Reliability, Validity, and Methodological Response to the Assessment of Physical Activity Via Self-Report

Patricia Patterson

Keywords: validity, reliability, physical activity, questionnaires

I have been asked to respond to a paper by Dr. Jim Sallis and Dr. Brian Saelens that discusses the status, limitations, and future directions of the self-report of physical activity (PA). Their paper is limited to instruments used in the 1990’s with the following additional inclusion criteria: 1) an objective method was used to validate the self-report instrument (e.g., doubly labeled water, accelerometer, direct observation, HR monitor); and 2) the recall time frame was 1 year or less. Given these criteria, 17 instruments were located for youth, 7 for young and middle-aged adults, and 4 for older adults. The authors present clear in table form the test-retest reliability coefficients of these instruments, along with validity coefficients and descriptions of each instrument’s content validity. The term “content validity”, as defined by the authors, is “the extent to which a test measures the full range of characteristics of the construct” (Sallis & Saelens, 2000, p. XX). Additionally, Sallis and Saelens discuss areas meriting additional attention including the following issues: reliability, content validity, validation criteria, absolute versus relative validity, the development of self-report measures, and differences across cultures, ethnicity, and demographics.

The review by Sallis and Saelens (2000) indicates that both validity and reliability coefficients are only moderate, although there is great variation in the size of the coefficients reported in the literature. My response will provide some suggested solutions to the issues raised by Sallis and Saelens, although in the area of validity this task is rather daunting. In particular, I will discuss the measurement issues of reliability and validity from a broad perspective. This approach is being taken given my view of the importance of being aware of one’s paradigm beliefs and how these beliefs influence the way we frame research questions and view evidence.

Reliability

Reliability is the consistency of a response either across multiple trials within a single administration, generally called internal consistency, or across days, generally called test-retest or stability reliability. In the case of self-report of physical activity, the literature reviewed by Sallis and Saelens (2000) indicates that reliability has been reported in a test-retest manner.

Most current measurement texts (e.g., Baumgartner & Jackson, 1999; Sallis & Jackson, 2005) address reliability issues in depth and in a paper by Morrow and Jackson (1993) provides an excellent review of the salient issues. I would like to highlight areas that merit attention in reliability estimation and reporting for self-report of physical activity.

As Sallis and Saelens point out, reliability should be estimated via an intraclass correlation coefficient obtained from an analysis of variance (ANOVA), not with a Pearson product moment correlation (PPMC). Using a PPMC for reliability estimation is limited to only two trials or days, but more importantly its use is inappropriate for two reasons. First, use of PPMC assumes a bivariate situation when reliability is univariate. Second, reliability is concerned with the consistency of scores across trials or days and PPMC is not able to detect trends. Thus, there could be a 10-fold increase in self-report of physical activity from day 1 to day 2 but if the order remained the same, r would be equal to 1.0, indicative of perfect reliability. Yet reliability can only be perfect when there is no change in a person’s score from day to day. It is also important that researchers indicate whether a one-way
ANOVA or two-way ANOVA model was employed, because the error term is defined differently for these two models. In a one-way ANOVA, all sources of variation other than differences among people are considered error. In a two-way ANOVA a researcher can decide that trial to trial variation should not be considered part of the error term and use only the interaction term as a source of error. This latter decision might occur when one expects differences from trial to trial or day to day and does not wish these differences to be considered error.

Test-retest reliability also requires that the time interval between test administrations be specified. While this has generally been done in the PA self-report literature, the time interval participants have been asked to recall is not only very long but often does not cover the same time period. Thus, what is reported as test-retest reliability includes true variability in physical activity and treats it as measurement error. Sallis and Saelens recommend that test-retest studies be conducted across time intervals that match the intended instrument's recall of activity. For example, a 1-day recall would be restated on the same day and a 7-day recall would be repeated within 7 days and include overlapping days. The overlapping days would provide the best estimate of reliability devoid of changes in physical activity behavior.

I would argue that in many of the self-report studies of PA test-retest reliability has not been examined in an appropriate way. Any real examination of test-retest reliability must have a very limited time interval, generally from one to three days, but not more than one week. To avoid the issue of true variation in PA being confounded into reliability estimation, one must design a study that asks participants to recall PA across a particular number of days, say the last 7 days. A couple of days later the participants are asked to recall PA across that same 7-day period. This design truly examines whether people can consistently recall PA. We might expect that PA will vary from day to day but asking people to recall the same time interval on more than one occasion allows a true examination of one's ability to recall PA on a consistent basis.

After reviewing many of the studies cited by Sallis and Saelens, I would like to address several additional problems with reliability estimation and reporting in physical activity research. These include the following: a) common use of significance testing when reporting reliability; b) lack of confidence interval reporting; c) lack of testing for significance across days in a test-retest format; d) lack of internal consistency reliability estimation of the instruments; and e) lack of reliability estimation for subgroups.

Significance Testing.

Most studies report probability values along with reliability coefficients. Significance testing is not relevant in reliability estimation because we are only interested in the actual value of the coefficient, not whether it is significantly different from zero.

Confidence intervals (CI).

Sample sizes vary tremendously across studies; thus, the precision of the reliability estimate will also vary, particularly in studies employing small samples. Use of 95% CI provide the reader with a better context in which to place the evidence for reliability.

Test-retest reliability.

This approach to reliability estimation is relevant when a self-report instrument is used as a measure of physical activity in an intervention study. Thus, high reliability is necessary in order to ensure that results are due to the intervention, not because of variation in scores due to low reliability. However, the reported reliability coefficient is the average of the number of days being evaluated. That is, if a self-report instrument is administered on two occasions, the reliability coefficient is the average of the two trials. The estimate that should be reported is the single trial reliability since that is how self-report instruments are generally used. It is well known that reliability is lower for shorter tests than longer tests; thus, the reliability coefficients reported in the literature are overestimates of single trial reliability. While it is possible that some researchers may be reporting single trial reliability, this should be explicitly stated in published literature to clarify this issue.

Because we are interested in the stability of physical activity recall in a test-retest situation, it is necessary to examine whether significant differences exist between the two recalls. That is, it is possible to have a reasonable level of reliability but still find significant differences between day 1 and day 2 reports. Thus, researchers should report these findings as well as the test-retest reliability coefficient.

Internal Consistency.

The literature reviewed also demonstrates the lack of information on the internal consistency of self-report instruments. That is, we need to know the reliability of the items within the instrument when given on a single day. Coefficient alpha should be employed for this purpose. This should be standard practice in instrument development because it allows an examination of the reliability of various components of physical activity like frequency, intensity, and duration as well as the reliability of the entire instrument. This information is essential to understanding which aspects of self-report of PA are more (or less) difficult to recall and can lead to instrument refinement.
Physical Activity Across Groups.

One final problem is the lack of reporting reliability across diverse groups. If one expects that there would be differences in PA between groups (e.g., gender, age) it is clear that combining these groups for reliability estimation will increase the range of PA scores, thus inflating reliability estimation. Reliability should be reported for appropriate subgroups when differences in PA are expected.

Validity

Most measurement specialists would agree that validity is the most important concept in measurement. Historically, validity has been categorized into three types: content, criterion-related, and construct (Cronbach & Meehl, 1955). More recently, however, validity is viewed as a unitary concept. In fact, measurement specialists (e.g., Messick, 1989; Pedhazur & Schmelkin, 1991; Shepard, 1993) would argue that construct validity is the only “type” of validity and that content-related and criterion-related validity are really only estimation strategies for construct validity rather than truly independent “types” of validity.

Validity is defined as “the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores” (Standards for Educational and Psychological Testing, American Psychological Association [APA], 1985, p. 9). This definition of validity makes it clear that one does not validate a test because scores can be used for multiple purposes and in varied contexts. For example, PA scores could be used to estimate caloric expenditure, predict relative risk of disease, or to classify individuals into activity levels for an exercise program. Giving this changing nature of test score use and the fact that it is the interpretations of the scores that are being validated rather than the test itself, validity must be viewed as a dynamic and ongoing process. A researcher’s ability to understand the importance of construct definition and its changing nature due to changes in context, culture, and intended use is the key to successful validation efforts.

Sallis and Saelens speak to a number of validity issues in their paper. I would like to begin with the importance of detailed descriptions of test development. The quality of the methods used to develop the instrument provides the foundation for how successful we will be in gathering additional evidence for the soundness of the interpretations we wish to make.

A discussion of test development must begin with a detailed discussion of construct validity and methods that can be employed to provide evidence of validity. These approaches include gathering content-related and criterion-related evidence and employing both qualitative and quantitative methods.

Content-Related Evidence.

There are numerous published resources delineating the steps one should follow for providing content-related evidence of a construct during test development (e.g., Baumgartner & Jackson, 1999; Haynes, Richard, & Kubany, 1995; Lynn, 1986; Safford & Wood, 1995; Wood, 1989). However, there are far fewer examples of these principles being followed in a systematic way in the actual practice of developing PA self-report instruments. Sallis and Saelens briefly outline the procedures in their own paper. However, given the importance of this process, the steps necessary to gather content-related evidence will be described in greater detail.

The first and most important step in construct validation is to define the targeted construct, in this case, physical activity. According to Caspersen, Powell, and Christenson (1985) PA is “any bodily movement produced by skeletal muscles that results in energy expenditure.” Total energy expenditure includes basal metabolic rate, thermic effect of food, and physical activity, with the latter being the most variable. Physical activity is comprised of activities of daily living, occupational activities, and sports and leisure activities (Kriske & Caspersen, 1997). Sallis and Saelens are more specific in their view of PA and state that the dimensions of physical activity include type, frequency, intensity, and duration. They also state two other criteria. First, that the instrument should “produce estimates related to public health guidelines”. Second, that the instrument should have the “ability to obtain scores for various contexts or purposes of PA (i.e., leisure, work, household, transportation)”.

It is clear that what at first might seem an easy task is fraught with difficulties as one considers that constructs may have different meanings in different contexts and cultures. These definitions differ in specificity and content and there could be several other definitions of physical activity depending on the intended purpose. In fact, Sallis and Saelens discuss the lack of inclusion of musculo-skeletal and flexibility exercises in self-report instruments as one example of how the construct of PA might differ from context to context. The lack of a clear definition of PA and its dimensions, or the recognition that its definition might differ according to intended use, is one of the reasons that validation work is so difficult. Providing evidence of construct validity is both complex and time-consuming (Pedhazur & Schmelkin, 1991) and researchers must view test development as a process that “may take years... not days or weeks” (Bohrnstedt & Borgatta, 1981; p. 14).

The next step involves generating items that measure the defined dimensions. This process should be grounded in the literature and may also be improved through carefully structured interviews with both experts and intended target samples (Haynes et al., 1995). These methods make it more likely that the items generated will
be both representative and relevant to the dimensions of the construct.

After initial item generation has been completed and the format for the instrument has been decided, judges should be employed to examine the items for relevance, representativeness, clarity, and specificity. There is no agreement as to the optimal number of judges but at least five judges is recommended and it is unlikely that employing more than 10 will be helpful (Haynes et al., 1995; Lynn, 1986). It is important that this step include explicit instructions to the judges regarding what they should evaluate. This requires that a detailed description of the construct be provided along with a clear explanation of its dimensions. Additionally, information about the intended use of the scores generated from the instrument is critical. The quality of information gathered at this stage of test development is greatly improved by the use of an ordinal rating scale. Some specialists recommend a 5 - 7 point Likert scale (Haynes et al., 1995) while others (Lynn, 1986) recommend use of a 4-point scale. The latter choice eliminates the inclusion of ambivalent middle ratings. Each judge then rates each item on this scale after which an index of content validity (CVI) can be calculated. Using a 4-point scale, the CVI for each item is the proportion of judges giving the item a rating of 3 or 4 (Waltz & Bausell, 1981). The CVI for the entire instrument is the proportion of total items judged to be content valid. In addition to rating the relevance of items, judges should also be asked whether any dimensions of the construct have been omitted. These quantitative indices can be supplemented with feedback from the judges regarding rewording. A final consideration is evaluating the proportional representation of the items to ensure that they accurately reflect the weightings of the dimensions of the construct (Anastasi, 1988). These procedures increase the likelihood that a dimension is not over - or under-represented, a problem that would bias both the obtained scores and any inferences made from these scores. In the case of self-report of PA, lack of attention to these issues could result in inaccurate prevalence rates and/or inaccurate interpretation of relationships between PA and public health concerns. Revisions, deletions, or additions of items requires another round of rating until one is confident that the “elements of an assessment instrument are relevant to and representative of the targeted construct for a particular assessment purpose” (Haynes et al., 1995; p. 258).

At this point, factor analytic techniques should be employed to examine the underlying structure of the instrument. While exploratory factor analytic techniques are useful when hypotheses about factors cannot be explicitly stated, it is highly likely that most researchers in the area of PA have some idea of the number and relationship of factors being measured. In this case, structural equation modeling, and in particular, confirmatory factor analysis (CFA) would be more appropriate because it is a theory driven approach. CFA has been used to assess, develop, and modify theoretical models (Anderson & Gerbing, 1988) and its use has increased with the availability of computer programs like AMOS and EQS (Hoyle, 1995).

CFA requires stating the proposed model a priori, estimating the parameters, and then determining how well the model fits the data. It is important to note that finding a model that has adequate fit is no assurance that it is the best model or the only model. The goal is to find a model that fits the data and provides an interpretable solution. Thus, adjustments to poor fit or model re-specification should be based on theory and not simply statistics (Anderson & Gerbing, 1988).

As with all statistical procedures the assumptions on which CFA is based must be checked prior to examining for fit and interpreting the data. Moreover, an adequate sample size and an adequate number of indicators per variable is necessary to improve the accuracy of parameter estimation (see Anderson & Gerbing, 1988 for suggestions).

CFA could also be used to examine the invariance of the model of a self-report instrument across groups, such as age, gender, ethnicity, or culture. Procedures described by Joreskog and Sorbom (1989) provide the steps necessary to examine these issues. Essentially, tests of equality for the covariance structure, factor model, factor loadings, error variances, and factor covariances should be nonsignificant to demonstrate that the same model can be used across different samples. Thus, CFA could be used to address the concerns Sallis and Saelens raise regarding instrument proliferation and finding instruments that can be successfully used across various ethnic or cultural groups.

Additional support for construct validity can be determined with convergent and discriminant evidence (Campbell & Fiske, 1959) and examining the relationships between similar constructs and different constructs. This step is an essential part of construct validity efforts because of the importance of examining the instrument’s relationship to other external variables (Goodwin, 1999). In the case of self-report of PA, evidence must be provided that the instrument is related to similar constructs but not related to dissimilar constructs.

**Criterion-Related Evidence.**

Sallis and Saelens discussed criterion-related validity evidence when a study employed an objective measure as the “criterion” (e.g., doubly labeled water, heart rate monitors, accelerometers, and direct observation). However, I hesitate to describe the correlation coefficients obtained from these studies as criterion-related evidence. I think they should be better viewed as providing convergent evidence toward the construct of PA. That is, all of these measures, even those viewed as more objective than others, are still not likely to be true "crite-
tion” measures of PA. For example, emotion, fitness level, caffeine, etc can affect heart rate. Accelerometers have not been validated across all ages, ethnic groups, or with all kinds of activities. Finally, the use of doubly labeled water does not allow information about frequency or intensity. We need to continually search for measures that provide the best (most precise) information about PA. For those measures that incorporate technology, such as accelerometers or HR monitors, we need to work to improve their precision and the assurance that the information garnered from these tools is relevant to the definition of PA. Sallis and Saelens make this important point when they discuss the selection of measures and the necessity of matching purpose and intention. For example, using a physical fitness measure as a criterion for a self-report of PA that includes all levels of intensity would not be expected to yield a high validity coefficient. In the mean time, researchers working in this area need to recognize the strengths and weaknesses of these measures when interpreting validity coefficients. Finally, studies that report correlation coefficients between multiple measures of PA may provide a broader perspective.

Sallis and Saelens also discuss the importance of absolute versus relative validity and state that most work has been conducted with relative validity in mind. That is, studies have reported correlations between measures. However, given that prevalence rates of PA are often based on self-report, the actual, or absolute, amount of PA must be estimated. In order to examine the over- or under-estimation of PA by self-report the measurement units from the self-report instrument must be the same as those gathered by another measure. Thus, methods to compute mean minutes of PA in various categories or mean number of kilocalories expended can then be compared with repeated measures analysis of variance with these same measures from a HR monitor or an accelerometer. While I support Sallis and Saelens’ belief that research should include this information as part of validity evidence, I would caution both researchers and practitioners regarding the accuracy of these estimates from either measure. Perhaps there is a desire to obtain information that is simply too specific for the precision afforded by self-report measures. This issue clearly needs additional examination. Research in this area could lead to instrument refinement to improve the measurement of dimensions of PA that have higher estimation errors. Research on this issue may also lead to more accurate information on not only prevalence rates of PA but also the relationship between PA and health.

In summary, I would like to reiterate the importance of having a clear and well-conceptualized definition of the construct of PA as the premise for any research conducted in this area. Until that is done, measurement evidence of the reliability and validity of self-report instruments will be hampered by the lack of specificity. I do not mean to imply, however, that construct specificity will result in a single self-report instrument. The complex nature of PA and the diversity of contexts and uses for the scores make this scenario highly unlikely. Moreover, it is important to keep in mind that any construct is dynamic. Changes in the definition of the dimensions of PA and/or the intended use of the instrument require that the construct be re-examined and revised to reflect these changes. Sallis and Saelens allude to this when they describe changes in work and leisure environments and continuing evidence of relationships between PA and health. The importance of this feature of validity cannot be underscored enough. Finally, it is important that we all recognize the need for an array of methods and approaches that can assist us in better matching appropriate methods for our hypotheses. Becoming aware of our own paradigms and how they shape the questions we ask and the methods we select to answer these questions is critical to being open to new and different ways of inquiring. Harris (1975) underscores this theme very well:

"...the research question should dictate the appropriate statistical analysis rather than letting the ready availability of a statistical technique generate a search for research paradigms which fit the assumptions of the technique". (p. 3)

References


