Invited Commentary: Another Perspective on Food Frequency Questionnaires

Gladys Block

As expected, Subar et al. (1) have produced an excellent study and a balanced report. Such work is important, so that we can continuously improve the instruments used in epidemiologic research, as well as those used in clinical or counseling situations or in nutrition policy decisions and recommendations.

It is important to note that Subar et al. decided to test the 1995 version of the Block food frequency questionnaire (FFQ) (Block95) instead of the 1998 version (Block98), which was available at the time. Block98 was developed from National Health and Nutrition Examination Survey (NHANES) III data, whereas Block95 had been based on NHANES II; Block98 also incorporates other design elements, such as additional low-fat food choices and portion-size pictures. The decision by Subar et al. to use the 1995 version was made to provide a greater contrast between the Diet History Questionnaire (DHQ) and an older FFQ method, since the 1998 version already incorporated many of the same cognitive enhancements as the DHQ. It is also regrettable that they chose to change the analysis options we use and recommend, including “Fruit-Adjust,” “Veg-Adjust,” and “Recalc.”

An earlier version of the manuscript observed substantially higher folate estimates by the Block95 than by the DHQ, Willett, and records. It should be noted that at the time we provided Subar et al. with the nutrient database and software in August 1998, the new folate fortification regulations had been in effect for 8 months. Our databases were updated for folate as of July 1998, since at that point anyone answering the questionnaire about the “past year” would have had a minimum of more than half a year of exposure to the higher folate content of grain products. (In fact, most manufacturers have already implemented the higher folate content in the flour.) Because the 24-hour recall collection ran from September 1997 through August 1998, it seems likely that we provided Subar et al. with the nutrient database and software in August 1998, the new folate fortification regulations had been in effect for 8 months. Our databases were updated for folate as of July 1998, since at that point anyone answering the questionnaire about the “past year” would have had a minimum of more than half a year of exposure to the higher folate content of grain products. (In fact, most manufacturers had already implemented the higher folate content in the flour.) Because the 24-hour recall collection ran from September 1997 through August 1998, it seems likely that only the Block FFQ provided accurate folate estimates.

Received for publication July 25, 2001, and accepted for publication September 14, 2001.

Abbreviations: Block95, the 1995 version of the Block food frequency questionnaire (Block98 defined similarly); DHQ, Diet History Questionnaire; FFQ, food frequency questionnaire; FFQenergy, energy as estimated by FFQ; FFQnutrient, nutrient as estimated by FFQ; NHANES, National Health and Nutrition Examination Survey; RecordsEnergy, energy as estimated by records.

From the School of Public Health, University of California, Berkeley, 426 Warren Hall, Berkeley, CA 94720-7360 (e-mail: gblock@uclink4.berkeley.edu). (Correspondence to Dr. Gladys Block at this address).

It would be useful to have some additional information about the analysis of the Block95 FFQ. As noted above, some calculation options were changed by Subar et al. Were there any changes to other options or the nutrient and portion size databases we provided? (The observed medians are 100–200 kcal lower than are usually seen with Block95.) Similarly, in any such analysis it is important to be assured that the 24-hour recall data were rigorously reviewed to exclude any vitamin supplement data that might have been accidentally included in the recall, since the data from the FFQs presumably represent nutrients from diet alone.

With regard to the speculation that the use of portion sizes in the DHQ and Block95 FFQ causes artifactually high correlations because of correlated errors in portion size estimation, this question could be resolved by Subar et al. simply by rerunning the DHQ and Block95 FFQ nutrient estimates, ignoring the portion size responses and simply assigning a standard medium, for the DHQ, or the age-sex-specific medium, for the Block95 FFQ. I would expect the resulting correlations to be lower, but not dramatically lower.

The whole issue of the appropriateness of adjusting for energy in all situations and for all nutrients has been resolved in the minds of many researchers, but not in mine. For micronutrients in particular, requirements are usually expressed in absolute terms rather than relative to energy intake. I really want to know whether a person’s vitamin C intake is 45 mg or 90 mg per day. I want to be able to infer that 45 mg/day is low, and I don’t want to infer that this level is normal or even high, simply because it is high in relation to a very low reported energy intake. Finally, the argument has been made that if a micronutrient is not highly correlated with energy, then energy adjustment would have little impact. It is troubling that vitamins C and E, for example, do have a large increase in correlations, since they are poorly correlated with energy ($r < 0.3$) in data from multiple diet records. I think we have much more to learn about the behavior and value of energy adjustment with different instruments and for different purposes.

The much greater effect of energy adjustment on Willett questionnaires as compared with other questionnaires has been seen before. Could large and correlated overestimates in energy and nutrients produced by a few people increase the range and thereby the correlation coefficient? Are there higher slopes in Willett FFQs compared with the slopes seen for other FFQs, in the regression of FFQenergy on RecordsEnergy, and/or in the regression of FFQnutrient on FFQenergy, which could influence the calculations? Could the greater effect of energy adjust-
ment be similar to what can happen with adjustment for measurement error, where very high intradividual variation (whether from true day-to-day variation or from recording errors) can inflate the “deattenuated” correlations to improbably high levels? Hopefully we can learn more about the behavior of these instruments in these computational situations in the future.

Another way of looking at agreement between a test instrument and a “gold standard” is by cross-classification and the percentage of agreement. This method is independent of distribution assumptions and potential artifacts caused by the use of a statistical formula. It would be helpful if researchers presented their data in this manner, as well as in terms of correlations. Similarly, I believe that all researchers should report the crude correlations, unadjusted for measurement error as well as unadjusted for energy, in addition to adjusted correlations. As mentioned above, correction for measurement error can produce unrealistic improvements in correlations, simply because of the extreme variability or error in the reference data. The numbers used for measurement-error correction are themselves imperfect estimates. All we can say for sure is that the correlations with “truth” are higher than the observed correlations; the actual degree of improvement is itself measured with error.

The major issues in this research, as I see it, are the generalizability and the implications for the various types of uses to which these instruments may be put. Generalizability is limited, in my view, and the correlations for all of these instruments should be used cautiously because they probably are representative of the optimal rather than the usual sample. Not only were 84 percent of the final respondents White, but 77 percent of them had education beyond high school. In contrast, only 49 percent of all Americans aged 18 years and above had any education beyond high school, in 1999, and only 41 percent of African Americans and 29 percent of Hispanics. In addition to these problematic educational differences, this validation study was designed in such a way that FFQ respondents had already completed four 24-hour recalls, had dedicated 1 year to the study and, for half the sample, had already completed one other FFQ. While the timing of the test and reference instruments is always a dilemma in validation studies, there is little doubt that such respondents would be much more attuned to their own diets and more conscientious in completing questionnaires than would typical study respondents receiving the FFQ for the first time. In the real world, careless, confused, and extreme responses to these FFQs are seen with a considerably higher frequency than is implied by the low percentages that had to be excluded on the basis of improbable calorie estimates in this study. I think we must conclude that all of these instruments would have poorer correlations with “truth” if administered for the first time in a less well-educated population.

Having said that, I hasten to add that correlations even lower than those seen here have been very useful in public health research, for “instruments” such as blood pressure measurement, skinfold measurements, and numerous other physiologic measures that have proven themselves valuable despite imperfections.

The other major issue involves the implications for the variety of types of uses to which these instruments may be put. One could conclude that they are all equally useful if the purpose is purely for epidemiologic research, providing that they are energy adjusted. However, FFQs are also widely used for several other purposes, for which the absolute, unadjusted level of nutrient intake is important and/or where energy adjustment is either clearly inappropriate, difficult to interpret, or too cumbersome for routine use. Some examples follow: first, screening for eligibility to enter a prospective or intervention study, where one may wish to enroll people with, for example, vitamin C intake below the Recommended Dietary Allowance or fat intake above the recommended level. For this, a reasonably accurate absolute level may be more useful than an energy-adjusted level. A second example is nutrition screening at the clinical level, in which a simpler instrument, such as a FFQ, can be used to give health care providers a preliminary look at people’s diets to know whether a more extensive nutrition workup is needed, or nutrition screening at the community or population level, in which people could assess their own intake in order to improve their own diets. Third, at the policy level, we need to advance beyond simply saying “people low on this nutrient are at increased risk” to saying “people with less than x mg of this nutrient are at increased risk.” None of these instruments is error free for these purposes. However, increasing our understanding of the usability of FFQs and improving their performance in a variety of settings are important goals. Subar et al. have contributed to this goal.

REFERENCES