professor, Food Innovation Center, Department of Agricultural and Resource Economics, Oregon State University. King is Professor, Department of Applied Economics, University of Minnesota.

This research was partially supported by the National Research Initiative of the Cooperative State Research, Education, and Extension Service, USDA, Grant #2005-35400-15240.

Kraft and Goodell (1993) who evaluated consumers' wellness orientations.

In this report we use our experiences in conducting a PCA of responses to questions on consumer attitudes collected in intercept surveys that were part of a larger study on the value of ecolabels in food marketing as the basis for a general review of PCA procedures. In the sections that follow, we first introduce the survey questions used in our analysis and present our PCA results. We then conclude with a discussion of the potential benefits from more widespread collaboration among researchers in developing standard sets of attitudinal questions and in using pooled data sets to conduct PCA of consumer responses to standard questions and construct component scores using commonly developed weighting coefficients. It should be noted at the outset that this is not a primer on PCA. For that the reader is referred to Garson's online introduction or to any of a large number of excellent texts on multivariate analysis (e.g., Johnson and Wichern 2008; Hair et al. 2006).

An Illustration of Principal Component Analysis

We illustrate PCA using data collected in consumer intercept surveys during the summer and fall of 2006 in Minnesota, Oregon, and Rhode Island. In each state, shoppers in a variety of market settings—supermarkets, farmers markets, and natural food stores—were asked to complete a survey on factors that influence food selection. The survey included questions on shopping behavior; preferences for conventional, organic, and ecolabel products across 15 product categories; preferences for buying locally produced foods and the definition of "local"; attitudes on environmental, health, and food-policy issues; and demographic information. It also included sequences of questions designed to elicit willingness to pay for ecolabel products. See Durham, King, and Roheim (2009) for a more complete description of the survey and a presentation of results for consumer definitions of "local" for fresh fruits and vegetables.

Following Garson, the first step in our PCA was to estimate an initial component matrix for which the number of components will equal the number of variables. At this point it was necessary to decide whether to include all 24 variables in a single

PCA or to separate the variables based on prior experience and *a priori* expectations. Including all of the variables in a single PCA has the advantage of letting the data reveal patterns that might not be expected, but this can also lead to spurious results if the data set is small. Creating several initial groups of variables and conducting the PCA with each set has the advantage of reflecting the design of the survey instrument and building on findings from previous studies, but this procedure may impose structure on the data that is not supported empirically. We tried both approaches.

We also needed to decide whether thirteen experimental questions that had not been used in previous studies should be included in the analysis. All of the questions were retained for the unified analysis for ungrouped variables. Four were eventually dropped from the analysis for distinct groups of questions. These were:

- I donate money to support farmland preservation.
- I make donations to wildlife protection organizations (e.g. Audubon Society, World Wildlife Fund).
- I am concerned about the welfare of domestic farm animals.
- I am an avid fisherman and/or hunter.

Decisions about whether to retain the experimental questions were based on evaluation of their contribution to the components representing the motivations they were designed to represent and their response distributions.

A single component matrix for the 20 questions in Table 1 plus the four deleted questions was estimated under the unified approach. Individual component matrices were estimated for each of the three sets of questions under the grouped question approach. These sets are shown in the following order in Table 1:

1. A set of questions related to non-personal beliefs about environmental concerns, biodiversity, and wildlife (first nine questions),
2. A set of questions representing personal beliefs about food and health (middle seven), and
3. A set of questions about farming and farm labor (final four).

Table 1. Attitudinal Questions Groupings and Rotated Component Loadings.

Questions (on five-point scale)		Type	Factor	
Wildlife preservation and environmentalism	I buy environmentally friendly products even if they are more expensive.	<u>A</u>	.797	.195
	I have switched products for environmental reasons.	<u>A</u>	.806	.258
	I have convinced family/friends not to buy env. harmful goods.	<u>A</u>	.783	.184
	I will not buy from a company if it is ecologically irresponsible.	<u>A</u>	.747	.251
	I have purchased products because they cause less pollution.	<u>A</u>	.809	.222
	I try to buy only products that can be recycled.	<u>A</u>	.673	.222
	I do not buy household products that harm the environment.	<u>A</u>	.688	.212
	Preserving all plant and animal species is important.	<u>A</u>	.235	.856
Food aficionado & health concerns	I would vote for referendums/initiatives to preserve wildlife habitat.	<u>A</u>	.255	.845
	I worry that there are harmful chemicals in my food.	<u>A</u>	.777	.164
	I avoid foods containing nitrates and preservatives.	<u>T</u>	.742	.143
	I am interested in information about my health.	<u>T</u>	.701	.167
	I'm concerned about my drinking water quality.	<u>A</u>	.668	.054
	I look for new types of food to try.	<u>T</u>	.163	.904
	I go out of my way for new food experiences.	<u>T</u>	.188	.879
	I enjoy magazines about food.	<u>T</u>	.092	.660
Farm preservation & farm labor	I'm concerned about wages received by farm laborers in other countries.	<u>A</u>	.891	.264
	I'm concerned about working conditions for farm laborers in the US.	<u>A</u>	.871	.305
	I would vote for referendums/initiative to preserve farmland.	<u>A</u>	.236	.852
	I'm concerned about the survival of family farms in the US.	<u>A</u>	.303	.801

T-type responses: always true, mostly true, sometimes true, rarely true, never true.

A-type responses: strongly agree, agree, neither agree or disagree, disagree, strongly disagree.

The second step in our analysis involved determining the number of components to retain. We used several criteria, including retaining components for which the eigenvalue is greater than one, retaining components up to the point where a Scree plot of eigenvalues flattens out, and retaining components that explain more than some cut-off percentage of the overall variance in the data. As Garson notes (2009, p. 10):

The eigenvalue for a given factor measures the variance in all the variables which is accounted for by that factor. . . . If a factor has a low eigenvalue, then it is contributing little to the explanation of variances in the variables

and may be ignored as redundant with more important factors.

In general, large component loadings cluster in a way that makes it possible to give each component a descriptive name, though clearly there is some subjectivity in the interpretation process.

The third step in a PCA involved re-estimation of the component matrix with a restricted number of components for all questions or for each group of questions. This was followed by a fourth step, rotation of the resulting component matrix or matrices. We used varimax rotation. As Garson notes (2009, p. 15):

Varimax rotation is an orthogonal rotation of the factor axes to maximize the variance of the squared loadings of a factor (column) on all the variables (rows) in a factor matrix, which has the effect of differentiating the original variables by extracted factor. Each factor will tend to have either large or small loadings of any particular variable. A varimax solution yields results which make it as easy as possible to identify each variable with a single factor. This is the most common rotation option.

The matrices produced from the varimax rotation were the basis for interpreting the components. The PCA based on *a priori* groups of questions yielded six factors, which were interpreted based on the questions contributing the highest rotated component loadings (shown with shading in Table 2) within each of the three groups of questions. They are "Environmentalism" and "Wildlife Preservation" from the first question grouping; "Health Concerns" and "Food Aficionado" from the second grouping; and "Farm Labor", and "Farm Land" from the final grouping. The PCA for the unified analysis that included all the questions yielded six factors that we interpret as being associated with "Environmentalism," "Political Activism," "Food Aficionado," "Willing to Donate," "Health Concerns," and "Hunter." The variables that contribute most highly to the Environmentalism, Health Concerns, and Food Aficionado components are identical for the two analyses. The questions that produced the Wildlife Preservation, Farm Land, and Farm Labor components from the analysis based on grouped sets of questions are all included in the Political Activism component of the analysis based on the entire data set along with the question about animal welfare that was deleted from the grouped question analysis. The Willing to Donate and Hunter components from the unified analysis are composed of the other three variables that were deleted from the grouped question analysis.

The final phase of our PCA analysis was to compute component scores to be used as explanatory variables in subsequent analysis. Under each approach, component scores were calculated for each consumer's set of responses by multiplying the standardized response for each question by a coefficient calculated from the rotated component

loading for the component to which it is assigned and then summing the resulting products across all the questions assigned in the PCA.

The results based on the PCA with questions clustered into three groups were selected for use in further analysis. These component scores fulfilled the needs of the ecolabel analysis, as they were designed to do. Furthermore, the components from this grouped analysis were consistent with those found in earlier work.

We had two important concerns about both sets of PCA results. First, each included some components composed of only one or two of the original variables. A general rule of thumb in PCA is that components should comprised at least three variables. As noted earlier, several of the questions included in this survey were "experimental" in the sense that they had not been used in other studies. They may not have been as clear and discriminating as questions that have been more thoroughly tested, and this can lead to problems. In preliminary analysis with grouped questions, the experimental donation questions did not work as expected since they correlated with each other rather than with their expected components. Similarly, the hunter question yielded a single variable component in the analysis that included all questions. We recommend that additional questions be designed and tested that expand on some areas of interest such as farm preservation.

Collaboration on the Design and Analysis of Questions on Consumer Attitudes

Our experience in conducting this PCA raised several questions that we believe have relevance well beyond the limits of a single study. These include:

- What and how many attitudinal questions should be included in consumer surveys?
- How stable will PCA results be across studies?
- How much *a priori* structure should be imposed at the start of a PCA?
- How should experimental questions be handled?

Thinking about these questions led to the conclusion that there may be great value from collaboration among researchers in the design of

attitudinal questions and in pooling response data sets for PCA.

The most obvious benefit from collaboration and cooperation in the design of attitudinal questions for consumer surveys is that it can greatly facilitate the comparison of results across surveys. When surveys include different sets of attitudinal questions, it is not possible to determine whether similarities and differences in findings reflect empirical reality or are the result of an omitted or additional question. Clearly the set of relevant attitudinal questions will evolve as new issues and concerns emerge and as theories of consumer choice develop. However, if questions can evolve slowly with as much consistency as possible being maintained across studies, cross-study comparability will be maximized. Collaboration on survey question design and on decisions about changes in a set of standard questions can be accomplished through a number of mechanisms, such as regional research projects and working groups in professional associations such as the Food Distribution Research Society.

There are also clear benefits from pooling survey response data across studies in order to make PCA results more robust. As is true for any statistical procedure, PCA results are subject to sampling error that declines as sample size increases. This will affect both the identification of components and the estimates of component loadings that are used to compute component scores. Replicating attitudinal questions across studies and conducting PCA on pooled data sets can reduce sampling error and can yield common component definitions and component loadings. This is especially important for small studies for which the sample size is not large enough to conduct a reliable PCA. In such cases, the component definitions, means and standard deviations for standardizing Likert questions, and component score coefficients from a large pooled data set can be used to construct component scores for responses to questions in the small study. This can only be done, however, if the small study uses a standard set of attitudinal questions and if results from a PCA conducted for a large sample of responses to the standard questions are available.

Pooling responses to a standard set of attitudinal questions across studies can be difficult. There are confidentiality and human subjects concerns, but these can be addressed if care is taken to separate

the attitudinal response data that will be pooled from other survey responses. There is no need for subject identifiers or other demographic information when conducting a PCA. However, it may be helpful to retain some identifier for source study for data to be pooled. This would make it possible to determine if PCA results are changing systematically over time or if they differ across regions. Another challenge to pooling attitudinal response data will be in determining who has the right to conduct the PCA, but this is an issue that can be resolved within a well-functioning collaborative group of researchers.

In conclusion, we believe this research report can be a starting point for further discussion on the important issue of collaboration and cooperation in survey design and data analysis.

References

- Durham, C. A. 2007. "The Impact of Environmental and Health Motivations on the Organic Share of Produce Purchases." *Agricultural and Resource Economics Review* 36:304–320.
- Durham, C. A., R. P. King, and C. A. Roheim. 2009. "Consumer Definitions of "Locally Grown" for Fresh Fruits and Vegetables." *Journal of Food Distribution Research* 40:56–62.
- Garson, G. D. 2009. "Factor Analysis" from *Statnotes: Topics in Multivariate Analysis*. <http://faculty.chass.ncsu.edu/garson/pa765/statnote.htm> on 10/31/2009.
- Hair, J. F., B. Black, B. Babin, R. E. Anderson, and R. L. Tatham. 2006. *Multivariate Data Analysis*, 6th ed. Upper Saddle River, NJ: Prentice Hall.
- Johnson, R. A. and D. W. Wichern. 2008. *Applied Multivariate Statistical Analysis*, 6th ed. Upper Saddle River, NJ: Prentice Hall.
- Johnston, R. J., C. R. Wessells, H. Donath, and F. Asche. 2001. "A Contingent Choice Analysis of Ecolabeled Seafood: Comparing Consumer Preferences in the United States and Norway." *Journal of Agricultural and Resource Economics* 26:20–39.
- Kraft, F. B. and P. W. Goodell. 1993. "Identifying the Health Conscious Consumer." *Journal of Health Care Marketing* 13(3):18–25.
- Roberts, J. A. 1996. "Green Consumers in the 1990s: Profile and Implications for Advertising." *Journal of Business Research* 36:217–232.