Journal of Physical Activity and Health, 2012, 9(Suppl 1), S44-S55

Sources of Validity Evidence Needed With Self-Report Measures of Physical Activity

Louise C. Mâsse and Judith E. de Niet

Background: Over the years, self-report measures of physical activity (PA) have been employed in applications for which their use was not supported by the validity evidence. ***Methods:*** To address this concern this paper 1) provided an overview of the sources of validity evidence that can be assessed with self-report measures of PA, 2) discussed the validity evidence needed to support the use of self-report in certain applications, and 3) conducted a case review of the 7-day PA Recall (7-d PAR). ***Results:*** This paper discussed 5 sources of validity evidence, those based on: test content; response processes; behavioral stability; relations with other variables; and sensitivity to change. The evidence needed to use self-report measures of PA in epidemiological, surveillance, and intervention studies was presented. These concepts were applied to a case review of the 7-d PAR. The review highlighted the utility of the 7-d PAR to produce valid rankings. Initial support, albeit weaker, for using the 7-d PAR to detect relative change in PA behavior was found. ***Conclusion:*** Overall, self-report measures can validly rank PA behavior but they cannot adequately quantify PA. There is a need to improve the accuracy of self-report measures of PA to provide unbiased estimates of PA.

Keywords: psychometrics, validation, questionnaire, reliability, physical activity assessment

The utility of self-report is often debated; however, self-report measures of physical activity (PA) (ie, interviewer- or self-administered measures) remain the most practical and economical method to assess PA in large studies.[1] Selecting the appropriate self-report measure for a specific application requires a careful review of the validation studies associated with the selected measure to determine whether the interpretation sought is supported by the validity evidence. However, it appears that self-report measures of PA have often been employed in applications for which their use was not supported by such evidence.

In the area of health outcomes assessment and more recently in PA, review checklists have been developed to evaluate the methodological rigor of validation studies [eg, the COSMIN (**CO**nsensus-based **S**tandards for the selection of health status **M**easurement **IN**struments) checklist and the Medical Outcomes Trust (MOT) criteria].[2-4] These checklists highlight the importance of weighting the evaluation criteria based on the interpretation sought with the measure. Unfortunately, little guidance is provided to researchers as to which criteria should be evaluated for certain applications.

Because many self-report measures of PA have been employed in applications that are not supported by the validity evidence, this paper will discuss the sources of validity evidence needed for specific applications in PA (eg, epidemiological, surveillance, and intervention studies). This paper will also focus on aspects of validity that influence the interpretations made with self-report measures and the utility of these measures for common applications in PA. The specific purposes of this paper are to 1) provide an overview of the sources of validity evidence that can be assessed with self-report measures of PA; 2) discuss the validity evidence needed to support interpretations commonly sought in epidemiological, surveillance, and intervention studies; and 3) conduct a case review of the 7-day PA recall (7-d PAR) to demonstrate which interpretations are supported by the measure and the utility of the measure for specific applications. For a more comprehensive overview of factors affecting the validity and reliability of self-report measures of PA, see references.[2,3]

Sources of Validity Evidence

Validity Operational Definition

This paper adopts the following definition of validity from the Standards for Educational and Psychological Testing:

> Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing and evaluating tests. The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations. It is the interpretation of test scores

The authors are with the Dept of Pediatrics, School of Population and Public Health, University of British Columbia, Vancouver, British Columbia, Canada.

> required by proposed uses that are evaluated, not the test itself. When test scores are used or interpreted in more than one way, each intended interpretation must be validated (page 9).[5]

This definition indicates that it is inappropriate to state that a "PA measure is valid" as the measure itself cannot be defined as valid. Instead, it is most appropriate to indicate that the measure is valid for a specific interpretation (eg, valid for ranking PA behaviors among a specific group of individuals). This perspective focuses on validity being a hypothesis driven process requiring one to specify, a-priori, the type of interpretations to validate with the PA measure. Such emphasis does not use the Trinitarian "C"s (ie, content-, construct-, and criterion-related validity) to define the sources of validity evidence; instead, it shifts the focus to describe the sources of evidence needed for a given interpretation.[5] As stated in the Standards for Educational and Psychological Testing:

> The sources of evidence may illuminate different aspects of validity, but they do not represent distinct types of validity. Validity is a unitary concept. It is the degree to which all of the accumulated evidence supports the intended interpretation of test scores for the intended purposes (page 11).[5]

This paper operationalizes the sources of validity evidence by incorporating this perspective and by tailoring these definitions for the assessment of PA. While it is recognized that self-report measures may be developed to assess various components of PA (eg, fitness, bone health, and flexibility), this paper focuses on measures aimed at quantifying the amount of PA (ie, minutes of PA).

Sources of Validity Evidence for PA Measures

Similar to the Standards for Educational and Psychological Testing,[5] Figure 1 lists 5 sources of validity evidence. The list differs slightly from the Standards as it takes into account the measurement properties of PA measures. As a result, validity evidence based on internal structure was eliminated as the Standards refers to methods (eg, factor analysis, Cronbach alpha, and item response theory) that are not appropriate for validating amounts of PA. Specifically, these methods assume that the questions within a measure are correlated. This assumption is unlikely met with measures that assess amount of PA as the various domains of PA are not expected to be correlated (eg, occupational activities are not expected to be correlated with leisure or household activities). Finally, validity based on the behavioral stability of responses was added to the list as self-report measures of PA are often used for surveillance studies (see details below).

Evidence Based on Test Content. This source of validity evidence consists of determining whether the content of the self-report measure comprehensively represents the domains associated with the construct.[5,6] In other words, does the measure assess all relevant domains of PA to reflect the construct assessed? Does the measure adequately cover the sources of activities performed by the targeted population and are example of activities, if provided, appropriate for the targeted population? Is the terminology employed understood by the respondents? Finally, is the measure appropriate or free of bias for various characteristics (eg, gender, age, and race/ethnicity) of the targeted population? Validity evidence based on test content is mostly gathered when the measure is initially developed through focus groups, cognitive interviews, and expert reviews.[5,6]

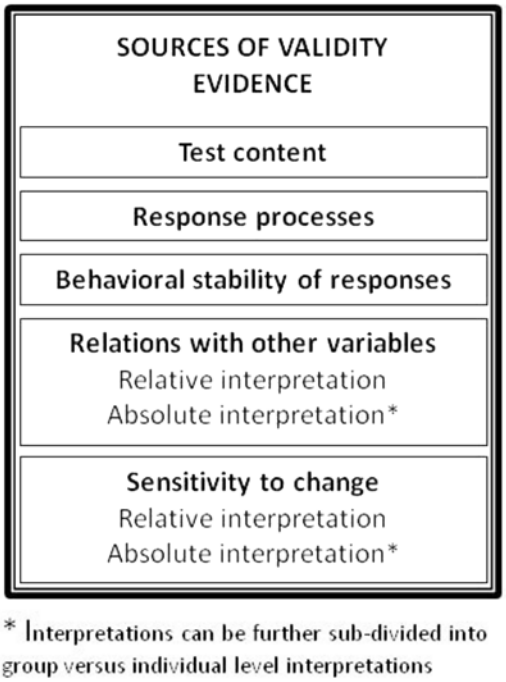

Figure 1 — Sources of validity evidence.

Evidence Based on Response Processes. This source of validity evidence examines the extent to which the responses provided by the respondents demonstrate that they understood what they were asked to recall.[5,6] Examples for this source validity of evidence may include examining whether; different groups interpret the questions similarly, respondents have the cognitive ability to recall the information, the format and wording elicit appropriate responses, and whether respondents are not providing socially desirable answers. Besides observation and feedback, cognitive interview methods[7] are often used to assess this source of validity evidence although a number of other qualitative and quantitative methods can be used.[5,6]

Evidence Based on the Behavioral Stability of Responses. This source of validity evidence is most often referred to as assessing the consistency of responses over time.[8] While many refer to this evidence as assessing the reliability of a measure; it is important to distinguish between reliability and behavioral stability. Basic

definitions of reliability emphasize the extent to which a measure is free of measurement errors (ie, systematic and random errors).[8] In contrast, behavioral stability assesses whether the behaviors recalled by a PA measure are stable over time (eg, 1 week, 2 weeks, or "x" weeks apart). While reliability is a prerequisite for validity,[8] assessing the reliability of a measure may not be enough to validate the assumption of behavioral stability. For example, readministering a measure on the same day to recall the same period of activities as the first administration can be used to assess the reliability of a measure, but will provide no information about behavioral stability as the recall period is identical. The concept of behavioral stability is particularly important for surveillance studies. Assessing whether the behaviors evaluated, for a given population, are stable over time is essential to support the utility of a measure when comparing levels of PA across reporting periods.

Evidence Based on Relations With Other Variables. This source of validity evidence is most often reported in PA validation studies. It consists of examining the relationship between the self-report measures of PA and an instrument that assesses the same construct or a construct associated with PA (eg, minutes of activities, fitness level, accelerometer counts, pedometer steps, and energy expenditure).[5,6] In some cases, group comparisons are employed to demonstrate this source of validity evidence, where the focus is to validate expected group differences.[6] The evidence has a *relative interpretation* when the ranking obtained with the self-report measure of PA is associated with the validation criterion. For example, if the self-report measure of PA was found to be correlated with accelerometer counts, one can infer that scores obtained with the measure provide valid rankings. The evidence has an *absolute interpretation* if the scores obtained with the self-report measure agree with the validation criterion (ie, it does *not* overestimate or underestimate the amount of PA assessed in comparison with the criterion). This source of validity evidence can only be demonstrated if the self-report measure assesses the behavior using the same units as the validation criterion and if the statistical method employed assesses agreement between measures. In the field of PA, the Bland and Altman method[9] has often been employed for assessing agreement although more advanced methods have been developed (eg, measurement error methods developed in the nutrition area).[10]

Evidence Based on Sensitivity to Change. This source of validity evidence aims to test whether the self-report measure is able to detect change in PA behavior.[5,6] This validity evidence is most often tested with an experimental design that manipulates the behaviors of interest.[6] Similar to the previous source of validity evidence, the evidence has a *relative interpretation* if the change detected with the self-report measure is related to the change found with the validation criterion. It becomes an *absolute interpretation* if the amount of change detected with the self-report measure agrees with the amount of change detected with the validation criterion.

Sources of Validity Evidence for Common Applications in PA

Table 1 highlights which sources of validity evidence are needed for epidemiological, surveillance, and intervention studies commonly conducted in the field of PA. The examples in Table 1 illustrate the need to understand which interpretations are supported by the validity evidence to determine the suitability of a measure for certain applications. The sections that follow will focus on the sources of validity evidence needed, above and beyond evidence of test content and response processes, as these sources of evidence are relevant for all applications. In general, this validity evidence assesses the extent to which the content of a self-report measure represents the PA domains targeted by the measure. In addition, this validity evidence provides the foundational knowledge to determine for whom the measure can be used. Although reliability is not emphasized in Table 1, reliability is required for any interpretations to be valid.

Epidemiological Studies—Associations Between PA and Health Outcomes

The first example from Table 1 discusses the validity evidence needed to assess the association between PA and health outcomes (eg, cardiovascular risk factors) in the context of epidemiological studies. Beyond evidence of test content and response processes, validity evidence based on relations with other variables leading to relative interpretations are also required (see Table 1, example 1). In this example, the scores should, at least, have a relative interpretation, as producing valid ranks is a required property of the scores for associations to be examined. While associations can be examined with self-report measures that overestimate minutes of PA, such measures may attenuate or even mask associations. Under such circumstances, it would be best to employ a measure that allows scores to have absolute interpretation (ie, accurately assessing minutes of PA).

Surveillance Studies—Population Level of PA

Table 1 presents 2 examples specific to surveillance studies (examples 2 and 3). Beyond evidence of test content and response processes, 2 other sources of validity evidence are listed with each example. Validating the behavioral stability of the responses is common to both examples (Table 1). Examining whether the PA levels are stable over time is relevant in many surveillance studies. It is particularly important to assess this validity evidence as many surveillance studies want to determine whether PA levels have increased or decreased for a given population. This is often achieved by comparing PA levels measured across reporting periods.

The major distinction between examples 2 and 3 is whether the evidence based on relations with other variables lead to relative or absolute interpretations. If the evidence has a relative interpretation, the minutes of

Table 1 Examples of Validity Evidence Needed for Studies Commonly Conducted in the Field of Physical Activity (PA)

Example 1: Epidemiological studies aimed at assessing association between PA and health outcomes	
Property of the measure	Produce valid rankings
Sources of scores[a]	• Evidence based on test content • Evidence based on response processes • Evidence based on relations with other variables providing a relative interpretation
Example 2: Surveillance aimed at monitoring levels of PA for a given population over time	
Property of the measure	Produce stable estimates of PA at the population level, although the estimate may be biased (ie, consistently overestimate minutes of PA)
Sources of validity evidence[a]	• Evidence based on test content • Evidence based on response processes • Evidence based on the behavioral stability of the responses • Evidence based on relations with other variables providing a relative interpretation
Example 3: Surveillance studies aimed at assessing the percentage of the population who meet current PA guidelines.	
Property of scores	Produce stable estimates of PA at the population level that are free of bias (ie, accurately estimate the minutes of PA)
Sources of validity evidence[a]	• Evidence based on test content • Evidence based on response processes • Evidence based on the behavioral stability of the responses • Evidence based on relations with other variables providing an absolute interpretation
Example 4: Intervention studies designed to assess relative changes in PA behavior.	
Property of scores	Detect relative change in PA behavior
Sources of validity evidence[a]	• Evidence based on test content • Evidence based on response processes • Evidence based on the stability of the responses[b] • Evidence based on relations with other variables providing a relative interpretation[b] • Evidence based on sensitivity to change providing a relative interpretation
Example 5: Intervention studies aimed at quantifying the magnitude of change in PA	
Property of scores	Quantify change in PA behavior
Sources of validity evidence[a]	• Evidence based on test content • Evidence based on response processes • Evidence based on the stability of the responses[b] • Evidence based on relations with other variables providing an absolute interpretation[b] • Evidence based on sensitivity to change providing an absolute interpretation

[a] While reliability is not emphasized in this table, it is assumed to be a prerequisite for valid interpretations.

[b] The extent to which such evidence is required will depend on whether it supports the sensitivity to change evidence.

PA estimated with the self-report measure are considered to be biased. Specifically, the self-report measure of PA may overestimate the amount of time the population is physically active. However, if the bias or the overestimation is consistent over time, it may be possible to use the self-report measures of PA in surveillance studies aimed at monitoring levels of PA for a given population over time (example 2 in Table 1). In this application, comparing PA behavior over time allows one to estimate the relative change in behavior. With regards to example 3, it is not appropriate to use a self-report measure that produces biased estimates of PA behavior (ie, overestimating minutes of PA performed). If the goal of the surveillance study is to determine what percentage of the population meets current PA guidelines, the self-report measure needs to produce an accurate estimate of PA behavior meaning that the scores need to have absolute interpretations (Table 1).

Intervention Studies—Detecting Change in PA Behavior

For intervention studies, it is essential to validate whether the self-report measure can detect change in PA behavior. The major distinction between examples 4 and 5 (see Table 1) is whether the change scores should have relative or absolute interpretations. If the validation study found the change in PA obtained with the self-report measure was associated with change in fitness level (for example), the scores would have a relative interpretation (example 4). If the magnitude of change in behavior is important to quantify (example 5), it becomes important to employ a criterion that assesses the magnitude of change using the same unit as the self-report measure of PA. The focus of this validation is to assess agreement between change scores. Evidence based on the behavioral stability and relations with other variables are also listed as important; however, the extent to which they are needed will depend on whether they support the utility of the measure for detecting change in PA.

Other Methodological Considerations

Identifying what interpretations are sought with self-report measures of PA will determine what sources of validity evidence should be emphasized. While this paper has focused on linking the sources of validity evidence with specific interpretations, the strength of the proposed interpretation will depend on 1) the evidence obtained, 2) the assumptions made to demonstrate the evidence, 3) the validation criteria used, 4) the methodological rigor of the validation study, and 5) whether interpretations have been replicated across studies and study populations. The reader should consult published evaluation checklists for validation studies to gain a broader understanding of how these issues affect the validity of interpretations.[2-4]

Selection of an appropriate validation criterion has validity implications that are unique to PA. As there are no agreed upon criteria to validate self-report measures of PA numerous methods have been employed, including, activity diaries, accelerometers, the Doubly Labeled Water methodology, and fitness tests.[11] While accelerometers are often regarded as the best method to objectively assess PA; they are conceptually not measuring the same outcomes as self-report measures of PA. As a result, our ability to validate the minutes of PA obtained with self-report measures is hampered by the lack of an appropriate validation criterion that assesses PA on the same units as self-report. This significantly limits our ability to validate measures for absolute interpretations.

Sources of Validity Evidence Examined—Case Review

To illustrate the concepts presented thus far, the validity evidence associated with the 7-d PAR was reviewed. The 7-d PAR was selected for illustrative purposes only, as other self-report measures could have been reviewed. The 7-d PAR is an interviewer administered self-report measure that asks participants to recall PA in the past 7 days.[12] The 7-d PAR was initially developed to recall leisure and occupational PA that are of moderate, hard, or very hard intensities.[12]

Validation studies were identified by conducting Medline searches as well as searching the Pittsburgh PA assessment website (www.parcph.org/subjPrimSrchRes.aspx). This review included all studies published in English, validation studies conducted in adults, studies that included an objective validation criterion or were validated against an activity diary, and studies that were published before February 2011. In total, 33 published papers have examined the validity and / or reliability of the 7-d PAR. Table 2 summarizes the validity evidence evaluated in these papers with further details provided in the Appendix. Validity evidence presented in previous reviews of self-report measures of PA was not included in the Appendix (ie, evidence based on behavioral stability and relations with other variables leading to relative interpretations).[1,13,14]

Evidence Based on Test Content and Response Processes

As observed in Table 2, none of the published studies provide validity evidence based on test content and response processes. Although it is tempting to conclude the evidence was not assessed, it is possible that such information was not reported.

Evidence Based on the Stability of Responses

Ten studies have examined whether the PA behaviors assessed with the 7-d PAR are stable over time and 1 study has examined the reliability of responses (Table 2). As reported in previous reviews, the behavioral stability of the 7-d PAR varied greatly with correlations ranging from .17 to .93.[1] Overall, there was no consistency as to which index from the 7-d PAR was reported and the studies varied greatly in the recall period reported, making it difficult to generalize the findings.

Evidence Based on the Relations With Other Variables—Relative Interpretation

Most of the validation studies have examined whether the scores obtained with the 7-d PAR are associated with a given criterion (Table 2). Accelerometers have been used most often as the validation criterion. From previous reviews, self-report measures of PA have been significantly associated with objective measures of PA and / or measures of fitness. However, interviewer-administered measures have higher associations (eg, correlations of .50 or higher) than self-administered measures (eg, correlations around .30).[13] Overall, there is a great deal of evidence that many self-report measures

Table 2 Review the Sources of Validity Evidence Commonly Used to Assess the Properties of Self-Report PA Measures—Case Analysis of the 7-Day Physical Activity Recall (7-d PAR)

Evidence based on	Criterion / description of evidence	Studies	Analytical methods
Test content		None	
Response processes		None	
Stability of responses			
Behavioral stability	Readministered 1 week, 2 weeks, or 1 month apart	24,27,29-35	Pearson correlation & Intraclass correlation
Test retest reliability	Recalling same days	26	Pearson correlation
Relations with other variables			Pearson correlation, Spearman correlation, Intraclass correlation, & Kendall tau correlation
Relative interpretation	Doubly Labeled Water (DLW)	17-19,21,22,36	
	Accelerometer	15,19,24-27,30-33,37-42	
	Pedometer	19,37,43	
	Activity diary	23,25,32-34,39,44	
	Energy intake	12,45	
	Cardiorespiratory test	18,27,28,32-34,44	
Absolute interpretation (group level)	DLW	17,18,19,21,22	Anova, *t* test, Chi-square test, & Cohen's kappa
	Vitalog (combined heart rate and motion sensor device)	23	
	Accelerometer	15,16,20,24	
Absolute interpretation (individual level)	DLW	17,18,19,21,22	Bland & Altman plots
	Accelerometer	15,16,20,24,26	
Sensitivity to change			
Relative interpretation	Change in behavior following an intervention—no criterion	27	Paired *t* test
	Change in $V0_2$ max, body fat, & treadmill time	12,28	Spearman correlation
Absolute interpretation (group level)	Change estimated from DLW	17	*t* test
Absolute interpretation (individual level)	Change estimated from DLW	17	Bland & Altman plots

of PA, including the 7-d PAR, can provide valid rankings of PA behavior. [1,13,14]

Evidence Based on the Relations With Other Variables—Absolute Interpretation

As can be seen in Table 2, 11 papers have examined the ability of the 7-d PAR to accurately estimate total and PA energy expenditures as well as minutes of moderate to vigorous PA.[15-25] Only 3 of the 11 studies include more than 25 subjects (Appendix).[15,16,22] One of these studies suggest that the 7-d PAR provides a valid estimate of total or PA energy expenditures at the group level.[22] The other 2 studies examined whether the 7-d PAR could validly classify subjects as meeting current PA guidelines and failed to provide such support.[15,16] In addition,10 papers demonstrated that the 7-d PAR could not accurately estimate levels of PA at the individual level.[15-22,24,26]

Evidence Based on the Sensitivity to Change

As shown in Table 2, the ability of the 7-d PAR to detect change in PA behavior has received little attention.[12,17,27,28] Four studies provide some evidence that the 7-d PAR can detect group level change in behavior following a PA intervention; however, these studies did not examine whether the magnitude of the change could be validly quantified (Appendix). As a result, the source of evidence is limited to a relative interpretation. Only 1 study examined whether the magnitude of the change could be validly quantified with the 7-d PAR, unfortunately that study was underpowered (n = 14) to evaluate such evidence.[17]

Highlights From the Case Review

While findings from the case review cannot be generalized to all self-report measures of PA, this section will review which interpretations are supported by the 7-d PAR. In addition, this section will also highlight the extent to which the validity evidence supports the use of the 7-d PAR in studies commonly conducted in PA (ie, those presented in Table 1). Note that the rigor of the methods employed in these validation studies were not examined in the case review. As a result, the interpretations provided below should be evaluated with this in mind. Published evaluation checklists can be used to fully evaluate the rigor of the methods and the extent to which it influences the validity of the interpretation presented below.[2-4]

Based on the review, the scores obtained with the 7-d PAR can be validly ranked to identify those who are more or less physically active—an interpretation supported by previous reviews of all self-report measures of PA.[1,13,14] In addition, there is some initial evidence that the scores obtained with the 7-d PAR are sensitive enough to detect change in PA behavior. However, the magnitude of the change cannot be validly quantified as there is no evidence supporting that minutes are accurately reported. As only 3 studies support the latter interpretation, further replications are needed. With this in mind, the 7-d PAR appears to have some validity evidence supporting its use in epidemiological studies as described in example 1 (see Table 1). In addition, initial validity evidence exists (albeit not enough) for its use in intervention studies as described in example 4 (see Table 1). The utility of the 7-d PAR in these applications will be valid, only, if the study is focused on measuring the domains of PA targeted by the 7-d PAR (ie, leisure and occupational PA that are of moderate, hard, or very hard intensities).[12]

Although the validity evidence was amassed primarily in adults, it remains unclear for which populations these interpretations are valid. The review indicated a complete lack of attention to providing evidence based on test content and response processes. This validity evidence evaluates the appropriateness of the content from a construct perspective. In addition, this evidence assesses whether the targeted population: understands the terminology employed; recalls the information based on their cognitive abilities; interprets the questions as intended and similarly across subgroups; and whether examples of activities (if provided) are appropriate for the targeted population. A lack of attention to this evidence makes it difficult to detail the specific characteristics of the population (eg, age, gender, race/ethnicity, and cognitive abilities) for which the 7-d PAR interpretations are valid.

While it was never the original intent of the 7-d PAR to be used in surveillance studies, it is important to highlight why the current evidence would not support its use in such studies. As previously mentioned, evidence of behavioral stability needs to be provided to demonstrate the utility of a measure for surveillance studies. Specifically, such evidence would examine whether the PA level estimated with the self-report measure is stable over time. With the 7-d PAR, the behavioral stability evidence was exclusively examined with correlational methods (Table 2). Unfortunately, correlational methods cannot test whether there is a significant shift in PA behavior over time. In fact, the correlation can still be high even if there is a shift in PA behavior. As a result, there is no evidence supporting the use of the 7-d PAR in surveillance studies, as the "behavioral stability" assumption was never validated. In the surveillance content, measurement error methods have improved the utility of self-report measures of dietary intake.[10] The extent to which such methods may improve the utility of self-report measures of PA for surveillance studies remains unknown.

Summary

This paper highlighted 5 sources of validity evidence that can be assessed to validate certain interpretations of self-report measures of PA. In addition, this paper discussed which sources of validity evidence are needed to support the utility of self-report measures of PA in epidemiological, surveillance, and intervention studies

commonly used in PA. A case review of the 7-d PAR suggested that the scores can be validly ranked. This supports the utility of the 7-d PAR for some epidemiological studies (eg, those focused on examining associations with health outcomes). Initial support was found for using the 7-d PAR to detect relative change in PA behavior; however more studies are needed to support its use in intervention studies. Finally, the review stressed that it is difficult to obtain valid estimates of behavior as the minutes of PA obtained with self-report measures are biased (eg, overestimating or underestimating PA). Striving for absolute interpretations (ie, unbiased estimate of PA) with self-report measures of PA continues to be a challenge for our field.

Acknowledgments

The authors would like acknowledge the contribution of Dr. Jeffrey I. Toward to this paper. Dr. Toward provided valuable editorial comments to this manuscript. Dr. Louise C. Mâsse received salary support from the Michael Smith Foundation for Health Research (Senior scholar award) and the Child, the Family Research Institute (Level 2 investigator scientist award), and the Sunny Hill Foundation to support this work. In addition, Dr. Mâsse received support from the National Cancer Institute at the National Institutes of Health to prepare this paper. Dr. Judith E. de Niet received a post-doctoral scholarship from the Child and Family Research Institute.

References

1. Shephard RJ. Limits to the measurement of habitual physical activity by questionnaires. *Br J Sports Med*. 2003;37(3):197–206.
2. Scientific Advisory Committee of the Medical Outcomes Trust. Assessing health status and quality-of-life instruments: attributes and review criteria. *Qual Life Res*. 2002;11:193–205.
3. Mokkink LB, Terwee CB, Knol DL, et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. *BMC Med Res Methodol*. 2010;10:22.
4. Hagstromer M, Ainsworth BE, Kwak L, Bowles HR. A Checklist For Evaluating the Methodological Quality of Validation Studies on Self-Report Instruments for Physical Activity and Sedentary Behavior. *J Phys Act Health*. 2012;9(Supp 1):S29–S36.
5. American Education Research Association, American Psychological Association, National Council on Measurement. *Standards for education and psychological testing*. Washington, DC: American Education Research Association; 1999.
6. Goodwin LD, Leech NL. The meaning of validity in the new standards for educational and psychological testing: implications for measurement courses. *Meas Eval Couns Dev*. 2003;36:181–191.
7. Willis GB. *Cognitive interviewing: a tool for improving questionnaire design*. Thousand Oaks: Sage Publications; 2005.
8. Streiner DL, Norman GR. *Health measurement scales: a practical guide to their development and use*. 4th ed. Oxford: Oxford University Press; 2008.
9. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;1(8476):307–310.
10. Dodd KW, Guenther PM, Freedman LS, et al. Statistical methods for estimating usual intake of nutrients and foods: a review of the theory. *J Am Diet Assoc*. 2006;106(10):1640–1650.
11. Montoye HJ, Kemper HC, Saris WH, Washburn RA. *Measuring physical activity and energy expenditure*. Champaign, IL: Human Kinetics; 1996.
12. Blair SN, Haskell WL, Ho P, et al. Assessment of habitual physical activity by a seven-day recall in a community survey and controlled experiments. *Am J Epidemiol*. 1985;122(5):794–804.
13. Sallis JF, Saelens BE. Assessment of physical activity by self-report: status, limitations, and future directions. *Res Q Exerc Sport*. 2000;71(2, Suppl):S1–S14.
14. Westerterp KR. Assessment of physical activity: a critical appraisal. *Eur J Appl Physiol*. 2009;105(6):823–828.
15. Johnson-Kozlow M, Rock CL, Gilpin EA, Hollenbach KA, Pierce JP. Validation of the WHI brief physical activity questionnaire among women diagnosed with breast cancer. *Am J Health Behav*. 2007;31(2):193–202.
16. Johnson-Kozlow M, Sallis JF, Gilpin EA, Rock CL, Pierce JP. Comparative validation of the IPAQ and the 7-Day PAR among women diagnosed with breast cancer. *Int J Behav Nutr Phys Act*. 2006;3:7.
17. Racette SB, Schoeller DA, Kushner RF. Comparison of heart rate and physical activity recall with doubly labeled water in obese women. *Med Sci Sports Exerc*. 1995;27(1):126–133.
18. Bonnefoy M, Normand S, Pachiaudi C, Lacour JR, Laville M, Kostka T. Simultaneous validation of ten physical activity questionnaires in older men: a doubly labeled water study. *J Am Geriatr Soc*. 2001;49(1):28–35.
19. Leenders NY, Sherman WM, Nagaraja HN, Kien CL. Evaluation of methods to assess physical activity in free-living conditions. *Med Sci Sports Exerc*. 2001;33(7):1233–1240.
20. Leenders NYJM, Sherman WM, Nagaraja HN. Comparisons of four methods of estimating physical activity in adult women. *Med Sci Sports Exerc*. 2000;32(7):1320–1326.
21. Conway JM, Seale JL, Jacobs DR, Jr, Irwin ML, Ainsworth BE. Comparison of energy expenditure estimates from doubly labeled water, a physical activity questionnaire, and physical activity records. *Am J Clin Nutr*. 2002;75(3):519–525.
22. Washburn RA, Jacobsen DJ, Sonko BJ, Hill JO, Donnelly JE. The validity of the Stanford Seven-Day Physical Activity Recall in young adults. *Med Sci Sports Exerc*. 2003;35(8):1374–1380.
23. Taylor CB, Coffey T, Berra K, Iaffaldano R, Casey K, Haskell WL. Seven-day activity and self-report compared to a direct measure of physical activity. *Am J Epidemiol*. 1984;120(6):818–824.
24. Soundy A, Taylor A, Faulkner G, Rowlands A. Psychometric properties of the 7-Day Physical Activity Recall questionnaire in individuals with severe mental illness. *Arch Psychiatr Nurs*. 2007;21(6):309–316.
25. Poudevigne MS, O'Connor PJ. Physical activity and mood during pregnancy. *Med Sci Sports Exerc*. 2005;37(8):1374–1380.
26. Hayden-Wade HA, Coleman KJ, Sallis JF, Armstrong C. Validation of the telephone and in-person interview

versions of the 7-day PAR. *Med Sci Sports Exerc*. 2003;35(5):801–809.
27. Dubbert PM, Vander Weg MW, Kirchner KA, Shaw B. Evaluation of the 7-Day Physical Activity Recall in urban and rural men. *Med Sci Sports Exerc*. 2004;36(9):1646–1654.
28. Young DR, Jee SH, Appel LJ. A comparison of the Yale Physical Activity Survey with other physical activity measures. *Med Sci Sports Exerc*. 2001;33(6):955–961.
29. Sallis JF, Haskell WL, Wood PD, et al. Physical activity assessment methodology in the Five-City Project. *Am J Epidemiol*. 1985;121(1):91–106.
30. Williams E, Klesges RC, Hanson CL, Eck LH. A prospective study of the reliability and convergent validity of three physical activity measures in a field research trial. *J Clin Epidemiol*. 1989;42(12):1161–1170.
31. Rauh MJ, Hovell MF, Hofstetter CR, Sallis JF, Gleghorn A. Reliability and validity of self-reported physical activity in Latinos. *Int J Epidemiol*. 1992;21(5):966–971.
32. Jacobs DR, Jr, Ainsworth BE, Hartman TJ, Leon AS. A simultaneous evaluation of 10 commonly used physical activity questionnaires. *Med Sci Sports Exerc*. 1993;25(1):81–91.
33. Richardson MT, Ainsworth BE, Jacobs DR, Leon AS. Validation of the Stanford 7-Day Recall to assess habitual physical activity. *Ann Epidemiol*. 2001;11(2):145–153.
34. Dishman RK, Steinhardt M. Reliability and concurrent validity for a 7-d re-call of physical activity in college students. *Med Sci Sports Exerc*. 1988;20(1):14–25.
35. Philippaerts RM, Lefevre J. Reliability and validity of three physical activity questionnaires in Flemish males. *Am J Epidemiol*. 1998;147(10):982–990.
36. Irwin ML, Ainsworth BE, Conway JM. Estimation of energy expenditure from physical activity measures: determinants of accuracy. *Obes Res*. 2001;9(9):517–525.
37. Motl RW, McAuley E, Snook EM, Scott JA. Validity of physical activity measures in ambulatory individuals with multiple sclerosis. *Disabil Rehabil*. 2006;28(18):1151–1156.
38. Johansen KL, Painter P, Kent-Braun JA, et al. Validation of questionnaires to estimate physical activity and functioning in end-stage renal disease. *Kidney Int*. 2001;59(3):1121–1127.
39. Miller DJ, Freedson PS, Kline GM. Comparison of activity levels using the Caltrac accelerometer and five questionnaires. *Med Sci Sports Exerc*. 1994;26(3):376–382.
40. Matthews CE, Freedson PS. Field trial of a three-dimensional activity monitor: comparison with self report. *Med Sci Sports Exerc*. 1995;27(7):1071–1078.
41. Hale LA, Pal J, Becker I. Measuring free-living physical activity in adults with and without neurologic dysfunction with a triaxial accelerometer. *Arch Phys Med Rehabil*. 2008;89(9):1765–1771.
42. Liu K, O'Brien E, Guralnik JM, et al. Measuring physical activity in peripheral arterial disease: a comparison of two physical activity questionnaires with an accelerometer. *Angiology*. 2000;51(2):91–100.
43. Wilkinson S, Huang CM, Walker LO, Sterling BS, Kim M. Physical activity in low-income postpartum women. *J Nurs Scholarsh*. 2004;36(2):109–114.
44. Sallis JF, Patterson TL, Buono MJ, Nader PR. Relation of cardiovascular fitness and physical activity to cardiovascular disease risk factors in children and adults. *Am J Epidemiol*. 1988;127(5):933–941.
45. Albanes D, Conway JM, Taylor PR, Moe PW, Judd J. Validation and comparison of eight physical activity questionnaires. *Epidemiology*. 1990;1(1):65–71.

Appendix

Summary of Validity Evidence Based on Relations With Other Variables and Sensitivity to Change—Case Review of the 7-Day Physical Activity Recall (7-d PAR)

Criterion	Study	Population	Evidence validated	Evidence supported[a]
Evidence based on relations with other variables—absolute interpretation (group level)				
Doubly Labeled Water (DLW)	[17]	14 obese women attending a 12-week weight reduction program	Total Energy Expenditure (TEE) (not weight adjusted)	Not supported
			TEE (weight adjusted)	Supported
			Physical activity energy expenditure (PAEE)	Partially supported
	[18]	19 healthy older men (Mean age = 73.4)	TEE	Supported
			PAEE	Supported
	[19]	13 healthy women (Mean age = 25.8; Mean Body Mass Index (BMI) = 23.5)	PAEE	Supported
	[21]	24 healthy men (Mean age = 41; Mean BMI = 25.1) attending a dietary study	TEE estimated with the DLW method	Not supported
			Physical Activity (PA) ratio (TEE/Basal Metabolic Rate)	Not supported
	[22]	46 overweight / obese adults [17 men (Mean age = 23.9) 29 women (Mean age = 23.3)]	TEE	Supported
			PAEE	Supported
Vitalog (heart rate and motion sensor)	[23]	23 men (age 34 to 64) 1/2 attended a cardiac rehabilitation program	TEE	Supported
TriTrac & CSA accelerometer & Yamax pedometer	[20]	12 healthy women (Mean age = 26.0)	PAEE	Not supported
Accelerometer Actigraph 7164	[16]	159 breast cancer women (Mean age = 56.6; Mean BMI = 26.9; 91% White)	Minutes / week of PA	Supported
Accelerometer Actigraph 7164	[15]	74 breast cancer women (Mean age = 55.0; Mean BMI = 26.9; 82% White)	Minutes / week of PA	Partially supported

(continued)

Appendix *(continued)*

Criterion	Study	Population	Evidence validated	Evidence supported[a]
RT3 accelerometer	[24]	10 men and 4 women with severe mental illness (Mean age = 52.9; Mean BMI = 29.5; 100% White)	TEE	Not supported
Activity diary	[25]	12 pregnant women (Mean age = 29.8) & 12 control women (Mean age = 30.7)	PAEE	Supported
Evidence based on relations with other variables—absolute interpretation (group level)				
Accelerometer Actigraph 7164	[16]	159 breast cancer women (Mean age = 56.6; Mean BMI = 26.9; 91% White)	Agreement with PA guidelines	Not supported
Accelerometer Actigraph 7164	[15]	74 breast cancer women (Mean age = 55.0; Mean BMI = 26.9; 82% White)	Agreement with PA guidelines	Not supported
Evidence based on relations with other variables—absolute interpretation (individual level)				
DLW	[17]	14 obese women attending a 12-week weight reduction program	TEE	Not supported
	[18]	19 healthy older men (Mean age = 73.4)	TEE & PAEE	Not supported
	[19]	13 healthy women (Mean age = 25.8; Mean BMI = 23.5)	PAEE	Not supported
	[21]	24 healthy men (Mean age = 41; Mean BMI = 25.1) attending a dietary study	TEE	Not supported
	[22]	46 overweight / obese adults [17 men (Mean age = 23.9) 29 women (Mean age = 23.3)]	TEE & PAEE	Not supported
TriTrac accelerometer	[26]	47 healthy women (Mean age = 31.9) & 27 healthy men (Mean age = 37.2)	Minutes / week of PA	Not supported
TriTrac & CSA accelerometer & Yamax pedometer	[20]	12 healthy women (Mean age = 26.0)	PAEE	Not supported
Accelerometer Actigraph 7164	[16]	159 breast cancer women (Mean age = 56.6; Mean BMI = 26.9; 91% White)	Minutes / week of PA	Not supported
Accelerometer Actigraph 7164	[15]	74 breast cancer women (Mean age = 55.0; Mean BMI = 26.9; 82% White)	Minutes / week of PA	Not supported
RT3 accelerometer	[24]	10 men and 4 women with severe mental illness (Mean age = 52.9; Mean BMI = 29.5; 100% Caucasian)	TEE	Not supported

(continued)

Appendix *(continued)*

Criterion	Study	Population	Evidence validated	Evidence supported[a]
Sensitivity to change—relative interpretation				
No criterion	27	220 male veterans with chronic health conditions (Mean age = 68.5) 29.2% African-American	Change following an exercise intervention (no control group)	Supported
VO_2max, body fat, energy intake, & treadmill time	12	45 male intervention participants	Change in physiologic variables (VO_2max & body fat) & energy intake	Supported except for body fat
		1561 men and women participating in a worksite intervention	Change energy expenditure between intervention participants and controls	Supported
		117 school teachers (73% women)—participating in a worksite exercise intervention	Change energy expenditure between intervention participants and controls	Supported
$V0_2$ max (ergometer test)	28	59 participants (78% women) (Mean age = 66.5; 32 non African-American; 27 African-American)	Change in VO_2max following an intervention	Not supported
Sensitivity to change—absolute interpretation (group level)				
DLW	17	14 obese women attending a 12-week weight reduction program	Change in TEE following a weight reduction program	Supported
Sensitivity to change—absolute interpretation (individual level)				
DLW	17	14 obese women attending a 12-week weight reduction program	Change in TEE following a weight reduction program	Not supported

[a] Reporting whether the validity evidence is supported or not supported is based on the authors' conclusions. The validity of the evidence may in some cases be questionable based on the rigor of the study but this aspect was not accounted in this review—see text.