Estimating activity energy expenditure: how valid are physical activity questionnaires?1–3
Heather K Neilson, Paula J Robson, Christine M Friedenreich, and Ilona Csizmadi

ABSTRACT
Activity energy expenditure (AEE) is the modifiable component of total energy expenditure (TEE) derived from all activities, both volitional and nonvolitional. Because AEE may affect health, there is interest in its estimation in free-living people. Physical activity questionnaires (PAQs) could be a feasible approach to AEE estimation in large populations, but it is unclear whether or not any PAQ is valid for this purpose. Our aim was to explore the validity of existing PAQs for estimating usual AEE in adults, using doubly labeled water (DLW) as a criterion measure. We reviewed 20 publications that described PAQ-to-DLW comparisons, summarized study design factors, and appraised criterion validity using mean differences (AEEPAQ – AEEDLW, or TEEPAQ – TEEDLW), 95% limits of agreement, and correlation coefficients (AEEPAQ versus AEEDLW or TEEPAQ versus TEEDLW). Only 2 of 23 PAQs assessed most types of activity over the past year and indicated acceptable criterion validity, with mean differences (TEEPAQ – TEEDLW) of 10% and 2% and correlation coefficients of 0.62 and 0.63, respectively. At the group level, neither overreporting nor underreporting was more prevalent across studies. We speculate that, aside from reporting error, discrepancies between PAQ and DLW estimates may be partly attributable to 1) PAQs not including key activities related to AEE, 2) PAQs and DLW ascertaining different time periods, or 3) inaccurate assignment of metabolic equivalents to self-reported activities. Small sample sizes, use of correlation coefficients, and limited information on individual validity were problematic. Future research should address these issues to clarify the true validity of PAQs for estimating AEE. Am J Clin Nutr 2008;87:279–91.

KEY WORDS Energy expenditure, motor activity, physical activity, metabolic equivalents, questionnaires, retrospective studies, doubly labeled water, validation studies, epidemiologic methods, adults

INTRODUCTION
The amount of energy expended during volitional and nonvolitional activities in humans is an emerging area of interest in the fields of disease prevention and health promotion. Activity energy expenditure (AEE), or activity thermogenesis (1), expands on the classic notions of physical activity and exercise in humans (2, 3) because it refers to thermogenesis from all activities associated with day-to-day living, not just planned and structured exercise.

Recently, higher levels of AEE have been reported to decrease the risk of all-cause mortality in elderly people (4) and blood pressure in younger adults (5). Higher levels of energy expenditure from nonexercise activities may also prevent weight gain (6, 7). However, owing to the inherent difficulties in assessing the duration, frequency, and intensity of all types of activities undertaken by free-living participants in large population studies (8, 9), the amount of AEE required for disease prevention and health promotion remains unclear.

In the continued absence of inexpensive, readily available, relatively noninvasive, valid and reliable technology for measuring AEE in large numbers of free-living humans, researchers may, by necessity, rely on estimates of AEE derived from physical activity questionnaires (PAQs; Figure 1). Although a number of PAQs have been designed to capture various activity parameters, many have shown limited reliability and validity (12). Moreover, it is not entirely clear whether or not any are valid for estimating AEE at the individual level or even at the group level.

The purpose of this review, therefore, was to explore the validity of existing PAQs for estimating AEE in free-living adult populations. We were particularly interested in the potential for PAQs to capture usual AEE in large, population-based, etiologic...
studies of chronic disease. Various criterion measures, such as cardiorespiratory fitness (13–22), motion sensors (13, 17, 23–28), heart rate monitors (21, 22, 29), activity records (13, 16, 26, 30, 31), 24-h physical activity recalls (32), and other PAQs (17, 30, 33) have been used to determine the relative validity of PAQs for capturing various aspects of activity. The most suitable approach for this task, however, is the doubly labeled water (DLW) method (9, 34). DLW measures total energy expenditure (TEE) at the individual level, and AEE is derived by subtraction: AEE = TEE – (resting metabolic rate under near basal conditions + dietary thermogenesis) (Figure 1). Resting metabolic rate under near basal conditions (RMR), also referred to in the literature as basal metabolic rate (BMR), can be estimated by using prediction equations based on weight or weight and height (35), whereas dietary thermogenesis is usually assumed to be 10% of TEE (36, 37). Although there are potential limitations to DLW studies (34, 38–42), the method is widely regarded as the “gold standard” for estimating TEE in free-living individuals (9, 34, 38). This review, therefore, is limited to those studies that have compared DLW estimates of AEE or TEE with measures derived from PAQs.

SUBJECTS AND METHODS

Publication search strategy

In October 2006 we searched the published literature using the electronic databases PubMed (National Institutes of Health, Bethesda, MD) and EMBASE: Excerpta Medica (1980 to week 42; 2006). To locate studies we combined search terms using the following 3 search strategies (indicates all variations of a word stem): 1) (valid* or predict*) and (isotop* or doubly) and (activity or energy expenditure or energy or “food frequency” or food intake or dietary intake), 2) (doubly) and (physical activity) and (self report* or questionnaire* or survey* or record* or recall*), and 3) (energy requirement* and energy balance or energy expenditure) and (physical activity). We searched the dietary literature to account for studies that validated activity and dietary questionnaires simultaneously. Reference lists from relevant publications were also examined to identify relevant studies.

We included studies that aimed to 1) validate a PAQ against DLW, or 2) predict DLW-derived energy expenditure using a PAQ. We defined PAQs as instruments requiring retrospective activity recall beyond 24 h. The search was limited to human studies published in full text, in English, and with no restrictions on publication year. All subjects needed to be adults (aged ≥19 y) studied under free-living conditions. We excluded studies that were exclusive to athletes, pregnant or lactating women, or individuals with acute or chronic disease. We also excluded studies of PAQs designed for, and studied in, one ethnic minority subgroup because we were interested in PAQs that could, in theory, be applied more broadly to the general adult population.

Appraisal of study design

Using criteria described by Rennie and Wareham (43), we summarized study design characteristics that could affect the quality of PAQ validation studies. To appraise study designs, we extracted the following information from each article: mean age and body mass index (BMI; in kg/m²) of the respondents, sample size, length of DLW phase (defined as the number of days between DLW administration and the last day of urine collection), mode of PAQ administration, and the timing of each PAQ relative to its corresponding DLW phase.

Appraisal of questionnaire design

We summarized PAQ attributes to assess face validity for estimating usual AEE. We extracted information on the types and parameters of activities ascertained, time periods recalled, PAQ format, and the outcome summary measures. If this information was not provided for a particular PAQ, we consulted additional publications to complete our data collection. Physical activity duration, frequency, and intensity were classified as “self-reported parameters” only if respondents were asked explicitly to report them. Although a variety of activities were queried across

FIGURE 1. Total energy expenditure (TEE) and activity energy expenditure (AEE) can be derived by using doubly labeled water (DLW) or, possibly, physical activity questionnaires. * Values for intensity (ie, metabolic cost of volitional activities) can be assigned by using the Compendium of Physical Activities (10, 11) or another similar source.
PAQs, we limited our discussion to major activity types, defined as occupational, household, and leisure-time. Using criteria described previously in the literature (8, 38, 44, 45), we classified the format of each PAQ as global, recall, or quantitative.

**Summary of analytic results**

We quantified PAQ criterion validity in terms of the magnitude of agreement and correlation between DLW and PAQ estimates of TEE (TEE<sub>DLW</sub>, TEE<sub>PAQ</sub>) or AEE (AEE<sub>DLW</sub>, AEE<sub>PAQ</sub>). Mean differences (e.g., AEE<sub>PAQ</sub> - AEE<sub>DLW</sub>) and Pearson’s or Spearman’s correlation coefficients (e.g., AEE<sub>PAQ</sub> versus AEE<sub>DLW</sub>) were reported as group measures of agreement in kcal/d. We chose these statistics because no single measure is without limitation (46, 47) and because they are often used to estimate the validity of physical activity assessment tools (48, 49). We also reported SEM differences, P values for correlation coefficients, and 95% limits of agreement (95% LOA), where 95% LOA = mean difference [i.e., AEE<sub>PAQ</sub> - AEE<sub>DLW</sub> or TEE<sub>PAQ</sub> - TEE<sub>DLW</sub>] ± 2 × SD of the mean difference; 50)] as a measure of between-subject variability. If these statistics were not reported by authors, or were not reported as kcal/d, we derived them if enough data were provided in the article.

**RESULTS**

**Study design**

A total of 20 publications (“DLW studies”) were identified as eligible for inclusion in our review. Of these, 36 separate comparisons were described between PAQs and DLW-derived energy expenditure. Characteristics of the various study designs are summarized in Table 1. It should be noted that Philippaerts et al (62) described several comparisons between TEE<sub>DLW</sub> and various indexes within the modified Stanford Usual Activity (Five City) Questionnaire; however, we focused on the 7-d index because its content was not limited to moderate or vigorous activities, and it attempted to estimate energy expenditure as opposed to a unitless score. Moreover, 2 publications by Lof et al (59, 60) described the same DLW study in the same respondents, with the exception of 3 women. Thus, for the purposes of this review, we referred to these publications in combination.

**Study population**

In Table 1, the largest study involved 80 women (51), and the smallest studies involved only 10 (63) or 13 (56, 57, 64) participants. Of 20 studies, 8 included both sexes (52, 56, 58, 65–67, 70, 71), 4 included only males (53–55, 62), and 8 included only females (51, 57, 59–61, 64, 68, 69). Six studies were conducted in the elderly, i.e., mean age >65 y (53, 56, 53, 65–67). Of all the studies, one (64) intentionally focused on obese individuals (i.e., mean BMI ≥ 30), whereas 13 (51, 52, 54, 55, 58, 61, 65–71) included generally overweight subjects (i.e., mean BMI of 25 to <30), but only 2 (69, 70) did so intentionally.

**PAQ administration**

Most studies (12 of 20) used interviewer-administered PAQs (53, 54, 56–60, 62, 63, 66, 67, 70, 71), whereas only 5 used self-administered PAQs (51, 55, 61, 68, 69). For 3 studies, the mode was not reported (52, 64, 65).

Only 4 of 36 comparisons (57, 59, 60, 62, 71) involved PAQs that covered exactly the same period of activity as the DLW phase, with the 4 periods ranging from 7 to 14 d (Table 1). Ten of 36 comparisons (in 7 publications) (52, 53, 55, 56, 62, 67, 69) involved past year PAQs, with corresponding DLW phases ranging from 10 to 14 d. Twenty-one of the 36 comparisons presented in Table 1 (in 10 publications: 51, 53–55, 57, 59, 60, 62, 64, 68, 71) involved PAQs administered at the end of the DLW phase, when the period of recall would have included the period of DLW measurement. The timing of PAQ administration was not reported for 3 of 36 comparisons (52, 56, 58).

**Questionnaire design**

Within the 20 publications included in our review, we identified 23 distinct PAQs (Table 2). Eleven of the 23 PAQs (51, 55, 58, 61, 62, 68, 69, 71, 76, 77) covered all major types of activity (i.e., occupational, household, and leisure time) to varying degrees, and 3 others (18, 60, 70) included all types of activity in terms of intensity. Regarding the time period recalled, 6 of 23 PAQs inquired about activity over the past year (32, 52, 62, 69, 75, 77). One other PAQ asked about occupational activities over the past year and leisure activities over the past month (55), whereas another PAQ asked about sports and leisure activities over the past year and other activities as they are usually performed (73). Of 23 PAQs, we classified 7 as quantitative (15, 51, 55, 65, 68, 69, 75, 77), 13 as recall (18, 32, 52, 60–63, 70–73, 76), and 2 as global (74, 78). We were unable to classify the format of one PAQ (58) because of lack of information. Fifteen of 23 PAQs were used to derive estimates of AEE or TEE (15, 18, 51, 52, 55, 56–62, 68–70, 73, 75, 77), and all were deemed to be recall or quantitative PAQs. Across all PAQs, intensity was assigned using information provided in the Compendium of Physical Activities (10, 11) or another approach (15, 18, 25, 61, 64, 73, 75, 77, 83, 84).

Four PAQs [Questionnaire d’Activité Physique Saint-Etienne (QAPSE; 77), Tecumseh Community Health Study (TCHS; 62), Tecumseh Occupational Activity and past month Minnesota Leisure-Time Questionnaire (MLTQ; 55), and Tecumseh Occupational Activity and past year MLTQ (69)] asked about all major types of activities, covered activities over the past year, and provided estimates of energy expenditure.

**Analytic results**

To appraise criterion validity, summary statistics are presented and arranged according to PAQ outcome measure and period of recall in Table 3. Across all 20 publications, a total of 10 comparisons were made between DLW and either a unitless PAQ score (53, 62, 63, 65) or the duration of activity derived from a PAQ (58, 71). Because the 10 PAQs involved in these comparisons estimated neither AEE nor TEE, we excluded these 10 comparisons from our appraisal, which left only 26 comparisons (16 publications) in Table 3.

**Group level agreement**

To assess the level of agreement between PAQs and DLW, we summarized the mean differences for 26 comparisons. As shown in Table 3, some comparisons were made in terms of AEE (i.e., AEE<sub>PAQ</sub> - AEE<sub>DLW</sub>), whereas others focused on TEE (i.e., TEE<sub>PAQ</sub> - TEE<sub>DLW</sub>, where TEE<sub>PAQ</sub> was derived by summation; Figure 1). Two studies compared AEE<sub>PAQ</sub> with AEE<sub>DLW</sub> and TEE<sub>PAQ</sub> with TEE<sub>DLW</sub>; however, we chose to report on only one comparison from each study: AEE in the first instance (70) and...
### TABLE 1
Study design factors for comparisons made between physical activity questionnaires (PAQs) and doubly labeled water (DLW)

<table>
<thead>
<tr>
<th>DLW study</th>
<th>Name of questionnaire</th>
<th>PAQ period of recall in DLW study</th>
<th>Length of DLW phase</th>
<th>Timing of PAQ relative to DLW</th>
<th>Mode in DLW study</th>
<th>Study population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adams et al, 2005 (51)</td>
<td>1) 7-d Physical Activity Recall</td>
<td>j) Past week</td>
<td>14</td>
<td>Mid or end</td>
<td>Self</td>
<td>80 F</td>
</tr>
<tr>
<td>2) 7-d Minnesota Leisure Time</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Barnard et al, 2002 (52)</td>
<td>Modifiable Activity, adapted</td>
<td>Past year</td>
<td>14</td>
<td>NR</td>
<td>NR</td>
<td>7 F, 8 M</td>
</tr>
<tr>
<td>Bonnefoy et al, 2001 (53)</td>
<td>1) Lipid Research Clinics</td>
<td>j) Usual</td>
<td>14</td>
<td>End, in random order</td>
<td>Int</td>
<td>19 M</td>
</tr>
<tr>
<td>2) Modified Baecke</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3) Stanford Usual Activity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4) Dallosso, modified</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5) Physical Activity Survey for the Elderly</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6) Yale PAQ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7) Minnesota Leisure Time, past year</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8) Harvard/College Alumni/Paffenbarger</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9) Questionnaire d’Activité Physique Saint-Etienne</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10) 7-d Physical Activity Recall</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conway et al, 2002a (54)</td>
<td>7-d Physical Activity Recall</td>
<td>Past week</td>
<td>14</td>
<td>End</td>
<td>Int</td>
<td>24 M</td>
</tr>
<tr>
<td>Conway et al, 2002b (55)</td>
<td>Tecumseh Occupational Activity and past month Minnesota Leisure Time</td>
<td>Past year and past month</td>
<td>14</td>
<td>Start, end (used mean)</td>
<td>Self</td>
<td>24 M</td>
</tr>
<tr>
<td>Goran and Poehlman, 1992 (56)</td>
<td>Minnesota Leisure Time, past year</td>
<td>Past year</td>
<td>10</td>
<td>NR</td>
<td>Int</td>
<td>6 F, 7 M</td>
</tr>
<tr>
<td>Leenders et al, 2001 (57)</td>
<td>7-d Physical Activity Recall</td>
<td>Past week</td>
<td>7</td>
<td>End</td>
<td>Int</td>
<td>13 F</td>
</tr>
<tr>
<td>Livingstone et al, 1991 (58)</td>
<td>Habitual</td>
<td>Habitual</td>
<td>15</td>
<td>NR</td>
<td>Int</td>
<td>16 F, 16 M</td>
</tr>
<tr>
<td>Lof et al, 2003 (59)</td>
<td>2-wk Recall</td>
<td>Past 2 wk</td>
<td>14</td>
<td>End</td>
<td>Int</td>
<td>34 F</td>
</tr>
<tr>
<td>Lof and Forsum, 2004 (60)</td>
<td>2-wk Recall</td>
<td>Past 2 wk</td>
<td>14</td>
<td>End</td>
<td>Int</td>
<td>37 F</td>
</tr>
<tr>
<td>Mahabir et al, 2006 (61)</td>
<td>1) Harvard/College Alumni/Paffenbarger</td>
<td>j) Typical week</td>
<td>NR</td>
<td>Start</td>
<td>Self</td>
<td>65 F</td>
</tr>
<tr>
<td>2) 7-d Physical Activity Recall</td>
<td></td>
<td>j) Past week</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3) CAPS 4-wk Activity Recall</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4) CAPS Typical Week Activity Survey</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Philippaerts et al, 1999 (62)</td>
<td>1) Stanford Usual Activity, modified; Five City Questionnaire (7-d index)</td>
<td>j) Past week</td>
<td>14</td>
<td>Mid, end (mean)</td>
<td>Int</td>
<td>19 M</td>
</tr>
<tr>
<td>2) Tecumseh Community Health Study</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3) Baecke</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reilly et al, 1993 (63)</td>
<td>Dallosso, modified</td>
<td>Typical day during DLW phase</td>
<td>14–18</td>
<td>Mid (typical day)</td>
<td>Int</td>
<td>10 F</td>
</tr>
<tr>
<td>Racette et al, 1995 (64)</td>
<td>7-d Physical Activity Recall</td>
<td>Past week</td>
<td>14</td>
<td>End</td>
<td>NR</td>
<td>13 F</td>
</tr>
<tr>
<td>Physical Activity Survey for the Elderly</td>
<td></td>
<td>Past week</td>
<td>14</td>
<td>Mid</td>
<td>NR</td>
<td>11 F, 10 M</td>
</tr>
<tr>
<td>Schuit et al, 1997 (65)</td>
<td>7-d Physical Activity Recall</td>
<td>Past week</td>
<td>14</td>
<td>Mid</td>
<td>NR</td>
<td>11 F, 10 M</td>
</tr>
<tr>
<td>Seale et al, 2002 (66)</td>
<td>7-d Physical Activity Recall</td>
<td>Past week</td>
<td>14</td>
<td>Mid (twice)</td>
<td>Int</td>
<td>13 F, 14 M</td>
</tr>
</tbody>
</table>

(Continued)
TABLE 1 (Continued)

<table>
<thead>
<tr>
<th>DLW study</th>
<th>Name of questionnaire</th>
<th>PAQ period of recall in DLW study</th>
<th>Length of DLW phase</th>
<th>Timing of PAQ relative to DLW</th>
<th>Mode of PAQ in DLW study</th>
<th>Study population</th>
<th>No. of subjects</th>
<th>Age</th>
<th>BMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Starling et al, 1999 (67)</td>
<td>Minnesota Leisure Time, past year</td>
<td>1) Past year</td>
<td>10</td>
<td>Start</td>
<td>Int</td>
<td>35 F; 32 M</td>
<td>67 ± 9 (F), 66 ± 11 (M)²</td>
<td>24.8 ± 3.9 (F), 25.7 ± 4.5 (M)²</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yale PAQ</td>
<td>2) Past month and typical week, past month</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Staten et al, 2001 (68)</td>
<td>Arizona Activity Frequency</td>
<td>Past week or past month</td>
<td>9</td>
<td>Pre (past month), end (past week)</td>
<td>Self</td>
<td>35 F</td>
<td>43.8 ± 9.2²</td>
<td>28.0 ± 8.1²</td>
<td></td>
</tr>
<tr>
<td>Walsh et al, 2004 (69)</td>
<td>Tecumseh Occupational Activity and Minnesota Leisure Time, past year</td>
<td>Past year</td>
<td>14</td>
<td>Start</td>
<td>Self</td>
<td>44 F</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

|                  | 1) Minnesota Leisure Time, past year                                                  | 1) Past year                     | 10                  | Start                        | Int                      | 35 F; 32 M            | 67 ± 9 (F), 66 ± 11 (M)² | 24.8 ± 3.9 (F), 25.7 ± 4.5 (M)² |
|                  | 2) Yale PAQ                                                                            | 2) Past month and typical week, past month |                     |                              |                          |                        |                                |

In 8 comparisons, the direction of bias [ie, AEE PAQ − AEE DLW (51) or TEE PAQ − TEE DLW (54, 55, 59, 61)] became more positive with increasing mean AEE [ie, average of AEE PAQ and AEE DLW (51)] or TEE [ie, average of TEE PAQ and TEE DLW (54, 55, 59, 61)]. In other words, there was a positive trend between the individual differences and the means. Conversely, a slight negative trend, in which the direction of bias became more negative with increasing AEE was reported for one comparison (70). Two other comparisons also showed negative trends (57, 67); however, in these instances the differences (AEE PAQ − AEE DLW) were plotted against AEE DLW rather than as the average of the 2 measures, which is recommended (86).

To assess validity further, we examined correlation coefficients reported for 19 comparisons in Table 3 (51, 53–56, 61, 62, 68–70). Coefficients ranged from 0.05 (7-d MLTQ in 80 middle-aged females; 51) to 0.83 (past year MLTQ in 13 elderly males and females; 56). Five coefficients were notably >0.60 (55, 56, 62, 68).

We also evaluated PAQ validity by considering mean differences and correlations, simultaneously, for 19 of the 26 comparisons in Table 3. Of these, only 3 comparisons resulted in mean percentage differences of ≤10% and a correlation of ≥0.60 [modified Stanford Usual Activity 7-d index (62), TCHS (62), and Tecumseh Occupational Activity and past month MLTQ (55)]. All 3 comparisons were conducted in middle-aged men with BMIs ranging from values indicating normal weight to overweight, on average. Two of these PAQs—the Tecumseh

TEE in the other (68) because more data were reported on TEE in this publication. Overall, the highest percentage difference in mean values was 113% for the 7-d MLTQ (51), whereas 8 other comparisons resulted in percentage differences of ≤10% (55, 59, 60, 62, 64, 66, 67, 70). In terms of kcal/d, the magnitude of the mean difference (AEE PAQ − AEE DLW) ranged from −10 kcal/d (Yale PAQ: 67) to 952 kcal/d (7-d MLTQ: 51). The mean difference (TEE PAQ − TEE DLW) ranged in magnitude from 17 kcal/d (0.7%) (66) to 989 kcal/d (31%) (64). However, the same questionnaire resulted in percentage differences of 10% (55, 59, 61) and ±5% (56), and, in the third study, −487 kcal/d (−17%) (57) and −752 kcal/d (−62%) for females and males, respectively (67). Conversely, 5 evaluations of TEE PAQ from the 7-d Physical Activity Recall questionnaire (53, 54, 61, 64, 66) showed group overreporting, with mean differences ranging from 17 kcal/d (0.7%) (66) to 989 kcal/d (31%) (54). However, the same questionnaire resulted in group underreporting in 2 (57, 70) of 3 studies (51, 57, 70) that compared this PAQ with AEE DLW.
<table>
<thead>
<tr>
<th>Name of PAQ (primary reference)</th>
<th>Reference to DLW study</th>
<th>Type</th>
<th>Units in DLW study</th>
<th>PAQ period of recall in DLW study</th>
<th>Self-reported variables</th>
<th>Types of activities</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-wk Recall (60)</td>
<td>Lof et al, 2003 (59); Lof and Forsum 2004 (60)</td>
<td>Recall kcal/d, kJ/d</td>
<td>Past 2 wk</td>
<td>x</td>
<td>x</td>
<td>s, very light, light, moderate, vigorous, very vigorous</td>
</tr>
<tr>
<td>7-d Physical activity recall, Stanford Five City Project (70)</td>
<td>Washburn et al, 2003 (70)</td>
<td>Recall kcal/d, MJ/d, kcal/kg/d</td>
<td>Past week</td>
<td>x</td>
<td>x</td>
<td>s, moderate, hard, very hard</td>
</tr>
<tr>
<td>Arizona Activity Frequency (68)</td>
<td>Staten et al, 2001 (68)</td>
<td>Quantitative kJ/d</td>
<td>Past week or past month</td>
<td>x</td>
<td>x</td>
<td>H, L, O, pc, s</td>
</tr>
<tr>
<td>Baecke (72)</td>
<td>Philippaerts et al, 1999 (62)</td>
<td>Recall Scores for multiple indexes</td>
<td>Usual</td>
<td>x</td>
<td>(sport)</td>
<td>x</td>
</tr>
<tr>
<td>Baecke, modified (32)</td>
<td>Bonnefoy et al, 2001 (53)</td>
<td>Recall Score</td>
<td>Past year</td>
<td>x</td>
<td>(L)</td>
<td>x</td>
</tr>
<tr>
<td>CAPS 4-wk Activity Recall (61)</td>
<td>Mahabir et al, 2006 (61)</td>
<td>Recall kcal/d</td>
<td>Past month</td>
<td>x</td>
<td>x</td>
<td>H, L, O, c, t, w (focus on moderate, vigorous)</td>
</tr>
<tr>
<td>CAPS Typical Week Activity Survey (61)</td>
<td>Mahabir et al, 2006 (61)</td>
<td>Recall kcal/d</td>
<td>Typical week</td>
<td>x</td>
<td>x</td>
<td>H, L, O, t, w (focus on moderate, vigorous)</td>
</tr>
<tr>
<td>Dallosso, modified (63)</td>
<td>Reilly et al, 1993 (63)</td>
<td>Recall Score</td>
<td>Typical day last week, typical week (53); Typical day during DLW phase (63)</td>
<td>x</td>
<td>(H, L, muscle-loading)</td>
<td>x</td>
</tr>
<tr>
<td>Harvard/College Alumni/ Paffenbarger (73)</td>
<td>Bonnefoy et al, 2001 (53); Mahabir et al, 2006 (61)</td>
<td>Recall kcal/wk</td>
<td>Usual and past year (53); typical week (61)</td>
<td>x</td>
<td>(L)</td>
<td>x</td>
</tr>
<tr>
<td>Lipid Research Clinics (74)</td>
<td>Bonnefoy et al, 2001 (53)</td>
<td>Global Score</td>
<td>Usual</td>
<td>x</td>
<td>(L)</td>
<td>x</td>
</tr>
<tr>
<td>Minnesota Leisure Time, past year (75)</td>
<td>Bonnefoy et al, 2001 (53); Goran and Poehlman, 1992 (56); Starling et al, 1999 (67)</td>
<td>Quantitative kcal/d, MJ/d</td>
<td>Past year</td>
<td>x</td>
<td>x</td>
<td>H, L, c, t, w</td>
</tr>
<tr>
<td>Minnesota Leisure Time, 7-d recall (51)</td>
<td>Adam et al, 2005 (51)</td>
<td>Quantitative kcal/kg/d</td>
<td>Past week</td>
<td>x</td>
<td>x</td>
<td>H, L, O, s, miscellaneous</td>
</tr>
<tr>
<td>Modifiable Activity, adapted (52)</td>
<td>Barnard et al, 2002 (52)</td>
<td>Recall MJ/d</td>
<td>Past year</td>
<td>x</td>
<td>(L, O, t)</td>
<td>x</td>
</tr>
<tr>
<td>MONICA Optional Study of Physical Activity (MOPSA) questionnaire, modified (71)</td>
<td>Yao et al, 2002 (71)</td>
<td>Recall min/d by intensity (METs); score for overall nonwork activity score (h/wk in moderate activity; x2 if more vigorous)</td>
<td>Past 8 d</td>
<td>x</td>
<td>(L)</td>
<td>x</td>
</tr>
<tr>
<td>NR (58)</td>
<td>Livingstone et al, 1991 (58)</td>
<td>Unable to discern</td>
<td>Habitual</td>
<td>x</td>
<td>Unable to discern</td>
<td>H, L, O</td>
</tr>
<tr>
<td>Physical Activity Scale for the Elderly (76)</td>
<td>Bonnefoy et al, 2001 (53); Schuit et al, 1997 (65)</td>
<td>Recall score</td>
<td>Past week</td>
<td>x</td>
<td>(L, O, s, w)</td>
<td>x</td>
</tr>
<tr>
<td>Questionnaire d’Activité Physique Saint-Etienne (77)</td>
<td>Bonnefoy et al, 2001 (53)</td>
<td>Quantitative kJ/d</td>
<td>Typical week in past year</td>
<td>x</td>
<td>x</td>
<td>H, L, O, pc, s, t</td>
</tr>
<tr>
<td>Stanford Usual Activity (78)</td>
<td>Bonnefoy et al, 2001 (53)</td>
<td>Global Score</td>
<td>Usual and past 3 mo</td>
<td>x</td>
<td>(L)</td>
<td>x</td>
</tr>
<tr>
<td>Stanford Usual Activity, modified/Five City Questionnaire (18)</td>
<td>Philippaerts et al, 1999 (62)</td>
<td>Recall kcal/d for 7-d index (kcal·kg⁻¹·wk⁻¹) in sports for 3-mo index; run-walk-jog index; sweat index</td>
<td>Past week (7-d index); past 3 mo (3-mo index); NR (other indexes)</td>
<td>x</td>
<td>(L, moderate vigorous; miles: bike, swim, run, walk, jog)</td>
<td>x</td>
</tr>
</tbody>
</table>

(Continued)
Occupational Activity and past month MLTQ (55) and the TCHS PAQ (62)—also showed the greatest potential for capturing AEE in our earlier appraisal of questionnaire design (Table 2). The former PAQ was self-administered (55), whereas the latter was administered in a face-to-face interview (62).

Individual level agreement

Inconsistent reporting precluded us from comparing individual level results across most studies and, thus, individual agreement was not summarized for this review. Only 7 of the 20 publications were explicit in reporting individual level agreement. Results were expressed as the proportions of positive and negative differences (ie, AEE\textsubscript{PAQ} − AEE\textsubscript{DLW}; 51), as the number of reporters within a certain percentage of their DLW estimate [within 10% (55, 70); TEE\textsubscript{PAQ} ≤5%, 5–10%, 10–20%, or >20% of TEE\textsubscript{DLW} (54)], or as tables of individual results (52, 56, 64).

We judged between-subject variability in terms of the 95% LOA for 22 of the 26 comparisons in Table 3. The widths of 95% LOA (ie, upper limit minus lower limit) for the mean difference (AEE\textsubscript{PAQ} − AEE\textsubscript{DLW}) spanned from 817 kcal/d (past year MLTQ (56) to 4096 kcal/d (7-d MLTQ; 51). For the mean difference (TEE\textsubscript{PAQ} − TEE\textsubscript{DLW}) the widths of 95% LOA ranged from 1133 kcal/d (Tecumseh Occupational Activity and past year MLTQ; 69) to 17 948 kcal/d (Cross-Cultural Activity Participation Study 4-wk recall; 61).

DISCUSSION

We found that the vast majority of PAQs previously compared with DLW did not have sufficient face validity for estimating usual AEE, judging from the types of activities, time period recalled, and summary measures derived from each PAQ. In a comparison of AEE\textsubscript{PAQ} with AEE\textsubscript{DLW} or of TEE\textsubscript{PAQ} with TEE\textsubscript{DLW}, the percentage difference in means was <10% in nearly one-third of the comparisons and about one-third of all correlations was >0.60. However, high correlations and small differences in means rarely coincided. Most validation study results were reported at the group level, with much less information on individual reporting. No study involved >80 subjects and many focused on overweight, female, and elderly subgroups. Most of the DLW experiments did not assess the same time period recalled in PAQs.

One common feature of the PAQs in this review was their inclusion of exercise, defined as planned, structured, and repetitious bodily movements intended to improve or maintain one or more components of physical fitness (87). In this review, however, our interest applied more broadly to AEE, or the energy expended from all exercise and nonexercise activities, both volitional and nonvolitional. We also wanted to explore PAQ validity in the context of population-based, etiologic studies of chronic disease. For this reason we were interested in “usual” AEE; in other words, relatively stable patterns of activity that, if prolonged, could contribute to disease risk.

Only 4 of the 23 PAQs in our review contained all of the basic design elements required for estimating usual AEE (Table 2): the QAPSE PAQ, TCHS PAQ, Tecumseh Occupational Activity and past month MLTQ, and the Tecumseh Occupational Activity and past year MLTQ. In addition to queries regarding the “major” activity types, these PAQs also inquired about personal care, climbing stairs, walking, transportation, or sedentary activities.
### Table 3
Summary of results from comparison studies between physical activity questionnaires (PAQs) and doubly labeled water (DLW)

<table>
<thead>
<tr>
<th>Name of questionnaire (reference to DLW validation study)</th>
<th>No. of subjects</th>
<th>Difference in means, PAQ – DLW kcal/d</th>
<th>Unadjusted correlation, PAQ vs DLW kcal/d</th>
<th>Individual differences, PAQ – DLW kcal/d</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AEEPAQ vs AEEDLW</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Past week to past month</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7-d Physical Activity Recall; Adams et al, 2005 (51)</td>
<td>80</td>
<td>106 ± 68 (13)</td>
<td>0.19 (kcal · kg⁻¹ · d⁻¹)</td>
<td>95% LOA: −1095 to 1306</td>
</tr>
<tr>
<td>7-d Physical Activity Recall; Leenders et al, 2001 (57)</td>
<td>13</td>
<td>−156 (–20)</td>
<td>NR</td>
<td>Range of differences: −633 to 280</td>
</tr>
<tr>
<td>Minnesota Leisure Time, 7-d; Adams et al, 2005 (51)</td>
<td>46</td>
<td>−53 ± 66 (–6)</td>
<td>0.12</td>
<td>95% LOA: −1043 to 937</td>
</tr>
<tr>
<td>7-d Physical Activity Recall; Washburn et al, 2003 (70)</td>
<td>80</td>
<td>952 ± 117 (113)</td>
<td>0.05 (kcal · kg⁻¹ · d⁻¹)</td>
<td>95% LOA: −1096 to 3000</td>
</tr>
<tr>
<td>7-d Physical Activity Recall; Racette, 1995 (55)</td>
<td>19</td>
<td>90 ± 64 (11)</td>
<td>0.18, r = 0.10</td>
<td>95% LOA: −464 to 645</td>
</tr>
<tr>
<td>7-d Physical Activity Recall; Seale et al, 1995 (64)</td>
<td>67</td>
<td>−10 ± 82 (–1), F¹, M¹</td>
<td>NR</td>
<td>95% LOA: −981 to 963 F; −1518 to 1310 M</td>
</tr>
<tr>
<td>Past year</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minnesota Leisure Time, past year; Bonneyfo, 2001 (53)</td>
<td>19</td>
<td>−313 ± 100 (–39)</td>
<td>0.23, r = 0.17</td>
<td>95% LOA: −1188 to 562</td>
</tr>
<tr>
<td>7-d Physical Activity Recall; Goran et al, 2002a (54)</td>
<td>13</td>
<td>−208 ± 57 (–37)</td>
<td>0.83²</td>
<td>95% LOA: −617 to 200</td>
</tr>
<tr>
<td>7-d Physical Activity Recall; Starling et al, 1999 (67)</td>
<td>67</td>
<td>−487 ± 59 (–56), F¹, M¹</td>
<td>NR</td>
<td>95% LOA: −1185 to 211 F; −1724 to 220 M</td>
</tr>
<tr>
<td>Other periods of recall</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Harvard/College Alumni/ Paffenbarger; Bonneyfo, 2001 (53) (usual and past year)</td>
<td>19</td>
<td>−240 ± 96 (–30)</td>
<td>0.39, r = 0.37</td>
<td>95% LOA: −1076 to 596</td>
</tr>
<tr>
<td><strong>TEEPAQ vs TEEELW</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Past week to past month</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7-d Physical Activity Recall; Bonneyfo, 2001 (53)</td>
<td>19</td>
<td>276 ± 155 (11)</td>
<td>0.37, r = 0.51⁷</td>
<td>95% LOA: −1075 to 1625</td>
</tr>
<tr>
<td>7-d Physical Activity Recall; Conway et al, 2002a (54)</td>
<td>24</td>
<td>989 ± 325 (31)</td>
<td>0.14 (r = 0.37)⁵,⁸</td>
<td>95% LOA: −2195 to 4174</td>
</tr>
<tr>
<td>7-d Physical Activity Recall; Mahabir et al, 2006 (61)</td>
<td>65</td>
<td>Using median: 410 (16) (IQR not reported)⁴</td>
<td>r = 0.47²</td>
<td>95% LOA: −1394 to 2860</td>
</tr>
<tr>
<td>7-d Physical Activity Recall; Racette, 1995 (64)</td>
<td>13</td>
<td>49 ± 87 (2)</td>
<td>NR</td>
<td>95% LOA during weight maintenance: 7² −579 to 679</td>
</tr>
<tr>
<td>7-d Physical Activity Recall; Seale et al, 2002 (60)</td>
<td>27</td>
<td>17 ± 183 (7.5) F¹, M¹; 304 ± 157 (10) M¹; used measured REE, F and M</td>
<td>NR</td>
<td>95% LOA: −1303 to 1336 F; −872 to 1479 M used measured REE</td>
</tr>
<tr>
<td>2-wk Recall; Lof et al, 2003 (59); Lof and Forsum, 2004 (60)</td>
<td>34 (59); 37 (60)</td>
<td>130 ± 79 (5) for predicted BMR (59); −160 ± NR (−6) for measured BMR (60)</td>
<td>NR</td>
<td>95% LOA: −793 to 1053 (59);⁵,⁶ NR (60)</td>
</tr>
<tr>
<td>Arizona Activity Frequency; Staten, 2001 (68)</td>
<td>35</td>
<td>−462 ± NR (–20), past month⁵,⁷, −441 ± NR (–19), past week⁵,⁷,⁸, using RMR from indirect calorimetry</td>
<td>r² = 0.40² (r = 0.63)⁴,⁸, past month; NR, past week; using RMR from indirect calorimetry</td>
<td>NR</td>
</tr>
<tr>
<td>CAPS 4-wk Activity Recall; Mahabir et al, 2006 (61)</td>
<td>65</td>
<td>Using median: 291 (11) (IQR not reported)⁴</td>
<td>r = 0.15</td>
<td>95% LOA: −7208 to 10740</td>
</tr>
<tr>
<td>CAPS Typical Week Activity Survey; Mahabir et al, 2006 (61)</td>
<td>65</td>
<td>Using median: −800 (−31) (IQR not reported)⁷</td>
<td>r = 0.16</td>
<td>95% LOA: −3372 to 2545</td>
</tr>
<tr>
<td>Stanford Usual Activity, modified; Five City Questionnaire (7-d index); Philippaerts et al, 1999 (62)</td>
<td>19</td>
<td>7-d index: −327 ± NR (−10)⁵,⁶,⁸</td>
<td>0.61¹</td>
<td>NR</td>
</tr>
<tr>
<td>Past year</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Modifiable Activity, adapted; Barnard et al, 2002 (52)</td>
<td>15</td>
<td>−800 ± 496 (−22), n = 5</td>
<td>NR, “no correlation”</td>
<td>95% LOA: −3939 to 2339</td>
</tr>
</tbody>
</table>

(Continued)
In theory, each activity could contribute to PAQ validity for estimating AEE. For instance, lower intensity activities (88) and posture during sedentary activities can influence AEE (7, 89, 90). Although all 4 PAQs inquired about sedentary activity, the Tecumseh Occupational Activity and past month MLTQ contrasted sitting versus standing activities, inquired about sleeping, and asked about “general activities” such as childcare, reading, and watching television (55). Incidentally, this PAQ also had higher criterion validity (Table 3) and was quantitative in format. In contrast, 3 separate evaluations of the past year Minnesota Leisure Time PAQ showed underestimation of AEE by 37% to 62% at the group level (53, 56, 67). Although we are uncertain why underestimation occurred, it is possible that the omission of key activities (eg, occupational) from this PAQ contributed to its discrepancies with DLW.

Face validity is important not only when interpreting PAQ validation studies, but also in etiologic research. Recently, the issue of PAQ validity was raised (91) when significant inverse associations were found between all-cause mortality and AEE estimated from DLW (4). Self-reported stair climbing and work for pay were more likely to be reported by participants with a higher AEE. In contrast, the proportion of individuals who self-reported high-intensity exercise, walking for exercise, or walking for reasons other than for exercise did not change significantly across AEE tertiles. It was speculated that self-report errors may be to blame for the latter (91). We propose, however, that these findings may have arisen in part because these activities were not significant contributors to AEE in this population.

Before discrediting self-reporting, therefore, we will need to determine whether or not PAQs have sufficient face validity for testing AEE-related hypotheses. Only then can we determine whether or not PAQs have sufficient face validity for testing AEE-related hypotheses.
that PAQs that appear to be valid (or not valid) under the conditions in which they were studied would perform differently in other contexts. Presumably, in some cases, small sample sizes were due to the prohibitive monetary cost of DLW experiments at the time the studies were undertaken. Regardless, the decision to “scale down” validation studies clearly comes at the expense of an inability to generalize results to other populations and provides less precise estimates of PAQ validity. Correlation coefficients, for example, become less precise when based on smaller samples because of an increased SE [ie, SE = [(1 − r^2) / (n − 2)]^(1/2) (47, 94)]. Unfortunately, SEs were rarely provided in the articles, whereas P values (H_0: p = 0) were common. It should be noted that less than half of the correlation coefficients in Table 3 differed significantly from zero, but this result may have arisen simply as a consequence of smaller sample sizes, at least in some instances. Even statistically significant correlations may be imprecise if based on very few observations. Thus, all correlations in Table 3 should be interpreted with caution.

Even with sufficient sample sizes, the limitations of correlation coefficients are well documented (47, 49, 50, 94), with the literature emphasizing 2 potential pitfalls. First, correlations depend on the degree of between-subject variability in a given study population, so an acceptable correlation found in one PAQ validation study may not apply to groups with a different range of energy expenditure levels (49). Second, correlations are measures of association as opposed to agreement. A method with known systematic bias can correlate quite strongly with an unbiased reference measure (47, 49, 50), thereby masking a lack of agreement between measures. Despite these limitations, correlation coefficients have thus far been the most commonly used statistics in the PAQ validation study literature (49). We recommend, as have others (47, 49, 94), that PAQ validity not be judged solely on the basis of correlations, but rather on several statistical methods that would each compensate for the other’s unique limitations.

Unfortunately, we found that important details of PAQ validation studies were sometimes omitted from the published literature, which made it difficult to generalize findings to other populations. In this review we noted missing information on the mode of PAQ administration (52, 64, 65), the BMI of validation study participants (53, 63), the timing of PAQs relative to DLW phases (52, 56, 58), and the length of DLW phase (61). Thus, authors are encouraged to be more specific when reporting on PAQ validation studies.

Bland-Altman plots, although not always presented by authors, were more informative. Some plots showed mean differences (ie AEE_PAQ − AEE_DLW or TEE_PAQ − TEE_DLW) or group level bias that increased with the level of energy expenditure, typically in the positive direction. These positive trends may imply that populations with higher levels of AEE or TEE tend to overreport their activities in PAQs. Although this tendency was apparent in some groups, it is noteworthy that not every individual with a high level of energy expenditure would overreport their activities. The group level, proportional bias we observed suggests systematic error in physical activity reporting that could be corrected, perhaps using regression calibration techniques (95), or with a modified PAQ design based on the determinants of misreporting, once they are better understood.

Alternatively, it is possible that the positive trends observed in several Bland-Altman plots were the result of random error. In a recent simulation study (49), a spurious, positive trend resulted in a Bland-Altman plot when one hypothetical unbiased measure of physical activity (M) was compared with a second, unbiased reference measure (R). The authors created the trend by simulating greater random error in measure M, thereby violating the assumption of equal variances between M and R which is inherent to Bland-Altman plots (86). This explanation seems plausible in the context of a PAQ validation study, in which DLW is a very objective, precise method with less random error than PAQs. Judging from Bland-Altman plots alone, it is unclear how much of the positive trend was attributable to random and systematic error, respectively. Regardless, this potential weakness of the Bland-Altman method could have important implications for PAQ validation research.

Across studies we observed an overall tendency to report agreement at the group level as opposed to the individual level. Individual validity cannot be inferred from group level validity (96) and in fact serves a different purpose. In this review, the mean differences across studies indicated that neither under- nor overreporting of activity was more common at the group level. This finding differs from dietary research in which underreporting is universally more common (96), which implies that different factors may be involved in misreporting of activities and diet, respectively. Because only a few articles reported results at the individual level, we were unable to evaluate fully the degree of individual level validity for many of the PAQs. However, the 95% LOA proposed by Bland and Altman (50) allowed for speculation by examining between-subject variability. The 95% LOA for the mean difference (AEE_PAQ − AEE_DLW) from one PAQ, for example, ranged from −464 to 645 kcal/d (Yale PAQ; 53). By definition (48, 97), this result implies that there is a 95% probability that an individual from the same population would underestimate AEE by no more than 464 kcal/d and overestimate AEE by no more than 645 kcal/d. If this level of error was not acceptable for practical purposes, the PAQ could not be used as a surrogate for AEE_DLW. In fact, no 95% limit in our review was within 100 kcal/d (10% if AEE = 1000 kcal/d) of the mean difference of AEE_PAQ − AEE_DLW and or within 250 kcal/d (10% if TEE = 2500 kcal/d) of TEE_PAQ − TEE_DLW, which suggests that the PAQs in this review may be of limited use for estimating individual AEE.

A related matter for concern is the validity of DLW for individuals. At the group level, the DLW method is widely accepted as the gold standard for estimating free-living energy expenditure in adults (9, 98–100). In a review on DLW validity (101), the percentage error in TEE estimation averaged ≈2% or 8%, depending on the equation used to calculate DLW results. This level of error is acceptable for evaluating PAQ validity at the group level, as we have assumed. For individuals, however, the precision of DLW (ie, the SD of individual percentage errors) in the same review article (101) was 8–9%, which meant that some individual estimates will deviate substantially from the average. Furthermore, in comparisons based on AEE, DLW estimates must be derived by subtraction (ie, AEE = TEE − RMR − TEF), which could introduce more error if RMR is based on prediction equations. In a review of prediction equations (35), the proportion of healthy adults with valid predicted RMR (ie, within 10% RMR measured by calorimetry) ranged from 45% to 81% in nonobese individuals and from 38% to 70% in obese subjects. Under a worst-case scenario, therefore, a PAQ validation study using one of these equations to predict AEE would inevitably suggest disagreement, which may be wrongly attributed to the
PAQ. Moreover, statistical measures of validity normally deemed “moderate” may be, under this scenario, as high as could be expected. Although an important consideration, of the 20 publications we reviewed, only 6 compared PAQs with DLW-derived AEE (10 PAQ comparisons in Table 3; 51, 53, 56, 57, 67, 70), and only one of those used an RMR prediction equation to derive AEE from DLW (51). Thus, for 34 of the 36 PAQ-to-DLW comparisons in our review [2 PAQs in Adams et al (51)] RMR-related error of this magnitude was probably not an issue. Otherwise, it is possible that some individual differences reflected in the 95% LOA may have arisen in part from DLW-related error.

Concurrence between PAQ and DLW administration is another factor to consider when evaluating PAQ validity (43). Ideally, in validation studies, the criterion measure and PAQ should observe the same time frame of reference. In this review, however, the majority of method comparisons (32 of 36) did not cover exactly the same time periods. Free-living energy expenditure, when assessed using objective measures, is known to vary over days of the week (102), over weeks (103), and over seasons (104, 105) in adults. In one study of habitual activity, Levin et al (104) periodically measured physical activity patterns in 77 adults in the United States over 1 y. On the basis of intradividual variability, they determined that six 48-h accelerometer sessions were needed to achieve 80% reliability in estimating mean annual physical activity in MET (metabolic equivalents)-minutes per day (104). Although analytic error does contribute to estimates of within-subject variability (102, 106), so does actual change in activity levels over time. The latter might partly explain some of the differences we observed between PAQ and DLW estimates of energy expenditure. In our review, 10 of 36 DLW comparisons involved past year PAQs and single 1- or 2-wk DLW phases. An alternative approach would have been to repeat DLW phases to coincide better with PAQs, but none of the 10 studies reported to have done this.

A final consideration for interpreting any PAQ-to-DLW comparison is the potential for error in converting self-reported physical activity into units of energy expenditure. In large epidemiologic studies, it is usually not feasible to measure energy costs of activities for each individual. Hence, the Compendium of Physical Activities (10, 11) has become a widely accepted extrapolation tool. The Compendium provides a convenient 5-digit coding scheme that can be used to classify activities according to rate of energy expenditure or METs. By definition, 1 MET is approximately equal to 1 kcal · h⁻¹ · kg body wt⁻¹ (11). Clearly, an assumed body weight or RMR will rarely reflect that of a given individual. One important limitation of the Compendium, therefore, is the reliance on group averages that may not apply to individuals (1, 10, 11, 107, 108).

In conclusion, despite the numerous validation studies already published, the validity of PAQs for AEE estimation remains unclear. Weaknesses in the design and reporting of studies, combined with a paucity of information on the original intent of many PAQ designers, mean that it is difficult to draw any firm conclusions about the validity of existing PAQs for the assessment of usual AEE in large population-based studies. Nevertheless, our review highlights some important considerations for scrutinizing PAQ validation studies. First, there is a need to consider each PAQ’s design and its expected level of agreement with DLW, which measures all activities (freely available PAQs would facilitate this; 109). Furthermore, if a PAQ is to be used to estimate individual AEE, then its validity must be supported by the appropriate statistical analyses. Results on individual level validity were generally lacking across the articles we reviewed. We speculate that some discrepancies found previously between PAQ and DLW estimates could have resulted, in part, because PAQs did not include key activities relating to AEE or, possibly, because the PAQ period of recall and DLW phase did not coincide. Issues related to small sample size, use of correlation coefficients, and conversion of self-reported activity into energy expenditure, all continue to be problematic. Future research and development efforts should address these issues to clarify the true validity of PAQs in this context.

We thank Pietro Ferrari for his helpful comments. The authors’ responsibilities were as follows—HKN: conducted the review and wrote the first draft of the manuscript; IC and PJR: helped write subsequent drafts of the manuscript; IC, PJR, and HKN: conceptualized the review and interpreted the results; and CMF: critically appraised the manuscript. None of the authors had a conflict of interest with respect to this manuscript.

REFERENCES
13. Ainsworth BE, Jacobs DR Jr, Leon AS, Richardson MT, Montoye HJ. Assessment of the accuracy of physical activity questionnaire occupa-
...


