Construct validity of a revised Physical Activity Scale and testing by cognitive interviewing

LISE G. ANDERSEN1, MOGENS GROENVOLD2,3, TORBEN JØRGENSEN1,2 & METTE AADAHL1

1 Research Centre for Prevention and Health, Glostrup University Hospital, Glostrup, Denmark, 2 Faculty of Health Science, University of Copenhagen, Denmark, and 3 Department of Palliative Medicine, Bispebjerg Hospital, Copenhagen, Denmark

Abstract

Aim: To validate the construct validity of a new version of a Physical Activity Scale (PAS 2) for measuring average weekly physical activity of sleep, work, and leisure time to determine whether a further criterion validation is justified.

Methods: The validity of responses to the questionnaire was evaluated by cognitive interviewing in 16 Danish men and women aged 21–70 years. Construct validity was validated in 342 men and women aged 35–66 years by assessing agreement between 24-h MET-scores obtained from average weekly physical activity measured by PAS 2 and a 24-h Physical Activity Scale (PAS 1), previously found to overestimate physical activity.

Results: Cognitive interviewing revealed few problems in the questions on physical activity in different domains. No problems regarding the structure of the questionnaire were identified. The agreement between PAS 1 and PAS 2 MET-scores was high among participants with a PAS 1 24-h MET-score < 45 MET-hours. Among participants with a PAS 1 24-h MET-score ≥ 45 MET-hours, the weekly-based scale, PAS 2, systematically estimated fewer MET-hours compared to the 24-h based scale, PAS 1. The difference increased proportionally with the average of the two MET-scores.

Conclusions: Few, small, lexical revisions were implemented into the new scale to improve the validity. As hypothesized, PAS 2 produced lower estimates of energy expenditure compared to PAS 1, indicating that the new scale may provide more valid measurements and that further validation against an objective criterion is justified.

Key Words: Cognitive interviewing, energy expenditure, MET-hours, physical activity, validity

Background

During the last decades strong evidence has been established that physical inactivity is a major risk factor for specific chronic diseases [1] and premature death [2]. Physical activity is a complex behaviour and consequently difficult to measure. The questionnaire is the most frequently used method in epidemiological surveys [3]. Although simple questionnaires give a relatively crude measure of overall physical activity, they are useful in larger cohort studies [4]. The validity of self-report measurements must, however, be assessed. Validation of an instrument refers to the process by which it is determined if an instrument measures what it is intended to measure. Construct validity covers the degree to which the questionnaire measures the unobservable construct it was designed to measure [5].

We previously developed a simple self-administered nine-step scale, Physical Activity Scale 1 (PAS 1), based on an original Swedish questionnaire developed by Lagerros et al. [6]. PAS 1 measures total physical activity during an average 24-h weekday [7] and the daily energy expenditure expressed in hours of metabolic equivalents (MET-hours [8]) can be estimated as a 24-h MET-score. PAS 1 was found to have acceptable validity compared to maximal oxygen uptake, but also a tendency to overestimate systematically, particularly among very active individuals, when compared with an activity diary and
accelerometry [7,9]. Significant associations between physical activity and various biological outcomes, e.g. body mass index (BMI) and high density lipoprotein (HDL) have been observed among individuals active <45 MET-hours/day, but not among individuals active at least 45 MET-hours/day when measured by PAS 1 [10]. These findings have not been replicated by others, suggesting that the instrument may be the underlying reason. Furthermore, reporting weekly sports participation on a one-day basis, and describing precisely 24 hours of activity seems to be difficult for respondents. Finally, weekends are not included in PAS 1. Therefore, we revised PAS 1 and developed a new physical activity scale, PAS 2 (see version 2.1 in Figure 1), which in contrast to PAS 1 measures physical activity during leisure time during an average week. PAS 2 comprises seven items corresponding to categories of activity such as sleep, work, and leisure time. We expected PAS 2 to estimate lower 24-h MET-scores compared to PAS 1, especially among highly active individuals. Clearly, it is important to investigate whether these revisions lead to the expected improvements in validity. The aim of this study was to validate the Physical Activity Scale, PAS 2, in order to determine whether a further validation against an objective criterion is justified.

Methods

Development of the physical activity scale

PAS 2 measures physical activity as daily hours and minutes of sleep, sitting, standing or walking, and heavy physical work, transportation to and from work, and TV-viewing/reading (see version 2.1 in Figure 1). In addition, it measures weekly hours and minutes of light physical activity, moderately strenuous activity, and strenuous activity. Each of these domains corresponds to a specific level of MET-intensity according to The Compendium of Physical Activity [8,11] (sleep = 0.9 MET, TV-viewing/reading = 1.0 MET, sitting work = 1.5 METs, standing/walking work = 2.0 METs, light activities = 3.0 METs, transportation = 4.0 METs, heavy work = 5.0 METs, moderately strenuous activity = 5.0 METs, strenuous activity = 6.0 METs). Daily MET-time was multiplied by 5 (work and transportation) or 7 (sleep and TV), and weekly hours not accounted for were weighted with a MET-value of 2 METs and added to derive the weekly MET-time. This adjustment allowed for indirect estimation of total physical activity during 24-h, and thus a 24-h MET-score for PAS 2, based on total weekly physical activity, which could be compared to the 24-h MET-score derived from PAS 1.

Cognitive interviewing

Validity in terms of meaningfulness of the questions to the respondents and the extent to which they answered in the intended way was evaluated by cognitive interviewing. Cognitive interviewing has been defined as “the administration of draft survey questions while collecting additional verbal information about the survey responses, which is used to evaluate the quality of the response or to help determine whether the question is generating the information that its author intends.” [12]. The purpose of cognitive interviewing is to generate verbal information to detect and repair problems with questions [13]. Cognitive interviewing pays explicit attention to the mental, cognitive processes the respondents use to answer the question [14] and divides the process of answering into four mental processes: 1) comprehension; 2) retrieval of relevant information from memory; 3) decision process; and 4) the response process [14,15]. Biases in the response can arise in each of these cognitive processes. Furthermore, structural defects, or logical problems, can be detected by this technique [14].

Population

Six men and 10 women aged 21–70 years were interviewed. The respondents were recruited among participants in a population-based, randomly sampled survey, Health2006 [16]. Health2006 was approved by the local ethics committee (KA-2000600011) and all participants gave informed written consent. The respondents were invited to participate when visiting the Research Centre for Prevention and Health for examination and were sampled strategically to get men and women of a wide range of age and occupational groups. None of the invited respondents declined the invitation to participate. Four respondents were sedentary, eight were lightly active, three were moderately active and one was overall highly active.

Data collection

Techniques used were a concurrent think-aloud interview during which the respondent was thinking aloud while completing PAS 2 [14], and probing of the respondent during which the interviewer took on a more proactive role to derive more information on the strategies behind the answers [14]. The interviews were audio-recorded.

Analysis

For each item we analyzed whether the response given by each respondent was in accordance with the intentions. We focused on problems encountered repeatedly across interviews, problems that could seriously threaten data quality [14], and whether apparent problems could be
### 9. Physical activity in everyday life

The following questions deal with your daily physical activity at work, at home and during your leisure time. Therefore, this includes more than sports and exercise. It includes all types of physical activity and inactivity that you perform in everyday life.

<table>
<thead>
<tr>
<th>Daily</th>
<th>How many hours and minutes do you sleep on an average weekday (include rest or naps during the day)</th>
<th>Hours</th>
<th>Minutes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Daily</td>
<td>In your work/studies*, how many hours and minutes per day do you engage in:</td>
<td>Hours</td>
<td>Minutes</td>
</tr>
<tr>
<td></td>
<td>Sedentary work?</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Standing or walking work?</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Heavy physical work? (for instance heavy lifting or climbing stairs)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Not working □</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Daily</td>
<td>How many hours and minutes per day do you ride a bicycle or walk for transportation to and from work?</td>
<td>Hours</td>
<td>Minutes</td>
</tr>
<tr>
<td></td>
<td>Not working □</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Daily</td>
<td>In your leisure time, how many hours and minutes per day do you spend with watching TV, sitting quietly, reading, and listening to music or the like?</td>
<td>Hours</td>
<td>Minutes</td>
</tr>
<tr>
<td>Weekly</td>
<td>In your leisure time, how many hours and minutes per week do you engage in light physical activity such as walking, light cleaning, raking lawn, or lightly strenuous exercise such as yoga, bowling or similar activities? (do not include transportation to and from work)</td>
<td>Hours</td>
<td>Minutes</td>
</tr>
<tr>
<td>Weekly</td>
<td>In your leisure time, how many hours and minutes per week do you engage in gardening, carrying loads upstairs or moderately strenuous sport such as gymnastics, swimming, bicycling, strength conditioning or similar activities? (do not include transportation to and from work)</td>
<td>Hours</td>
<td>Minutes</td>
</tr>
<tr>
<td>Weekly</td>
<td>In your leisure time, how many hours and minutes per week do you engage in strenuous sport and conditioning exercise such as running, jogging, soccer, tennis, aerobics or similar activities? (do not include transportation to and from work)</td>
<td>Hours</td>
<td>Minutes</td>
</tr>
</tbody>
</table>

*Underlining mark revision from PAS 2 to PAS 2.1.

Figure 1. Physical activity scale version 2.1. The revised vision of PAS 2 describes total amount of physical activity in an average week. Translated into English from the Danish version of the physical activity scale. Revisions from PAS 2 to PAS 2.1 are marked by underline.
logically ascribed to characteristics of the question [12]. As suggested by Conrad and Blair, problems were, in a respondent problem matrix, coded into types of lexical, temporal, inclusion/omission, logical and computational character occurring during the comprehension process, the retrieval and decision process, or the response process, respectively [17].

**Construct validity**

**Population** The study population was recruited from the five-year follow-up of a population-based, randomized, intervention survey, Inter99, previously described in detail [18]. The Inter99 study was approved by the local ethics committee (KA 98155) and all participants gave informed written consent. The population comprised 342 consecutively sampled participants aged 35–66 years among whom 46% were males. In order to get a study sample who filled in both PAS 1 and PAS 2 the last 342 participants from the Inter99 five-year follow-up were asked to fill out the PAS 2 questionnaire when meeting at the research centre for their health examination, in addition to filling in PAS 1 that had already been mailed to them and completed at their home before meeting at the centre.

**Analysis** The agreement between 24-h MET-scores obtained from PAS 1 and PAS 2 was analyzed by the limits of agreement (LoA) method as suggested by Bland and Altman [19]. The bias and variability between the two measurements estimated by the mean difference and the standard deviation (SD) of the differences were used to assess agreement by 95% LoA, between which 95% of the differences are expected to lie, provided that the differences follow a normal distribution. This was tested in a histogram. To test the applicability of constant, parallel LoA, the difference was modeled by means of linear regression as a function of the average of the two measurements. Since the residuals of this fitted line were independent of the range of measurement, the SD of the adjusted differences, the residual SD (SDres), was estimated as the SD of the difference. Regression-based 95% LoA were obtained as 

\[ b_0 + b_1A \pm 1.96SD_{\text{res}}, \]

where \( b_0 \) is the systematic constant difference, \( b_1 \) is the systematic proportional difference, and \( A \) is the average of the two measurements [19].

The population was sub-divided according to PAS 1 24-h MET-score (< vs. \( \geq 45 \) MET-hours/day) to investigate the nature of agreement in these subgroups. The average daily hours of activity and MET-time were calculated on a question-by-question basis and compared between PAS 1 and PAS 2, to investigate causes of systematic disagreement. All analyses were performed in SAS Enterprise Guide version 4.1.

**Results**

**Cognitive interviews**

Cognitive interviewing did not detect any problems in the response of the question on sleep. Although the questionnaire presupposed that the respondents had a five-day working week, most respondents reported occupational and commuting physical activity performed during a single day and not as an average. The reason for this appeared to be that the respondents misinterpreted the reference period. Problems encountered specifically by students were identified. They did not know how to complete the questions on occupational physical activity and TV-viewing/reading, had difficulties estimating the hours spent on their studies, and therefore proposed specific instructions for students.

The questions on light, moderately strenuous, and strenuous physical activity caused several problems (Table I). The meaning of the term “heavy housework” was not clear to some respondents. Gardening in terms of light activities such as watering and raking the lawn was mistakenly filled in by some respondents as moderately strenuous activity or not filled in at all. Both young and some middle-aged respondents had difficulties answering the question on duration of moderately strenuous activity, but for different reasons. The middle-aged had an estimation problem concerning gardening because the duration of the activity varies by season. The young respondents could not relate to the question because the examples were not part of their everyday life. Additionally, some respondents misplaced or forgot cycling during leisure time and swimming, perhaps because they did not know at which level to include them.

No significant problems relating to the form of the questionnaire, the distinction between the three categories of leisure time physical activity, or the interpretation of the term strenuous were detected. However, one elderly respondent misinterpreted the term strenuous as “something you dislike doing”. The respondents mainly based their answers on the examples of activities.

**Construct validity**

The mean total 24-h MET-score (SD in brackets) estimated from PAS 2 and PAS 1 were 40.86 (4.81) and 45.66 (10.03) MET-hours/day, respectively.
Among men, the respective values were 41.45 (5.37) and 47.11 (11.21) MET-hours/day, and among women 40.32 (4.22) and 44.41 (8.37). Thus, the level of physical activity estimated from PAS 2 was about 10% lower than from PAS 1 with the largest difference seen in men, and the SD was about twice as large for PAS 1.

Figure 2 shows the disagreement between measurements from PAS 2 and PAS 1. Among lightly active participants the differences between PAS 1 and PAS 2 deviated in both directions, whereas PAS 2 clearly estimated lower 24-h MET-scores compared to PAS 1 among highly active participants. The mean difference increased with increasing range of measurement as illustrated in Figure 3, whereas the variability of the differences were constant across the range of measurement as indicated by independence between the residuals and the range of measurement ($p = 1.0$ for all parameters). The regression-based difference was significantly and positively associated with the range of average 24-h MET-score ($p < 0.0001$ for $b_0$ and $b_1$). The regression-based 95% LoA were given by: 

$$-30.95 + 0.83 \times \text{PAS 1 - PAS 2}/2 \pm 1.96 \times 5.30 \text{MET-hours/day}.$$ 

Even though the assumption of normally distributed differences was not fulfilled, 93% of the differences lay within the 95% LoA, i.e. ±10.39 MET-hours from a specific mean difference.

The associations between the differences and the range of measurement in the subgroups with a PAS 1 24-h MET-score < or ≥45 MET-hours/day were significant and similar to the relationship seen in the total population, although the relationship both in terms of the mean difference and the variation around the mean difference was much weaker among participants with a PAS 1 24-h MET-score <45 MET-hours/day (data not shown). In this subgroup, the differences were approximately equally distributed around a difference of 0 MET-hours/day with a mean difference and SD of 0.37 (3.53) MET-hours/day, while the differences among participants

### Table I. Examples of coding of problems detected by cognitive interviewing.

<table>
<thead>
<tr>
<th>Problem type</th>
<th>Understanding</th>
<th>Retrieval/decision</th>
<th>Response formatting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexical</td>
<td>The meaning of the term “heavy housework” was not clear.</td>
<td>Gardening is understood on a continuum from very light to heavy activity and could explain why watering and raking the lawn was misplaced or even omitted.</td>
<td></td>
</tr>
<tr>
<td>Inclusion/omission</td>
<td>Gardening does not fit in the category moderately strenuous activity since gardening covers all levels of effort.</td>
<td>Bicycling and walking transportation during leisure time was forgotten.</td>
<td></td>
</tr>
<tr>
<td>Temporal</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logical</td>
<td></td>
<td>Young respondents have difficulties identifying themselves in the category of moderately strenuous activity because the examples do not apply to them.</td>
<td>Seasonal variation cannot be reported.</td>
</tr>
<tr>
<td>Computational</td>
<td></td>
<td>It is difficult for middle aged respondents to estimate moderately strenuous activity because of seasonal variation in gardening.</td>
<td></td>
</tr>
</tbody>
</table>
with a PAS 1 24-h MET-score ≥45 MET-hours/day were mainly positive with a mean difference and SD of 10.86 (7.85) MET-hours per day meaning lower PAS 2 MET-scores. This relationship was more pronounced with increasing activity level.

Inspection of the daily hours and MET-hours performed at each activity level revealed little difference between PAS 1 and PAS 2. However, for sitting work and the activity levels of ≥3.0 METs fewer hours were reported in PAS 2 (data not shown). A similar pattern was seen among men and women.

**Discussion**

**Cognitive interviews**

PAS 2 was constructed in order to solve the problems encountered in PAS 1. By means of cognitive interviewing, we found that the respondents did not experience the problems with the questionnaire structure that they encountered with PAS 1. According to the respondents, the shift between daily and weekly performed activities worked well, and the reference period seems to give a fairly reliable estimate of the total weekly activity. Therefore, the structure appears superior to the structure of PAS 1.

All identified problems could be classified according to the response stages and five types of problems. Hence, the respondent problem matrix provided a useful basis for analyzing the identified problems. To address these issues, PAS 2 was revised after this testing, leading to version PAS 2.1 (Figure 1). Any revision of a questionnaire may result in benefits and drawbacks. Therefore, not all identified problems were judged to qualify for a revision. The problems encountered by students have been handled by specifying that work includes studies. Raking the lawn has been added as an example of light physical activity; the term heavy housework has been replaced with “carrying loads upstairs”; and dancing has been replaced with “swimming” and “bicycling”. These examples now introduced in PAS 2.1 are all frequent and well-known in the Danish population and 19 to 60-year-old Danes can assess the intensity of these activities fairly accurately [7]. The seasonal variation in gardening has not led to any revision because a distinction between activity level during summer and winter would significantly complicate the completion of the questionnaire and would pose a great risk for other problems. The misinterpretation of the reference period in the questions on work and transportation could have been handled by adding a sentence saying, “Please report the average per day for a five-day working week”. This was abandoned, because it would introduce a lot of new text with potential for other problems. Additionally, the bias of the current design is likely to be small, since a five-day working week is the most common.

PAS 2 differentiates between the three levels of intensity of leisure time physical activity by using the term “strenuous”, which could result in misclassification if misinterpreted. However, the respondents based their categorization on the examples of activities given rather than on the term strenuous. This study suggested that PAS 2 might not work among very old individuals with possible functional limitations, because of misinterpretation of the term strenuous. Gender and occupational status did not seem to affect the applicability of PAS 2. Therefore, it seems suitable for the target group that comprises 18 to 69-year-old Danes.

Durante and Ainsworth suggested use of cognitive interviewing for development, improvement, and administration of instruments to obtain self-reports on physical activity [20]. However, we have not identified any studies that have used cognitive interviewing in the validation of a self-administered questionnaire for measuring the overall activity level. Cognitive interviewing was a useful method for validation, as the verbal reports gave a rich account of the underlying thought processes [21] that provided insight into the consequences of the design of the questionnaire and guidance to possible revisions [12]. The implemented revisions are rather small and all of lexical character.

Cognitive interviewing has been criticized for its subjectivity in data collection and analysis [14,17]. We used Conrad and Blair’s respondent problem
matrix as a basis for systematic quantification of respondents’ problems by means of objective criteria of problems. Hereby we tried to increase the objectivity and reliability of the analyses [17,21]. Clearly, the validity of the verbal reports is compromised if they indicate problems that actually do not exist. The analysis was therefore focused on problems that could logically be ascribed to the design of the questions. The validity of cognitive interviewing relies on the assumption that the thought processes are available to the respondents and that they can be accurately reported [12]. Physical activity is a complex behavior and it can be difficult to describe how one went about answering the question and estimated a specific number of hours. We supplemented the think-aloud interviews by probing to gain information that the respondents were conscious of, but for some reason did not report. According to Willis et al., 12–15 interviews are sufficient in a round of interviews [14]. The strategic sampling makes it probable that the respondents represent a wide range of the subjects of the population in which the questionnaire will be used as a survey instrument. Therefore, qualitative representativity is presumably fulfilled, and we have no reason to believe that important problems that could have been detected with this technique have been overlooked.

Construct validity
As expected, we found systematic disagreement between PAS 1 and PAS 2, which was mainly caused by the disagreement among participants with a PAS 1 24-h MET-score \( \geq 45 \) MET-hours/day. The difference increased by increasing the range of measurement, while the random variability was independent of the range of measurement. The lower SD of PAS 2 was expected since PAS 2 is designed differently than PAS 1 and reflects the fact that PAS 2 is less skewed, which was due to lower MET-scores among the highly active individuals. The underlying reason for the systematic difference was the combination of fewer hours of activities at intensities of \( \geq 3 \) METs reported in PAS 2 and lower MET-values assigned to PAS 2. PAS 2 measures leisure time physical activity on a weekly basis while PAS 1 measures daily activity. Thereby PAS 2 probably more precisely estimates moderately strenuous and strenuous activity performed on a weekly basis.

Schmidt et al. validated an interviewer-administered 24-h physical activity questionnaire by assessing agreement with a more comprehensive questionnaire [22]. They found a constant difference and therefore constant and parallel 95% LoA. Schmidt et al. based the assessment of acceptable agreement for replacement of methods on the interquartile range in MET-hours [22]. As we have designed PAS 2 to measure physical activity differently than PAS 1, by which the differences depend on the range of measurement, it is not meaningful to use constant levels, nor the interquartile range of PAS 1 as a criterion for acceptable 95% LoA. Rather, our hypotheses and the results from the validation studies of PAS 1 [7,9] suggested that the systematic bias would be positively associated with the range of measurement while the random variability should be the same for all participants given the average of the two MET-scores. The confirmation of these hypotheses indicates that PAS 2 may give relatively more valid measurements than PAS 1.

From a validation perspective, a limitation of this study is that one subjective instrument has been validated against another subjective instrument. It is likely that the errors of the two self-report methods are correlated and that agreement is therefore overestimated. Obviously, additional validation of PAS 2.1 against objective criteria such as movement sensors and biological measurements is warranted. However, the confirmation of the hypothesis that PAS 2 estimates lower 24-h MET-scores compared to PAS 1, and especially among highly active individuals is unlikely to be due to a specific bias. Therefore, the comparison provided useful information. Strengths of this study are the consecutive sampling of the study population and the participation of all invited.

Conclusion
PAS 2 is a simple instrument for collecting information about physical activity. It is not time consuming, requires few instructions, and describes the questions in both pictures and text. As PAS 2 measures physical activity on a continuous scale, it provides detailed information about the duration of physical activity at various intensities in various domains. At the same time, it provides a measure of total physical activity.

The cognitive validation of PAS 2 and validation against PAS 1 were necessary as to show whether PAS 2 leads to improvements in validity. The study showed that a few, small, lexical revisions should be made to PAS 2 (leading to the version PAS 2.1) to improve the validity in terms of the extent to which the respondents answer the questions in the intended way. The agreement between PAS 1 and PAS 2 confirmed our hypotheses. Thus, the gratifying results of this study to a great extent indicate that PAS 2 provides relatively more valid measurements than PAS 1. Obviously, the next step will be to
validate PAS 2.1 against objective criteria such as movement sensors and biological markers.

Acknowledgments

The Inter99 was initiated by Torben Jørgensen (PI), Knut Borch-Johnsen (co-PI), Hans Ibsen and Troels F. Thomsen. The steering committee comprises the former two and Charlotta Pisinger. The Health2006 steering committee comprised Allan Linneberg (PI) and Torben Jørgensen (co-PI).

Conflicts of interest

None.

Funding

None.

References