The authors investigated the role of food frequency questionnaire (FFQ) design, including length, use of portion-size questions, and FFQ origin, in ranking subjects according to their nutrient intake. They also studied the ability of the FFQ to detect differences in energy intake between subgroups and to assess energy and protein intake. In a meta-analysis of 40 validation studies, FFQs with longer food lists (200 items) were better than shorter FFQs at ranking subjects for most nutrients; results were statistically significant for protein, energy-adjusted total fat, and energy-adjusted vitamin C. The authors found that FFQs that included standard portions had higher correlation coefficients for energy-adjusted vitamin C (0.80 vs. 0.60, \( p < 0.0001 \)) and protein (0.69 vs. 0.61, \( p = 0.03 \)) than FFQs with portion-size questions. However, it remained difficult from this review to analyze the effects of using portion-size questions. FFQs slightly underestimated gender differences in energy intake, although level of energy intake was underreported by 23% and level of protein intake by 17%. The authors concluded that FFQs with more items are better able to rank people according to their intake and that they are able to distinguish between subpopulations, even though they underestimated the magnitude of these differences.

Food frequency questionnaires (FFQs) are widely used to assess the dietary intake of large populations. The popularity of FFQs stems from their ease of administration, ability to assess dietary intake over an extended period of time, and low costs (1). They are therefore often used in epidemiologic studies to investigate the relation between diet and disease. For some purposes, information about level of intake is very important, for example, to set recommendations for nutrient intake. For most epidemiologic studies, FFQs must be able to classify individuals correctly according to their dietary intake. However, Bingham and others have argued that, probably because of misclassification, FFQs are not always able to detect weak associations (2, 3). Because of this debate and their established role in epidemiologic research, FFQs need to be further characterized and subsequently improved.

FFQs differ in the way they are developed and show large variations in design characteristics, such as number of items or inclusion of portion-size questions. Such variations could affect reported intakes (4–6).
In this study, we first aimed to provide an overview of FFQ validation studies and the validity of FFQs in classifying subjects relative to their intake in relation to number of items in the FFQ, use of portion-size questions, origin of the FFQ, and administration mode. A second aim was to provide an overview of the validity of FFQs in assessing absolute energy and protein intakes as determined in studies using recovery biomarkers and to establish whether FFQs can detect known differences in energy intake between men and women.

**MATERIALS AND METHODS**

**Search strategy and study selection**

We searched MEDLINE (National Library of Medicine, Bethesda, Maryland) for validation studies of FFQs that assessed respondents’ habitual dietary intake and were published between 1980 and December 2006. An FFQ was defined as a questionnaire with a food list and a frequency-response section where subjects report how often each food item is consumed (7, p. 75). Search terms used were the Medical Subject Headings (MeSH) “nutrition assessment,” “questionnaires,” “evaluation studies” OR “reproducibility of results” AND keywords “food frequency questionnaire” OR “FFQ” AND “validity” OR “validation” OR the publication type “validation studies.”

Studies that met all of the following inclusion criteria were included in the review:

1) Describing FFQs developed for epidemiologic purposes
2) Addressing habitual diet by reporting at least energy intake but preferably also intake of other nutrients including total fat, carbohydrates, protein, alcohol, calcium, and vitamin C
3) Studying adult populations in the age range 18–82 years with a westernized diet but not FFQs validated exclusively among those older than age 60 years
4) Validating FFQs with one of the following reference methods: 24-hour dietary recalls, food records, diet history interview, or recovery biomarkers

Studies that assessed only specific nutrients or food groups such as fruit and vegetable consumption were excluded from this review.

**Data extraction and classification**

The following design characteristics of the FFQ were extracted: number of items in the FFQ (ranging from 44 to 350), use of portion-size questions versus predefined standard portion sizes, and “origin” of the FFQ, for example, the Willett type (8), the Block type (9), the European Prospective Investigation into Cancer and Nutrition (EPIC) type (10), and “other FFQs.” We extracted the following validation study characteristics: study size (N), gender (males, females), FFQ administration mode (interview or self-administration), reference method (food record, 24-hour dietary recall, or diet history interview), and total number of days over which the reference method was applied—categorized as a short period of 1–7 days, a medium period of 8–14 days, and a long period of 15 days or more. We used 8–14 days as the reference category in the meta-regression analyses because the 24-hour dietary recall method was not applied for 15 days or more.

We analyzed nutrients that covered different aspects of a habitual diet. We included energy and all energy-providing macronutrients (i.e., protein, carbohydrates, fat, and alcohol). We added vitamin C and dietary fiber to represent fruit and vegetables and some other food components, for example, calcium, to represent other specific foods.

**Statistical analysis**

**Ranking.** We extracted crude and energy-adjusted correlation coefficients and, if available, gender-specific and/or gender-adjusted correlation coefficients between FFQs and reference methods to compare studies with respect to ranking of subjects. Correlation coefficients were first converted into a standard normal metric by using Fisher’s r-to-Z transformation, expressed in equation 1 (11), in which \( r_i \) is the correlation coefficient from study \( i \).

\[
Z_{ri} = \frac{1}{2} \log_e \left( \frac{1 + r_i}{1 - r_i} \right) .
\] (1)

The transformed effect sizes were then used to calculate an initial pooled average, mean \( Z_{ri} \), in which each correlation coefficient was weighted by the variance, \( 1/(n_i - 3) \). Then, to identify confounders to be used for adjustments in meta-regression analyses, we stratified these correlation coefficients by sex, type of reference method, category of number of days that this method was applied, and administration method. The weighted averages, mean \( Z_{ri} \), were also used to perform the Cochrane \( Q \) test for heterogeneity between studies for the assessed nutrients, expressed in equation 2 (12).

\[
Q = \sum_{i=1}^{k} \frac{(n_i - 3)(Z_{ri} - \text{mean}Z_r)}{n_i - 3} .
\] (2)

This test had a \( p \) value of <0.0001 for all nutrients, which indicated that the correlation coefficients from the validation studies were heterogeneous.

Following the heterogeneity test, random-effects meta-regression was conducted to explain this heterogeneity by FFQ design characteristics using the restricted maximum likelihood approach, as per Thompson and Sharp (13). For each nutrient, number of items as a continuous variable, use of portion-size questions or standard portions, FFQ origin—Willett, Block, EPIC, or other—and several potential confounders (gender, reference method, and number of days over which this method was applied) were regressed on the transformed correlation coefficients \( Z_{ri} \). Weights were assigned based on the variance \( 1/(n_i - 3) \). Results of the meta-regression are presented as predicted values of \( Z \), retransformed to \( r \), using a model that included an intercept, a reference period of 8–14 days, an average value of 0.5 for the indicator variable for sex, and similarly so for the

Am J Epidemiol 2007;166:1468–1478
reference method. All data were analyzed by using STATA 8 software (Stata Corporation, College Station, Texas).

Validity of absolute intake. To assess energy intake, validity was defined as the difference between the mean levels of energy intake assessed by the FFQ minus the mean levels of energy expenditure determined by the doubly labeled water method. For protein intake, it was defined as the difference between protein intake assessed by the FFQ and protein intake estimated from 24-hour urinary nitrogen excretion. If only nitrogen excretion was reported in the paper, we estimated protein intake by assuming that urinary nitrogen was excreted as a constant proportion of 80 percent of total nitrogen intake (14), and 16 percent of protein is nitrogen (15). Thus, protein intake was estimated from the following formula:

$$\text{protein} = 6.25 \times \left( \text{urinary N}/0.80 \right).$$

(3)

Validity of gender differences in energy intake. To evaluate the extent to which gender differences in energy intake could be detected by FFQs, we extracted gender-specific mean energy intake, including standard deviations (if not available, we assumed it was 3 MJ because this mean of standard deviations was reported in 31 other included studies) and $N$, or the standard error (SE) for mean energy intake.

<table>
<thead>
<tr>
<th>First author, year (reference no.)</th>
<th>Origin†</th>
<th>No. of subjects</th>
<th>No. of FFQ items</th>
<th>FFQ administration method</th>
<th>Use of portion-size questions</th>
<th>FFQ reference period</th>
<th>Reference method</th>
<th>Repetitions reference method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Andersen, 1999 (26)</td>
<td>Other</td>
<td>125 men</td>
<td>180</td>
<td>Self</td>
<td>Yes</td>
<td>Unknown</td>
<td>FR*</td>
<td>14 d in 5 w*</td>
</tr>
<tr>
<td>Andersen, 2003 (53)</td>
<td>Other</td>
<td>17 women</td>
<td>180</td>
<td>Self</td>
<td>Yes</td>
<td>1 y</td>
<td>DLW*</td>
<td>1 (10 d)</td>
</tr>
<tr>
<td>Bingham, 1997 (27)</td>
<td>EPIC*</td>
<td>156 women</td>
<td>130</td>
<td>Self</td>
<td>No</td>
<td>Unknown</td>
<td>FR</td>
<td>4 × 4 d in 1 y</td>
</tr>
<tr>
<td>Block, 1990 (28)</td>
<td>Block</td>
<td>102 women</td>
<td>60</td>
<td>Self</td>
<td>Yes</td>
<td>Unknown</td>
<td>FR</td>
<td>2 × 7 d</td>
</tr>
<tr>
<td>Block, 1990 (9)†</td>
<td>Block</td>
<td>102 women</td>
<td>94</td>
<td>Self</td>
<td>Yes</td>
<td>6 m*</td>
<td>FR</td>
<td>3 × 4 d in 1 y</td>
</tr>
<tr>
<td>Block, 1992 (29)</td>
<td>Block</td>
<td>85 men and women</td>
<td>98</td>
<td>Interview</td>
<td>Yes</td>
<td>1 y</td>
<td>FR</td>
<td>4 × 3 d in 1 y</td>
</tr>
<tr>
<td>Bohlscheid-Thomas, 1997 (56)</td>
<td>EPIC</td>
<td>49 men, 55 women</td>
<td>104</td>
<td>Self</td>
<td>Yes</td>
<td>1 y</td>
<td>24-HR*</td>
<td>12× in 1 y</td>
</tr>
<tr>
<td>Boucher, 2006 (17)</td>
<td>Block</td>
<td>166 women</td>
<td>126</td>
<td>Self</td>
<td>Yes</td>
<td>Unknown</td>
<td>24-HR</td>
<td>2×</td>
</tr>
<tr>
<td>Brunner, 2001 (30)</td>
<td>Other</td>
<td>457 men, 403 women</td>
<td>127</td>
<td>Self</td>
<td>No</td>
<td>1 y</td>
<td>FR</td>
<td>7 d</td>
</tr>
<tr>
<td>Callmer, 1993 (31)</td>
<td>Other</td>
<td>57 men, 50 women</td>
<td>250</td>
<td>Self</td>
<td>Yes</td>
<td>1 y</td>
<td>FR</td>
<td>6 × 3 d in 1 y</td>
</tr>
<tr>
<td>Engle, 1990 (32)</td>
<td>Other</td>
<td>16 men, 34 women</td>
<td>120</td>
<td>Self</td>
<td>Yes</td>
<td>3 m</td>
<td>FR</td>
<td>7 d</td>
</tr>
<tr>
<td>Feunekes, 1993 (50)</td>
<td>Other</td>
<td>95 men, 96 women</td>
<td>104</td>
<td>Interview</td>
<td>Yes</td>
<td>1 m</td>
<td>DH*</td>
<td>1 × 1 m</td>
</tr>
<tr>
<td>Fidanza, 1995 (33)</td>
<td>Other</td>
<td>11 men, 35 women</td>
<td>93</td>
<td>Self</td>
<td>Yes</td>
<td>1 y</td>
<td>FR</td>
<td>7 d</td>
</tr>
<tr>
<td>Flagg, 2000 (18)</td>
<td>Block</td>
<td>216 men, 223 women</td>
<td>114</td>
<td>Self</td>
<td>Yes</td>
<td>1 y</td>
<td>24-HR</td>
<td>4× in 1 y</td>
</tr>
<tr>
<td>Goldbohm, 1994 (34)</td>
<td>Other</td>
<td>59 men, 50 women</td>
<td>150</td>
<td>Self</td>
<td>Yes</td>
<td>1 y</td>
<td>FR</td>
<td>3 × 3 d in 1 y</td>
</tr>
<tr>
<td>Hartwell, 2001 (35)</td>
<td>Other</td>
<td>16 men, 9 women</td>
<td>162</td>
<td>Self</td>
<td>No</td>
<td>1 y</td>
<td>FR</td>
<td>2 × 4 d</td>
</tr>
<tr>
<td>Jain, 2003 (19)</td>
<td>Other</td>
<td>151 men, 159 women</td>
<td>166</td>
<td>Self</td>
<td>Yes</td>
<td>1 y</td>
<td>24-HR</td>
<td>3× in 1 y</td>
</tr>
<tr>
<td>Johansson, 2002 (20)</td>
<td>EPIC</td>
<td>96 men, 99 women</td>
<td>84</td>
<td>Interview</td>
<td>Yes</td>
<td>Unknown</td>
<td>24-HR</td>
<td>10× in 1 y</td>
</tr>
<tr>
<td>Katsouyanni, 1997 (21)</td>
<td>EPIC</td>
<td>42 men, 38 women</td>
<td>190</td>
<td>Self</td>
<td>Yes</td>
<td>1 y</td>
<td>24-HR</td>
<td>12× in 1 y</td>
</tr>
<tr>
<td>Kroke, 1999 (51)</td>
<td>EPIC</td>
<td>75 men, 59 women</td>
<td>146</td>
<td>Self</td>
<td>Yes</td>
<td>1 y</td>
<td>DLW§, 24-HR</td>
<td>14 d, 12× in 1 y</td>
</tr>
<tr>
<td>Larkin, 1989 (36)</td>
<td>Other</td>
<td>228 men and women</td>
<td>116</td>
<td>Interview</td>
<td>Yes</td>
<td>1 y</td>
<td>FR</td>
<td>16 d in 1 y</td>
</tr>
<tr>
<td>Longnecker, 1993 (37)</td>
<td>Willett</td>
<td>64 men, 74 women</td>
<td>116</td>
<td>Self</td>
<td>No</td>
<td>1 y</td>
<td>FR</td>
<td>3 × 2 d or 2 × 2 d</td>
</tr>
<tr>
<td>Mannisto, 1996 (38)</td>
<td>Other</td>
<td>152 women</td>
<td>110</td>
<td>Self</td>
<td>No</td>
<td>1 y</td>
<td>FR</td>
<td>2 × 7 d in 3 m</td>
</tr>
</tbody>
</table>
intake. We subtracted the mean level of energy intake of the women estimated by each FFQ (FFQwomen) from that of the mean level of the men (FFQmen), and we did the same for the reference method (Refwomen and Refmen).

Gender difference FFQ:

\[ \text{FFQ}_{\text{men}} - \text{FFQ}_{\text{women}} = \sqrt{\text{SE}^2_{\text{FFQ}_{\text{men}}} + \text{SE}^2_{\text{FFQ}_{\text{women}}}} \]  (4)

Gender difference reference method:

\[ \text{Ref}_{\text{men}} - \text{Ref}_{\text{women}} = \sqrt{\text{SE}^2_{\text{Ref}_{\text{men}}} + \text{SE}^2_{\text{Ref}_{\text{women}}}} \]  (5)

We tested whether there was a difference between both results by using an independent \( t \) test.

RESULTS

Description of studies

The search procedure resulted in 40 papers (table 1) describing 42 FFQs that matched the inclusion criteria. The majority of FFQs were validated against 24-hour dietary recalls (16–25) or food records (8, 9, 26–49). One FFQ was validated against a diet history method (50), two FFQs against 24-hour dietary recalls and doubly labeled water (51, 52), and one FFQ against doubly labeled water only (53). Six FFQs (8, 24, 37, 41, 46, 49) were developed from the Willett FFQ (8). For this FFQ, an extensive food list was shortened by removing infrequently eaten items and including items contributing most to between-person variance.
using data from Nurses’ Health Study participants. Willett FFQs included on average 113 (range, 61–131) items and asked respondents to report their frequency of consumption of a given reference portion size in a table format. Another six FFQs (9, 17, 18, 24, 28, 29) were developed from the Block FFQ (9). This FFQ was developed by using food items that contributed over 90 percent of the total population intake of energy and several nutrients in the Second National Health and Nutrition Examination Survey database (54). These Block FFQs consisted of an average of 100 items (range, 60–126), and all asked portion-size questions.

Within the EPIC, project country-specific FFQs were developed including items that cumulatively contributed most to between-person variance (21, 23, 40, 48, 51, 55, 56) or to total nutrient intake (22); of these FFQs, the method of development was not described (20, 25, 45). For the EPIC FFQs, we found 11 validation studies performed in nine countries (16, 20–23, 25, 27, 40, 45, 48, 51, 56); two were conducted in the United Kingdom (27, 40) and two in Germany (16, 51, 56). We analyzed them as separate studies. The FFQs validated in the EPIC studies consisted of an average of 154 items (range, 47–350). Nine of these FFQs included portion-size questions, and two assigned standard portion sizes. Although FFQ design between EPIC FFQs varied a lot, the design of their validation studies (57) was carefully standardized, except for the United Kingdom, Denmark, and Sweden because they joined the EPIC project at a later stage (27, 45, 48, 57). Three other EPIC validation studies did not match the inclusion criteria (58–60).

The “other FFQs” were also developed by including items contributing most to between-person variance or to total population intake. We included 19 FFQs as “other FFQs” (19, 24, 26, 30–36, 38, 39, 42–44, 47, 50, 52, 53). They consisted of an average of 139 items (range, 44–250); 15 of them included portion-size questions, and four assigned standard portion sizes.

### Between-study differences in ranking subjects

For all nutrients, pooled correlation coefficients between FFQ and reference methods ranged from 0.45 for energy and protein to 0.74 for alcohol (table 2), and energy-adjusted correlation coefficients were 0.02–0.08 higher for most nutrients, except for vitamin C (0.05 lower). There were differences between studies due to gender and the reference method used, although they were not statistically significant (table 2). As expected, for most nutrients, correlation coefficients were significantly higher when the reference method was used for 8–14 days than for 1–7 days (table 2). Correlation coefficients did not increase further when the reference method was used for 15 days or more. For all nutrients, we also looked at the number of consecutive days on which a food record was kept and found that correlation coefficients were lower when the reference method consisted of food records kept for more than 5 days consecutively. After energy adjustment, these differences became less pronounced or even reversed.

We observed no statistically significant differences in correlation coefficients between the interviewer- (19, 20, 29, 36) and self-administered FFQs regarding the nutrients considered (table 2).

### Heterogeneity by FFQ design characteristics

In the meta-regression analyses, we observed that FFQs with longer food lists (200 items) had 0.01–0.17 higher correlation coefficients than FFQs with shorter food lists (100 items) for most nutrients (table 3). Correlation coefficients were even higher for longer food lists for crude protein (0.56 for 200 items vs. 0.46 for 100 items, \( p = 0.002 \)), energy-adjusted protein (0.68 vs. 0.51, \( p < 0.001 \)), energy-adjusted total fat (0.68 vs. 0.59, \( p = 0.02 \)), and energy-adjusted vitamin C (0.68 vs. 0.51, \( p = 0.001 \); table 3). The diet history method was used in only one study and therefore was excluded from main analyses.

FFQs with portion-size questions had much higher correlation coefficients for energy-adjusted alcohol than FFQs with predefined standard portions (0.76 vs. 0.61), although they were not statistically significant (\( p = 0.23 \)). On the contrary, FFQs with portion-size questions had significantly lower correlation coefficients for energy-adjusted protein (0.61 vs. 0.69, \( p = 0.03 \)) and for energy-adjusted vitamin C (0.60 vs. 0.80, \( p < 0.001 \)) than FFQs with predefined standard portions. For other nutrients, correlation coefficients were 0.08 lower to 0.08 higher for FFQs with portion-size questions compared with standard portions.

Regarding origin of the FFQ, we observed that correlation coefficients for most crude macronutrients were higher for Block and EPIC FFQs than for Willett and “other” FFQs. For calcium and vitamin C, Willett FFQs performed better, and, after energy adjustment, other correlation coefficients improved for Willett FFQs.

### Validity in absolute intake

Absolute energy intake estimated by FFQs was validated against energy expenditure estimated with the doubly labeled water method (51–53). Two FFQs in small European studies underestimated energy intake by 11 percent and 19 percent (51, 53). In one large study conducted in the United States, energy intake was underestimated by 34 percent for men and 36 percent for women (52), and protein intake was underestimated by 32 percent for men and 29 percent for women. Five EPIC FFQs were also validated against level of protein intake estimated from urinary nitrogen excretion (22, 27, 40, 45, 51, 52) (table 4). In these studies, estimation of protein intake varied from an underestimate of 23 percent to one study that overestimated protein intake by 18 percent for men and 25 percent for women (61). In the latter study, the longest FFQ with 350 items was applied.

### Gender differences in energy intake

Because the goal of FFQs is to distinguish subpopulations that differ with respect to nutrient intake, we tested whether FFQs are able to detect a “known” difference in energy intake between men and women. On average, the gender difference in energy intake was smaller according to FFQs (2.09 MJ for men minus women, 95 percent confidence interval: 1.62, 2.56) than according to the reference methods.
<table>
<thead>
<tr>
<th>No. of FFQs†</th>
<th>Energy intake (range)</th>
<th>Total fat (range)</th>
<th>Protein (range)</th>
<th>Carbohydrate (range)</th>
<th>Alcohol (range)</th>
<th>Calcium (range)</th>
<th>Vitamin C (range)</th>
<th>Dietary fiber (range)</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>0.45 (0.16–0.77)</td>
<td>0.53 (0.18–0.88)</td>
<td>0.45 (0.14–0.70)</td>
<td>0.53 (0.25–0.77)</td>
<td>0.74 (0.29–0.90)</td>
<td>0.55 (0.20–0.75)</td>
<td>0.58 (0.16–0.82)</td>
<td>0.46 (0.25–0.74)</td>
</tr>
</tbody>
</table>

**TABLE 2.** Pooled correlation coefficients† for energy and nutrients between FFQs and reference methods stratified by characteristics of the validation studies

**No. of FFQs†**

<table>
<thead>
<tr>
<th>Reference method</th>
<th>No. of days that reference method was applied</th>
<th>Crude r (range)</th>
<th>Energy-adjusted r (range)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diet history</td>
<td>1–7</td>
<td>0.77 NA§</td>
<td>0.56 (0.27–0.82)</td>
</tr>
<tr>
<td>24-HR§</td>
<td>1–7</td>
<td>0.42 0.54</td>
<td>0.49 (0.16–0.76)</td>
</tr>
<tr>
<td>FR§</td>
<td>1–7</td>
<td>0.53 0.56</td>
<td>0.56 (0.18–0.88)</td>
</tr>
<tr>
<td>FR &lt; 6 consecutive days</td>
<td>1–7</td>
<td>0.56 0.59</td>
<td>0.51 (0.14–0.70)</td>
</tr>
<tr>
<td>FR ≥ 6 consecutive days</td>
<td>1–7</td>
<td>0.42 0.51</td>
<td>0.46 (0.16–0.88)</td>
</tr>
</tbody>
</table>

**Validation study characteristics**

<table>
<thead>
<tr>
<th>Reference method</th>
<th>No. of days that reference method was applied</th>
<th>Crude r (range)</th>
<th>Energy-adjusted r (range)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diet history</td>
<td>1–7</td>
<td>0.77 NA§</td>
<td>0.56 (0.27–0.82)</td>
</tr>
<tr>
<td>24-HR§</td>
<td>1–7</td>
<td>0.42 0.54</td>
<td>0.49 (0.16–0.76)</td>
</tr>
<tr>
<td>FR§</td>
<td>1–7</td>
<td>0.53 0.56</td>
<td>0.56 (0.18–0.88)</td>
</tr>
<tr>
<td>FR &lt; 6 consecutive days</td>
<td>1–7</td>
<td>0.56 0.59</td>
<td>0.51 (0.14–0.70)</td>
</tr>
<tr>
<td>FR ≥ 6 consecutive days</td>
<td>1–7</td>
<td>0.42 0.51</td>
<td>0.46 (0.16–0.88)</td>
</tr>
</tbody>
</table>

**Population characteristics**

<table>
<thead>
<tr>
<th>Gender</th>
<th>Crude r (range)</th>
<th>Energy-adjusted r (range)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>0.44</td>
<td>0.56 (0.27–0.82)</td>
</tr>
<tr>
<td>Women</td>
<td>0.38</td>
<td>0.49 (0.16–0.76)</td>
</tr>
<tr>
<td>Both, unadjusted</td>
<td>0.48</td>
<td>0.56 (0.18–0.88)</td>
</tr>
<tr>
<td>Interviewer</td>
<td>0.50†</td>
<td>0.49† (0.14–0.70)</td>
</tr>
<tr>
<td>Self</td>
<td>0.43</td>
<td>0.51 (0.16–0.70)</td>
</tr>
</tbody>
</table>

† Results pooled as Z-transformed values, weighed by variance, retransformed to correlation coefficients between food frequency questionnaires (FFQs) and reference methods.

‡ Maximum number of FFQs in this category; not all FFQs included all nutrients.

§ NA, not available; 24-HR, 24 hour dietary recall method; FR, food record.

Based on <3 FFQs.
(2.62 MJ, 95 percent confidence interval: 2.24, 3.00). Thus, on average, FFQs underestimated the gender differences in energy intake compared with the reference methods (figure 1) by $-0.53 \text{MJ} \ (95 \text{ percent confidence interval:} -1.13, 0.07, p = 0.09 \text{ using an independent } t \text{ test}).$ Exceptions were the two longest FFQs (31, 45), consisting of 250 and 350 items, respectively; they overestimated the difference in energy intake between men and women. The three FFQs that did not include portion-size questions (24, 30, 40) found on average a smaller gender difference than FFQs that asked portion-size questions (18–24, 31, 45, 48).

**DISCUSSION**

This quantitative review of studies validating FFQs shows that the number of items in the food list is the major
determinant in ranking subjects with respect to their intake (for 100 extra items, correlation coefficients increased by 0.01–0.13). In general, portion-size questions and FFQ origin influenced ranking of subjects only slightly.

An important point of discussion is comparability of the studies included in this review. We aimed to address differences in characteristics of FFQ design and not of study design. We increased comparability by restricting our analyses to FFQs developed to cover the complete diet and validated among adults with Western food habits, and we adjusted for potential confounders such as the reference method and the total number of days this method was used. There were differences in the design of the validation study: correlation coefficients were lower when food records were used for more than 5 days consecutively. However, this finding did not influence our results because there were also numerous studies included with another design.

In this study, we were not able to adjust for differences in energy needs or underreporting related to body weight and physical activity because these data were available for only

### TABLE 4. Mean energy intake reported by the FFQs* compared with mean energy expenditure determined by doubly labeled water, and mean protein intake reported by the FFQs compared with protein intake calculated from nitrogen excreted in urine

<table>
<thead>
<tr>
<th>First author, year (reference no.)</th>
<th>No. of subjects</th>
<th>Mean intake reported by the FFQ</th>
<th>Reference method</th>
<th>Repetitions reference method</th>
<th>Underestimation in FFQ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doubly labeled water</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Andersen, 2003 (53)</td>
<td>17 women</td>
<td>8.28</td>
<td>9.23</td>
<td>1 (10 d&lt;sup&gt;#&lt;/sup&gt;)</td>
<td>11</td>
</tr>
<tr>
<td>Kroke, 1999 (51)</td>
<td>28 men and women</td>
<td>9.05</td>
<td>11.23</td>
<td>1 (14 d)</td>
<td>19</td>
</tr>
<tr>
<td>Subar, 2003 (52)</td>
<td>245 men, 206 women</td>
<td>7.90 for men, 6.11 for women</td>
<td>11.92 for men, 9.53 for women</td>
<td>1 (11–14 d)</td>
<td>34 for men, 36 for women</td>
</tr>
<tr>
<td>Protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kroke, 1999 (51)</td>
<td>75 men, 59 women</td>
<td>75</td>
<td>97</td>
<td>4× in 1 y&lt;sup&gt;*&lt;/sup&gt;</td>
<td>23</td>
</tr>
<tr>
<td>Ocke, 1997 (22)</td>
<td>46 men, 43 women</td>
<td>92 for men, 71 for women</td>
<td>99 for men, 81 for women</td>
<td>4× in 1 y</td>
<td>7 for men, 12 for women</td>
</tr>
<tr>
<td>McKeown, 2001 (40)</td>
<td>57 men, 77 women</td>
<td>89 for men, 78 for women</td>
<td>100 for men, 77 for women</td>
<td>6× in 9 m&lt;sup&gt;#&lt;/sup&gt;</td>
<td>18 for men, 12 for women</td>
</tr>
<tr>
<td>Riboli, 1997 (45)</td>
<td>29 men, 24 women</td>
<td>95 for men, 76 for women</td>
<td>78 for men, 62 for women</td>
<td>8× in 1 y</td>
<td>−18 for men, 25 for women</td>
</tr>
<tr>
<td>Bingham, 1997 (27)</td>
<td>156 women</td>
<td>82</td>
<td>77</td>
<td>4 × 2 d in 1 y&lt;sup&gt;*&lt;/sup&gt;</td>
<td>−6</td>
</tr>
<tr>
<td>Subar, 2003 (52)</td>
<td>202 men, 150 women</td>
<td>71 for men, 55 for women</td>
<td>104 for men, 77 for women</td>
<td>2× in 14 d</td>
<td>32 for men, 29 for women</td>
</tr>
</tbody>
</table>

* FFQs, food frequency questionnaires; d, days; y, year; m, months.

![Figure 1](image-url)
four studies. To account for this limitation, we evaluated both crude and energy-adjusted correlation coefficients because energy adjustment leads to a focus on the relative composition of the dietary pattern (62), and it has been suggested that it reduces correlated errors between the FFQ and reference methods (63).

In addition, variation in FFQ design was limited: the FFQs varied in the number of items and the use of portion-size questions, but differences in the reference period and the administration method were limited, prohibiting conclusions regarding the latter. Finally, we accounted for unknown between-study differences originating from study design and population by using a random-effects model in the meta-regression.

Our analyses showed that FFQs with a longer food list (200 items) were better at ranking people than FFQs with a shorter food list (100 items). These findings were clearest for protein and total fat, which are calculated from many different food sources. In the development phase of an FFQ, similar items are grouped together into items whose composition can be heterogeneous; an example is 20 different meat items combined into two items on the FFQ. Sometimes, items that contribute not much to total intake are omitted although they were important in explaining between-person variance. In summarizing, our results regarding the number of items should be used as an argument not to reduce the length of the food list too much when developing FFQs to rank persons according to nutrient intake.

Results of the meta-regression analyses showed that inclusion of portion sizes did not consistently affect the ranking of different nutrients. Ranking was worse for protein and vitamin C determined by FFQs that used portion-size questions instead of standard portions, and ranking improved for alcohol when FFQs used portion-size questions. An explanation for this unexpected finding might be that, for some foods such as vegetables, it is difficult to indicate how much was eaten, especially when they are part of mixed dishes (64). It might be easier to quantify the number and amount of alcoholic drinks; alcohol intake is ranked relatively well compared with other nutrients (65).

An important disadvantage of using standard portions is that interindividual variance decreases (66, 67). However, two validation studies in Denmark and the Netherlands found only small differences when analyzing FFQs using information from portion-size questions compared with analyzing the same FFQs using standard portions (68, 69). These small differences may reflect that quantification of portion sizes is of minor importance compared with frequency, that the relevant individual portion sizes were not estimated correctly (69), or that portion sizes listed do not correspond well with portions actually consumed. For example, actual portion sizes (e.g., super size) are probably much larger than standard portions used by US Department of Agriculture and the Food and Drug Administration (70, 71). Portion sizes were also estimated in different ways in the FFQs analyzed by including photographs, descriptions, and household measures such as spoons. Thus, it must be taken into account that portion-size questions do not always improve the performance of FFQs or that methods to estimate portion sizes should be improved.

The novelty of this review compared with previous reviews of FFQs (7, 67, 68, 72) is that we specifically analyzed the association between design characteristics of FFQs and their validity. Three other studies specifically compared the validity of Block- and Willett-type FFQs (24, 73, 74). A limitation of our review is that we could not disentangle the effects of type of questionnaire—Block or Willett FFQs—from the effects of use of portion-size questions and number of items. We did not have enough power to do so because only six Block and six Willett FFQs were included in the models. In general, we found that the Block FFQ performed better than the Willett FFQ, but, after energy adjustment, results regarding the different types were more comparable. This finding was also observed previously (24).

Apart from ranking subjects for etiologic studies, FFQs are sometimes used to assess absolute level of intake, for example, to calculate the percentage of a population that meets recommended dietary intake guidelines. We found that FFQs validated against recovery biomarkers underestimated the level of energy intake on average by 20 percent and the level of protein intake by 11 percent; thus, they are not suitable to assess levels accurately. However, FFQs are able to distinguish between subpopulations, as indicated by the analyses of the gender differences in energy intake. This difference was very similar to the gender difference found in a review that used doubly labeled water to estimate mean energy expenditure (75). These results showed that the average difference in energy intake was much smaller when standard portions were used. This is an argument to use at least sex-specific standard portions.

Our review shows that the number of items on the FFQ should not be reduced just because of the length of the food list; doing so might reduce the validity of the FFQ. In addition, portion-size questions do not improve FFQs for all nutrients. We should pay attention to this factor in the development process or by improving methods to estimate portion size. In addition, our review shows that FFQs are able to distinguish between subpopulations, although the magnitude of these differences is underestimated.

ACKNOWLEDGMENTS

This study was supported by the Netherlands Organization for Health Research and Development (ZonMw, grant 91104005).

Conflict of interest: none declared.

REFERENCES


Am J Epidemiol 2007;166:1468–1478


39. McKeown NM, Day NE, Welch AA, et al. Use of biological markers to validate self-reported dietary intake in a random sample of the European Prospective Investigation into Cancer
1478  Molag et al.


