Published in final edited form as:

Med Sci Sports Exerc. 2008 July ; 40(7 Suppl): S544–S549. doi:10.1249/MSS.0b013e31817c6651.

How Many Days Was That? We're Still Not Sure, But We're Asking the Question Better!

TOM BARANOWSKI¹, LOUISE C. MÂSSE², BRIAN RAGAN³, and GREG WELK⁴

¹ Baylor College of Medicine, Houston, TX

² University of British Columbia, Vancouver, BC, CANADA

³ University of Northern Iowa, Cedar Falls, IA

⁴ Iowa State University, Ames, IA

Abstract

Unreliable measures limit the ability to detect relationships with other variables. Day-to-day variability in measurement is a source of unreliability. Studies vary substantially in numbers of days needed to reliably assess physical activity. The required numbers of days has probably been underestimated due to violations of the assumption of compound symmetry in using the intraclass correlation. Collecting many days of data become unfeasible in real-world situations. The current dilemma could be solved by adopting distribution correction techniques from nutrition or gaining more information on the measurement model with generalizability studies. This would partition the variance into sources of error that could be minimized. More precise estimates of numbers of days to reliably assess physical activity will likely vary by purpose of the study, type of instrument, and characteristics of the sample. This work remains to be done.

Keywords

PHYSICAL ACTIVITY; MEASUREMENT; ASSESSMENT; VALIDITY; RELIABILITY

Physical activity is a complex behavior whose measurement is characterized by considerable inter- and intraindividual variability (14). Research aimed at understanding and promoting physical activity requires methods that can provide valid and reliable indicators of physical activity in an efficient and effective manner. Reliability provides an upper limit for validity coefficients.

Poor reliability limits the conclusions that can be drawn from any type of research. When dealing with continuous measures (e.g., average activity counts), lower reliability “attenuates” any correlation of this measure with other variables (19). That is, any correlation of this variable with other variables will be lower than the true correlation; and how much lower will be inversely proportional to the level of reliability (19). When dealing with a binary variable (e.g., meeting physical activity guideline criterion or not), unreliability leads to greater misclassification error (5), which can bias epidemiological studies using logistic regression or lead to inaccurate surveillance statistics.

Classical test theory states that any measure has true variability, systematic error, and random error. Reliability, which is concerned with the replicability of assessment, is most typically associated with random error. There are two main sources of random error that influence physical activity measurements: one reflects normal variability in human behavior (biological or behavioral error) and the other reflects analytical variability due to the way in which the data were collected, processed, or reported. Analytically unreliable measurement procedures need to be fixed or removed, whereas inherent behavioral variability needs to be carefully studied and overcome using statistical procedures.

This article will briefly review factors influencing time-related (usually day to day) reliability, identify limitations with existing statistical procedures, and describe methods that may help elucidate or overcome these limitations.

FACTORS INFLUENCING RELIABILITY

Researchers are generally interested in capturing habitual or usual levels of physical activity either because they want to test whether these are related to health-related variables or because they are trying to influence physical activity through intervention. Because days organize human experience, researchers are usually interested in habitual activity per day. Typically, they measure activity over multiple days and then compute an average across days to reflect habitual activity. In regard to reliability, they want to know whether the average from the number of days available equals the average across all the possible days (if one could possibly know that). Investigators typically want to maximize reliability while minimizing participant burden; so ascertaining the minimum numbers of days required for acceptable reliability has been the focus of physical activity research.

The number of days needed to get reliable data is dependent on variability that exists in the data. Highly reliable data tend to have lower variability or higher consistency within each individual (see Fig. 1) and more variability between individuals (e.g., lots of parallel lines). If variability for most individuals is high (i.e., consistency is low) and so the patterns for each person overlap with those for other people, then more days of monitoring are needed to capture the usual level of physical activity. A group within the high consistency group (see Fig. 1) with some differences between people could have their usual activity assessed with only 1 d of monitoring due to the exact equivalence of activity across days. A group of people in the medium consistency group might require 2 d; if there is some difference between people in the low consistency group, for example, people 1 and 2 would need many days due to the high overlap. People 1 and 3, however, would need fewer days due to the lack of overlap.

The amount of variability that exists across days is dependent in part on the sample and the time frame being assessed. In sedentary office workers, there may not be much variability in day-to-day activity. If each office worker was highly consistent (see Fig. 1) and they differed somewhat among themselves, high reliability could be attained if each office worker is highly consistent, but if they have the same level of activity, that is, they overlap, then low reliability will be obtained. In farmers or warehouse laborers whose activity is subject to weather and other influences and thus little consistency (see Fig. 2), when patterns overlap, more days may be needed.

The intraclass correlation (ICC) is the most common way of summarizing the consistency of measures across days. The theoretical equation for the ICC (13) is

$$ICC = \sigma_b^2 / [\sigma_b^2 + (\sigma_w^2)]$$

where σ_b^2 is the between-subject variance and σ_w^2 is the within-subject variance (intraindividual variability).

The within-person variability component is considered error variance. This generic equation is for a single-day ICC. The ICC is very sensitive to the numbers of days of assessment because the within-subject variance is divided by the numbers of days. The above formula can be rewritten as follows

$$ICC = \sigma_b^2 / [\sigma_b^2 + (\sigma_w^2/n)]$$

where n is the number of days of assessment.

As can be observed from the above formula, as the numbers of days of assessment increases, the within-subject term decreases. This allows us to understand the impact of number of days on the ICC.

SUMMARY OF RESEARCH ON RELIABILITY OF PHYSICAL ACTIVITY

Many studies examined the reliability of different physical activity indicators. The standard typically used to designate acceptable reliability is an ICC of 0.8; so many studies have sought to determine the number of days needed to obtain this degree of reliability. Number of days has generally been estimated from the ICC with the Spearman brown formula (19). Reliability estimates, with the associated numbers of days necessary, have varied substantially across different methods of data collection, including pedometers, HR monitors, accelerometers, observation, and self-report. This is a brief summary of this literature.

Pedometers

One study using pedometers with 90 middle-aged adults who provided 7 d of assessment from a sample of more than 375 participants obtained a single-day ICC of 0.72 (21). Any combination of 3 d obtained ICC that ranged from 0.86 to 0.91. Although ANOVA revealed that numbers of steps were significantly lower on Sundays than all other days of the week, they concluded that the difference was small and negligible (21).

HR monitoring

One study of 3- or 4-yr-old children revealed that approximately 4.3 d of HR monitoring was necessary to assess habitual activity using any of these indicators (mean HR, percent of minutes above 120 bpm) (6). When the same children became 5, 6, or 7 yr, the numbers of days increased to 8.4 d for mean HR but lowered to 4.0 d for numbers of minutes above 120 bpm (6). The increase for mean HR was surprising because younger children are thought to be more random in their movements, but perhaps mean HR was more sensitive to factors other than physical activity (e.g., the emotions of dealing with peers, teachers, and other elements in school). The reduction in days for percent of minutes above 120 bpm suggests that the reduced time to be physically active with the advent of school consistently led to fewer possible minutes with an elevated HR. Great variability in the numbers of days of assessment across indicators of activities (e.g., 1.9 d for percent of minutes that were 25% or more above resting HR; 6.2 d for percent of minutes that were 50% or more above resting HR) suggests that a careful selection of indicator needs to be made that best reflects the intent and the need of the proposed research.

Accelerometry

One study using Caltrac accelerometers with children revealed that 5 d of assessment were necessary to assess the frequency of vigorous activity with a reliability of 0.8 and 6 d to assess average movement counts and frequency of moderate activity (9). A study among adults graphically documented how ICC values asymptoted to 1.0 as the number of days of recording increased (12), and approximately 3 d of assessment were needed to attain a reliability of 0.8 among men for activity counts, about 4 d for minutes of moderate to vigorous activity, and 7 d for minutes of inactivity (12). Only slight differences were obtained among women (12).

Observations

Among 3- or 4-yr-old children, the number of days to attain a reliability of 0.8 varied by indicator from 15.0 to 18.7 d (7). This is the largest number of days reported and may reflect that the data were obtained over a 3-yr period, which may have encompassed different developmental stages with inherent differences in activity and likely included seasonality differences.

Self-report

Three weeks of a physical activity diary kept by elementary school teachers revealed that 2 wk of diary recording were necessary to attain a reliability of 0.8 (1). In a sample of Hispanic and African American women aged 45 yr or older, between 8 and 14 d were needed to get a reliability estimate of 0.8 using generalizability theory (G-theory) methods. Alternatively, a single-item indicator attained a single-day reliability of 0.85; however, the 3-d sweat recall was not able to achieve a reliability of 0.8 with six replicates, and the 3-d aerobic recall achieved a reliability of 0.8 after four replicates (9). Although the single-item PA rating achieved a high reliability with only one replicate, the validity of this scale was not analyzed or reported. These results demonstrate that the reliability estimate can vary based on the type of instrument and the indicator used.

LIMITATIONS OF CURRENT APPROACHES TO RELIABILITY

Assumptions underlying the ICC may not be appropriate for the data collected and may account for these differences. Conceptually, ICC methods assess the correlation “among measures of a common class, which means that they share both their metric and variance” (13). ICC estimated with ordinary least squares (OLS) methods within the general linear model assume compound symmetry, that is, the variances and the covariances among multiple days of measurement are the same. In other words, the correlations among days are assumed to be similar. Violating the compound symmetry assumption would threaten the validity of the reliability obtained as one of the ICC assumptions is that the reliability obtained provides an overview of the correlation that is observed in the data. Violating the compound symmetry assumptions suggests that there is too much day-to-day variability; therefore, we cannot summarize the reliability with one number as it would be difficult to interpret (22). The alternative generalized least square (GLS) models for estimating ICC, within the linear mixed models, can model different variance–covariance structures and thus handle variability in the correlations among multiple time points. It is unclear if the articles based on the OLS methods have used such methods as this information is rarely reported; however, it remains that modeling the variability still does not solve the problem of how meaningful it is to interpret the single reliability estimate when there is tremendous variability between days. Enough evidence (1,10,11) suggests frequent violations of the compound symmetry assumption, which questions many of the estimates of number of the days published in the literature.

The compound symmetry assumption was tested using the GLS methods of estimation in a sample of 165 adult teachers who completed a 7-d physical activity record (1). The assumption

of compound symmetry was violated for the day-to-day data. Although this is the only study that specifically examined the correlational structure of physical activity data, other evidence suggests this assumption was violated in both children (20) and adults (1,12). Adults had higher levels of inactivity on Sunday and greater levels of activity on Saturday. The correlation among weekdays was higher than weekend days, and there was a slightly higher correlation between Sunday and Monday. Elementary school children were more active on weekends than weekdays, and this pattern reversed among high school adolescents whose patterns begin to resemble those of adults (20). These data provide some evidence that day-to-day variability was not constant, and thus any estimate based on the compound symmetry assumption would underestimate the number of days needed to obtain a given reliability. Conceptually, violating compound symmetry suggests that the smallest unit of measurement to capture individual patterns of physical activity should be a week (1). From a practical perspective, however, requiring participants to wear a monitoring device for 7 d has proven to be problematic, so these two solutions may be difficult to implement in practice.

Nutrition researchers have dealt with similar issues of day-to-day variability in diet intake when collecting multiple 24-h recalls. However, many dietary recalls are also not feasible in practice. Although it is beyond the scope of this article to provide an extensive description of these methods (4,8), statistical methods were developed to model the within-person variability and correct the distribution obtained with one 24-h recall per participant. This has been done primarily for group level estimates. The usefulness of these methods for modeling physical activity behaviors is unknown and depends on the ability to model the within-person (day-to-day) variability. Given the difficulties in having participants comply with a 7-d data collection protocol, these alternative methods are attractive.

Another issue that has not received sufficient attention is how to capture patterns of activity over an extended period (e.g., months or years). Failure to adequately capture variability across weeks or seasons (16,17) likely explains why point estimates of activity often do not generalize to typical activity patterns over longer periods of time. Another issue not addressed in activity measurement is collecting consecutive versus random days. Consecutive days tend to be negatively correlated. That is, if a person overdoes activity on 1 d, he or she may go slow the next day to compensate. Thus, consecutive days of assessment are not statistically independent and thereby do not meet that assumption of statistical analysis. Collecting data on single days overcomes the correlated days issue, but not seasonality. More research is warranted on these issues.

APPLICATIONS OF G-THEORY TO ADVANCE RESEARCH ON RELIABILITY

G-theory provides a framework to provide more detailed information on the sources of variance in physical activity research models. In G-theory, reliability is based on the defined universe, and the influence components have on the error is quantified. This is a more flexible approach to reliability than more traditional methods. In ICC, the error component of the reliability analysis is not well defined and singular. Brennan (2) reported that reliability was not a property of a test but rather a set of scores. He further clarified that the reliability can be quite different for different objects of measurement and different universe of generalization (design of the data collection process). The reliability or standard error of measure is dependent on the sources of error that are incorporated into the design, and there is no definitive way to identify what sources of error should be included in the analysis. To achieve a reliable measure, it is very likely that multiple measures are needed to achieve a stable score that is called a universe score. The average of multiple scores will always be more stable than a single measurement.

All scores contain some degree of error caused by sources of error in the measurement procedure (previously called the analytic variability) (3,15). In G-theory, alternatively,

components of the error variance are defined and examined. This is accomplished by defining conditions of the measurement procedure, such as number of days, number of devices, and type of day, which are called facets, with corresponding sources of variation built into the design of the study. G-theory thus goes a step further than ICC by examining the error variances.

The researcher can use this information to design a better measurement procedure to reduce the error and achieve the desired reliability. An example of a basic reliability study design is illustrated in Figure 2 by a Venn diagram. In G-theory, there are two types of studies: one that estimates the variance components of the facets to identify contribution to the error (G studies), and the other examines measurement procedures to select the model with acceptable reliability (D studies). G-theory can be used to determine how many measurement occasions are needed or the design of the measurement procedure that allows for dependable scores.

Generalizability studies

There are four main steps in the G study: a) identification of facets, b) performance of a repeated-measures ANOVA, c) calculation of the variance for each facet and interaction, and d) calculation of the relative contributions of each variance component. Possible relevant facets that could be included in a study are persons (P), days (D), and types of devices (T). The model used in this example is one of many possible measurement models that could be appropriate depending on the circumstances of the research question. A repeated-measures ANOVA provides the mean square values for the facets and the interactions with a residual error term. From the mean squared values, the relative contributions of each variance component to the overall measurement error can be calculated using the variance of each facet and interaction. The percentage of the variance for P, D, T, $P \times D$, $P \times T$, $P \times D \times T$, and e indicates which components are influencing the error the greatest. This information can then be used in D studies.

G-theory does suffer from issues with calculating variance estimates like ICC. These issues include problems with unbalanced mixed models and sampling variability (2). There have been numerous methods developed to address these issues that included maximum likelihood (18) and bootstrapping techniques. For further information on methods and issues with variance component calculation, please see Brennan's generalizability theory (2).

Decision studies

Two reliability coefficients are produced with D studies: the *G*-coefficient and the Phi (Φ). The *G*-coefficient is a norm-referenced reliability measure, whereas the Phi coefficient is a criterion referenced measure. These coefficients reflect the parameters of the design. For a facet in a G-study that was a major source of error, like the number of days a participant was measured (e.g., 2 d), an investigator could use the variance components to calculate the *G*-coefficient reliability over four, six, or any number of days the researcher wished to consider. The *G*-coefficient is norm referenced in that it includes the relative error variance, which is defined as any interaction with the unit of measure (people usually). The relative error term in this example is defined as

$$\sigma_{\text{Rel}}^2 = \sigma_{pd}^2 + \sigma_{pt}^2 + \sigma_{pdt,e}^2$$

where the variance components and interactions only dealing with the object of measure (people) are included. The formula for the *G*-coefficient is

$$G = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{\text{Rel}}^2}$$

The Phi coefficient uses the absolute error variance as in criterion-referenced standards. This is the reliability coefficient that should be used with an absolute value, a benchmark, or cut score. The difference between the relative, and absolute variance is with the absolute error variance all variance not unique to the unit of measure (i.e., people) are included. The absolute error term in this example is:

$$\sigma_{\text{Abs}}^2 = \sigma_{pd}^2 + \sigma_{pt}^2 + \sigma_{pdt,e}^2 + \sigma_d^2 + \sigma_t^2 + \sigma_{dt}^2$$

where all sources of variance are included in the error term other than the variance for the object of measure. Phi-coefficient formula is

$$\Phi = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{\text{Abs}}^2}$$

Defining facets and measuring the percent of the variance each facet and its interactions contribute can be useful in designing efficient data collection procedures. Designs can be examined to see if procedures satisfy the practical and statistical concerns.

After the measurement model has been examined, reliability can be further investigated by calculating the root mean square error (RMSE). This measure is differences between the scores from different days squared, averaged, and square rooted. A smaller RMSE represents a more reliable physical activity measure. The RMSE is another important issue in reliability that is often overlooked. This statistic is critical in interpreting whether reliability is acceptable or high due to heterogeneity in the sample. If reliability is artificially high due to the heterogeneity of the sample, then the point estimates for physical activity may be misleading due to unchecked intraindividual variability. Although RMSE is unrelated to G-theory, it is extremely important that researchers examining reliability of any measure include an examination of RMSE for full understanding.

CONCLUSION

The number of days necessary to assess physical activity at a reliability level of 0.8 or higher has varied substantially across methods and types of participants. The usual method for assessing reliability in these studies, the ICC, involves assumptions not ordinarily met in this research, thereby likely underestimating the true number of days needed. Statistical methods correcting for within-person variability need to be tested in physical activity research. G-theory provides a set of methods for better understanding factors (e.g., days, types of methods) contributing to unreliability and thereby providing accurate answers to the issue of number of days of assessment. Research in this area needs to evolve to use G and D method designs before true answers can be ascertained.

References

1. Baranowski T, Smith M, Thompson WO, Baranowski J, Hebert D, de Moor C. Intraindividual variability and reliability in a seven day exercise record. *Med Sci Sports Exerc* 1999;31(11):1619–22. [PubMed: 10589866]
2. Brennan, RL. Generalizability Theory. New York: Springer-Verlag; 2001.
3. Brennan, RL. Some applications of generalizability theory to the dependability of domain-referenced tests. Paper presented at the Annual Meeting of the National Council on Measurement in Education; April 1979; San Francisco, California.
4. Carriquiry AL. Estimation of usual intake distributions of nutrients and foods. *J Nutr* 2003;133(2): 601S–8S. [PubMed: 12566510]
5. de Moor C, Baranowski T, Cullen KW, Nicklas T. Misclassification associated with measurement error in the assessment of dietary intake. *Public Health Nutr* 2003;6:393–9. [PubMed: 12795828]
6. DuRant RH, Baranowski T, Davis H, et al. Reliability and variability of indicators of heart rate monitoring in five, six or seven year old children. *Med Sci Sports Exerc* 1993;25(3):389–95. [PubMed: 8455456]
7. DuRant RH, Baranowski T, Puhl J, et al. Evaluation of the Children's Activity Rating Scale (CARS) in young children. *Med Sci Sports Exerc* 1993;25(12):1415–21. [PubMed: 8107551]
8. Freedman LS, Midthune D, Carroll RJ, et al. Adjustments to improve the estimation of usual dietary intake distributions in the population. *J Nutr* 2004;134(7):1836–43. [PubMed: 15226478]
9. Janz KF, Witt J, Mahoney LT. The stability of children's physical activity as measured by accelerometry and self-report. *Med Sci Sports Exerc* 1995;27(9):1326–32. [PubMed: 8531633]
10. Levin S, Jacobs DR, Ainsworth BE, Richardson MT, Leon AS. Intra-individual variation and estimates of usual physical activity. *Ann Epidemiol* 1999;9(8):481–8. [PubMed: 10549881]
11. Matthews CE, Freedson PS, Hebert JR, et al. Seasonal variation in household, occupational, and leisure time physical activity: longitudinal analyses from the seasonal variation of blood cholesterol study. *Am J Epidemiol* 2001;153(2):172–83. [PubMed: 11159163]
12. Matthews CE, Hebert JR, Freedson PS, et al. Sources of variance in daily physical activity levels in the seasonal variation of blood cholesterol study. *Am J Epidemiol* 2001;153(10):987–95. [PubMed: 11384955]
13. McGraw K, Wong S. Forming inferences about some intraclass correlation coefficients. *Psychol Methods* 1996;1:30–46.
14. McMurray RG, Ring KB, Treuth MS, et al. Comparison of two approaches to structured physical activity surveys for adolescents. *Med Sci Sports Exerc* 2004;36(12):2135–43. [PubMed: 15570151]
15. Morrow, JR. Generalizability theory. In: Safrit, M.; Wood, T., editors. *Measurement Concepts in Physical Education and Exercise Science*. Champaign, IL: Human Kinetics; 1989. p. 73-96.
16. Pivarnik JM, Reeves MJ, Rafferty AP. Seasonal variation in adult leisure-time physical activity. *Med Sci Sports Exerc* 2003;35(6):1004–8. [PubMed: 12783049]
17. Plasqui G, Westerterp KR. Seasonal variation in total energy expenditure and physical activity in Dutch young adults. *Obes Res* 2004;12(4):688–94. [PubMed: 15090638]
18. Searle, SL. *Linear Models for Unbalanced Data*. New York: Wiley; 1987.
19. Traub, RE. *Reliability for the Social Sciences, Theory and Applications*. Vol. 3. Thousand Oaks, CA: Sage; 1994.
20. Trost SG, Pate RR, Freedson PS, Sallis JF, Taylor WC. Using objective physical activity measures with youth: how many days are needed. *Med Sci Sports Exerc* 2000;32(2):426–31. [PubMed: 10694127]
21. Tudor-Locke C, Burkett L, Reis JP, Ainsworth BE, Macera CA, Wilson DK. How many days of pedometer monitoring predict weekly physical activity in adults? *Prev Med* 2005;40(3):293–8. [PubMed: 15533542]
22. Ugrinowitsch C, Fellingham GW, Ricard MD. Limitations of ordinary least squares models in analyzing repeated measures data. *Med Sci Sports Exerc* 2004;36(12):2144–8. [PubMed: 15570152]

FIGURE 1.
Pattern in days of physical activity by levels of consistency.

FIGURE 2.

Venn diagram of a study including the people's (P), days (D), and number of devices (T) facets. The interactions of people and days are represented by $P \times D$, the interaction between people and number of devices are represented by $P \times T$, the interaction between days and number of devices is $D \times T$, and people, days, and the number of devices include undefined error ($P \times D \times T, e$).