Physical activity questionnaires
- A critical review of methods used in validity and reproducibility studies

Elin Bandmann

GIH - THE SWEDISH SCHOOL OF SPORT AND HEALTH SCIENCES
Graduate Essay 17:2008
Sport Science and Health Science: 2005-2008
Supervisor: Peter Schantz
Abstract

Aim. The purpose of this paper was to investigate physical activity questionnaires, and to examine studies of their reliability and validity, focusing on the variable frequency. The research questions were following:

1. What are the methods and the results of the investigated questionnaires’ validity and reproducibility studies?
2. How is the frequency of physical activity assessed and validated in different PA questionnaires?

Method. The 16 (out of 24) first questionnaires concerning individuals from 18-69 years were selected from a collection of physical activity questionnaires. The topical International Physical Activity Questionnaire (IPAQ), both the long and the short version were also included in the investigation. Additional searches for articles were done in PubMed. Information about the design of the in total 18 questionnaires were structured in a scheme to clarify similarities and differences. How the frequency of physical activity (PA) was assessed was of particular interest. To evaluate methods and statistics the first validity and reliability study from each questionnaire’s reference list was selected and reviewed. This based on certain statistical criteria, regarding within-subject variation, test-retest, changes of mean, linear regression analysis, and criterion/construct methods for the validation.

Results. The common approach is to compute correlation coefficients (n=18). One validity study out of 18 investigates systematic changes. No study presents results of changes of mean. In many studies, within-subject variation is calculated (n=14). Common validity measures are VO2 max (n=8), PA record (n=7), accelerometer (n=9) and to validate the questionnaire against other questionnaires (n=7). The correlation coefficient is at best r =0.30. Direct assessment of accuracy of reported frequency of PA was not measured in any study. One questionnaire assesses seasonal variations of physical activity.

Conclusions. The accuracy of the investigated PA questionnaires is low. There are areas for improvements in the construct and administration of PA questionnaires, and the assessment of PA will need a combination of measurement instruments to cover all aspects of PA. Reliability and validity studies must focus on detecting the order effect and the systematic error of the questionnaire. Until then interpretations of reported physical activity data will be misleading. Current physical activity recommendations may have to be redesigned.
TABLES AND FIGURES

Table 1: Characteristics of the physical activity questionnaires................................. 16
Table 2: Characteristics of physical activity questionnaires........................................ 17
Table 3: Validity and reliability scheme........................................................................ 18
Table 4: Validity and reliability scheme........................................................................ 19

Figure 1: Kriska and Caspersen’s illustration of the computation of summary estimates of physical activity .................................................................................. 5
Figure 2: Scattergram describing the linear regression of the interview-based 7-d recall with a concurrent 7-d diary.......................................................... 47
Figure 3: Scattergram describing the linear regression of the self-administered 7-d recall with a concurrent 7-d diary.......................................................... 47
Figure 4: Scatterdiagram describing the linear regression of the 7-d recall with VO2max. ........................................................................................................... 47
1 Introduction

Assessing physical activity is of importance in epidemiological studies to examine the relationships between inactivity and development of diseases. The results are used in health prevention and the forming of physical activity recommendations. Since the 50’s researchers have tried to measure physical activity for these matters. Initially, focus was held on assessing occupational activity. As jobs became less physically demanding, an interest for leisure time physical activity, mainly sport and recreation, increased.¹

Today, physical inactivity is responsible for a large worldwide burden of the disease and health care costs. This was concluded in the 1996 Surgeon General’s Report and highlighted at the International Congress on Physical Activity and Public Health 2006.² For this reason, the World Health Organisation (WHO) has raised focus on national monitoring and surveillance of physical activity. They see an urgent need for accurate methods for large-scale surveillances.³ The most common and easiest method to assess physical activity on large populations has so far been a subjective measure; self-administered or interview based questionnaires. Questionnaires are both economical and sufficient for large groups. Physical activity is however complex to assess. It consists of several components and dimensions, and occurs in many different arenas. How are physical activity questionnaires formed, and how reproducible and valid are they? They contribute to the forming of physical activity recommendations, but by which methods are their validity and reliability evaluated? The aim of this review is to investigate this subject.

1.1 Physical activity

In 1985 the epidemiologist Carl J Caspersen defined physical activity as “any bodily movement produced by skeletal muscles that results in energy expenditure”\(^4\). It is a broad definition, as it theoretically includes every move from scratching your toes, to running a Marathon. The research field of assessment of physical activity is broadening; new aspects, dimensions and areas have been surveyed. Neville Owen (et. al), head of the School of Human Movement at Deakin University, Australia, has recognized five arenas in which physical activity can be categorized into. They are:

- Leisure-time physical activity (sports) for recreational purposes, condition and competition.
- Leisure-time physical activity such as gardening and household.
- Physical active transportation and commuting
- Physical activity as a scheduled school subject or as a paid exercise during work-time.
- Non-scheduled physical activity as seen in kinder garden or school, or in a physical active job.\(^5\)

The report from the 2006’ International Congress on Physical Activity and Health, confirms a change from an interest in measuring leisure-time and sports-related physical activity, to an interest in assessing moderate-intensity physical activity in other domains in which physical activity occur. Household activity, occupational activity and active transportation are exemplified. One recently published instrument that has been developed for international surveillance of several physical activity domains in adult population is the International Physical Activity Questionnaire (IPAQ).

The report also discusses new emerging techniques to assess domain-specific physical activity. Measurement methods from other disciplines can possibly be used as assessment


tools. For instance, instruments from transport research could be used for the assessment of physically active commuting.6

1.2 Assessing physical activity levels

Figure 1: Kriska and Caspersen’s illustration of the computation of summary estimates of physical activity7

M. J. Lamonte and B. E. Ainsworth, researchers in movement and health, explain physical activity as a construct of body movement that can be assessed either as the behaviour, or as the energy cost of the movement.8 Andrea M. Kriska and Carl J. Caspersen at the Department of epidemiology, University of Pittsburgh, say that the two most common ways to estimate questionnaire data in physical activity questionnaires or forms are to summarize frequency with duration and intensity; or to multiply total time with intensity. Energy expenditure of physical activity is achieved by expressing the intensity variable as a MET-value9 (see figure 1). The estimates are valuable in relative terms and can be used to rank individuals or groups of subjects within a population from the least to the most active. The ranking is compared with physiologic parameters and disease outcomes.10 In a recently published review of methods for physical activity assessment in epidemiological

7 Kriska M.A, Caspersen, CJ.p. 7
9 1 MET represents the metabolic rate of an individual at rest, and equals approx. 1 kcal/kg/h. An activity with an intensity of 5 MET would require 5 times the resting metabolic rate.
10 Kriska M.A, Caspersen, CJ,. pp. 6-7
research, Ylva Trolle Lagerros and Pagona Lagiou from the Unit of Clinical Epidemiology at the Karolinska Institute conclude that physical activity questionnaires should include questions about all three components (frequency, duration, and intensity). Questionnaires inquiring solely one component cannot be generalized and are not easily be converted to public health recommendations.\textsuperscript{11} Furthermore, Kriska and Caspersen suggest that an importance when examining correlations between physical activity and disease is to focus on the \textit{health-related dimension} (or dimensions). That is, to focus on the physical activity most likely to be associated with the specific disease or condition. Health-related dimensions are defined as caloric expenditure, aerobic intensity, weight bearing, flexibility and strength.\textsuperscript{12}

As physical activity questionnaires are subjective methods; the results are highly dependant on the respondent’s cognition. It is also influenced by interviewer bias, the day of the week, the sequence of administration, and if the questionnaire is combined with other physical activity measures.\textsuperscript{13} Peter T. Katzmarzyk and Mark S. Tremblay, researchers in epidemiology and health, has published a recent discussion about the subject’s perception. They say as benefits of physical activity are becoming more published, social desirability of reporting “healthy” behaviours may have increased over the last two decades. This may result in over-reporting of physical activity. In addition, the promotion of physical activity has emphasized to count activities such as walking, gardening and yard work (which previously were not viewed as “exercise”) as physical activity. It is possible that people did a fair amount of walking 1980, but that they did not consider it as “exercise”. In more recent surveys respondents may be more aware of reporting walking, resulting in an imaginary increase of physical activity.\textsuperscript{14}

\textbf{1.2.2 Frequency}

In this paper, extra focus will be put on frequency as an aspect of physical activity. Little is known about the validity of reported frequency of physical activity.

\textsuperscript{11} Lagerros TY, Lagiou P. \textit{Assessment of physical activity and energy expenditure in epidemiological research of chronic diseases}. Eur J Epidemiol 2007 22:353-362
\textsuperscript{12} Kriska M.A, Caspersen, CJ. p. 5
\textsuperscript{13} Ibid. p. 7
By frequency means, with what regularity a certain activity is performed. This can be expressed in several ways, for example as how many times a day, a week, a month, or a year an activity is performed.\textsuperscript{15} The physiological effects of physical activity in adult population are not stable and need to be maintained with a certain frequency. Jan Henriksson at the department of physiology and pharmacology, Karolinska Institute, wrote in 1998 a review of physical activity’s positive effects on blood pressure, cholesterol, and insulin sensitivity. However, the physiological effects were not durable. In one study where eight athletes had a break from exercise for 10 days, the break resulted in a 70% decrease in insulin sensitivity.\textsuperscript{16} In another study by Peter Schantz et.al., healthy subjects exercised their arms for 8 weeks. The training resulted in a 40-100% increase in skeletal muscle fibres, capillaries and enzyme activity in m. triceps brachii, but after 6 weeks of detraining, the levels had decreased to the pre-training point.\textsuperscript{17}

As with assessing total physical activity, accurately measuring frequency is multifaceted. P. Tucker and J. Gilliand’s literature review of 37 studies (published 1980-2006) illustrate how the activity pattern over the year differs between nations, due to climate and altitude.\textsuperscript{18} A Swedish study by Peter Schantz and Erik Stigell examined the frequency of physically active commuting in 1120 women and 573 men. They found the frequency to vary substantially over the year, and therefore using spot data for this variable is misleading.\textsuperscript{19} For example, assessing frequency of physical activity by a 7-day recall questionnaire does not take weekly- and seasonal variations in concern. Also, the level of physical activity can easily be misinterpreted if the subject is ill during the measured week (or if the week in other ways differ from the subject’s normal living). Due to these aspects, long time surveys would be preferable. However, the issue with longer time frame is the increase of recall bias.\textsuperscript{20}

\textsuperscript{15} Lagerros TY, Lågiou P. p. 355
\textsuperscript{16} Henriksson J. “Forskning om sambandet kroppsförskåder och hälsa”. Svensk Idrottsforskning 1998 (3) 42-45
\textsuperscript{19} Schantz P, Stigell, E. Frequency of bicycle trips per week and bicycling days per year as input data in cost-benefit analyses.
\textsuperscript{20} Kriska M.A, Caspersen, CJ. p. 7
These findings complicate the assessment of physical activity further. Therefore, how continuity and frequency of physical activity is assessed and evaluated will be investigated in this paper.

1.4 Reliability and Validity

What methods are there to test reliability and validity of a physical activity questionnaire? Will G. Hopkins at the department of physiology, University of Otago, New Zealand, has examined measures of reliability in sports medicine and science. He explains reliability as the reproducibility of a measurement in repeated trials on the same individuals.²¹ The procedure, the instrument itself and the subject all affect reliability. According to Hopkins, there are three important types of measures to study reliability: within-subject variation, systematic changes in the mean, and test-retest correlation.

Within-subject variation is known as the standard error of measurement, which is the standard deviation of an individuals repeated values. It is defined as a coefficient of variation, and expressed as percentage of the mean. For many physiological measurements, this error becomes bigger as the value of the measure gets bigger. Another form of within-subject variation is the limits of agreement, which is expressed as the 95% likely range of change between test 1 and 2. Instead of using the standard deviation, the range in which a subject’s different scores would fall most (95%) of the time is calculated.

Changes of mean, also referred to as order effect, refers to a change in results between consecutively repeated trials. The change is due to either a random change or a systematic change. A random change in the mean is due to random error of the measurement. The random change is therefore smaller with larger study populations. An example of systematic change is learning effect; the test is more familiar the second time it is performed. The subject may also recall answers from the first test, have a change in motivation, or an intention to improve the result from the first test.

The changes of mean are easiest studied by *Paired T*-tests, a method described elsewhere by Hopkins.\(^{22}\) To study the order effect, he supports that at least three trials should be performed with a minimum of 50 individuals.\(^{23}\)

A *Test-retest* correlation is computed by calculating the *correlation coefficient* between a first and a second test. If the data is parametric, a Pearson correlation is computed. Spearman correlation is used for non-parametric data and ordinal data. The closer the correlation is to 1.0 the higher correlation between the two measures.\(^{24}\) Whereas the typical error is a measure of within-subject variation, the correlation coefficient illustrates the reproducibility of the rank order of all the subjects on a retest. Thus, the individual’s (absolute) values can change from one trial to another without a change in correlation, as correlation coefficients only measures rank order. In a high correlation the subjects will mostly remain on the same place in the rank between tests, whereas a low correlation means they will have changed the rank. Another issue with test-retest correlation is that the correlation value is sensitive to the spread of values between participants. If the spread is illustrated in a scatter plot one may detect how plots that differs from the mean dramatically affects the correlation coefficient. Hopkins therefore points out the importance of studying within-subject variation.\(^{25}\) However, a third type of correlation, the intra class correlation (ICC) can indicate systematic changes, which the Pearson and Spearman methods do not detect.\(^{26}\)

Timo A Lakka and Jukka T Salonen have also discussed the performing of test-retest reliability of physical activity questionnaires. They found that several studies reported stronger short-term than long-term test-retest correlations for physical activity measurements.\(^{27}\)


\(^{23}\) Hopkins, W.G. 2000, p.11


\(^{25}\) Hopkins, W.G. 2000. pp. 2-6

\(^{26}\) Discussion with Peter Schantz, peter.schantz@gih.se June 23, 2008

Validity can be defined as the accuracy or precision of an instrument: Does it measure what it intends to measure? There are many different ways to assess validity. In the field of physical activity, criterion validity (concurrent validity) and construct validity are the most investigated aspects. In this context, criterion validity is when a method is contrasted to another method, which is set as criterion (golden standard) for the variable that both the instruments intend to measure. Construct validity indicates the consistency between the activity instrument and a physiological variable related to physical activity, for example maximum aerobic capacity or resting heart rate.

A method to study systematic changes, both in tests of reliability and validity, is to fit a *linear regression* to the scatter plot of the assumed correlation between two measures. In the linear regression, two variables (in this context *methods* or trials) are compared. One method is considered to be a valid method (criterion) and the other is the method of investigation. The slope of the line can reveal major *systematic changes* between the measures, although the *correlation coefficient* between two measures is high. Thus, solely calculating correlation coefficients when studying validity is not enough. The most common method for a linear regression line is the method of *least squares*. Another method to discover systematic changes is to do a Bland-Altman test, but due to time restraints this test will not be described further.

Measurements can be reliable without being valid, but a valid measurement is always reliable. As illustrated, there are several important statistical methods that should be used to accurately assess reliability and validity in studies of physical activity questionnaires.

---

2 Aim

The purpose of this paper was to investigate physical activity questionnaires, and to examine studies of their reliability and validity, focusing on the variable frequency.

2.1 Research questions

1. What are the methods and the results of the investigated questionnaires’ validity and reproducibility studies?
2. How is the frequency of physical activity assessed and validated in different PA questionnaires?

3 Methods

To gather information about the background of the topic and its field today searches for articles concerning validity and reliability of physical activity questionnaires, the specific frequency variable, and statistical methods used in sports medicine and science was done in the web-based database PubMed. For details about key words used in searches, see appendix 1. Additional articles regarding the frequency variable as well as physical activity and health was provided from Peter Schantz, dr med sc, associate professor at The Research Unit for Movement, Health and Environment at The Åstrand Laboratory in Stockholm, Sweden.

The primary base for this study was a collection of physical activity questionnaires published in 1997 by Kriska and Caspersen\textsuperscript{31}. The 16 (out of 24) first Questionnaires concerning individuals from 18-69 years were selected from the compilation. The chosen questionnaires are diverse in design, and have been used in studies which a lot of today’s knowledge about preventive health and physical activity recommendations rely on. Due to its topicality, The International Physical Activity Questionnaire (IPAQ), both the long and the short version were also included in the investigation. Information about the design of the in total 18 questionnaires were structured in a scheme to clarify similarities and differences. How the frequency of physical activity was assessed was of particular interest. Thereafter, references to validation and reproducibility studies

\textsuperscript{31} Kriska M.A, Caspersen, CJ. Introduction to a Collection of Physical Activity Questionnaires, Med Sci Sports Exerc 1997 Jun;29(6) Supplement: 3-201
retrieved from the collection were chosen based on two criterions: The first was to systematically select the first article(s) in the reference list of each questionnaire, and the second that the full versions of the articles were available through the library at the Karolinska Institute. The Karolinska Institute library is a well-recognized base for international publications of this type and is located in Stockholm, Sweden, where this paper was written.

In the validity and reproducibility studies, a search for prevalence of validity and reproducibility testing of reported frequency of physical activity was made, and the methods of the studies were investigated. The results were illustrated in tables, and each study was also further explained in text. Finally, the field of new methods for the assessment of physical activity was examined and methodological improvements were discussed.

3.1 Evaluation of statistical methods

In the evaluation of validity and reliability, the methods and the chosen statistical techniques in the studies was criticised based on following questions:

- Is a linear regression analysis presented?
- Has change of mean been studied?
- What methods have been used to determine validity?
- Is there an estimation of within-subject variation (computation of the standard error of measurement, or limits of agreement)?
- How is test-retest performed?
- What are the characteristics and size of the study population?
4 Results

• What are the methods and the results of the questionnaires’ validity and reproducibility studies?

A scheme of methods and instruments to determine reliability and validity of each questionnaire is presented in table 3 and 4 (page 18-19). To test reliability, test-retest correlation is the primary method. Time between tests vary from 3 days\textsuperscript{32} to approximately 2 1/2 year\textsuperscript{33}. No study has performed an investigation of changes of mean. One study (IPAQ)\textsuperscript{34} reports a learning effect between measures, but it is not reported how the authors have come to this conclusion. In many studies, within-subject variation is calculated (n=14).

Details vary in the description of how the questionnaires has been validated and re-tested. Less detailed are studies where the validation and reliability procedure of the questionnaire is not the main focus, but the investigation of physical activity’s correlation to specific diseases. The administration of the reliability and validity procedure is in these studies less informative, and the validity measures are often referred to as described elsewhere. The validity of the physical activity history questionnaire\textsuperscript{35} used as criterion in several studies is not discussed. A few studies are very descriptive both in describing the reliability and validity procedures. For example the Minnesota Questionnaire and the KIHD study.

Reliability is overall reported to be good or very good among all questionnaires. They generally refer this to the test-retest correlation coefficient. Change of mean or a linear regression of reliability coefficients is not investigated in any study. Validity is also presented as correlation coefficients, with a median around 0.30. One study (the 7-day

\textsuperscript{34} Craig C. L., A. L. Marshall, M. Sjöström, et.al. pp.1381-1390
\textsuperscript{35} See the validity scheme of studies
recall) has fitted linear regressions to its validity coefficients. Although the questionnaires vary both in complexity and length, the reliability and validity results are fairly equal among several of the investigated questionnaires. One study suggests that it is not the length or attention to detail of a questionnaire that leads to higher validity. More important seems to be the logic of how the questions are constructed.

Common construct validity methods are instruments assessing VO$_2$max and body fat. Accelerometers, physical activity log books, and to validate the questionnaire against other physical activity questionnaires are common criterion methods. VO$_2$max as a construct criterion is favourably for questions regarding higher intensities such as sport activities. Several questionnaires show the highest correlations in questions about vigorous bouts (for example running and swimming), rather than light and moderate activities. Walking is the least reliable activity in several studies. This is for example illustrated in the studies of the Modifiable activity questionnaire, the Five-City project, and the College Alumni questionnaire.

The majority of the studies investigated, conclusively support and recommend the tested questionnaire as a useful measurement. Often the results are stated to be comparable to results from other studies of questionnaires.

- **How is the frequency of physical activity assessed and validated in different PA questionnaires?**

A majority of the questionnaires ask questions about, or include the frequency of physical activity. A few exclude the dimension and ask for total duration spent on an activity a week and multiply it with intensity for an estimation of total physical activity. A further detailed scheme of characteristics of the investigated questionnaires, including how frequency is assessed, is presented in table 1 and 2 (page 16-17).

As mentioned in the introduction, spot data does not give an accurate illustration of an individual’s physical activity over time. A few questionnaires ask for the frequency of

---

38 For example studies of Baecke Questionnaire, CARDIA, Framingham, Godin and Minnesota.
physical activities over a year. The KIHD 12-Month Physical Activity Questionnaire asks for monthly participation in leisure-time physical activities, and total energy expenditure is summarized month-by-month. This questionnaire is aspiring to illustrate variations in physical activity patterns during the year. However, a long time frame in a questionnaire does not necessarily reveal seasonal variations: The Modifiable Activity Questionnaire asks for physical activity month-by-month during the past year. Nevertheless all activities are summed to total energy expenditure over the year, and averaged to energy expenditure a week, and therefore no information about continuity of physical activity is given. The CARDIA physical activity history asks for participation, mainly in sports, during the past 12 months. There is however no distinction between months; thus the questionnaire does not measure any variation in frequency of physical activity.

The frequency variable has not been directly validated or tested for reproducibility in any of the investigated studies. The unit of interest to validate is total energy expenditure, not the components of it (frequency, duration, and intensity). The studies also validate energy expenditure by separating physical activity into intensity levels (light, moderate, and vigorous intensity). Two studies out of 18 have examined construct validity of reported frequency in activities with higher intensity levels: The Aerobic Centre Questionnaire study has studied the accuracy and reproducibility of reported frequency of sweating; how many times a week the subject is physically active enough to make up a sweat. This question is indirectly validated by tread mill time (VO₂max). It result in a correlation coefficient of 0.51, and in a multiple linear regression the same question result in $\beta=0.35^*$ (CI=29.40 - 62.95). This means, 0.35 of the obtained VO₂max result is related to the question about frequency of sweat. The Godin Leisure-Time Exercise Questionnaire has validated a similar question against accelerometer ($r=0.31$), VO2max ($r=0.57$), and a four-week physical activity history questionnaire ($r=0.52$). The procedure for this is not clear though, and the unit of the accelerometer and the four week-physical activity history is energy expenditure, not frequency. In addition only correlation coefficients are computed in this study.

A further presentation of each questionnaire and its validity and reproducibility studies is found in appendix 1.
### Table 1: Characteristics of physical activity questionnaires

<table>
<thead>
<tr>
<th>Test</th>
<th>IPAQ short</th>
<th>IPAQ long</th>
<th>Aerobic Center</th>
<th>Baecke</th>
<th>Bouchard</th>
<th>CARDIA</th>
<th>Framingham</th>
<th>Godin</th>
<th>HIP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Type of activity assessed:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leisure Time</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Occupational</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Household</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td><strong>Transport</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Questions about sedentary time</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td><strong>Time of frame:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24 hours</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>3 days</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>7 days (past week)</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 months</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>12 months</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td><strong>Usual activity</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td><strong>Recall of specific life periods</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Administration:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interview</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Self-administered</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Frequency assessed as:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Never/seldom/sometimes/often et.c</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Days a week</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Times a day</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Times a week</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Times a year</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x &gt;15 min</td>
<td>x</td>
</tr>
<tr>
<td>Times per month (jan feb et.c)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Months a year</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>assessed as part of duration</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>
### Table 2: Characteristics of physical activity questionnaires, continue.

<table>
<thead>
<tr>
<th>TEST:</th>
<th>Historical</th>
<th>KIHD 7-days</th>
<th>KIHD 12mos</th>
<th>KIHD 24hrs</th>
<th>Lipid Resrch</th>
<th>Minnesota</th>
<th>Modifiable</th>
<th>Paffenbarger</th>
<th>7 day recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type of activity assessed:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leisure Time</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Occupational</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Household</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Transport</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Questions about sedentary time</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time of frame:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24 hours</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>3 days</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>7 days (past week)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>3 months</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>12 months</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Usual activity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Recall of specific life periods</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Administration:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interview</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Self-administered</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frequency assessed as:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Never/seldom/sometimes/often et c</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Days a week</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Times a day</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Times a week</td>
<td>x (≥ 3 times)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x ≥ 1 time</td>
</tr>
<tr>
<td>Times a year</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Times per month (jan feb et c)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Months a year</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>assessed as part of duration</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
</tbody>
</table>
### Table 3: Validity and reliability scheme

<table>
<thead>
<tr>
<th>Test</th>
<th>IPAQ short</th>
<th>IPAQ long</th>
<th>Aerobic Center</th>
<th>Baecke</th>
<th>Bouchard</th>
<th>CARDIA</th>
<th>Framingham</th>
<th>Godin</th>
<th>HIP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validity:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study sample</td>
<td>International</td>
<td>International</td>
<td>Caucasian men</td>
<td>elder women</td>
<td>children/adults</td>
<td>educated</td>
<td>urban/rural men</td>
<td>educated</td>
<td>men/women</td>
</tr>
<tr>
<td>n=</td>
<td>30-257</td>
<td>30-257</td>
<td>375</td>
<td>246</td>
<td>150+150</td>
<td>78</td>
<td>9 824</td>
<td>78</td>
<td>75</td>
</tr>
<tr>
<td>Frequency validated</td>
<td>0.51 (sweating/wk)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correlation coefficients</td>
<td>Spearman</td>
<td>Spearman</td>
<td>Pearson</td>
<td>x</td>
<td>x</td>
<td>Spearman</td>
<td>Pearson</td>
<td>x</td>
<td>Pearson</td>
</tr>
<tr>
<td>Linear regression</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within-subject variation</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Criterion/Construct validity methods:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treadmill test performance (VO2max)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accelerometer: CALTRAC, LSI or CSA</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>The Paffenbarger PA questionnaire</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other PA questionnaire(s)</td>
<td>x (IPAQ)</td>
<td>x (IPAQ)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Physical activity log book</td>
<td>x (IPAQ)</td>
<td>x (IPAQ)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean of 14 4-wk PA histories</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Caloric intake</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ergometer (PWC/kg)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Body Fatness (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Resting heart rate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Vital capacity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Grip strength</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VO2max (method non-specified)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MET of highest intensity activity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Reliability:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study sample</td>
<td>19-257</td>
<td>19-255</td>
<td>246</td>
<td>150+150</td>
<td>78</td>
<td>9 824</td>
<td>78</td>
<td>75</td>
<td></td>
</tr>
<tr>
<td>n=</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within-subject variation</td>
<td>x</td>
<td>x</td>
<td>246</td>
<td>150+150</td>
<td>78</td>
<td>9 824</td>
<td>78</td>
<td>75</td>
<td></td>
</tr>
<tr>
<td>Changes of mean</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test-retest correlation coefficient</td>
<td>Spearman</td>
<td>Spearman</td>
<td>x</td>
<td>ICC</td>
<td>x</td>
<td>Pearson</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time between test 1 and 2</td>
<td>3-7 days</td>
<td>3-7 days</td>
<td>3 mos</td>
<td>6-10 days</td>
<td>1 mos</td>
<td>2.5 / 3 yrs</td>
<td>1 mos</td>
<td>1 mo</td>
<td></td>
</tr>
<tr>
<td>Reproducibility of frequency variable</td>
<td>x (sweating/wk)</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear regression</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 4: Validity and reliability scheme, continue.

<table>
<thead>
<tr>
<th>TEST:</th>
<th>Historical</th>
<th>KIHD 7-days</th>
<th>KIHD 12mos</th>
<th>KIHD 24hrs</th>
<th>Lipid Research</th>
<th>Minnesota</th>
<th>Modifiable</th>
<th>College Alumn</th>
<th>7 day recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validity:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study sample</td>
<td>elderly women</td>
<td>middle aged men</td>
<td>middle aged men</td>
<td>middle aged men</td>
<td>men/women</td>
<td>men/women</td>
<td>Pima indians</td>
<td>men/women</td>
<td>uni students</td>
</tr>
<tr>
<td>n=</td>
<td>223</td>
<td>2,492</td>
<td>2,492</td>
<td>2,492</td>
<td>28+50</td>
<td>28+50</td>
<td>17</td>
<td>28+50</td>
<td>80+78</td>
</tr>
<tr>
<td>Frequency validated</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correlation coefficients</td>
<td>Spearman</td>
<td>ICC</td>
<td>ICC</td>
<td>ICC</td>
<td>?</td>
<td>x</td>
<td>Spearman</td>
<td>x</td>
<td>Pearson</td>
</tr>
<tr>
<td>Linear regression</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within-subject variation</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Criterion/Construct validity methods:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treadmill test performance (VO2max)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accelerometer: CALTRAC, LSI or CSA</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>The Paffenbarger PA questionnaire</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other PA questionnaire(s)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Physical activity log book</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean of 14 4-wk PA histories</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caloric intake</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ergometer (PWC/kg)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Body Fatness (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Resting heart rate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vital capacity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grip strength</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VO2max (method non-specified)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MET of highest intensity activity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reliability:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study sample</td>
<td>elderly women</td>
<td>Men 54 yrs</td>
<td>Men 54 yrs</td>
<td>Men 54 yrs</td>
<td>men/women</td>
<td>men/women</td>
<td>Pima indians</td>
<td>men/women</td>
<td>uni students</td>
</tr>
<tr>
<td>n=</td>
<td>23</td>
<td>51</td>
<td>51</td>
<td>51</td>
<td>28+50</td>
<td>28+50</td>
<td>29</td>
<td>28+50</td>
<td>91</td>
</tr>
<tr>
<td>Within-subject variation</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x(11)</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Changes of mean</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test-retest correlation coefficient</td>
<td>Pearson</td>
<td>ICC</td>
<td>ICC</td>
<td>ICC</td>
<td>Pearson</td>
<td>x</td>
<td>Spearman</td>
<td>Pearson</td>
<td>ICC</td>
</tr>
<tr>
<td>Time between test 1 and 2</td>
<td>3 mos</td>
<td>52-58 wks</td>
<td>52-58 wks</td>
<td>52-58 wks</td>
<td>1 mo</td>
<td>1 mo, 1 yr</td>
<td>1-3 wks</td>
<td>1 mo/8-9 mo</td>
<td>3 &amp; 7 wks</td>
</tr>
<tr>
<td>Reproducibility of frequency variable</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear regression</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
5. Discussion

The purpose of this paper was to investigate physical activity questionnaires, and to examine studies of their reliability and validity, focusing on the variable frequency.

Since the investigated studies in this paper were performed from 1978-1992, an assumption would be that there have been methodological developments in today’s studying of validity and reproducibility of physical activity questionnaires. However, M. E Schmidt and K. Steindorf at the Unit of Environmental Epidemiology, Heidelberg, Germany investigated in 2006 statistical methods used in validation studies for questionnaires. By a literature review, own simulations and an own performed validation study they studied limitations, advantages and new aspects of the methods. The literature review revealed that correlation coefficients still are the common approach. This was found in 41 of the 46 reviewed publications. In their own simulations they showed that serious bias in questionnaires could be revealed by Bland-Altman plots but may be hidden when using correlation coefficients.39

Another study, conducted by L.M. Mackay, G.M Schofield and P.J Schutler at the Center for Physical Activity and Nutrition Research, Auckland University of Technology, New Zealand, was designed to validate a self-administered questionnaire by using both correlation coefficients and Bland Altman methods. The correlation coefficients were ranked as moderate and supported the use of the questionnaire, whereas the results from the Bland Altman methods indicated large discrepancies between the measures. The authors conclude that these findings illustrate both the limitations of correlation coefficients in validation studies, and the inaccuracy of physical activity self-report questionnaires.40

In the investigated literature, information of frequency of physical activity had not been directly validated. Focus is on total energy expenditure and to validate different physical activity intensity levels. In addition, none of the questionnaires investigated meets the requirements of including both several physical activity domains and seasonal variation, not even the topical survey IPAQ. Although IPAQ assesses the major physical activity

domains, it measures physical activity from the past week. IPAQ would need to be administered several times during a year to gather information about seasonal variation and variation in physical activity. Almost none of the validity and reproducibility studies perform accurate statistical methods. Still, many of them promote the questionnaires to be used in surveys. There is also a general lack of self-criticism in the discussions. The recent survey IPAQ however, point out that further statistical analysis of this international questionnaire need to be performed.

These findings raise questions about the researcher’s objectivity. Is the purpose of the study to critically test the accuracy of a questionnaire, or to support its use? Results are often stated to be as good as results in other studies of questionnaires. But does this make the questionnaire more valid?

An interesting notice is that the collection of physical activity questionnaires, which the references to the validity and reproducibility studies are briefly presented in and gathered from, focuses on what type of correlation coefficient is calculated (and its result). There is no information or discussion about whether there has been an investigation of changes of mean or systematic errors. This may illustrate a common idea and a tradition in the field; that correlation coefficients are a sufficient method, and the prime method of interest. This tradition is confirmed by other authors.41

Using questionnaires in large-scale surveys are feasible and economic, but their wide use and construction needs to be discussed. Physical activity questionnaires and physical activity logs may reflect structured activity such as sport and physically active transport fairly adequately. Strenuous physical activities seem to be easier to recall than light activities. As high intensity activities positively affect VO₂max, this may be one reason to why questions about sports are more valid than other questions when tested against physical fitness - this is concluded in several studies. However, physical fitness does not equal physical activity. But what single measurement does? The challenge seems to be to find a combination of measurement instrument that can cover as much as possible of a person’s physical activity pattern.

The new approach to measure all physical activity and not solely in the domain of exercise and sport activities, increases the responsibility of the subject and the demand of its ability

to recall, and respond. As a questionnaire is a subjective instrument, to assess spontaneous physical activity and unconscious movements, such as non-exercise activity thermogenesis (NEAT)\(^\text{42}\), is a challenging issue. It becomes obvious that a physical activity questionnaire cannot assess all types of activity.

Regarding the facts that people are affected by the social context in which they live in, other people should not be the norm that the subject is asked to relate to when answering questions about physical activity. This type of question is seen in The Lipid Research Clinic’s Questionnaire. The answers will depend on the person’s social network, not the frequency, duration and intensity of activities the subject performs. Perhaps a questionnaire as a measurement of physical activity cannot be expected to be highly valid, as a subject not affected by recall bias and social desirability does not exist. Despite this, can a physical activity questionnaire still be useful?

The administration of a questionnaire raises focus on the self and the question “how physically active am I?” This may initiate behavioural changes within the subject. This is something positive; a questionnaire could possibly serve an intervention in itself. In addition, this effect is potential in all types of physical activity measures, also in “objective” measures such accelerometers and pedometers. In a test-retest reliability situation however, this is an undesired effect. The increased physical activity report on the second test would indicate low reliability within the measure, when it in fact is the subject that has changed his or her way of living. The length of the time between test 1 and 2 may also affect reliability results. This illustrates the complexity in assessing physical activity. Important is to give clear guidelines to the subject; they should not alter their activity patterns and illness should be reported. Change of mean and systematic changes must be studied. Also, the interviewer’s technique must be evaluated, especially if there are several interviewers. Literature written by W.G Hopkins is recommended to any student or researcher who wishes to perform reliability studies.

To make a result applicable on a whole nation, the study sample must be representative for the nation’s population. None of the investigated studies result can be generalised. Words and meanings may be interpreted differently between subjects, age groups and between

cultures. This challenge is briefly mentioned in the Baecke study, and the IPAQ survey is reports to have this in mind. Physical activity patterns vary among nations due to many factors. These aspects must be taken in concern when constructing questions for an international questionnaire.

There is no golden standard measurement for physical activity in large-scale populations. Accelerometer has been suggested to be a good instrument for assessing physical activity. This measure is however not highly correlated to any questionnaire in the investigated studies. Reasons for this can be poor validity in the questionnaires, or methodological issues. For example: if accelerometer reading is validated against reported “usual” physical activity (or past year physical activity), the instrument will not measure the same thing - as the accelerometer measures actual activity. The study of the Modifiable questionnaire reports instrumental failure and missing data due to incorrect usage of the accelerometer. This illustrates the potential administrative problems there are when using accelerometers, and that the human factor’s influence is crucial also in objective measurements. Studies using accelerometers as validity measures should suggestively report and evaluate the administration of the instrument, as in the study of the Modifiable questionnaire.

A correlation cannot explain cause and effect. In studies where a questionnaire is set as criterion method, it is impossible to analyse which is the better measure, as no questionnaire is reported to be valid. The comparison between questionnaires is therefore an inadequate method.

One way to avoid seasonal variations and illness could be to ask for usual physical activity and not a specific week. But asking for usual physical activity would be the same as asking for average physical activity. This would create an imaginary continuity. Regarding all types of physical activity measures (questionnaires, accelerometers, logs et cetera) an issue with the data collection is that it is spot data collected at a certain time. As illustrated, physical is complex and vary over time. Suggestively, repeated physical activity measures, such as once a month, could provide more accurate data of a person’s physical activity, and the variances in physical activity.

The author of this report previously had limited practical experience of analysing statistical methods. To increase the understanding for the specific statistical methods was necessary,

---

43 Kriska, A.M., Knowler, W. C., LaPorte, R.E. et. al. p. 404
and with information from supervisor and by literature the knowledge was deepened. This has been the precondition when writing this report and may be a potential limitation to have in mind when reading the results and discussion. However, in order to minimise the risk of the author misinterpret the findings due to uncertainties, there has been consultations with expertise within the field.

In conclusion it is likely that physical activity recommendations will be modified further, as they rely on the investigated PA questionnaires. To do this, future reproducibility and validity studies must focus on detecting order effect and systematic error of the questionnaire, until then interpretations of reported physical activity data will be misleading.
References


Kohl, H. W., Blair, S. N., Paffenbarger Jr, R.S., et.al.,”A mail survey of physical activity habits as related to measured physical fitness”. Am J. Epidemiol. 127:1228-1239, 1988


Lakka, T.A, and Salonen J.T., ”Intra-Person Variability of Various Physical Activity Assessments in the Kuopio Ischaemic Heart Disease Risk Factor Study”, Int. J. Epidemiology. 21(3) 467-472. p. 467-468


Schantz P., Stigell, E., "Frequency of bicycle trips per week and bicycling days per year as input data in cost-benefit analyses”. In press *


**Electronical sources:**

Appendix 1

4.1 IPAQ short and long version

The International Physical Activity Questionnaire (IPAQ) was developed for cross-national monitoring of physical activity and inactivity in adults, aged 18-65 year. The short version consists of 9 items and provides information on the time spent walking, in vigorous- and moderate-intensity activity and in sedentary activity, and participants are instructed to refer to all arenas of physical activity. The long version (31 items) asks for information within the domains of household activities and gardening, occupational activity, self-powered transport, leisure-time physical activity and time spent sedentary. Frequency is assessed as “how many days a week”, but the variable in itself has not been validated in the validity and reproducibility study investigated. In the study, IPAQ forms was translated and adapted by standard methods to suit the 12 different countries participating. The samples in each country ranged between 28-210 individuals. They were mostly convenience samples but collectively they represented a wide range of age, education, income, and activity level.

Criterion validity was examined with accelerometer (CSA model 7164) as criterion. The subjects wore the CSA during 7 days and data were stored in 1-min intervals, and contrasted to the estimated total physical activity (expressed as MET x min x wk⁻¹) from the IPAQ. The Spearman correlation coefficient for total physical activity was 0.33 (95% CI 0.26-0.39) for the long forms and 0.30 (95% CI 0.23-0.26) for the short forms. A wider range of correlations was associated with the long version. Concurrent validity was measured by comparing the data from two different IPAQ forms administered during the same day. The correlation between the long and short version resulted in an agreement of 0.67. Comparison between short forms resulted in a coefficient of 0.58 (0.51-0.64) In the Canadian study a comparison between telephone and self-administered modes of data collection was performed, with no major differences in correlation coefficients between the methods. Some development countries reported a preference for self-administration as telephones were not sufficient available.

Test-retest correlation coefficients were assessed within a week, and computed for total physical activity and total sitting time. Coefficients for the long version ranged from 0.96 (USA) to 0.46 (South Africa, rural sample) but most results were around 0.80 (95% CI 0.79-
0.82). It is unclear whether change of mean was investigated; some of the countries administered the forms a third time, 3 days after the second visit. In the discussion it is mentioned that both the long and short form showed evidence of a learning effect over time; subjects showed improvements over time in reliability and in concurrent validity. How they conclude it to be an effect of learning is however not presented. The question regarding walking and cycling pace made little contribution to reliability and validity and was removed from both questionnaires, but questions about occupational physical activity in the long form may have contributed to the absolute differences between the long and short version.

There is no computed linear regression in the study. The authors state that further work requires an examination of the absolute validity, especially between the CSA and self-reported IPAQ data. They also conclude that the assessing of multiple domains of activity leads to higher prevalence rates of physical activity, and suggests that new cut-points for health may need to be explored.  

4.2 Aerobic Centre Longitudinal Study Physical Activity Questionnaire

This questionnaire measures leisure-time and household physical activity for the last three months. Subjects are asked to quantify their weekly participation in different activities (performed at least once a week, mainly sports) and to report miles and duration of the activities. Participants who report walking, running or jogging are asked to provide the number of workouts per week (frequency). The last question asks for frequency as “how many times a week do you engage in vigorous physical activity long enough to work up a sweat?”

This validity study was conducted 1988 in Dallas, the US, by the preventive medicine clinic The Cooper Clinic. 375 men (mean age 47.1 years) served as subjects. They were chosen from a follow-up mail survey, which included all patients who had had at least one examination at the clinic. All subjects were Caucasian and reported a current smoking habit substantially lower than the national prevalence.

Construct validity was assessed: the Subject’s responses to the physical activity recall questionnaire were compared to maximal treadmill test performance. (VO$_2$max).

---

Age-adjusted Pearson correlation coefficients were computed, where frequency as workouts per week resulted in 0.29, and frequency of sweating per week resulted in 0.51. Sweating per week was the question with the strongest correlation to treadmill performance. A multiple linear regression with all subjects included was made, where the sweat frequency significantly resulted in $\beta=0.35^* \ (CI=29.40 - 62.95)$. The result was similar ($\beta=0.35^* - 0.36^*$) also when subjects are divided into age groups. However, no linear regression to investigate systematic changes is performed.

They comment that physical fitness and physical activity are not exactly comparable. Also, they mention the genetic component of physical fitness, which weakens its relation to physical activity. Though, conclusively they state physical fitness as a good “surrogate instrument”, and that one must expect low to moderate correlations between the estimates of physical activity and physical fitness.

Unfortunately, no study of reliability was listed in the reference list for the Aerobic Centre Longitudinal Study.\textsuperscript{45}

\textbf{4.3 Baecke Questionnaire}

The Baecke questionnaire asks questions regarding occupational, leisure and sport activities, all based on frequency (how many times) per week. The subject answers on a nominal scale “never”, “seldom”, “sometimes”, “often”, or “always”. For questions about sport and exercise, perceived intensity is included (as light, moderate or strenuous). In the validation study, population was based on 246 white, postmenopausal women who were randomised into either an intervention or a control group. The women were part of an ongoing clinical trial on the effect of moderate weight-bearing activity on adult bone loss.

The intervention group were asked to walk 7-9 miles a week during 12 months. Five measurement techniques were used as criterions to validate the questionnaire: the Paffenbarger survey of Harvard alumni, a modified Paffenbarger, the large-scale integrated activity monitor (accelerometer) and caloric intake. The reliability of each of these

\textsuperscript{45} Kohl, H. W., S. N. Blair, R.S. Paffenbarger Jr, et.al."A mail survey of physical activity habits as related to measured physical fitness”. \textit{Am J. Epidemiol.} 127:1228-1239, 1988
instruments was tested prior to the study. Inter-correlation among these measurements was also examined.

Solely correlation coefficients were used as statistical method in the validity study. Correlations were computed for the work-leisure-time and sport index separately (method not specified). The accelerometer was carried for 3 “usual” days and caloric intake was calculated from three-days food records. With the assumption that people with stable body weights expend as much energy as they consume, the data was compared to the calculated energy expenditure from the questionnaire. Caloric intake was however not a good index of physical activity in this population (correlation coefficients ranged from -0.31 to -0.08). According to the authors the reason for this may be that the individual’s metabolic rate must be taken into account if caloric intake is to be used as a measure of caloric expenditure. This was not done in this study. The correlation between the Baecke and the Paffenbarger questionnaire was 0.06 (work index), 0.19 (leisure index), and 0.48 (sport index). The Accelerometer correlations ranged between 0.07 and 0.16. No linear regression or other method measuring systematic bias was calculated. The study mention that a qualitative scale as in the Baecke questionnaire lacks precision in estimating activity and bias can occur as the words may not mean the same thing in each population.

In the reliability-study, Dutch men (n=139) and women (n=167) from various socio-economic classes with an age between 22 and 32 year served as subjects. Subjects were invited by mail to complete the questionnaire at home and then to visit a mobile research unit that was stationed for 7 days in each section of the town. The questionnaire was checked for completeness and anthropometric measurements were checked. After approximately three months the participants were visited at home and requested again to fill in the questionnaire. Product-moment correlation coefficients were calculated to study the test-retest reliability. The resulting correlation coefficients were 0.80 - 0.90 for the work index and sport index, and 0.74 for the leisure-time index. Standard deviation for the results are not computed, only a standard error of the mean (SEM). The Change of mean was not investigated, and so was not a linear regression.

---

4.4 Bouchard Three-Day Physical Activity Record

This questionnaire is a three-day record, where one of the three days is a weekday. It measures activities in 15-min bouts, where frequency is seen as a part of duration and not specifically mentioned in any question. The examined validation and reproducibility study is based on families living in the Quebec area, Canada. They were recruited through the media and totally 150 adults and 150 children were randomly selected from as many families. Age ranged from 10 to 50 year and the sample had a diverse socioeconomic background, although participation rates were better for middle and upper socioeconomic classes. Test-retest reliability of the activity record was performed; 61 of the subjects (from 16 different families) repeated the test, with 6 to 10 days between test 1 and test 2. The intra-class reliability was 0.91 for children and 0.97 for adults, and the results were significant for both groups. Standard deviations were calculated for the 3-day record, the PWC/150 (kpm x min^{-1}), PWC150/kg and percentage body fat. Change of mean is not investigated.

Validity is measured by investigating the relationship between energy expenditure (calculated from the questionnaire), physical working capacity (expressed as PWC150 and PWC150/kg), and body fitness. Activities from the questionnaires were divided into categories of different intensity. The reported frequency of activities with high intensity had a significant correlation coefficient of 0.22-0.32* when compared with physical working capacity/kg and -0.23* when being correlated to percentage of body fat. Mean expenditure per day/kg and PWC150/kg had a correlation of 0.31*. When excluding the variable weight, energy expenditure expressed as kcal/day was highly correlated with PWC150 (0.70*). Based on this, the study concludes that the 3-day activity record is suitable to estimate energy expenditure in population studies.47

4.5 CARDIA physical activity history questionnaire

The CARDIA physical activity history questionnaire measures leisure, job and household activities over an entire year, where frequency is expressed as times per week (with minimum duration set to 60 minutes). The one-year time frame specifies time as ‘last 12 months’. The subject is asked to answer how many months the activity was performed, but not the specific months.


32
The investigated validation and reproducibility study had a study population of 78 individuals (28 men and 50 women), aged 20-59. They were recruited by advertisement from the local university community and were to 94% Caucasians with college or graduate degrees (71%). Test-retest reliability, with at least a 1-month interval was performed, resulting in a correlation coefficient of 0.88 in total.

Validation of the questionnaire was conducted against five different criterions; treadmill exercise performance, vital capacity, body fatness, the average of 14 4-wk physical activity histories and the average of 14 2-day accelerometer readings. Sex-specific means and standard deviations for the measurements were computed. Reproducibility of the accelerometer and the treadmill exercise test was examined after 1 month and after about 1 year, and had a test-retest correlation of 0.69 or greater. The relationship between the questionnaire and treadmill exercise performance (V\text{O}_2^{\text{max}}) were 0.08 for moderate activities and 0.63 for heavy physical activity. With accelerometer as criterion method (compared MET values), the correlation was 0.31 for heavy intensity and 0.11 for moderate intensity. Percent of body fat had a correlation coefficient of -0.35 to high intensity activity. A correlation of 0.83 was found between high intensity score in CARDIA and the high intensity score from the four-week history. It was concluded that vital capacity in the normal state was not a useful validation standard (resulted in a coefficient of 0.15 for high-intensity activities). The method used was Spearman correlations. Further detailed analysis of change of mean and linear regression was not made.48

4.6 Framingham Physical Activity Index

Framingham Physical Activity Index is an interviewer-administered short questionnaire, which briefly assesses usual occupational and leisure activity components over the course of a 24-h day. The interviewer asks the individual about the average hours of participation in sedentary, light, moderate and heavy activities.

The examined validity and reliability study aims to investigate correlations between physical activity and cardiovascular diseases rather than to report the accuracy of the Framingham Physical Activity Index. The study population consists of one rural group and one urban group. Totally 9, 824 men between 45 and 64 year participated. As physical activity in this

questionnaire is computed as $duration \times intensity$, no information or investigation of frequency is presented. Repeatability of the questionnaire was tested three times, $2 \frac{1}{2}$ to 3 years apart, with result in Pearson test-retest correlations between 0.30 and 0.59*. Change of mean is not investigated, neither is within-subject variation between the trials. Standard deviations are only computed for the results from examination 1.

Validity of the questionnaire was examined by two criterions: resting heart rate and the MET’s of the highest intensity activity. Pearson correlation coefficients were investigated. The correlation between physical activity index and heart rate was inversed: -0.17 for rural men and -0.11 for urban men (age 45-54 for both groups). For the age group 55-64 year the results were -0.21 for the urban sample and -0.15 for the latter. The relationship between MET of highest intensity activity and physical activity index was 0.63 for rural men, and 0.55 for the urban group. How this comparison is made is not clearly described. No further statistical methods than Pearson correlations are presented. The study concludes the Framingham physical activity index to be inversely associated with most known coronary risk factors.49

4.7 Godin Leisure-Time Exercise Questionnaire

Godin’s questionnaire is self-administered and measures usual leisure-time physical activity during no specific time frame. The subject simply estimates a usual 7-day period. Frequency is measured as number of 15-min bouts a week.

In the investigated validity and reproducibility study, the study population consisted of 78 Americans (28 males and 50 females) between the ages of 20 and 59 year. They were recruited by advertisement from the local university community and had college or graduate degrees, or had administrative or professional positions. The study evaluated totally 10 commonly used physical activity questionnaires and the Godin Leisure-Time Exercise Questionnaire was one of them. It is a challenge to distinct the specific procedure for this form.

Reliability of the questionnaire was tested by test- retest, with “at least” one month between test 1 and 2. Means and standard deviations were computed at one of the tests, men and women separately. The test-retest correlation for light activity was 0.24*, moderate activity was 0.36* and strenuous activity was 0.84*. The reproducibility of the frequency of high

---

intensity activities was tested: Retest correlation of the question “how many times per week do you usually engage in activities, long enough to work up a sweat?” was 0.69*.

Statistical method to examine validity was correlation coefficients, with the motive that a large volume of data efficiently could be presented with this method. Means and standard deviations for the validation measures were computed, but further detailed analysis was explained to be “beyond the span of the study”. Thus no investigation of systematic changes is performed. Criterion and construct validity methods used in the validation procedure were treadmill exercise performance, vital capacity, body fatness, the average of 14 4-week physical activity histories and the average of 14 2-d CALTRAC accelerometer readings. Treadmill exercise performance was performed with a direct oxygen measurement. Vital capacity was measured at several times throughout the study period. Underwater weighing and skin-fold measures assessed body fatness. However, it is not clear whether one of the methods or both were used in the study of the Godin questionnaire. The four-week activity history questionnaire was derived from the Minnesota Leisure-time Physical Activity Questionnaire. For the administration of the accelerometer they were instructed to record energy expenditure from the display every four hours while awake for the 2-day period.

The questionnaire’s correlation to treadmill time was 0.57* and 0.52* for the specific sweat question. Vital capacity had no relationship at all to the questionnaire, but percent of body fat showed an inverse correlation of -0.43*. The average of the four week physical activity histories ranged from 0.31 to 0.36*. The accelerometer showed no correlation at all when measured in calories/day. Accelerometer readings expressed as MET-min/day had a correlation coefficient of 0.32*, and 0.29* for the sweat-question. VO2 max had a correlation coefficient of 0.56* for the leisure score, and 0.57* for usual sweat. The authors conclude that maximum aerobic capacity and body fatness are commonly used as in validity studies of physical activity questionnaires. Their data suggest though, that these measures mainly correlate with heavy intensity activity and should therefore not be used as the only validation standards. Godin Leisure-Time Exercise Questionnaire probes almost exclusively structured physical activity (sports).50

4.8 Health Insurance Plan of New York (HIP) Activity Questionnaire

The Health Insurance Plan questionnaire is a self-administered instrument. It consists of six questions about time spent in sitting and walking at work, transportation to and from work (number of blocks walked), and the frequency of heavy lifting at work. It also includes six questions about physical activity off work, which cover most arenas outside the occupational. The study population consisted of 75 American men and women (age 23 to 59), working in administrative jobs. The method is well explained, more detailed than the majority of other examined studies. Test retest-reliability was accomplished with one month between test 1 and test 2, with the result of a Pearson product-moment correlation of 0.86*. Means and standard deviations were calculated in the first trial.

Data from the questionnaire were validated against an occupational record book, and data from other occupational questionnaires. Before each study visit, study participants completed physical activity records for a 48-hour period. In the diaries they noted a general and specific description of the physical activity performed, an estimate of intensity and duration, and if the activity was occupationally related. At each visit, a trained interviewer edited the record for clarity in the presence of the participant, and trained coders transferred the data into MET values. To take daily variation in leisure time and occupational physical activity in account, subjects completed the activity diary on both weekends and weekdays. Physical activity scored zero, were excluded. Pearson correlations were calculated and analyses were adjusted for age and gender due to the lower amount of male subjects. The correlation between Physical activity index and activity records was 0.10. Correlations to other occupational physical activity questionnaires ranged from 0.12 to 0.38* where the highest value was for the Baecke Questionnaire. Means and standard deviations from the validation measures were computed and presented. Changes of mean or linear regression were not computed. 51

4.9 Historical Leisure Activity Questionnaire

This is an interviewed-based questionnaire, which divides the life span into four time periods; 14-21, 22-34, 35-50 and 50+ years. It asks about participation in leisure time physical activities for each period (expressed as numbers/year, months/year, and hours/week). The subject is also asked to circle all sport activities (from a list) performed more than 10 times

during lifetime. It includes bicycling, and *walking for exercise* but exclude walking as a way of transportation. Each activity is, based on estimated intensity level, and converted into kilocalories of energy expenditure.

The aim of the validity and reproducibility study was to determine the association between historical physical activity and bone loss. Therefore, statistics and results mainly focus on this relationship. 220 American postmenopausal women (mean age 54 years) formed the population in the study. They were considered to be fairly inactive. In average, the women were slightly overweight. In order to determine test-retest reliability of the questionnaire, 10 % of the women (n=23) randomly filled in the test two to three months after the initial test. Kappa statistics (which measure the level of agreement between the test and the retest while accounting for chance) and Spearman rank correlation coefficients were calculated for the question that regards frequency (“How often do you participate in sports and leisure time physical activity”). The result ranged from 0.39 to 0.47, depending on time period. Correlation coefficients for total summary of estimates were 0.69 (age period 14-20) and 0.85 (age period 50+). Change of mean in not investigated, and no linear regression is computed.

Construct validity was evaluated by comparing the historical physical activity to bone mass measurements, where total physical activity resulted in a Pearson correlation of 0.19*, walking included. Criterion validity is limited to the most recent time period (50+). The kilocalories of activity expended/week during this period were compared to data from an accelerometer (LSI), the Paffenbarger Survey and grip strength. Pearson correlation coefficients resulted in 0.22* (mean blocks walked/day), 0.46* (sport index) and 0.41* (kilocalories expended/week) for the Paffenberger survey, 0.12* for the accelerometer (day counts/hour), and 0.19* for grip strength. It is not presented how day counts per hour from the accelerometer are compared with calories/week from the questionnaire. Neither is any further analysis of the correlation coefficients.

### 4.10 KIHD (24 hour record, 7 day physical activity recall, and 12 month-physical activity history)

The 24-hour record measures leisure-time and occupational activity in 30-min bouts, including sedentary activity and sleep. The 7-day physical activity recall is interview-

administered and assesses leisure-time from the past week. The 12-month physical activity history is mainly self-administered and asks for common leisure-time activities (mostly sports). In the history the subject is asked to estimate frequency as “how many times per month”, and the duration and intensity class of each activity. Each month of the year is represented. These questionnaires were made for a survey with the purpose of investigating risk factors for Ischemic heart disease and carotid artherosclerosis in middle-aged Finnish men. From this survey, 63 men were invited to the reproducibility study where totally 51 served as subjects. 37% lived in a rural settlement, 73% were married or engaged.

The 12 months and the 24-hour recording were handed out at the first visit, to be completed and returned 7 days later at the second visit. The 7-day past week physical activity-recall interview was performed at the second visit. To avoid seasonal variation, the subjects were invited to a re-examination after approximately 12 months. The test-retest period ranged from 52-58 weeks and the second test was done on the same weekday as the first test. The same person did not administrate all the retests. A nurse with long experience in epidemiological studies performed 13 of the re-interviews, and another nurse with similar experience accomplished the rest of the re-interviews. The second interviewer was trained by the first in practice for 1 week and was able to consult the first concerning the assessments. This study is the first to be this detailed with the procedure of the test-retest.

The means and standard deviations of the physical activity indexes at first and second visit were computed and compared with paired t-tests. The difference between test 1 and 2 in relation to their mean was used as an estimator of agreement between baseline and re-examination values. Intra-class correlation coefficients were also used to describe percentage of the mean (absolute test-retest difference). The change of interviewer was estimated and tested with a multivariate least squares regression analysis, and the change did not explain the differences in test-retest. Intra-class correlation coefficients for the 24-hours total physical activity record, total 7-day recall, and 12-months questionnaire were 0.43, 0.35, and 0.58 respectively. No significance is reported. The within-subject test-retest difference against test-retest for the 24-hour total activity recording, as well as the 12-month activity history mean, is presented in a figure. The 12-month activity history was reported to have relatively small within-person variability. Reproducibility of the frequency variable or change of mean was not investigated. Neither was a linear regression analysis. They found a bias towards either over-estimation of activity in the retest or underestimation at the first test. They suggest that
subjects may have paid attention to physical activity habits or tried to please the interviewer by giving higher responses in the re-examination.

In the validity study, the study population consisted of 2,492 randomly selected Finnish men, aged 42-60. The self-administered questionnaires were sent 4 weeks in advance, and checked in an interview at the first visit. The subjects administered the 24-hour total activity recording the hours before the second visit, which was 7 days after the first visit. To this second appointment they had completed the 12-month physical activity history and the 7-day leisure time activity recall was completed during the 7 days between the visits. No frequency of activities was validated. All data was summarized into total METs. The questionnaire’s total results were compared to each other and to maximal oxygen uptake (VO$_2$ max). The comparisons between forms resulted in correlation coefficients between 0.07* and 0.13* for the 24-hour physical activity recall, and a correlation of 0.45* between the 12-month history and the 7-day recall. Means and standard deviations were computed for the questionnaire’s total results. All questionnaires had age- and examination year-adjusted Pearson correlation coefficients of about 0.17* to VO$_2$ max. A linear regression analysis is not presented.  

4.11 Lipid Research Clinic Questionnaire

This is a self-administered questionnaire, which measures usual leisure time and occupational physical activity with no specific time component. It consists of four simple questions, and is in the introduction of the validity and reproducibility study described as a global assessment tool. The first two questions concern occupational activity level and leisure time activity, and the subject is asked to compare his or her activity level with others of their age and sex on a qualitative scale. The two last queries are closed questions, where the first ask if the subject regularly engages in strenuous activity or hard physical labour, and the last question assesses frequency, as “Do you exercise or labour at least three times a week?”

58 subjects (28 men and 20 women) age 21-59 year, participated in the reliability and validity studies. They were recruited from advertisements on bulletins at the University. Some of the requirements for participating were ability to keep detailed records of physical activity and diet, and willingness to comply with the study protocol for 14 months. How ability was

---

controlled is however not presented. The data was derived from a larger study (SAFE—the Survey of Activity, Fitness, and Exercise). It included 14 clinic visits, approximately 26 days apart.

The questionnaire was indirectly validated against VO₂max (treadmill graded exercise test), body composition (hydrostatic weighing) and estimated energy expenditure from the average of a four-week interview-based physical activity questionnaire, assessed at each study visit. Also, the average of 14 2-days Caltrac accelerometer readings was set as validation method. 48 hours before each clinic visit, participants recorded all physical activity and food intake on a recording form designed for the study. Additionally they wore the accelerometer, and the subject recorded its obtained energy expenditure scores every 4 hour in the same record form as the recorded physical activity. Statistical methods to study validity were to calculate least square means and standard errors (SEM, not SD) of the results from the validation instruments and scores from the questionnaire. This is not clearly described. A multiple linear regression analysis (not to confuse with linear regression) to evaluate the amount of variance among all validation instruments in contrast to the Lipid Research Clinic Questionnaire was performed. The table for these results is however titled “linear regression”, which is confusing. The study highlights the partial r² values for the questionnaire and VO₂max (0.29) and the partial r² value for percent body fat (0.17). They suggest that the Lipid Research Clinic Questionnaire mainly reflects physical activity patterns that increases aerobic capacity and reduces body fat.

Reliability of the questionnaire was conducted by 1-month test-retest, with a Pearson correlation coefficient of 0.88 for the total study sample. There was no larger discrepancy between results for men and women. The ratio of the within-person variance to between-person variance was calculated using variance ratio. This resulted in a 13.6% variance. The particular question regarding frequency is not validated. Neither is change of mean or linear regression a linear regression.⁵⁴

4.12 Minnesota Leisure-Time Physical Activity Questionnaire

This questionnaire measures leisure and household activity. It is interviewer-administered and concerning activity the past 12 month. Totally 63 activities are divided into 8 categories: walking and miscellaneous, conditioning exercise, water activities, winter activities, sports, lawn and garden activities, home repair activities, fishing and hunting, and the interview is estimated to take 10-20 minutes. Frequency of the activities is assessed month-by-month, as “average numbers of times per month”. Each physical activity has its own intensity code (MET). The intensity code and the duration of exercise in minutes for a year results in an index expressed as \( \text{MET} \times \text{min} \times \text{day}^{-1} \). Hence the month-by-month information of possible seasonal variations of physical activity is not further investigated.

The validation and reproducibility study of the Minnesota questionnaire solely focuses on validate and test reliability of the questionnaire, not the inverse relationship between PA and common diseases. Prior to the study, a supplement with household chores was added to the list. Subjects in the study initially were 103 healthy men and women, age 20-59 years, recruited predominately from the University of Minnesota. Almost all of them were non-smokers and Caucasian. They were screened in an attempt to stratify the study population with respect to age, gender, and subjective self-assessment of habitual PA status. They were informed about the requirement to be able to keep detailed records of PA and food intake, and to comply with regular clinic visits for 12-months. Benefits and risk of the study were made clear, and the participants received 200 dollars upon completion of the study. They were also given the results of their physiological testing. 78 participants (76% of the initial group, 28 men and 50 women) completed the entire study. The dropout was mainly due to time constraints.

Reliability was tested both after one month and one year, with age and gender adjusted test-retest correlation coefficients and significant results. One-month test-retest resulted in a correlation of 0.92* and one year test-retest had a correlation 0.69*. Chosen method for this procedure is not explained. Means and standard deviations were calculated at one test. No change of mean is investigated, and no linear regression is fitted.
Criterion validity measures in the study were a physical activity record, Caltrac accelerometer readings, and a 4-week physical activity history, repeated at each study visit\(^5\). Means and standard deviations from the physical activity indices of each instrument were computed. Construct validity was assessed by comparing results from the questionnaire to maximal oxygen consumption (VO2max), and percent body fat. The methods are clearly and well descript. The criterion methods are translated into the same unit: (MET x min x day\(^{-1}\)) and activity is validated for light, moderate and heavy, household and total activity. Correlations were adjusted for age and gender in all comparisons.

The questionnaire’s correlation to the accelerometer was 0.23* at best. This result was seen in moderate activity. In contrast, the 4-week physical activity history questionnaire had a stronger correlation with the accelerometer. The 4-week history’s household indices’ correlation to Minnesota LTPA Questionnaire was 0.77*. Reported heavy physical activity in the PA record contrasted to the Minnesota questionnaire resulted in 0.55*. Validity is high, for total (0.78*) heavy (0.90*) and light (0.77*) activity, when compared to the 4-week history questionnaire.

The associations between the questionnaire and the criterion measures were determined for men and women separately, however the study reports similar results for both genders. Two exceptions were the relationships between the Minnesota LTPA Questionnaire and the PA record’s total activity, which had a correlation coefficient of 0.58* for men and 0.36* for women. The relation between the accelerometer and the Minnesota LTPA Questionnaire also differed between the genders; 0.58* for men and 0.20 (non significant) for women.

This study suggest the PA diary to be a good criterion method for PA questionnaires\(^5\) and that recall bias, especially for assessment of light and moderate intensity activities, is an issue with the questionnaire. No relationship was observed between the PA record and the questionnaire regarding these intensities. The 4-week questionnaire had correlations of 0.70* and 0.72* for moderate and light activities in the Minnesota LTPA Questionnaire.\(^7\)

No further statistical analysis of systematic changes was performed.

---

\(^5\) The criterion methods were administered in the same way as in the validity study of the Godin Leisure-Time Physical Activity questionnaire


\(^7\) Ibid. p. 276
4.13 Modifiable Activity Questionnaire

This interviewer-administered questionnaire was constructed to assess the activity patterns of Native Americans to evaluate the relationship between physical activity and diabetes. It measures past-year and past-week leisure and occupational activity to obtain a general estimate of how physically active an individual was during his/her past, and how physically active he/she is currently. Transportation to and from work is included in the occupational activity section, and inactivity is also measured in the questionnaire. Past-year activities are reported month-by-month, which could reveal information about seasonal variations of physical activity. However, all activities over the year are summed and divided into an average of physical activity per week.

Subjects for the validity and reproducibility study were 29 Pima Indians aged 21-36 years. The population engaged in physically demanding occupations such as farming, therefore assessment of the domain occupational activity was necessary. Test-retest reliability ranged from 0.62 to 0.96, where past year physical activity had higher correlations that past week physical activity. According to past year, an average of physical activity over the year was tested, not month-by-month. If the interviewer subjectively determined that the participant was not capable to correctly answer the questions, the interviewer was instructed to record this unreliability. Values were assessed with Spearman rank-order correlation coefficients, with 1 to 3 weeks between the tests. Authors motivate the use of nonparametric statistics as the data were highly skewed. How skewed is however not presented. Only median hours per week were assessed, no means and standard deviations are presented. The answers concerning past-year leisure activity were more reliable than the answers about the past week. This could be explained by short-term variability of physical activity. They suggest that it is necessarily is to evaluate the degree to which the previous week was typical or representative, when examining physical-activity levels over a short period such as 1 week. When excluding the question about walking reliability was higher, which indicates that the amount of walking was both difficult to recall and varied day to day in this population. Two interviewers performed the test-retest procedure and the agreement between them were

59 Ibid., p. 402
60 Ibid. pp.403-404
61 Ibid., 406
62 Ibid., p.404
examined and found to be high in all leisure-activity variables (0.78 to 0.94). Agreement was lacking between the interviewers in the occupational activity estimation, where they did not ask questions quite the same way. The authors conclude that interviewer testing is important in interview-based surveys to ensure that queries are administered in the same way.

Validity was assessed by compare the questionnaire’s indices to data from the Caltrac accelerometer. 21 Individuals from the test-retest study volunteered to the validity study and were asked to wear the activity for 7 days during the time between the two activity interviews. Of these, 17 subjects wore the monitor correctly for several days (5-7 days). One individual had 4 days of data due to instrument failure. Correlations between the activity monitor’s counts per hour (average over the week) and the reported hours per week of physical activity for both test and retest of the questionnaire were calculated. The questionnaire data from the first test reported the correlations with the accelerometer. Past week leisure time physical activity were more strongly related to the accelerometer than reported past year leisure time physical activity.\textsuperscript{63}

The author’s conclude that although this population has a lower socio-economic status than other study samples, the reliability of the questionnaire is comparable to that of other questionnaires. They also compare the reliability between this questionnaire and the Minnesota questionnaire and find similar reliability, despite the longer time between test and retest in the Minnesota questionnaire (5 weeks, compared to 1-3 weeks)\textsuperscript{64}. However, analysis of within-subject variation, changes of mean or systematic changes have not been made.

\textbf{4.14 College Alumnus Physical Activity Questionnaire}

In the origin survey students from Harvard college alumni were asked to fill in the number of flights of stairs climbed per day, city blocks walked per day, usual walking pace, and the frequency and duration of sports and recreational activities performed during the past year. Also, number of years of participation is requested, and the subject fills in the usual level of exertion. One question asks if the subject engage in regular activity at least once a week long enough to work up a sweat. If the answer is yes, the subject fills in number of times per week. The responses were summed and transferred into energy cost independent of body weight, expressed as MET-min \cdot \text{wk}^{-1}.

\textsuperscript{63} Ibid., p. 404-405  
\textsuperscript{64} Ibid., p. 405
The examined validity and reproducibility study aimed to replicate the method of the original survey. Data was collected from the SAFE study, and so details of protocols and subject eligibility requirements were reported elsewhere. 28 men and 50 women (21-59 yr) recruited by advertisements in the University’s academic and hospital facilities completed the study. The questionnaire was mailed to participant’s homes at the beginning and during the middle of the survey. Participants completed the questionnaire at home and brought it with them to the next clinic visit (totally 14 appointments). The questionnaire was collected without any editing, as in the original study of Paffenbarger et al. 65

Two test-retests were conducted to examine reproducibility, and Pearson product-moment correlation analysis adjusted for age in men and women was computed. Short-term (1 month) reliability ranged from 0.31 to 0.88*, where the latter result was for “flight of stairs climbed”. Long-term reliability (8 to 9 months) ranged from 0.01 (sport and recreation) to 0.63* (flight of stairs). The low test-retest correlation over the 8 and 9 months may, according to the authors, reflect seasonal differences in PA as well as variability in physical habits over time. 66

It is also possible that it is due to recall bias. The Means MET-min · wk\(^{-1}\) and standard deviations of three forms (three clinic visits) were computed, where the total College Alumnus physical activity index, the sport and recreation index, city blocks walked, and flights of stairs climbed were presented separately. The scores were generally higher at the first clinic visit. This result must not be confused with an estimation of change of mean, as the trials were non-consecutive (i.e. 8 months between test 1 and 2). Further computation of means and standard deviations of the test-retest was not made.

Validity was evaluated against measures of maximum aerobic capacity, body fatness, accelerometer, and a physical activity record. The procedure is well described in the text.

All physical activity was recorded by participants on a physical activity record during 48-hours before each clinic visit. Each day of the week, including weekends were recorded at least twice during the study. Subjects recorded a general and specific description of physical activity, estimated the intensity and duration of activity in minutes, and noted if the activity was related to occupation. A trained interviewer edited the 48-hour record for clarity together

66 Ainsworth, B. E., A. S. Leon, T. Richardson, D. R. Jacobs, Jr. and R. S. Paffenbarger, pp.1408-1409
with the participant. Trained coders transferred data into MET’s, and energy expenditure was computed by multiplying the MET levels with the duration of each activity. Means and standard deviations from three physical activity records were computed, and contrasted to the College Alumnus Questionnaire.

The Caltrac accelerometer reported energy expenditure both in kcal/day (where height, weight, age and sex was included in the estimation) and in MET’s/day. Caltrac\textsubscript{MET’s} scores were averaged over all 14 study visits to represent a measure of daily physical activity during the study year and thereby “reflect seasonal variability in physical activity habits” (How is it possible to reflect seasonal variation by computing a mean?). A Beckman Metabolic Measurement Cart assessed maximum aerobic capacity during a treadmill test. Body composition was measured by hydrostatic weighing and converted to percent body fat by the Siri equation. Residual lung volume was measured by the Wilmore method. All these tests are referred to as described in detail elsewhere.\textsuperscript{67}

Means and standard deviations were calculated for the validation realms and the data from questionnaire was separated for men and women. Skewed data from the questionnaire and the physical activity records were normalized by decimal logarithmic transformations. 1.0 was added to each value to account for the possibility of zero values for some variables. Average questionnaire- results from three visits (nr 2, 8 and 9) were averaged to one physical activity index. Validity was examined in a “gender-specific, age-adjusted linear regression analysis” to identify the amount of variability (r\textsuperscript{2}) in the validation realms explained by the questionnaire and its components (flight of stairs climbed, city blocks walked, and sports and recreational activities). However, there is no linear regression presented, neither in the result nor the discussion. Correlation coefficients were computed by taking the square root of the r\textsuperscript{2} obtained from the regression analyses. Walking and stair climbing appeared to be underestimated by the questionnaire in both men and women. Sports and recreation scores were lower on the questionnaire when compared to the physical activity records. This led to under reporting of physical activity in the questionnaire. In men, VO\textsubscript{2}max was significantly correlated to questions about total- and heavy-intensity leisure-time physical activity (0.69*) whereas light activities resulted in 0.08. The accelerometer and the College Alumnus Questionnaire had a correlation coefficient of 0.29*. The physical activity records reported city blocks walked had a correlation of 0.64* to city blocks walked in the College Alumnus Questionnaire. Furthermore, they found correlations in heavy intensity activities between

\textsuperscript{67} Ainsworth, B. E., A. S. Leon, T. Richardson, D. R. Jacobs Jr, R. S. Paffenbarger, p. 1407
these two instruments (0.69*). The table for these results is however confusing. It has separated the results for men and women, but has it has two section of results for men with different results. 68

4.15 7-day recall

Subjects in the investigated validity and reproducibility study were students at the University of Georgia. They came from a physical education service program, and were representative for the undergraduate population of the University (55% male, 90% Caucasian)

The frequency variable of physical activity was not investigated, neither in the reliability nor the validity study.

In the validity study the 7-day recall was tested both as an interview-administered instrument, and as a self-administered measurement: During the first week of the spring and the winter quarters, subjects were given a physical activity diary to note their physical activities during the next 7 days. Instructions were given in its use. After the diaries were returned at the end of the 7-day period, each subject completed the self-administered version of the 7-day recall. This was followed by the interview-based form, which was administered by one trained interviewer. Subjects had been unaware they would be asked to complete the 7-day recall after the week’s diary recording. All participants were able to complete the physical activity diary and the follow up recall within a week’s time of each other.

The physical activity diary was set as criterion measure to measure concurrent validity. In a multiple linear regression, the interview-administered recall and the self-administered recall were set as predictors, to determine the degree to which the two forms accurately estimated the diary record. One group completed both the interview-based recall and the self-administered recall, whereas one group completed only the self-recall. They then examined whether the prediction equation generated in each group could provide an accurate estimate of the results from the criterion diary. The regression equation showed a high correspondence between the 7-day recall and the physical activity diary, for both forms (see figure 2 and 4).

68 Ibid. p. 1406
Means, standard deviations and standard errors of the mean for the energy expenditure estimates were computed for each instrument. A Pearson correlation revealed a high correlation (0.82) between all three measures; the interview recall, the self-recall and the diary. Although, measurement error was apparent when the scales were represented along linear intervals.

Based on the results, they made a discriminate classification to categorize the students into groups of “high active” “low active” and “inactive”. They found it difficult to distinct the low active from the inactive, indicating that the recall is most effective for discriminating high active versus inactive.

The 7-day recall was also compared to \( \text{VO}_2\text{max} \) (ergometer test). 24 subjects first completed the self-administered form, followed by a past-year activity history questionnaire that assessed type, frequency, and duration of physical training. The subjects were also asked to categorize themselves as trained or untrained, based on aerobic conditioning during the past year. \( \text{VO}_2\text{max} \) was next determined. Pearson correlation for the past-year activity questionnaire and \( \text{VO}_2\text{max} \) was 0.83* and 0.61* respectively. A linear regression of the 7-day recall with \( \text{VO}_2\text{max} \) was performed, as seen in figure 4.

**Figure 2:** Scattergram describing the linear regression of the interview-based 7-d recall with a concurrent 7-d diary.

**Figure 3:** Scattergram describing the linear regression of the self-administered 7-d recall with a concurrent 7-d diary.

**Figure 4:** Scatterdiagram describing the linear regression of the 7-d recall with \( \text{VO}_2\text{max} \).
To examine the reliability of the 7-day recall, the self-administered version was completed in a quiet settlement five and nine weeks after the first completion (totally three times during the semester). Energy expenditure from the forms was then calculated and compared. Five week test-retest reliability was 0.58*, and reliability between 5th and 9th week was 0.63*. Intra-class correlations resulted in 0.89 and 0.90 for total and vigorous physical activity recall. Change of mean is not investigated.69

69 Dishman, R.K, Steinhardt, M. pp.14-24
Appendix 2

KÄLL- OCH LITTERATURSÖKNING

Frågeställningar:

1. What are the methods and the results of the investigated questionnaires’ validity and reproducibility studies?
2. How is the frequency of physical activity assessed and validated in different PA questionnaires?
3. Is there a need for methodological improvements and if so: what could be improved and how?

VAD?
Vilka ämnesord har du sökt på?

<table>
<thead>
<tr>
<th>Ämnesord</th>
<th>Synonymer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency, physical activity, validity, validation, reliability, IPAQ, seasonal variation*, Bland-Altman, statistical methods, physical activity questionnaire,</td>
<td>Reproducibility</td>
</tr>
</tbody>
</table>

VARFÖR?
Varför har du valt just dessa ämnesord?

Ämnesorden är relevanta för ämnesområdet. Linear regression gav inga bra träffar, därav testades bland-altman som ämnesord.

HUR?
Hur har du sökt i de olika databaserna?

<table>
<thead>
<tr>
<th>Databas</th>
<th>Sökräng</th>
<th>Antal träffar</th>
<th>Antal relevanta träffar</th>
</tr>
</thead>
<tbody>
<tr>
<td>PubMed</td>
<td>Validation IPAQ</td>
<td>16</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Bland-Altman physical activity questionnaire</td>
<td>113</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Systematic error physical activity questionnaire</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Reliability validity “physical activity questionnaire”</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>physical activity seasonal variation</td>
<td>43</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>24</td>
<td>1</td>
</tr>
</tbody>
</table>

KOMMENTARER:
Pubmed’s sökaleternativ “related article” och “review” har varit ett väldigt användbart verktyg. Att inkludera sökordet “frequency” gav många irrelevanta träffar på artiklar gällande alkohol- mat- och drogmisbruk. Signalerar att denna variabel inte är vidare utforskad inom detta område. Att använda synonymen “reproducibility” ökade antalet träffar något vid sökning. Söksträngen validation + IPAQ ger god information om studier gjorda på IPAQ

“Frequency AND physical activity AND validity” och “Frequency AND physical activity AND reproducibility” gav få relevanta träffar.
Handledare har bistått med artiklar (i källhänvisningarna markerade med *) som i sin tur har innehållit användbara referenser. Majoriteten av källorna har hittats på detta sätt.

De analyserade frågeformulärens kommer ursprungligen från “A collection of physical activity questionnaires”. För att kunna ta del av fullängden av dessa artiklar behövs lösenord till e-tidsskrifter. Det har man per automatik om sökningarna sker på Karolinska Institutets bibliotek.