# Data-based Questionnaire Design: the NCI Survey

Gladys Block, Ph.D.

We at the National Cancer Institute have developed a dietary assessment questionnaire, making use of a large national survey and database as an integral part of the development approach. My purpose here is not to make a case for a particular instrument, but to demonstrate the value of the data-based approach (1). This approach is not unique. Others have used the percent contribution of a nutrient to guide food item selection, notably Jean Hankin (2), Pickle and Hartman (3), and Buzzard (4).

Several purposes guided the choice and development of the dietary assessment method. First and foremost, it must be a valid reflection of an individual's usual diet, since we want to make associations with individual health outcomes. This clearly means that 24-hour recalls are inappropriate, and in my view also means that dietary intake over a few consecutive days is inadequate, as well. And it must be brief, since only then will a dietary assessment be acceptable to many investigators. These requirements made a quantified frequency questionnaire the appropriate assessment method.

There are two broad aspects of a diet questionnaire: 1) the components of the questionnaire itself -- the food list and quantitations; and 2) the human response to it -- the ability to estimate frequency, for example. I will address the second briefly at the end of the paper. Most of what I will present will concern the first aspect, the development, improvement and testing of the components of the instrument itself. When these two aspects are kept separate and are tested separately, it is possible to maximize the potential performance of the instrument itself, before the vagaries of the human response are introduced. In this way it is possible to identify the sources of error, and target improvements more appropriately.

Within aspect #1 above, the instrument itself, there are three main components the food list, the nutrient composition associated with each food on the list, and the portion size assumption associated with each food on the list. Each of these three components was developed in a data-based manner.

The food list was selected by identifying those foods which were the main contributors to population intake, in the NHANES II survey. That survey collected 24-hour recalls from approximately 12,000 adults. For calories and each of the 17 nutrients on the NHANES II database, we summed the total nutrient consumed by the entire adult population, and then determined what percent of the total each food contributed. This work has been published in detail elsewhere (5,6). We then identified a list of foods which represents at least 90 percent of each of the 17 nutrients and energy on the database. For example, the list contains foods which represent 95 percent of the vitamin C, 97 percent of dietary cholesterol, and 93 percent of total calories consumed by the U.S. adult population. By capturing calories and 17 other major nutrients adequately, and by keeping the food items reasonably distinct, it is likely that other nutrients not now calculated by the program (or perhaps not even discovered) will be able to be assessed by this food list.

As indicated above, the list is comprehensive, that is, represents a high percentage of each of the nutrients. However, comprehensiveness is not enough. In order to make accurate nutrient estimates, precision is also necessary. For example, an item, "pasta", may be comprehensive but is not precise enough to enable you to distinguish between retinol (in pasta with cheese sauces), carotenes (in pasta with tomato sauce) or neither (in plain pasta). To as great a degree as possible consistent with time constraints, foods were kept separate and were grouped only if conceptually similar and similar in nutrient content per usual serving.

The nutrient composition of each listed food item was selected as follows. All of the NHANES II foodcodes which comprised a food item were grouped together. For example, there were 11 different codes of green beans, each of which could have a different content of, e.g., protein per 100 grams. How is one to choose among the different protein contents, in order to determine a single protein value to be used with the questionnaire line item, "green beans"? Clearly it is not appropriate simply to take the mean or the median of the 11 values, since that is greatly influenced by what happens to have been analyzed. The value must take into account the frequency of consumption of each of the items. We used the frequency-weighted median, that nutrient composition for which half of the consumers ate green beans with a higher protein content, and half ate green beans with a lower protein content. This approach minimizes the error more than does a weighted mean, and frequently chooses the value of the food eaten by the greatest number of respondents.

In addition to the food list and the nutrient composition of each food, the third component of a dietary assessment instrument is the portion size assumptions for each food. These were also developed in a data-based way, by examining the actual reported portion sizes among the 12,000 adults in NHANES II. The NHANES II survey is an invaluable source for such data, since three-dimensional models were used by the respondents to estimate portion sizes. We thus have large-population data using three-dimensional models, a unique resource. In addition, because the NHANES II population is so large we were also able to develop age- and sex-specific portion sizes for the foods on our questionnaire. This greatly improves the precision of our nutrient estimates, particularly for men and for older women whose usual portions differ quite substantially from "standard" portions.

To take account of individual food preferences and variability, the questionnaire also asks the respondent to indicate whether his usual portion is small, medium or large with respect to a stated medium portion. This feature, unique to this instrument, adds additional precision in nutrient estimates.

The resulting questionnaire consists of a 98-item food list (printed on the front and back of a single page), and the nutrient composition and portion size arrays as discussed above. There is also an open-ended question which permits the respondent to identify other commonly-eaten foods which are not on the list; and questions on restaurant foods and on the frequency and types of fat added in cooking and at the table, all of which are used in the nutrient calculations.

We evaluated and perfected this list and its quantitations by testing how well it could approximate the values resulting from a detailed diet record. For this evaluation we used 50 one-day diet records calculated in the usual way, and then used our food list and quantitations as a brief scoring system. Nutrients were calculated by both methods, and the two estimates compared.

This approach is a test strictly of the instrument itself, before the addition of the uncertainty of the human ability to estimate frequency. It enabled us to evaluate the adequacy of the list, since if foods could not be coded on the food list the nutrient estimates would be low and the correlations poor. It also enabled us to evaluate the nutrient composition and portion size assumptions; if our quantitations were not adequate to reflect the diversity of foods and portion sizes in actual diets, again the correlations would be poor.

The results indicate that the list, nutrient composition and portion size assumptions are robust in the face of the diversity of the normal free-living human diet. When used as a brief coding system for the 50 diet records, the nutrient estimates correlated with those of the more detailed method at $r = 0.7$ to 0.9. Figure 1 shows scatterplots of the relationship between the record and food-list methods, for calories and vitamin A.

As indicated previously, the above evaluations test the degree to which the food list and quantitations could produce accurate estimates, if respondents could respond accurately about their diets. The performance of the questionnaire when actually administered to respondents has been tested in three field validations. In one, the reference method was a detailed diet history; in another, it was a seven-day diet record; and in the third, questionnaire results were compared with a known diet actually administered to the subjects. Results of these validations will be reported elsewhere (7-9). Preliminary results, however, indicate correlations of 0.7 or better in each of these validations, for factors as diverse as vitamin A, calcium, and percent of calories from fat.

These results suggest that a systematic data-based approach to questionnaire development may permit us to enhance the accuracy of nutrient assessments in some situations.

## Bibliography


1. Block G, Hartman A, Dresser CM, Gardner L, Gannon J. Data-based diet questionnaire design and testing. Submitted for publication.

2. Hankin JH, Rawlings V, Nomura A. Assessment of a short dietary method for a prospective study on cancer. Am J Clin Nutr 1978; 31:355-9.

3. Pickle LW and Hartman AM. Indicator Foods for vitamin A assessment. Nutr Cancer 1985; 7:3-23.

4. Buzzard M, Personal communication; 1983.

5. Block G, Dresser CM, Hartman AM et al. Nutrient Sources in the American Diet: Quantitative data from the NHANES II Survey. I, Vitamins and Minerals. Am J Epidemiol 1985; 122:13-26.

6. Block G, Dresser CM, Hartman AM et al. Nutrient Sources in the American Diet: Quantitative data from the NHANES II Survey. II, Macronutrients and Fats. Am J Epidemiol 1985; 122:27-40

7. Cummings SR, McHenry K, Block G, et al. Validation of two check-list methods for measuring dietary calcium intake of elderly women. Am J Clin Nutr, Submitted.

8. Block G, Jones Y, Hartman AH. Agreement of a quantified frequency questionnaire with known dietary intake. In preparation.

9. Burgess M, Block G, Bowen P, Hartman AH. A comparison of serum, diet interview and self-administered questionnaire for carotenoids and total vitamin A. In preparation.

Figure 1