Published in final edited form as:

Med Sci Sports Exerc. 2017 July ; 49(7): 1473–1481. doi:10.1249/MSS.0000000000001237.

Calibration of Self-Report Measures of Physical Activity and Sedentary Behavior

Gregory J. Welk¹, Nicholas K. Beyler², Youngwon Kim^{1,3}, and Charles E Matthews⁴

¹Department of Kinesiology, Iowa State University, Ames, IA

²Department of Data Science and Statistics, Mathematica Policy Research, Washington, DC

³MRC Epidemiology Unit, University of Cambridge School of Clinical Medicine, United Kingdom

⁴Division of Cancer Epidemiology & Genetics, Nutritional Epidemiology Branch, National Cancer Institute, Rockville, MD

Abstract

Introduction—Calibration equations offer potential to improve the accuracy and utility of self-report measures of physical activity (PA) and sedentary behavior (SB) by re-scaling potentially biased estimates. The present study evaluates calibration models designed to estimate PA and SB in a representative sample of adults from the Physical Activity Measurement Survey (PAMS).

Methods—Participants in the PAMS project completed replicate single day trials that involved wearing a Sensewear armband (SWA) monitor for 24 hours followed by a telephone administered 24-hour physical activity recall (PAR). Comprehensive statistical model selection and validation procedures were used to develop and test separate calibration models designed to predict objectively-measured SB and moderate to vigorous PA (MVPA) from self-reported PAR data. Equivalence testing was used to evaluate the equivalence of the model-predicted values with the objective measures in a separate holdout sample.

Results—The final prediction model for both SB and MVPA included reported time spent in SB and MVPA, as well as terms capturing sex, age, education, and BMI. Cross-validation analyses on an independent sample exhibited high correlations with observed SB ($r = 0.72$) and MVPA ($r = 0.75$). Equivalence testing demonstrated that the model-predicted values were statistically equivalent to the corresponding objective values for both SB and MVPA.

Conclusion—The results demonstrate that simple regression models can be used to statistically adjust for over or underestimation in self-report measures among different segments of the population. The models produced group estimates from the PAR that were statistically equivalent to the observed time spent in SB and MVPA obtained from the objective SWA monitor; however additional work is needed to correct for estimates of individual behavior.

Corresponding Author: Name: Gregory J. Welk, Mailing Address: Department of Kinesiology, Iowa State University, 257 Forker Building, Ames, IA, 50011, Telephone: 1 - 515-294-3583, Fax: 1 - 515 - 294 - 8740, gwelk@iastate.edu.

Conflicts of Interest

The authors declare no conflicts of interest.

Keywords

calibration; physical activity; sedentary behavior; self-report

Introduction

The development of more accurate assessments of physical activity (PA) has been established as an important research priority since it impacts our ability to advance research in a number of other areas. A specific challenge in PA research is estimating the distribution of PA levels for a group of individuals in the context of population surveillance. Considerable attention has been focused on the refinement of objective monitoring tools, such as accelerometry-based activity measures (3, 11, 36), but parallel advances are also needed to improve the utility of self-report measures, which can provide novel insights about the nature and context of PA in the population (6, 33).

Many researchers have avoided and/or dismissed self-report instruments due to concerns about accuracy and bias (20). This has been fueled by numerous studies that have reported large differences between reported and observed levels of PA when both are directly compared (32, 34). Researchers place trust in objectivity so the logical assumption has been that all self-report measures are inherently flawed; however, recent studies using a 24-hour/previous day recall approach demonstrate improved accuracy and less bias than other recall formats and survey approaches (8, 12, 19). The utility of 24-hour recall instruments might be further improved by combining their results with that of an objective measure in a calibration framework, an under-utilized method that enables the self-report responses to be re-scaled so that they more closely match responses from objective measures.

Calibration is a widely used measurement technique in all branches of science and it has been widely adopted in nutritional epidemiological research to improve accuracy of self-reported food intake. Specifically, calibrated self-reported data of food intake have been used to more accurately quantify the amount of food consumed (13, 25) and to determine unbiased relationships with various health outcomes (24, 31). Calibration models linking self-reported PA data to objectively assessed PA data have shown promise in both children (26) and adults (35). Calibration models to correct for biases in estimating energy expenditure (EE) have also been developed from a study by Neuhouwer et al. (21). However, a fundamental gap in the literature is the lack of calibration equations to improve the accuracy and utility of self-report measures for estimating time spent in sedentary behavior (SB) and PA.

The Physical Activity Measurement Study (PAMS) was an NIH-funded study (R01HL091024) designed to specifically model, quantify and adjust for measurement error in PA self-report measures. We have previously reported on a preliminary set of measurement error models for estimating EE from self-report data (22). We also quantified sources of error in estimates of SB and moderate to vigorous PA (MVPA) using estimates provided by the Sensewear armband monitor as the criterion (17, 36). In the current paper, we build on this work by developing and validating two sets of correction models, one which calibrates time spent in SB and another which calibrates time spent in MVPA.

Method

Participants

The PAMS study employed a multi-stage stratified sampling technique to recruit a representative sample of Iowan adults from four target counties (2 rural and 2 urban). A random sample of potential participants (selected from Survey Sample International) was contacted through random digit dialing. To obtain a diverse sample which included Blacks and Hispanics, two co-strata were created in the four target counties based on the relative proportions of Blacks and Hispanics, thereby increasing the likelihood of obtaining similar numbers of Blacks and Hispanics across the four counties. To be eligible for the study, participants had to be between 20–75 years old and capable of walking and performing given surveys either in English or Spanish. Each participant provided written informed consent before participation. The study design and protocol of the PAMS project were approved by the local Institutional Review Board. Additional details about the sample characteristics are provided in the Results section.

Instruments

24-Hour Physical Activity Recall (PAR)—The 24-hour PAR was selected for calibration in PAMS because it systematically and efficiently captures valuable information about type, intensity and duration of PA (8, 18). It also provides an ideal complement to the established 24h diet recall time instrument currently used in the National Health and Nutrition Examination Survey. Validation studies have favored the 24h diet recall format over other longer recall formats (28) and studies have demonstrated improved utility when re-scaled with measurement error models (9, 10, 23).

SenseWear Armband Mini (SWA)—The SWA (BodyMedia®, Pittsburgh, PA; now merged and acquired by Jawbone) is a non-invasive activity tracking monitor that is equipped with multiple biological sensors (i.e. heat flux, galvanic skin response, skin temperature, and near body temperature), as well as a tri-axial accelerometer. The SWA is designed to be worn on the back of an upper arm. It provides aggregated signals from each sensor as well as an overall estimate of EE developed using proprietary pattern recognition algorithms that are refined over time. Numerous studies have documented the validity of the SWA for estimating EE. For example, a series of doubly labelled water studies on adult populations (7, 14) documented mean absolute percent error (MAPE) values of approximately 10%. The accuracy for EE estimation has been attributed to the use of multiple sensors and pattern recognition technology that is now commonly used with other contemporary devices. A recent study (16) quantitatively documented stronger accuracy for estimates of EE across a range of different intensities when compared to the ActivPAL and Actigraph. We used the latest software (v8.0) coupled with the v5.2 algorithms to download and process the data collected in the PAMS project since these were the most well supported in the literature.

Data Collection

Replicate data from both the SWA and PAR were collected from the recruited participants across eight 3-month quarters over a 2-year time span. For each measurement trial,

participants were required to wear a SWA for a full 24-hour period, and then to perform a PAR interview with our staff over the telephone the following day to recall activities of the previous day (i.e. the same day as the SWA monitoring day). Field staff members visited each participant's house to distribute the SWA monitor and to provide instructions. Participants were asked to remove the SWA when doing aquatic activities (including showers, swimming), and to record the duration and types of activities performed while the SWA was not worn. The following day, trained interviewers called the participants to complete the guided PAR interview that allowed them to recall details of the previous day. The PAR interviews were performed through a computer-assisted telephone interview (CATI) system programmed with the Blaise Trigram methodology. The interviewer asked respondents to report on the previous day's activities across four 6-hour time windows (i.e. midnight–6am, 6am–noon, noon–6pm, and 6pm–midnight) in bouts of at least 5 minutes. The interviewers guided the participants through the segments and helped to ensure that each 6 hour block was fully coded with activities before moving to the next section. This ensured that each participant had accounted for a full day (i.e. 24 hours $\times$ 60 minutes = 1440 minutes). Every PAR interview was supervised by project staff to ensure adherence to the interviewing protocol. Spanish-speaking interviewers were available for Spanish-speaking participants. The PAR interviews took approximately 20 minutes to complete with ranges from 12 to 45 minutes.

Data Processing

The PAR data included types, minutes, and contextual information (i.e. location and purpose) of each reported activity. Reported activities were assigned MET scores based on the corresponding activity codes in the reduced version of the Compendium of PA (1) refined by the project team to avoid redundancies and misclassification. Assigned MET values were used to categorize each reported activity as sedentary (MET 1.5) or MVPA (MET 3.0). Total daily sedentary time and MVPA time were calculated by adding up the reported minutes of the respective intensity categories.

The SWA data were processed using the software to obtain MET values for every minute. The same standard MET-derived criteria were used to classify every minute as sedentary (MET 1.5) or MVPA (MET 3.0). We aggregated up the classified minutes of SB and MVPA to obtain total daily SB time and MVPA time, respectively. MET values for non-wear time periods were determined by linking the activities recorded on the participant log with the MET scores from the reduced PA Compendium, and then added to the aggregated SB time and MVPA time for SWA.

A wide variety of socio-demographic variables (collected during the PAR interview) was merged with the PAR and SWA data. The specific variables included were sex (female or male), age (between 20 to 71yrs), body mass index (BMI), education background (a high school diploma/some high school, some college, or a college degree or higher), race/ethnicity (white or non-white), income level (less than \$25,000, between \$25,000 and \$75,000, or more than \$75,000), town size (urban or rural), employment status (full time, part time, or others), and marital status (married/living as married, divorced/separated/widowed, or single/never married).

Statistical Analyses

The study employed comprehensive statistical model selection and validation procedures to develop and test calibration models to improve the accuracy of estimates of SB and MVPA from the PAR. The response variable for the SB model was the daily time spent engaged in SB from the SWA (*SWA_{sed}*) while the response variable for the MVPA model was the daily time spent engaged in MVPA from the SWA (*SWA_{MVPA}*). Both the SB and MVA models were evaluated in minutes/day. The primary explanatory variables were the associated self-reported values of SB and MVPA from the PAR (*PAR_{sed}* and *PAR_{MVPA}*, respectively). However, preliminary evaluations determined that reported MVPA was an important predictor of SB and reported SB was an important predictor of MVPA. Thus, they were both included in the two models.

Additional explanatory models included sex, scored as an indicator for being female vs. male; age, ranging from 20 to 71yrs; education level, with three categories coded as high school diploma or some high school, some college, or a college degree or higher; and BMI. We also considered the fifteen two-way interactions between each of these variables. Other explanatory variables considered during preliminary analyses included race/ethnicity (white vs. non-white), income level, town size (urban vs. rural), employment status, and marital status. None of these variables were significant in the preliminary models after accounting for the explanatory variables mentioned above and were dropped from final model selection procedures.

Different models were required for SB and MVPA due to differences in the distribution of the data and the associated model fitting assumptions. For the SB outcome, we constructed the prediction models using multivariate linear regression models (based on SAS's PROC REG procedure). For the MVPA outcome, we constructed the prediction models using log-Gamma generalized linear models (using SAS's PROC GENMOD model fitting procedure). A log-Gamma generalized linear model fit better to the data than a traditional multiple linear regression model because the MVPA data were heavily skewed, more closely resembling a Gamma distribution than a Normal distribution and because there was evidence of a relationship between the mean and variance (i.e., there was no evidence of constant variance across mean outcomes). During preliminary analyses we also checked to make sure there were no differences between individuals with two days of data and individuals with one day of data (i.e. those that did not respond during one of their two measurement days). Measures of *SWA_{sed}*, *PAR_{sed}*, and *PAR_{MVPA}* for individuals with two days of data were averaged for the model fitting procedures.

Once the final model was selected, we evaluated the model assumptions and the model's validity and predictive capacity by performing a number of statistical tests and examining a series of visual analysis plots. We also checked for outliers, high leverage points, and multicollinearity in the model variables. We compared the final model to alternative models, including a full model with additional interaction variables and a model which accounted for the complex survey design used to collect the study data.

For cross-validation analyses, we predicted objective SB and MVPA for a hold-out subsample (20% of original sample) and compared their model-predicted values with the

objective estimates of SB and MVPA from the SWA. Individual agreement was evaluated by computing mean absolute percent error (MAPE); group level agreement was examined using equivalence testing. Equivalence zones were defined as $\pm 10\%$ of the mean values of the objectively estimated SB or MVPA from the SWA. Calculated 90% confidence intervals of predicted SB or MVPA values were compared against the pre-specified equivalence zones. Specifically, if 90% confidence intervals fell completely within equivalence zones, predicted and observed objective estimates were considered statistically equivalent to each other on average. All statistical tests were conducted using SAS. All calculations for graphical illustrations were performed using R statistical software.

Results

Results are considered separately for the SB and PA calibration models.

Sedentary Behavior Calibration Model—SB data were available for 1458 survey participants. A random subsample of 80% of these survey participants ($n = 1166$) were used for model development. The remaining 20% were included in a hold-out subsample ($n = 292$) for cross-validation analysis. Table 1 provides characteristics of the two subsamples, including summary statistics for the model variables (*SWAsed*, *PARsed*, *PARMVPA*, *Female*, *Age*, *Education*, *BMI*). In both subsamples, over half of the participants are female (58%).

The final model selected for prediction of SB is presented in Table 2 (top half) along with an alternative model that excludes *Education*, and the interaction between *Age* and *Education*, as model variables. The regression coefficients for both models are provided along with their corresponding standard errors. All regression coefficients were statistically significant at an alpha level of 0.01. The R-square for the final prediction model was about 0.55, indicating that the model explained about 55% of the variability in the response variable, *SWAsed*. An alternative model that excluded *Education*, and the interaction between *Age* and *Education*, yielded a similar R-squared value of 0.54 and was selected as an alternative model since it would be more practical to use by other researchers using the PAR in general applications.

In both SB models, more reported SB (*PARsed*) and less reported MVPA (*PARMVPA*) were indicators for more objective SB (*SWAsed*), holding all other variables constant. Being female, being older (*Age*), having a higher education level, and having a higher BMI (*BMI*) were all indicators for having higher amounts of objectively assessed SB. However, reporting more SB while also having higher BMI (*PARsed* $\times$ *BMI*) was associated with lower objective SB, holding all other factors constant. Similarly, being older and having a higher education level (*Age* $\times$ *Education*) was associated with lower objective SB, holding all other factors constant.

Additional checks were conducted to evaluate assumptions and fit of the model. A plot of residuals based on the final and alternative prediction model represented in the top portion of Table 2 yielded a random scatter of points about the line at 0, suggesting no evidence of a poor model fit (results not shown). Additional residual plots comparing residuals to the specific model variables also showed no patterns or outliers, which would suggest a poor

model fit (results not shown). The basic assumptions for the modeling were met (i.e. no evidence of multicollinearity, outliers or highly influential observations, and non-constant error variances); however, several additional models were developed with slightly different approaches. One model was based on a stepwise model selection procedure that included all the main model variables (*PARsed*, *PARMVPA*, *Female*, *Age*, *Education*, and *BMI*) and four interaction terms (*PARsed* × *BMI*, *Age* × *BMI*, *Age* × *Education*, *Female* × *BMI*). An F-test statistic comparing the full model with interaction terms to the reduced, final model (top half of Table 2) was 2.12 on (2, 1155) degrees of freedom, yielding a p-value of 0.12. This suggests that the reduced model is not significantly different than the full model. Another model used the same variables as those in Table 2, but was fit using weighted least squares (instead of ordinary least squares) so that the survey weights could account for the complex survey design used to select study participants. We tested the difference between the model fit with ordinary least squares to one fit with weighting least squares to account for the survey design using an F-test procedure (Fuller 2009, Chapter 6; Beyler 2010). The F-test statistic was 1.39 on (9, 1148) degrees of freedom, yielding a p-value of 0.19. Hence, there is little evidence to suggest that the design-based model is different than the model fit with ordinary least squares (not accounting for the survey design). These results support the utility of the proposed model.

For cross-validation, we generated model-predicted values of objective SB using the final model for participants in the holdout sample ($n = 292$). The Pearson correlation coefficient comparing the model-predicted values to the observed objective SB values, measured from the SWA, was 0.72 and the MAPE value was 21 percent. Figure 1a provides a plot of these comparisons.

Tests of statistical equivalence were used to more quantitatively evaluate the group level agreement. The mean *SWAsed* (criterion method) in the holdout sample was 679 minutes/day, yielding a zone of equivalence ranging from 611 to 747 minutes/day. The 95% confidence interval for the model-predicted values (test method) ranged from 659 to 693 minutes/day, so the narrower 90% confidence interval was within the zone of equivalence. We also fit the same model to the hold-out sample and used an F-test to compare the estimated regression coefficients for the hold-out sample to those for the final model presented in Table 2. There was no significant difference in the regression coefficients across samples according to the test ($p\text{-value} = 0.62$). Collectively, these results suggest that the final model accurately predicted objective SB for the holdout sample. Similar results were obtained for the alternative model presented in the top half of Table 2 (not shown).

MVPA Calibration Model—Similar approaches were used to develop and test the MVPA model but, in this case, we restricted the sample to survey participants with average daily MVPA, in minutes per day, greater than 0. Data were available for 1406 survey participants; another 52 individuals had 0 minutes of daily MVPA and were excluded from these analyses to make model development more user-friendly (including individuals with 0 MVPA would require a two-step modeling procedure which is beyond the scope of this paper). A random subsample of 80% of these survey participants ($n = 1125$) were used for model development. The remaining 20% were included in a hold-out subsample ($n = 281$) for cross-validation

analysis. The characteristics of both subsamples are similar to those presented in Table 1 for the SB model development.

The final model selected for prediction of MVPA is presented in the bottom half of Table 2 along with an alternative model that excludes the education variables (i.e. *Education*, and *Age* $\times$ *Education*). We used similar model building procedures as we did for building the SB model, comparing full and reduced models and comparing weighted and un-weighted models to check if there was evidence of correlation between the model errors and the survey weights. The regression coefficients are given in the log scale, which is common practice for a log-Gamma generalized linear model. All regression coefficients were statistically significant at the 0.01 alpha level for the final and alternative models with two exceptions (the p-value for the regression coefficient on *Age* $\times$ *Education* was 0.017 in the final prediction model and the p-value for the regression coefficient on *PARsed* $\times$ *BMI* was 0.023 in the alternative model). More reported MVPA (*PARMVPA*) and less reported SB (*PARsed*) were both indicators for more objective MVPA (*SWAMVPA*), holding all other variables constant. The *Age* and *BMI* terms were both negatively associated with more objective MVPA.

The residual plots for the final and alternative MVPA models yielded a random scatter of points around 0, suggesting little evidence of a poor model fit. Additional goodness-of-fit statistics, including Pearson Chi-Square, AIC values, and Cook's distance metrics, presented in the SAS output also suggested reasonable model fit. We found a few observations with large Cook's distance statistics, suggesting that those points were highly influential on the model fit. However, when we refit the model after removing these observations the regression coefficients were very similar.

For cross-validation, we generated model-predicted values of objective MVPA from the final model for participants in the holdout sample ($n = 281$), similar to analyses for SB. The Pearson correlation coefficient comparing the model-predicted values to the observed objective MVPA values was 0.75 but there was considerable individual error as evidenced by a MAPE value of 142 percent. The equivalence test for group estimation of MVPA was conducted in the log scale (due to the use of the log-gamma GLM model). The mean estimate of MVPA from the SWA in the holdout sample was 4.0 minutes per day (log scale), yielding a zone of equivalence of 3.6 to 4.4 minutes. The 95% confidence interval for the model-predicted values in the log scale yielded a range from 4.15 to 4.34, so the narrower 90% confidence interval is within the zone of equivalence. We also fit the same generalized linear model to the hold-out sample and compared the estimated regression coefficients for the hold-out sample to those for the final model presented in Table 2 (bottom). The coefficients were generally similar across samples. Collectively, these results suggest that the model provided predicted MVPA values that were comparable with observed objective MVPA values for the hold-out sample.

A plot of the observed and model-predicted values for the holdout sample is provided in Figure 1b. The points were more scattered for larger values, suggesting that there is more uncertainty in such predictions (and that there is a relationship between the mean and variance, which is accounted for using a log-Gamma generalized linear model; when means

are larger the corresponding variances are expected to be larger as well. This, however, does not apply to general multiple linear regression models that have an assumption of constant variance). The cross-validation results were similar for the alternative model, which excludes Education variables (not shown).

Discussion

Numerous studies have documented the limitations of self-report measures for estimating PA or energy expenditure; however, it is surprising that there has been relatively little work done to try to improve these predictions. The PAMS was conducted to specifically quantify and adjust for error in the PAR self-report measure so that it can produce estimates that more closely match those from the SWA. We previously reported significant overestimation of MVPA (MAPE = 162 percent) and significant underestimation of SB (MAPE = 120 percent) when raw estimates from the PAR were related to the SWA (17, 36). In the present study, we demonstrated the utility of simple calibration equations to reduce this error and to improve the accuracy of the estimates. The models for both SB and MVPA were found to produce group estimates that were statistically equivalent to group means obtained from the SWA.

The proposed linear multiple regression models help to address differences in perception of SB and MVPA in older individuals and those with higher BMI levels. This is typically labelled as “error”, but this is an over-simplification considering the important distinctions between relative and absolute indicators of PA. For example, an overweight person would likely categorize the same bout of activity as being more vigorous than a normal weight person because it is harder for them on a relative basis. An objective monitoring device categorizes all activities the same regardless of fitness since it scores PA primarily using absolute acceleration data. Few studies have systematically examined possible differences in perception of PA in typical adults, but it must be considered when interpreting results from self-report measures when they are related to objective measures. A recent study on the International Physical Activity Questionnaire (IPAQ) demonstrated similar effects with greater likelihood of overestimation observed in less fit individuals compared to more fit individuals (27).

The inclusion of fitness and other variables in a calibration model would likely further improve the predictive utility of this type of calibration model; however, it was not available in the present study. However, the potential value of further reducing error would also have to be weighed against the additional burden or challenge of collecting additional data. In our models, we decided to exclude *Education* (and its corresponding interaction terms) in the alternative models since it often proves challenging to even obtain these demographic variables from participants. Thus, we proposed more parsimonious models for both SB and MVPA (the “alternative” models in Table 2) that could be utilized with only *Age*, *Sex* and *BMI* as the primary predictor variables (See Supplemental Digital Content for specific details about how the final and alternative models can be used to make predictions of objective SB and MVPA). To illustrate the impact of the calibration models on the estimates, we calculated group mean response predictions and corresponding 95% confidence bands using R statistical software. The mean response predictions for SB estimations from two hypothetical sets of 50-year old females is plotted in Figure 2a using the final prediction

models from Table 2. The only difference between these sets of individuals is that one has a BMI in the normal weight range ($BMI = 23$) and the other has a BMI in the obese weight range ($BMI = 35$). The solid blue line represents mean predicted objective SB for the normal-weight individuals, while the solid red line represents the same estimates for the obese individuals. The dotted lines give the 95% confidence bands around the model-based mean predictions. Predicted SB is consistently lower for the normal-weight individuals compared to the obese individuals. Based on the model, individuals who report lower amounts of SB are predicted to have more SB than they report (illustrated by the red and blue lines above the dashed identity line). Similarly, individuals who report more SB are predicted to have less SB than they report (illustrated by the red and blue lines falling below the dashed identity line). The discrepancies between both the red and blue lines and the dashed identity line suggest that without the calibrations, obtaining accurate measurements of SB based on the PAR alone is difficult.

A similar plot was created for predictions of MVPA (Figure 2b). In this case, the predicted MVPA is consistently higher for the normal-weight individual (solid blue line) compared to the obese individual (solid red line). The confidence bands are wider for more reported MVPA, suggesting more uncertainty in predictions of more active individuals. The reported MVPA values in this case ranged from 200 to 700 minutes/day. However, no model-predicted MVPA values are above 200 minutes/day, represented by the horizontal dashed line in Figure 2b, suggesting that there is consistent over-reporting of MVPA relative to what is objectively measured from the armband monitor for all individuals (according to the model). Similar results were found by Beyler et al. (5) who developed prediction models for MVPA using NHANES data.

The models provide a way to re-score self-reported PAR data to adjust for known tendencies for over or underestimation in different segments of the population (36). The results of the equivalence testing demonstrate that the models can produce group estimates that are statistically equivalent to group means obtained from the SWA; however, it is important to note that there is still considerable error for estimating individual behavior. The individual error for SB was reasonable ($MAPE = 20$ percent) but the error for MVPA was still considerable ($MAPE = 142$ percent). The larger MAPE value for MVPA is expected given the larger differences in observed and model-predicted MVPA values for the holdout sample (Figure 1) and the wider confidence bands for the models (Figure 2). Closer examination reveals that the large MAPE values are primarily explained by the larger errors at the tail ends of the distribution. If we restrict the sample to more “typical” values of MVPA (e.g. between 30 and 120 minutes per day), the MAPE is 53.9%. It is important to note that the models were not intended to directly addressing underlying measurement error in the instruments. Using measurement error models to account and adjust for measurement errors and biases in both the PAR and SWA could improve the statistical precision in the model-based calibration equations (22).

Comprehensive measurement error modeling procedures have been previously developed to understand the measurement error structure in dietary assessment instruments (13, 24). Kipnis et al. (17), for example, proposed an extended measurement error model that includes group-specific biases related to true nutrient intake and individual-specific biases that can be

correlated across days for the same individual and for both instruments. This approach offers similar potential for addressing the problem of biases and day-to-day variability in PA assessments. Preliminary models have been developed from the PAMS sample for capturing EE (22), but additional work is needed to develop parallel measurement error models for MVPA and SB, since the distributions differ for these outcome measures. The evidence in this study that person-specific biases in self-report vary with the true level of MVPA and SB of the individual (based on the SWA) suggest that this needs to be taken into account in future models based on this instrument and sample.

The calibration models developed in this study offer potential to reduce error in self-reported estimates of PA and SB. However, it is important to consider some potential limitations of the methods when interpreting the results and considering applications. First, the participants of the PAMS project were randomly recruited only from Iowa, USA, so the findings may not be generalizable to other states and/or countries. It is also important to emphasize that the reported data are calibrated to equate with estimates from the SWA. Thus, the accuracy of the estimate is directly dependent on the precision and accuracy of the criterion measure. Past work has supported the advantages of the SWA for overall estimations of EE (14) and MVPA,(16) but there is likely some compounding of error when estimating a less than perfect criterion measure. The lack of precision may be especially evident with the estimation of SB due to greater limitations of the SWA for assessing SB.(16) The SWA cannot technically categorize activities as “sedentary” since it cannot distinguish sitting from standing. Both activities have MET levels that are below the 1.5 threshold for a sedentary activity, but standing obviously doesn’t meet the postural requirement used to define SB (2). However, previous research (15) using the same PAMS data reported that only 1 of the top 20 most frequently reported sedentary activities included standing in its definition (e.g., “Stand quietly, stand waiting in line”), and the average time spent on this activity was only 19.3min/day. Thus, the potential impact of misclassifying non-SB as SB on the calibration estimates would be minimal.

Another limitation of our method is that we did not take “bouts” of MVPA or SB into account. Public health guidelines focus on bouts of MVPA greater than 10 minutes and evidence suggests that breaks in SB are an important indicator in addition to total SB time. Our calibration method focused on total time in MVPA and SB so the estimates may not directly equate with values needed to directly inform public health guidelines for MVPA or quantification of SB patterns. Additional work with related models could make it possible to capture additional metrics from the same data to provide a more complete profile as advocated by some researchers (29, 30).

A final consideration is to appreciate the difference between individual and group level error. As described above, the equations have good accuracy for group level estimates but the utility for estimating individual behavior is limited. The incorporation of more robust measurement error models would likely further improve the precision but these equations show good utility for capturing group estimates of PA and SB. We chose to calibrate the 24 PAR in the PAMS project because of its widespread use and because of potential applications for population surveillance (similar to how the 24 hour recall is used for dietary intake). The online Physical Activity 24-Hour Recall (ACT24) assessment available through

the National Cancer Institute uses a similar time sequencing and coding method as the PAMS interview protocol so direct applications with ACT24 data may also be promising.

Overall, the results of this study provide new insights into the potential utility of calibration methods to improve the accuracy of self-reported PA data. The current equations are specific to the 24 hour PAR survey format but similar methods can be used to improve the utility of other instruments. Calibration is an inherent part of standardizing any measurement instrument and this certainly applies to the various self-report tools used to capture PA and SB.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The present study was supported by the National Institute of Health grant (R01 HL91024-01A1). The authors declare no conflicts of interest. The findings of the current study do not constitute endorsement by the American College of Sports Medicine. The results of the study are presented clearly, honestly, and without fabrication, falsification, or inappropriate data manipulation.

References

1. Ainsworth BE, Haskell WL, Herrmann SD, et al. 2011 Compendium of Physical Activities: a second update of codes and MET values. *Med Sci Sports Exerc.* 2011; 43(8):1575–81. [PubMed: 21681120]
2. Barnes J, Behrens TK, Benden ME, et al. Letter to the Editor: Standardized use of the terms “sedentary” and “sedentary behaviours”. *Appl Physiol Nutr Me.* 2012; 37(3):540–2.
3. Bassett DR, Rowlands A, Trost SG. Calibration and Validation of Wearable Monitors. *Med Sci Sport Exer.* 2012; 44:S32–S8.
4. Berntsen S, Hageberg R, Aandstad A, et al. Validity of physical activity monitors in adults participating in free-living activities. *Br J Sports Med.* 2010; 44(9):657–64. [PubMed: 18628358]
5. Beyler N, Fuller W, Nusser S, Welk G. Predicting objective physical activity from self-report surveys: a model validation study using estimated generalized least-squares regression. *Journal of Applied Statistics.* 2015; 42(3):555–65.
6. Bowles HR. Measurement of active and sedentary behaviors: closing the gaps in self-report methods. *J Phys Act Health.* 2012; 9(Suppl 1):S1–4. [PubMed: 22287442]
7. Calabro MA, Kim Y, Franke WD, Stewart JM, Welk GJ. Objective and subjective measurement of energy expenditure in older adults: a doubly labeled water study. *Eur J Clin Nutr.* 2015; 69(7):850–5. [PubMed: 25351651]
8. Calabro MA, Welk GJ, Carriquiry AL, Nusser SM, Beyler NK, Mathews CE. Validation of a computerized 24-hour physical activity recall (24PAR) instrument with pattern-recognition activity monitors. *J Phys Act Health.* 2009; 6(2):211–20. [PubMed: 19420399]
9. Carriquiry AL. Estimation of usual intake distributions of nutrients and foods. *J Nutr.* 2003; 133(2): 601s–8s. [PubMed: 12566510]
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):1640–50. [PubMed: 17000197]
11. Freedson P, Bowles HR, Troiano R, Haskell W. Assessment of Physical Activity Using Wearable Monitors: Recommendations for Monitor Calibration and Use in the Field. *Med Sci Sport Exer.* 2012; 44:S1–S4.

12. Gomersall SR, Olds TS, Ridley K. Development and evaluation of an adult use-of-time instrument with an energy expenditure focus. *J Sci Med Sport*. 2011; 14(2):143–8. [PubMed: 20932797]
13. Huang Y, Van Horn L, Tinker LF, et al. Measurement error corrected sodium and potassium intake estimation using 24-hour urinary excretion. *Hypertension*. 2014; 63(2):238–44. [PubMed: 24277763]
14. Johannsen DL, Calabro MA, Stewart J, Franke W, Rood JC, Welk GJ. Accuracy of armband monitors for measuring daily energy expenditure in healthy adults. *Med Sci Sports Exerc*. 2010; 42(11):2134–40. [PubMed: 20386334]
15. Kim Y, Welk GJ. Characterizing the context of sedentary lifestyles in a representative sample of adults: a cross-sectional study from the physical activity measurement study project. *BMC Public Health*. 2015; 15:1218. [PubMed: 26646428]
16. Kim Y, Welk GJ. Criterion Validity of Competing Accelerometry-Based Activity Monitoring Devices. *Med Sci Sports Exerc*. 2015; 47(11):2456–63. [PubMed: 25910051]
17. Kim Y, Welk GJ. The accuracy of the 24-h activity recall method for assessing sedentary behaviour: the physical activity measurement survey (PAMS) project. *J Sports Sci*. 2017; 35(3): 255–61. [PubMed: 27019092]
18. Matthews CE, Ainsworth BE, Hanby C, et al. Development and testing of a short physical activity recall questionnaire. *Med Sci Sports Exerc*. 2005; 37(6):986–94. [PubMed: 15947724]
19. Matthews CE, Keadle SK, Sampson J, et al. Validation of a previous-day recall measure of active and sedentary behaviors. *Med Sci Sports Exerc*. 2013; 45(8):1629–38. [PubMed: 23863547]
20. Matthews CE, Moore SC, George SM, Sampson J, Bowles HR. Improving self-reports of active and sedentary behaviors in large epidemiologic studies. *Exerc Sport Sci Rev*. 2012; 40(3):118–26. [PubMed: 22653275]
21. Neuhouser ML, Di C, Tinker LF, et al. Physical activity assessment: biomarkers and self-report of activity-related energy expenditure in the WHI. *Am J Epidemiol*. 2013; 177(6):576–85. [PubMed: 23436896]
22. Nusser SM, Beyler NK, Welk GJ, Carriquiry AL, Fuller WA, King BM. Modeling errors in physical activity recall data. *J Phys Act Health*. 2012; 9(Suppl 1):S56–67. [PubMed: 22287449]
23. Nusser SM, Carriquiry AL, Dodd KW, Fuller WA. A semiparametric transformation approach to estimating usual daily intake distributions. *J Am Stat Assoc*. 1996; 91(436):1440–9.
24. Prentice RL, Huang Y. Measurement error modeling and nutritional epidemiology association analyses. *The Canadian journal of statistics = Revue canadienne de statistique*. 2011; 39(3):498–509. [PubMed: 23606784]
25. Prentice RL, Tinker LF, Huang Y, Neuhouser ML. Calibration of self-reported dietary measures using biomarkers: an approach to enhancing nutritional epidemiology reliability. *Current atherosclerosis reports*. 2013; 15(9):353. [PubMed: 23881548]
26. Saint-Maurice PF, Welk GJ. Web-based assessments of physical activity in youth: considerations for design and scale calibration. *J Med Internet Res*. 2014; 16(12):e269. [PubMed: 25448192]
27. Shook RP, Gribben NC, Hand GA, et al. Subjective Estimation of Physical Activity Using the International Physical Activity Questionnaire Varies by Fitness Level. *Journal of Physical Activity and Health*. 2016; 13(1):79–86. [PubMed: 25898394]
28. Subar AF, Kipnis V, Troiano RP, et al. Using intake biomarkers to evaluate the extent of dietary misreporting in a large sample of adults: The OPEN Study. *American Journal of Epidemiology*. 2003; 158(1):1–13. [PubMed: 12835280]
29. Thompson D, Batterham AM. Towards integrated physical activity profiling. *Plos One*. 2013; 8(2):e56427. [PubMed: 23437131]
30. Thompson D, Peacock O, Western M, Batterham AM. Multidimensional physical activity: an opportunity, not a problem. *Exerc Sport Sci Rev*. 2015; 43(2):67–74. [PubMed: 25607280]
31. Tinker LF, Sarto GE, Howard BV, et al. Biomarker-calibrated dietary energy and protein intake associations with diabetes risk among postmenopausal women from the Women's Health Initiative. *Am J Clin Nutr*. 2011; 94(6):1600–6. [PubMed: 22071707]
32. Troiano RP, Berrigan D, Dodd KW, Masse LC, Tilert T, McDowell M. Physical activity in the United States measured by accelerometer. *Med Sci Sports Exerc*. 2008; 40(1):181–8. [PubMed: 18091006]

33. Troiano RP, Pettee Gabriel KK, Welk GJ, Owen N, Sternfeld B. Reported physical activity and sedentary behavior: why do you ask? *J Phys Act Health*. 2012; 9(Suppl 1):S68–75. [PubMed: 22287450]
34. Tucker JM, Welk GJ, Beyler NK. Physical activity in U.S.: adults compliance with the Physical Activity Guidelines for Americans. *Am J Prev Med*. 2011; 40(4):454–61. [PubMed: 21406280]
35. Wareham NJ, Jakes RW, Rennie KL, et al. Validity and repeatability of a simple index derived from the short physical activity questionnaire used in the European Prospective Investigation into Cancer and Nutrition (EPIC) study. *Public Health Nutr*. 2003; 6(4):407–13. [PubMed: 12795830]
36. Welk GJ, Kim Y, Stanfill B, et al. Validity of 24-h Physical Activity Recall: Physical Activity Measurement Survey. *Med Sci Sports Exerc*. 2014; 46(10):2014–2024. [PubMed: 24561818]

Figure 1. Scatterplot of observed and model-predicted values for Sedentary Behavior (top) and Moderate-to-Vigorous Physical Activity (bottom) for the respective holdout samples.

Figure 2.

Model-Predicted Values of Objective Sedentary Behavior (top) and Moderate-to-Vigorous Physical Activity (MVPA) (bottom). Note: The solid blue line represents predictions for a normal weight 50-year old female with some college. The solid red line represents predictions for an obese 50-year old female with some college. The dotted blue and red lines represent 95% confidence bands for the predictions. The dashed line is the diagonal identity line.

Table 1

Characteristics of the analytic samples used for developing the prediction model for objectively assessed sedentary behavior (n = 1166) and for model cross-validation (n = 292).

Characteristic	Model Development (%)	Cross-Validation (%)
Gender		
<i>Female</i>	677 (58.1%)	171 (58.6%)
<i>Male</i>	489 (41.9%)	121 (41.4%)
Age (yrs)		
<i>20 – 29</i>	76 (6.5%)	19 (6.5%)
<i>30 – 39</i>	187 (16.0%)	47 (16.1%)
<i>40 – 49</i>	263 (22.6%)	66 (22.6%)
<i>50 – 59</i>	312 (26.8%)	78 (26.7%)
<i>60 – 71</i>	328 (28.1%)	82 (28.1%)
Ethnicity		
<i>White</i>	1028 (88.2%)	261 (89.4%)
<i>Black</i>	98 (8.4%)	23 (7.9%)
<i>Other</i>	40 (3.4%)	8 (2.7%)
BMI		
<i>Normal weight (< 24.9)</i>	272 (23.3%)	64 (21.9%)
<i>Overweight (25.0 – 30.0)</i>	369 (31.7%)	97 (33.2%)
<i>Obese (>30.0)</i>	525 (45.0%)	131 (44.9%)
Education		
<i>High school diploma or lower</i>	331 (28.4%)	83 (28.4%)
<i>Some college</i>	392 (33.6%)	99 (33.9%)
<i>College degree or higher</i>	443 (38.0%)	110 (37.7%)
SB (minutes/day)		
<u><i>SWA:</i></u>		
<i>Mean (SD)</i>	679.5 (191.0)	679.3 (201.5)
<i>Min</i>	87.0	121.5
<i>Max</i>	1286.0	1150.5
<u><i>PAR:</i></u>		
<i>Mean (SD)</i>	456.5 (170.7)	464.7 (178.2)
<i>Min</i>	72.5	70.0
<i>Max</i>	967.5	992.5
MVPA (minutes/day)		
<u><i>SWA:</i></u>		
<i>Mean (SD)</i>	90.0 (99.7)	88.5 (100.8)
<i>Min</i>	0.0	0.0
<i>Max</i>	726.0	573.5

Characteristic	Model Development (%)	Cross-Validation (%)
<i>PAR:</i>		
<i>Mean (SD)</i>	139.8(132.8)	145.7 (144.7)
<i>Min</i>	0.0	0.0
<i>Max</i>	765.0	652.5

Abbreviations: BMI – Body Mass Index; SB – Sedentary Behavior; MVPA – Moderate-to-Vigorous Physical Activity; SWA – SenseWear Armband; PAR – 24-hour Physical Activity Recall

Table 2

Linear Regression Model Coefficients and Standard Errors for Final Prediction Model and Alternative Models with No Education Variable (Separate Models for Sedentary Behavior and Moderate-to-Vigorous Physical Activity)

Response Variable	Model Variable	Final Prediction Model	Alternative Model
SB (SWA)	Intercept	−235.574 (63.702)	−73.510 (48.541)
	<i>PARsed</i>	0.585 (0.088)	0.575 (0.088)
	<i>PARMVPA</i>	−0.296 (0.035)	−0.318 (0.035)
	<i>Female</i>	68.166 (8.318)	67.101 (8.366)
	<i>Age</i>	4.556 (0.825)	2.291 (0.317)
	<i>Education</i>	70.210 (19.734)	NA
	<i>BMI</i>	17.574 (1.383)	17.136 (1.387)
	<i>PARsed x BMI</i>	−0.010 (0.003)	−0.010 (0.003)
	<i>Age x Education</i>	−1.096 (0.383)	NA
MVPA (SWA)	Intercept	13.614 (1.686)	12.504 (1.680)
	<i>PARMVPA (log scale)</i>	0.207 (0.016)	0.211 (0.016)
	<i>PARsed (log scale)</i>	−1.150 (0.268)	−1.117 (0.271)
	<i>Female</i>	0.595 (0.238)	0.537 (0.239)
	<i>Age</i>	−0.028 (0.005)	−0.017 (0.002)
	<i>Education</i>	−0.397 (0.125)	NA
	<i>BMI</i>	−0.199 (0.053)	−0.183 (0.054)
	<i>PARsed x BMI</i>	0.026 (0.009)	0.024 (0.009)
	<i>Age x Education</i>	0.005 (0.002)	NA
	<i>Female x BMI</i>	−0.046 (0.008)	−0.045 (0.008)

Note: Values in parentheses indicate standard errors unless otherwise indicated.

Abbreviations: BMI – Body Mass Index; SB – Sedentary Behavior; MVPA – Moderate-to-Vigorous Physical Activity; SWA – SenseWear Armband; PAR – 24-hour Physical Activity Recall