Using Factor Analysis and Cronbach's Alpha to Ascertain Relationships Between Questions of a Dietary Behavior Questionnaire

Eric Grau
Mathematica Policy Research, 600 Alexander Park, Princeton, NJ, 08543

Abstract

Obesity and other dietary problems make it necessary to have a better understanding of dietary behavior and more effective nutrition education. A dietary behavior questionnaire was developed to measure outcomes of nutrition education as part of an effort to develop a standardized, flexible data collection tool. This questionnaire, which was separated into modules according to food groups, was field tested by Mathematica for internal consistency of responses to survey questions and the performance characteristics of individual and sets of questions. The field test data analysis identified questions that performed well and should be retained and some that performed poorly and should be either dropped or need further study. In this paper, we discuss the use of factor analysis and Cronbach's alpha to decipher the internal consistency of and relationships between questions within modules.

KEY WORDS: factor analysis, Cronbach's alpha, dietary behavior

1. Introduction

A draft questionnaire, called a Dietary Behavior Questionnaire (DBQ), was developed by Abt Associates for the purposes of assessing respondents' adherence to the Dietary Guidelines for Americans, as developed by the U.S. Department of Agriculture (USDA).[1] Mathematica Policy Research (MPR) was responsible for assessing how well the questionnaire worked in a real world setting, determining: (1) the ease with which interviewers were able to administer the instrument; (2) the ease with which respondents were able to complete the questionnaire; (3) the clarity and accuracy of the questionnaire instructions; and (4) the length of the questionnaire. Part of the process of answering these questions involved determining the level of redundancy in the questionnaire, which is the focus of this document. Factor analysis was used to identify common components among sets of items within the questionnaire. Cronbach's alpha is a coefficient that describes how well a group of items focuses on a single idea or construct (Cronbach, 1951). Cronbach's alpha is calculated among the set of variables used in the factor analysis to determine the reliability of those questions for measuring a single construct.

In order to assess this questionnaire, MPR conducted a field test, administering the instrument to 453 white, African American, and English-speaking Hispanic women who received food stamps from urban, rural, and suburban areas across the United States.

This document describes the application of factor analysis and Cronbach's alpha on the data from this field test. Section 2 describes the DBQ and how the questions in the DBQ are categorized into groups to be analyzed. The methodologies used, factor analysis and Cronbach's alpha, are described in Section 3. Results for intake modules are presented in Section 4, and for supplemental modules dealing with intake and behavior related to food, diet, and health in Section 5. Results presented in Sections 4 and 5 are limited to those modules where redundancies are apparent or other noteworthy patterns are observed. Section 6 concludes with a discussion of the results.

2. Questionnaire Description

In the DBQ, questions were asked about diet choices for all food groups. These questions specifically asked about consumption of various food groups either in the past seven days on a per-week or per-day basis, or on an ordinal scale of usually, sometimes, rarely or never. For the questions on a per-week or per-day basis, the respondent had the option of answering on a per-week or per-day basis; per-day responses were converted to total weekly consumption. Other questions asked about behavior and attitudes associated with foods. To ensure comparisons on the same scale, all variables were rescaled to have a zero mean and unit variance.

Questions about consumption of various foods were grouped into modules, within which it is desired to determine the relationship between the questions. Supplemental modules included questions about

[1] See http://www.health.gov/dietaryguidelines/ for more information on the Dietary Guidelines for Americans.

attitudes and behaviors associated with food, diet, and/or health. The layout of the modules, with the subject matter for each question within each module, is given in Table 1.

3. Methodology

3.1 Factor Analysis

The basic goal of factor analysis is to describe a set of p random variables in terms of a smaller number ($m<p$) of unobserved constructs, called "factors," which are determined by interpreting coefficients in a factor model, called "loadings."

The common factor model can be described as

$$X_i = a_{i1}F_1 + a_{i2}F_2 + ... + a_{i\,m}F_m$$

where X_i is the ith variable, a_{ij} is the jth factor loading for the ith variable, and F_1, F_2, ... F_m are the uncorrelated common factors. The square of the factor loading a_{ij} is the proportion of the variance of X_i that is explained by the factor F_j. The variance of ith variable can be split into two components, one corresponding to the variance specific to that variable (the "specific variance" or "unique variance") and a variance that is common to all variables ("the common variance"), in the form of the m factors. The estimate of this second component is the "communality," the sum of the squared factor loadings across the m factors for the variable in question.

The partial pairwise correlations between the variables after controlling for all other variables should be small compared to the original correlations; this indicates that the common factor model can do a good job of explaining the overall variation. However, if the partial pairwise correlations differ little from the original correlations, or worse are actually larger in absolute value, then this could be an indication that the common factor model is not appropriate for the data. Variables must be removed and/or added to the set included in the factor analysis to improve the factor model. Kaiser's Measure of Sampling Adequacy (MSA) is a summary of how much smaller the partial correlations are from the original correlations (Kaiser 1970; Kaiser and Rice 1974; Cerny and Kaiser 1977). This measure is calculated for each variable, and overall. For values under 0.7, which is not uncommon for these data, we review the MSAs for individual variables to determine if any of the included variables are candidates for exclusion from the common factor model.

Table 1. Subjects within Module

Module	Subjects within Module
Fruit, vegetables, and French Fries	Fruit, unsweetened fruit juice, all vegetables, potatoes, French Fries, dark green vegetables, orange vegetables
Fruit and vegetables	Fruit, unsweetened fruit juice, all vegetables, potatoes, dark green vegetables, orange vegetables
Dairy and calcium-enriched foods	Milk as beverage, milk on cereal, yogurt, cheese, calcium fortified juice or soy milk
Whole grain foods	Processed whole grain breakfast cereal, whole grain bread, brown rice/whole grain pasta
High protein foods	Poultry, red meat/pork, deli meats, fish, eggs, peanut butter, dry beans
Discretionary fats	Butter/margarine used on vegetables, skim milk, butter/margarine used on bread/pasta/tortillas, fried poultry, remove skin from poultry, trim fat from red meat/pork, drain fat when cooking hamburger, fat used when cooking, French fries, chips/cheese puffs/pork rinds
High sodium foods	Chips/cheese puffs/pork rinds, crackers/pretzels, salt as seasoning
Weight control	Fruit as dessert, fruit and vegetables as snacks, sweetened fruit drinks, soda, fast food, healthier diet past 12 months, lose weight past 12 months, snack or eat meals in front of television, eat breakfast
Shopping	Plan meals before shopping, use list when shopping, look at nutrition labels when buying product
Attitudes	Health status, healthy food too expensive, too busy to eat healthy, healthy food tastes bad, family says healthy food tastes bad, born to be fat or thin
Availability	Available at home: fruit, dark green vegetables, orange vegetables, chips/cheese puffs/pork rinds, candy, skim milk, soda/sweetened fruit drinks, whole wheat bread, whole grain cereal

There are several methods used to estimate the factors, each of which are discussed in detail in Johnson and Wichern (1981). The results presented in this document are taken from a principle factor analysis; results from maximum likelihood estimates of the factors are also compared. With each of these methods, factor analysis does not provide unique solutions. Hence it is possible to rotate the factors to obtain more easily interpretable factor loadings. In all cases with the analyses here, an orthogonal rotation was sufficient to obtain easily interpretable factors. In effect, an orthogonal transformation of the factor loadings corresponds to a rotation of the coordinate axes. The communalities and specific variances will remain unchanged. In these analyses, the VARIMAX rotation (Kaiser, 1958) was used, whereby the variances of the column vectors (corresponding to the squared loadings of the each factor) of the factor matrix are maximized. This forces coefficients to be either large or negligible in any column (associated with a given factor) of the rotated loadings matrix

The factor analysis model assumes that the variables to be investigated have a multivariate normal distribution. With the positive skewness that is evident with the weekly consumption variables (which make up the majority of questions within each module), and the four-category responses for the usually-sometimes-rarely-never questions, this assumption will obviously be violated. However, transforming the weekly consumption variables to dampen the skewness will give us variables that more closely approximate the normal distribution. In particular, a square-root transformation was used to dampen the positive skewness. The four-category response variables were in an ordinal scale; truly an approximation to a normal distribution cannot be obtained for these variables. Each of these variables will be discussed on a module-by-module basis.

3.2 Cronbach's Alpha

Cronbach's alpha is generally used as a measure of the reliability of a set of questions in a survey instrument. It measures the interrelatedness of a set of items, although a high value for alpha does not imply unidimensionality (where the items measure a single latent construct). It was first named as alpha by Cronbach (1951). The formula for Cronbach's alpha, as parameterized by Cortina (1993) is

$$N^2\left(\frac{Mean(Cov)}{Sum(Var/Cov)}\right)$$

where N is the number of items in the scale, Mean(Cov) is the mean inter-item covariance, and Sum(Var/Cov) is the sum of all the elements in the variance-covariance matrix. Standardized alpha is equivalent to the above, except that the average inter-item correlation replaces the average covariance and the sum of the correlation matrix replaces the sum of the variance-covariance matrix. In the research discussed in this document, Cronbach's alpha and its standardized form are equivalent, since all variables are rescaled to have zero mean and unit variance.

A level of alpha that indicates an "acceptable" level of reliability has traditionally been 0.70 or higher, although interpretation of alpha in specific contexts is generally more complicated than that. In particular, a high alpha is possible even though the item responses are multidimensional (Schmitt, 1996), and level of alpha is also related to the number of items being tested. Cortina (1993) showed that how the value of alpha varied according to the number of items being tested, and how alpha generally declined as the number of dimensions increased. He did indicate that, although a high level of alpha does not guarantee unidimensionality, nor does it necessarily indicate high average item intercorrelations, a low level of alpha is often associated with multidimensional data.

4. Results for Intake Modules

The intake modules consist of the fruit and vegetable modules (with and without French fries), dairy and calcium-enriched products, whole grain foods, high protein foods, discretionary fats (with and without French fries and chips/cheese puffs/pork rinds), and high sodium foods. For all of the intake modules, the factor analysis, using a VARIMAX rotation, provided a clear pattern where underlying constructs could be discerned. However, many of these constructs, while interesting, did not lead to conclusions that redundancies existed in the given module. Discussion here is limited to fruits and vegetables (with and without French Fries)

4.1 Fruit, Vegetables, with and without French Fries

The fruit and vegetables module, when the French Fries variable is included, contains 7 variables, pertaining to 6 foods: fruit, unsweetened juice, all vegetables, potatoes (not French Fries), French Fries, dark green vegetables, and orange vegetables. In the fruit, vegetables, and French Fries module, a common factor model with 2 factors was used. Kaiser's Overall MSA is 0.73, indicating that the partial correlations are relatively small compared to the original correlations. Two factors explain 86% of the common variance according to a principle factor analysis. One rotation of the factor pattern using a principle factor analysis is given in Table 2. Note that the factor loadings in Table 2 represent the correlation between the original variable and its factor.

Table 2. Factor Pattern, Fruit, Vegetables, and French Fries

Variable	Factor 1	Factor 2
Fruit	0.58	0.08
Unsweetened Juice	0.32	0.09
All vegetables	0.68	0.25
Potatoes (except French fries)	0.39	-0.26
French fries	-0.05	-0.29
Dark green vegetables	0.59	0.40
Orange vegetables	0.56	0.00

As is apparent from this rotation of the factor pattern, the first factor is a measure of all healthy foods, particularly raw fruit and dark green and orange vegetables. For this factor, French fries has a coefficient near zero. The second factor provides a contrast between green vegetables and potatoes. A similar result can be seen with a maximum likelihood (ML) factor analysis, where 89% of the common variance in the 7 fruit, vegetable, and French Fries variables was explained with the first two eigenvalues. A statistical chi-squared test is available with the ML factor analysis that tests whether 2 factors are sufficient to explain the common variance. With an alternative hypothesis that 2 factors were not sufficient for these data, a p-value of 0.08 was obtained for a common factor model with 2 factors. There is insufficient evidence to suggest that more factors are needed. (A similar test with 1 factor indicated that at least 2 factors were needed.)

The value of Cronbach's alpha was 0.58, which simply indicates that there is a high level of error variance for the items to be considered reliable for a single construct scale.

When the French fries variable is excluded from the factor analysis, the Kaiser's Overall MSA statistic remained fairly static (0.74), but the construction of factors changes somewhat. As the factor pattern in Table 3 indicates, there is no longer a "healthy vs. unhealthy" factor since none of the variables left are indicative, in and of themselves, of unhealthy eating habits. Rather, the first factor could be described as an indicator of the green-ness of the vegetable, with dark green vegetables having the highest value. The second factor describes anything not green, whether vegetable or not. Cronbach's alpha increased to 0.68, which is close to the cutpoint used that indicated sufficient reliability within a single construct. Removing this single item indicates (not surprisingly) that French Fries does not fit with the rest of the variables in the fruit and vegetables module.

Table 3. Factor Pattern, Fruit and Vegetables (excluding French Fries)

Variable	Factor 1	Factor 2
Fruit	0.22	0.46
Unsweetened Juice	0.28	0.08
All vegetables	0.63	0.15
Potatoes (except French fries)	-0.05	0.46
Dark green vegetables	0.80	-0.12
Orange vegetables	0.31	0.31

Even though these results give an interesting insight into the relationship between Cronbach's alpha and factor analysis, there is no clear indication that any of these variables are actually redundant.

4.2 High Protein Foods

The high protein foods module contains 7 variables: poultry, red meat/pork, deli meat, fish, eggs, peanut butter, and dry beans. The Kaiser's Overall MSA is 0.63, indicating that ideally more variables should be used to better define the factors in the common factor model. The smallest MSA among the variables is 0.57 for peanut butter, which is a large enough difference that we may want to consider reformulating the problem without peanut butter. A common factor model with 3 factors was used; three eigenvalues accounted for all of the common variance. One rotation of the factor pattern using a principle factor analysis is given in Table 4.

Table 4. Factor Pattern, High Protein Foods

Variable	Factor 1	Factor 2	Factor 3
Poultry	0.09	0.51	-0.06
Red meat/pork	0.65	0.02	0.09
Deli meats	0.62	0.22	0.07
Fish	0.09	0.47	0.24
Eggs	0.24	0.12	0.29
Peanut butter	0.03	0.04	0.46
Dry beans	0.10	0.41	0.28

As is apparent from this rotation of the factor pattern, the first factor is a measure of the "unhealthy proteins", with high loadings on red meat and deli meats. The second appears to be a measure of the "healthy proteins", with moderately high loadings on poultry, fish, and dry beans. Finally, the third factor is what is left over in the common variance, with a fairly high loading for peanut butter and moderate loadings for eggs and dry beans and, to a lesser extent, fish. Identifying a cluster across the three factors is real evidence of similarity: clearly, beans and fish cluster together very closely across all three factors. Poultry is also similar to beans and fish, but the similarity breaks down with the third factor; red meat/pork and deli meat are similar to each other, but they differ in the second factor, with deli meats "healthier" than red meat/pork. The similar result can be seen with a maximum likelihood (ML) factor analysis. Three eigenvalues accounted for all of the common variance in the variables of this module. A statistical chi-squared test is available with the ML factor analysis that tests whether 3 factors are sufficient to explain the variation in the fruit and vegetable variables. With an alternative hypothesis that 3 factors were not sufficient for these data, a p-value of 0.48 was obtained for a common factor model with 3 factors. There is insufficient evidence to suggest that more factors are needed. (A similar test with 2 factors indicated that at least 3 factors were needed.)

Cronbach's alpha for this module has a value of 0.56, indicating a high level of error variance for these items to be considered reliable for a single construct scale. Perhaps this is simply a result of 3 apparent dimensions in the data, which adds to the apparent high level of error variance.

5. Results for Supplemental Modules

The supplemental modules included questions that were not necessarily measuring food intake for specific foods, but included questions about attitudes and behaviors associated with food, diet, and/or health. These modules include modules with questions about weight control (with and without binary variables), shopping, attitudes, availability, and physical activity. As with the dietary modules, for all of the behavior modules, the factor analysis produced easily interpretable factors using the VARIMAX rotation. However, only the weight control module will be presented here, since that was the only module where the analysis indicated a possible redundancy.

5.1 Weight Control

The original definition for weight control includes 9 variables. Five refer to consumption of specific foods: fruit as dessert, fruit and vegetables as snacks, sweetened fruit drinks, soda, and fast food. The other four are behavior variables: change to a healthier diet in the past 12 months, attempts to lose weight in past 12 months, snack or eat meal in front of television, and eat breakfast. To facilitate interpretation, the levels of the variables are recoded so that smaller numbers correspond to attempts to control weight and larger numbers correspond to no attempt to control weight.

The variables for sweetened fruit drinks, soda, fast food, and eating breakfast are all per-week consumption variables. As with other per-week consumption variables, a square root transformation is used to more closely approximate a normal distribution, with outliers removed as discussed in Section 2. Since smaller numbers correspond to attempts to control weight, the transformed values from sweetened fruit drinks, soda, and fast food can all be taken directly. For the sake of interpretation, however, the negative of the transformed breakfast variable will be used, since larger values (i.e., eating breakfast more often) are considered conducive to weight control.

The variables for fruit as dessert, fruit and vegetables as snacks, and snack or eat meal in front of television, have an ordinal response. As noted earlier, an ordinal response is problematic for the multivariate normal assumption, since we are required to assume the same distance between levels, and that too few levels are available to approximate a normal distribution very well, particularly if any skewness occurs. For the fruit-as-dessert variable, two levels correspond to a real attempt to lose weight: eat fruit when eating dessert, or not eating dessert at all. Among respondents who usually eat fruit for dessert or do not eat dessert at all, it would be difficult to determine which should have higher value in terms of losing weight. Since so few respondents (only three) do not eat dessert, these two levels will be combined and set to the value "1", and the other levels will be assigned in increasing order from there. The fruit-and-vegetables-as-snacks variable will also be assigned with "usually" taking the value of "1" and other levels assigned in increasing order after that. However, if one usually snacks or eats meals in front of the television, this is considered detrimental to weight control, so for interpretation's sake "usually" will be assigned a value of "4" and the other levels assigned in decreasing order from there.

The variables for behavioral changes in the past year (change to healthier diet, attempting to lose weight) are binary variables and therefore do not even remotely resemble normally-distributed variables. These variables could provide misleading results in the factor analysis for this reason; they will, however, still be included for this module. The code for "yes" will be set to "1" and the code for "no" set to "2".

A 3-factor model was considered for these data, the loadings for which are presented in Table 5.

Table 5. Factor Pattern, Weight Control Variables

Variable	Factor 1	Factor 2	Factor 3
Fruit as dessert	0.03	0.09	0.48
Fruit/vegetables as snacks	0.10	0.12	0.46
Sweetened fruit drinks	0.11	0.26	0.15
Soda	0.07	0.43	0.16
Fast food	0.02	0.45	0.08
Switched to healthier diet	0.69	0.11	0.06
Attempted to lose weight	0.71	0.09	0.00
Snack or eat meals at TV	0.05	0.35	-0.11
Eat breakfast in the morning	-0.03	-0.02	0.30

For these data, the Kaiser's Overall MSA was only 0.56, which is rather low, particularly for the number of variables included. Recall that the inclusion of the binary variables could be considered problematic, and it could be argued that these binary indicators are not like the other variables. Indeed the individual MSAs for the overall health indicator and the losing weight indicator are 0.53 and 0.52 respectively, so that we may want to consider adjusting the set of variables used.

In spite of the issues observed in the previous paragraphs, some useful patterns are evident in these data. The first factor is a measure of actions in the past 12 months that were intended to improve health; whether these are borne out in the past-week data is less clear. The second factor is a measure of unhealthy eating habits: drinking soda, eating fast food, snacking or eating in front of the television, and, to a lesser extent, eating sweetened fruit drinks. The third factor is a measure of healthy eating habits: eating fruit as dessert, eating fruit and vegetables as snacks, and, to a lesser extent, eating breakfast each morning. A number of clusters are apparent in the data, indicating some redundancy. The two binary variables are very close together across all three factors, indicating a high degree of similarity, as are the two variables relating to eating fruit (eating fruit as dessert and eating fruit and vegetables as snacks). Although less evident than the other two clusters, the loadings associated with fast food, sweetened fruit drinks, and soda are also somewhat close together in value. Similar results were evident with ML estimation. With an alternative hypothesis that 3 factors were not sufficient for these data, a p-value of 0.08 was obtained for a common factor model with 3 factors. There is insufficient evidence to suggest that more factors are needed. (A similar test with 2 factors marginally indicated that at least 3 factors were needed, with a p-value of 0.001.)

The value of Cronbach's alpha for these data was 0.48, indicating a high level of error variance for the items to be considered reliable for a single construct scale. As with the high protein foods example, this is probably due to the high dimensionality in the data, which is more often associated with low values for Cronbach's alpha, while not necessarily indicating that the questions are not adequate or reliable.

6. Discussion

Not only did the factor analysis indicate underlying constructs that the dietary variables measured, it also pointed to redundancies in the data. In particular, it did not appear to be necessary to ask about both fish and dry beans. The two variables associated with the consumption of these two foods were correlated with other variables in roughly equivalent ways. The two variables associated with switching to a healthier diet and attempting to lose weight appeared to be redundant, as did variables associated with eating fruit as a dessert, or eating fruit or vegetables as snacks.

Factor analysis was useful for assessing what underlying constructs the items in each module were measuring, and where redundancies might occur. Cronbach's alpha was less useful in this setting, however, since the multidimensionality in the data made it difficult to ascertain what the generally low values for the alpha actually meant.

7. Acknowledgements

This work was done as part of a contract MPR had with the Economic Research Service (ERS) of the USDA. The project was called "Development of a Questionnaire on Dietary Behavior for Use in Low-Income Populations." The project officer at ERS was David Smallwood; the MPR project director was Rhoda Cohen.

8. References

Cerny, B. A. and H. F. Kaiser. "A study of a measure of sampling adequacy for factor-analytic correlation matrices." Multivariate Behavioral Research 12, 1977, pp. 43-47.

Cortina, J. "What is coefficient alpha? An examination of theory and methods." *Journal of Applied Psychology* 78:1, 1993, pp. 98-104.

Cronbach, L. J. "Coefficient alpha and the internal structure of tests." *Psychometrika* 22:3, 1951, pp. 297-334.

Johnson, R. A. and D. W. Wichern. *Applied Multivariate Statistical Analysis.* Englewood Cliffs, NJ: Prentice-Hall, Inc., 1981.

Kaiser, H. F. "The VARIMAX criterion for analytic rotation in factor analysis." *Psychometrika* 23, 1958, pp. 187-200.

Kaiser, H. F. "A second generation little jiffy." *Psychometrika* 35, 1970, pp. 401-415.

Kaiser, H. F. and J. Rice. "Little jiffy, mark IV," *Educational and Psychological Measurement* 34, 1974, pp. 111-117.

Schmitt, N. "Uses and abuses of coefficient alpha," *Psychological Assessment* 8:4, 1996, pp. 350-353.