Reproducibility of objectively measured physical activity: reconsideration needed

Eivind Aadland[1], Ada Kristine Ofrim Nilsen[1], Einar Ylvisåker[1], Kjersti Johannessen[1], Sigmund Alfred Anderssen[1,2]

[1]*Western Norway University of Applied Sciences, Faculty of Education, Arts and Sports, Department of Sport, Food and Natural Sciences, Campus Sogndal, Norway.*

[2]*Norwegian School of Sport Sciences, Department of Sports Medicine, Oslo, Norway.*

Corresponding author

Eivind Aadland

Department of Sport, Food and Natural Sciences, Faculty of Education, Arts and Sports,
Western Norway University of Applied Sciences, Campus Sogndal, Box 133, 6851 Sogndal,
Norway. Phone: +47 5767 6086; Email: eivind.aadland@hvl.no

Word count main text: 3836; word count abstract: 194

Abstract

Reliability of accelerometer-determined physical activity (PA), and thus the required length of a monitoring period, appears to depend on the analytic approach used for its calculation. We compared reliability of objectively measured PA using different resolution of data in a sample of 221 Norwegian 2-6-year-old children providing 2–3 valid 14-day periods of accelerometer monitoring (ActiGraph GT3X+) during September–October, January–February, and May-June 2015–2016. Reliability (intra-class correlation, ICC) was measured for 1–14 days of monitoring across the measurement periods using linear mixed effect modelling. These results were compared to reliability estimated using different resolution of data using the Spearman Brown formula. The measured reliability improved only marginally with increased monitoring length and levelled-off after 5–6 days. Estimated reliability differed substantially when derived from different resolution of data: 3.9–5.4, 6.7–9.2, 13.4–26.7, and 26.3–87.7 days of monitoring was required to achieve an ICC = 0.80 using an hour-by-hour, a day-by-day, a week-by-week, and a period-by-period approach, respectively. Reliability could not be correctly estimated from any single resolution of data. We conclude that reconsideration is needed with regard to how reproducibility of objectively measured PA is analyzed and interpreted.

Keywords: Test-retest; Reliability; Intra-class correlation; Measurement error; Accelerometry

Introduction

Procedures used to analyze accelerometry data and criteria applied to define what constitutes a valid physical activity (PA) measurement varies extensively (1). Because behavior vary greatly over time, an important aspect of accelerometer measurements is how many days or periods of measurement that should be included to obtain reproducible estimates of habitual PA levels. Arguably, the "true" habitual PA level would be superior to a short snapshot, as random error in measurements will increase the likelihood of type II errors and thus invalidate study conclusions (2).

Although findings vary between studies in both adults (3-7) and children (8-20), most evidence suggest that a reasonable reliability (i.e., intra-class correlation (ICC)) of ~ 0.70–0.80 are achieved with 3–7 days of monitoring. However, most previous estimates are derived from the Spearman Brown prophecy formula applied to measurements conducted over a single 7-day period. This procedure estimates the number of measurement periods (usually days) needed to obtain a sufficient reliability level, often considered to be an ICC = 0.80, based on variance components and ICC estimates for a single period. Unfortunately, these study designs have received critique for being likely to underestimate the number of monitoring days needed, and their conclusions should therefore be interpreted with caution (21-24). In comparison, studies that have determined the reliability for several periods of measurement over the course of 2 weeks up to a year, have shown considerable intra-individual variation over time (25, 26, 23, 27, 28, 24). Specifically, studies including several seasons have resulted in reliability estimates of ~ 0.50 for one week monitoring in children (26, 23, 24). These findings agree with studies showing substantial seasonal variation in PA in children and adolescents (29-31), which are obviously not captured when relying on a single measurement period.

Beyond seasonal variation, there is also differences in reliability between the analytic approaches applied (24, 23). When using a day-by-day approach (estimating reliability from single days of measurement), reliability is estimated from a correction of the residual (within-subject) variance by dividing by the number of scores to be averaged (i.e., the number of monitoring "units" (k), for example days, weeks, etc.) (32). This procedure leads to an underestimation of the residual variance compared to actually measured residual variance over the period (24, 23). Aadland et al (24) determined reliability over a week in a large sample of schoolchildren over 2 seasons, and found a systematic underestimation of residual variance and resulting overestimation of ICCs (0.64–0.77 vs. 0.49–0.63; 14 to 31% difference after controlling for season) using the day-by-day compared to a week-by-week approach. This finding is consistent with findings showing that the reliability of different numbers of monitoring hours per day and days per week is rather similar using a week-by-week approach (27, 24, 28), whereas an increased number of monitoring days inherently will improve reliability when estimated over an increased number of days. Thus, there appears to be a difference in reliability depending on whether the number of measurements needed is estimated from single days and then extrapolated (i.e., using a day-by-day approach) or actually measured over several weeks or periods (i.e., using a week-by-week approach). We infer from this finding that the resolution of data might be fundamentally important for determining reliability. Thus, the resolution of data should be systematically altered, including both higher (i.e., using an hour-by-hour approach) and lower resolution (i.e., using week-by-week and period-by-period approaches) than traditionally applied to thoroughly investigate this hypothesis.

The aim of the present study was to extend our previous findings comparing a day-by-day and a week-by-week approach over 2 seasons (24), using a dataset having 2–3 separate 14-day periods of monitoring over different seasons in preschool children. Using different resolution of data, we will compare an hour-by-hour, a day-by-day, a week-by-week, and a period-by-period approach to calculate reliability using the same dataset. We hypothesized that reliability for longer periods (up to 14 days) would be overestimated when estimated from higher resolution data (hour-by-hour and day-by-day) compared to using accumulated data over longer measurement periods (week-by-week and period-by-period).

Methods

Participants

The present analysis is based on data obtained in preschool children from the Sogn og Fjordane Preschool Physical Activity Study (PRESPAS) (33), conducted in Norway during 2015–2016. Physical activity was measured with accelerometry over one 14-day period in 1340 children (September 2015 to June 2016) and over 3 separate 14-day periods in a subsample of 376 children from 3 municipalities (September to October 2015, January to February 2016, and May to June 2016). In the present study, we included all available children for a comparison of "short-term" reliability over 2 consecutive weeks (*cross-sectional sample*), and the subsample having repeated measurements for comparison of "long-term" reliability over 2–3 separate periods of measurement (*longitudinal sample*).

Our procedures and methods conform to ethical guidelines defined by the World Medical Association's Declaration of Helsinki and its subsequent revisions. The Norwegian Centre for Research Data approved the study protocol. We obtained written informed consent from each child's parents or legal guardian prior to all testing.

Procedures

Physical activity was measured using the ActiGraph GT3X+ accelerometer (Pensacola, FL, USA) (34). During all measurements, participants were instructed to wear the accelerometer at all times over 14 consecutive days, except during water activities (swimming, showering) or while sleeping (at night). Units were initialized at a sampling rate of 30 Hz. Files were analyzed at 10 second epochs using the KineSoft analytical software version 3.3.80 (KineSoft, Loughborough, UK). Data was restricted to daytime (i.e., hours 06:00 to 23:59). In all analyses, consecutive periods of ≥ 20 minutes of zero counts were defined as non-wear time (35, 1). Results are reported for overall PA level (cpm), as well as minutes per day spent SED (< 100 cpm), in light PA (LPA) (100–2295 cpm), in moderate PA (MPA) (2296–4011 cpm), in vigorous PA (VPA) (≥ 4012 cpm), and in moderate-to-vigorous PA (MVPA) (≥ 2296 cpm), determined using the previously established and validated Evenson et al cut points (36, 37). Data were analyzed with wear requirements of ≥ 8 hours/day and ≥ 3 weekdays + ≥ 1 weekend day/week for each separate week. We required 2 valid weeks of measurement for the cross-sectional sample and 4–6 valid weeks for the longitudinal sample (≥ 2 periods). As reproducibility is marginally affected by wear hours per day (≥ 6 to ≥ 12 hours/day (27, 24, 28), we did not analyze sensitivity to this wear criteria herein.

Statistical analyses

Children's characteristics were reported as frequencies, means and standard deviations (SD). Differences in PA levels between the 3 measurement periods was tested using a mixed effect model including random intercepts for children and including wear time as a covariate.

We calculated reliability using 4 approaches based on different resolution of data; 1) hour-by-hour, 2) day-by-day, 3) week-by-week, and 4) period-by-period. Approaches 1, 2 and 3 were applied to the cross-sectional dataset, whereas approaches 1, 2, 3, and 4 was applied to the longitudinal dataset. Reliability for single hours (hour-by-hour approach), single days (day-by-day approach), single weeks (week-by-week approach), and single periods (period-by-period approach) of measurement (ICC_s) were calculated using variance partitioning applying a one-way random effect model not controlling for season (i.e., determining agreement based on an absolute definition) in both samples, whereas a two-way mixed effect model controlling for season (i.e., determining agreement based on a consistency definition) additionally were applied in the longitudinal sample (32). All models were adjusted for wear time by adding wear time as a covariate because wear time has a strong association with PA and SED estimates and also impact reliability (28), and since most studies control for wear time.

We directly determined (*"MEASURED"*) reliability for 1–7 monitoring days across 2 consecutive weeks in the cross-sectional dataset (using a week-by-week approach) and for 1–14 monitoring days across 2–3 separate 14-day periods in the longitudinal dataset (using a period-by-period approach). Thus, these analyses is based on the actual variance components for different numbers of monitoring days across weeks and periods. Contrary to this prosedure, we also extrapolated (*"ESTIMATED")* reliability for average measurements (ICC_k = between-subject variance/(between-subject variance + residual variance/k)) and the number of measurements needed using the Spearman Brown prophecy formula ($N = ICC_t/(1-ICC_t)*((1-ICC_s)/ICC_s)$, where ICC_t = the desired level of reliability, and ICC_s = the reliability for single measurement) (3, 32). N was rescaled to days (N_{days}) for ease of comparison across approaches using mean values of wear hours per day (11.8 hours in cross-sectional dataset; 11.7 hours in the longitudinal dataset), wear days per week (6.3 days in both datasets), and wear days per period (12.6 days in the longitudinal dataset only). The number of measurements (N) needed to obtain a reliable measurement were estimated using an ICC_t = 0.80.

In the week-by-week and period-by-period analyses (longitudinal dataset), we additionally calculated 95% limits of agreement (LoA) and coefficients of variation (CV) from the residual variance (i.e., within-subjects) error term based on the variance partitioning models (LoA = √residual variance *√2*1.96; CV = √residual variance/mean values) (38).

All analyses were performed using IBM SPSS v. 24 (IBM SPSS Statistics for Windows, Armonk, NY: IBM Corp., USA). A p-value < .05 indicated statistically significant findings.

Results

Of the 1340 children included in PRESPAS, 1308 children provided accelerometer data for the cross-sectional analyses, of whom 873 children (52% boys) fulfilled the wear criterion for 2 consecutive weeks and were included in the present analysis (Table 1). Of the 376 children included in the longitudinal subsample, 372 provided accelerometer data, of whom 221 (53% boys) had ≥ 2 valid measurement periods and were included in the present analysis.

The longitudinal analyses included 144 children having 2 measurements and 77 children having 3 measurements across seasons. In general, PA levels were highest during the summer and lowest during the winter (Supplemental Table 1). The greatest differences were seen for VPA (up to 67% difference), overall PA (up to 21% difference), and MVPA (up to 13% difference) (all p < .001), whereas smaller and less consistent differences over seasons were found for other intensities.

Reliability across 2 consecutive weeks – cross-sectional sample

Table 2 shows the reliability of single measurements (ICC) and the ESTIMATED number of monitoring days needed to achieve a reliability of 0.80 (N) using an hour-by-hour, a day-by-day, and a week-by-week approach. The 3 approaches relying on different resolution of data yielded different results; whereas 2.2–4.3 days was needed using an hour-by-hour approach, 4.1–7.7 days was needed using a day-by-day approach, and 4.1–14.1 days was needed using a week-by-week approach.

Table 3 shows the MEASURED reliability over an average of 1 to 7 days of monitoring using a week-by-week approach. Although the pattern of improvement was somewhat different across variables, in general, reliability improved up to a number of 5–6 monitoring days, after which reliability levelled off.

Reliability across 2–3 separate 14-day periods – longitudinal sample

Compared to the results shown for 2 consecutive weeks (Table 2 and 3), the reliability decreased and the required number of monitoring days increased when values were estimated and measured over several seasons (Table 4 and 5). Similar to results based on 2 consecutive weeks, different resolution of data yielded substantially different ESTIMATED values (Table 4); whereas 3.9–5.8 days was needed using an hour-by-hour approach, 6.7–10.2 days was needed using a day-by-day approach, 13.4–32.5 days was needed using a week-by-week approach, and 26.3–111.2 days was needed using a period-by-period approach. In contrast to the estimated reliability, MEASURED reliability increased marginally over the first 5–6 days, after which is levelled off (Table 5), similar to the findings in the cross-sectional dataset. Figure 1 shows the estimated (day-by-day, week-by-week, and period-by-period) and measured reliability for 1–14 days of monitoring. The figure shows that the measured reliability is not estimable by the different approaches.

Supplemental Figure 1 shows variance components for MVPA. Compared to actually measured variances, the residual variance is underestimated for long monitoring periods by the day-by-day approach and overestimated for short monitoring periods by the period-by-period approach. Increasing the length of the monitoring period also reduced the measured between-subject variance, whereas between-subject variance is kept constant in estimation models.

Controlling for season had in general a minor influence on the results, although it influenced reliability for overall PA, VPA and MVPA, for which the seasonal differences were most prominent (Table 2).

Agreement for 1 week and 1 period of measurement – longitudinal sample

Supplemental Table 2 shows 95% LoA and CV for 1 week (week-by-week approach) and 1 period (period-by-period approach) of measurement, indicating to what extent these monitoring periods are capable of capturing PA levels representing one-year habitual activity levels (1 out of 4–6 weeks and 1 out of 2–3 periods, respectively). Results were essentially similar for 1 week and 1 period of measurement; CVs were 9–42% across variables, whereas differences up to 332–385 cpm, 91–94 minutes/day of SED, 33–37 minutes/day of MVPA, and 17–22 minutes/day of VPA should be expected between monitoring periods over a year.

Discussion

The present study aimed to determine and compare the reproducibility of accelerometer-determined PA using different analytic approaches based on different resolution of data over the short-term (2 consecutive weeks in the cross-sectional dataset) and long-term (2–3 separate monitoring periods over different seasons in the longitudinal dataset). Our main finding was that reliability of PA as a function of monitoring length, and thus the required number of monitoring days, is not estimable by extrapolation using any single resolution of data. Our findings show that estimation of reliability applying the much-used Spearman Brown formula is invalid, and that reconsideration is needed with respect to the analysis and interpretation of reliability of accelerometry-derived PA measurements.

Most previous studies investigating reliability and the required number of accelerometer monitoring days have estimated reliability based on day-by-day analyses using a single 7-day monitoring period (8, 13, 14, 38, 15, 16, 19, 17, 18, 9-12). In general, these studies conclude that 3–7 monitoring days are sufficient in children. In contrast, studies comparing several monitoring periods captured over different seasons, have yielded substantially lower reliability estimates in both adults (25) and children (26, 23, 24), concluding that longer and/or several monitoring periods is needed. Mattocks (26) determined reliability over 4 separate 7-day periods over approximately one year using the Actigraph 7164 accelerometer in 11–12-year-old children and found a reliability of 0.45 to 0.59 across variables. Similarly, Wickel & Welk (23) found an ICC of 0.46 over 3 separate 7-day periods to assess steps for the Digiwalker pedometer in 10-year-old children. Finally, Aadland et al (24) found a reliability of 0.29–0.67 across 2 separate periods 3–4 months apart using the Actigraph GT3X+ accelerometer in a large sample of 10-year-old children.

The reliability estimates based on a single monitoring period versus several separate periods differ in 2 important ways. Obviously, separate periods are based on measurements collected over a longer time frame, possibly influenced by seasonality, which increase the likelihood of capturing changes in individuals' PA levels over time. These changes over time also cause differences in variance between the monitoring periods, which will attenuate ICCs as the model assumes compound symmetry and the ICC are sensitive to asymmetry (24, 32). Moreover, the statistical analyses are based on different resolution of data; a day-by-day approach for single period data and a week-by-week approach for multiple (weeklong) periods of data. Our results suggest both these differences are influential for the resulting reliability. First, comparable analytic approaches led to lower reliability in the longitudinal dataset than in the cross-sectional dataset (mean ESTIMATED ICC = 0.07 vs. 0.10 for an hour; 0.33 vs. 0.42 for a day, 0.57 vs. 0.77 for a week, respectively; mean MEASURED ICC = 0.51 vs. 0.77 for a week, respectively). These findings show that reliability decreases when more variation is added to the data when capturing a longer time frame with greater variation in behavior. Thus, our findings show that long-term reliability is underestimated when estimated from a single short measurement period. Even more important, our findings suggest that different resolution of data has a major influence on reliability estimates. Adding to the day-by-day (8, 13, 14, 38, 15, 16, 19, 17, 18, 9-12) and the week-by-week approach (26, 23-25) as applied previously, we extended our analysis to include data using higher (hour-by-hour) and lower (period-by-period) resolution, to obtain an even better picture of how data resolution influence reliability. These approaches led to substantially different reliability estimates and numbers of required monitoring days in both samples, particularly in the longitudinal dataset where the number of monitoring days to achieve an ICC = 0.80 based on the hour-by-hour and period-by-period approach varied from (mean) 4.5 to 49.7 days.

The differing findings among the analytic approaches based on differing resolution of data result from erroneous estimation of variance components across resolutions. The ICC is calculated from these variance components, which will vary by resolution. The estimated ICC using the Spearman Brown formula will thus be fully dependent on their correct estimation across different resolutions to obtain correct reliability estimates. However, compared to actually measured variances, the residual variance is underestimated for long monitoring periods using high-resolution data and overestimated for short monitoring periods using low-resolution data. Moreover, whereas between-subject variance is kept constant in estimation models, it decreased when stability of data improved over a longer monitoring period. To this end, both variance components underlying the resulting reliability was erroneously estimated compared to those measured when including 1–7 (cross-sectional dataset) and 1–14 (longitudinal dataset) monitoring days in a week-by-week and period-by-period analysis, respectively. These results shows that the correct variance components and thus reliability of objectively measured PA as a function of monitoring length is not estimable from any single resolution of data.

Previous studies using long-term measurements (i.e., more than a week) have suggested that periods longer than a week and/or several periods are necessary to determine PA reliably (25, 27, 24, 28, 26, 23). The findings herein are consistent with these studies in terms of the modest long-term reliability found for a single week (ICC = 0.35–0.64, mean 0.51) and period (ICC = 0.36–0.66, mean 0.52) of measurement. Taken together, our findings and those of others using several separate monitoring periods suggest a typical 3–7-day period of accelerometer monitoring result in a reliability of 0.29–0.67 across variables in children (23, 24, 26). Of great importance though, reliability did not improve beyond 5–6 days when measured over 1–14 days. This pattern contrasts reliability estimates derived from the Spearman Brown formula/ICC for average measurements, which are inherently predicted to improve when the number of measurements increase. Thus, our findings indicate a single 7-day measurement protocol would be the best choice in future research, as it maximize reliability and minimize participant and researcher burden. This recommendation is also in line with the results shown for agreement (LoA and CVs), which was similar for a 7-day and a 14-day period. Reliability could possibly be increased by including several separate monitoring periods for each individual, but such an approach would clearly be less feasible for participants as well as researchers.

As noise in exposure (x) variables will lead to attenuation of regression coefficients (regression dilution bias), and noise in outcome (y) variables will increase standard errors (2), unreliable measures weaken researchers ability to make valid conclusions in epidemiology. We argue that, in most cases, researchers are interested in the long-term “true” habitual PA level, rather than activity during the most recent days. Although some health characteristics, as for example insulin resistance, lipid metabolism and blood pressure, might change with acute increases or decreases in PA (40), a child’s level of fatness, aerobic fitness, or motor skills takes months or years to develop. For such stable traits, association analyses (using PA as an exposure variable) will inherently suffer from regression dilution bias if relying on an insufficient snapshot of children’s habitual activity level. For studies evaluating intervention effects (using PA as the outcome variable), low reliability will decrease power. Thus, in both situations, low reliability increase the likelihood of type II errors (2).

Strengths and limitations

The main strength of the present study is the inclusion of a large and representative sample of children and the use of 2 different datasets (cross-sectional and longitudinal) in which 14-day monitoring where used throughout. This allowed for calculation of short- (2 consecutive weeks) and long-term (3 seasons separated by approximately 9 months) reliability using different resolution of data. As reliability estimates depend on the sample variation (41, 38), the validity of the estimated ICCs presented herein should be generalizable to other contexts, including large-scale population studies. Importantly, the use of 14-day monitoring periods allowed for calculation of actual variance components for accumulation of 1–7 and 1–14 days of measurement over the short- and long-term, respectively, and the comparison of these measurements with estimation and extrapolation of these variance components across different periods. Thus, our findings extend those of Aadland et al (24), who directly compared the reliability of children’s objectively measured PA using a day-by-day and a week-by-week approach. Importantly, the hour-by-hour approach was included only to test the hypothesis that reliability improved with higher resolution; we find this approach of little practical importance for researchers.

Norway has profound seasonal differences in weather conditions and daylight, which may cause changes in PA levels and types across measurement periods. These characteristics might limit generalizability to areas with less pronounced seasonality. Still, as discussed above, our findings are consistent with previous studies when comparing similar approaches for determination of reliability (8, 13, 14, 38, 15, 16, 19, 17, 18, 9-12, 24, 26, 23). Importantly, this seasonal variation will not influence the comparison across the different analytic approaches, as they are based on the same underlying data. Finally, we could have extended our findings by reporting variance partitioning of multiple components (e.g., participant, day, and season) as shown previously (23), however, such analyses was out of scope for the present paper.

Conclusion

We conclude that reliability of objectively measured PA as a function of monitoring length, and thus the required number of monitoring days, is not estimable by extrapolation using any single resolution of data. Our findings suggest the estimation of reliability applying the much-used Spearman Brown formula to a day-by-day approach provide overly optimistic reliability estimates and is invalid for estimating reliability over multiple days or periods. Hence, we caution against this practice and recommend future studies measure reliability over separate monitoring periods. Nevertheless, because our results show that reliability levels off after 5–6 monitoring days, they support the use of a 7-day measurement protocol. However, the long-term reliability for this protocol in terms of representing the habitual PA level of children across an extended period, is considerably lower than estimated by most previous studies (mean ICC = 0.51–0.52 for 7–14 days of monitoring). These findings strongly indicate reconsideration is needed with respect to the design, analysis, and interpretation of reliability of accelerometry-derived PA measurements.

Acknowledgement

We thank all children, parents and staff at the participating preschools for their excellent cooperation during the data collection. We also thank the Masters students Katrin Oppheim, Janne Øvstetun, Oddbjørn Holmøy, and Edvard Veivåg at the Western Norway University of Applied Sciences (formerly Sogn og Fjordane University College) for their assistance during the data collection.

Funding

The study was funded by the County Municipality of Sogn og Fjordane and the Western Norway University of Applied Sciences (formerly Sogn og Fjordane University College). None of the funding agencies had any role in the study design, data collection, analyzing or interpreting data, or in writing the manuscripts.

Competing interests

The authors declare that they have no competing interests.

Data availability

The datasets used in the present study are available from the corresponding author on reasonable request.

References

1. Cain KL, Sallis JF, Conway TL, Van Dyck D, Calhoon L. Using Accelerometers in Youth Physical Activity Studies: A Review of Methods. J Phys Act Health. 2013;10(3):437-50.

2. Hutcheon JA, Chiolero A, Hanley JA. Random measurement error and regression dilution bias. BMJ. 2010;340. doi:10.1136/bmj.c2289.

3. Trost SG, McIver KL, Pate RR. Conducting accelerometer-based activity assessments in field-based research. Med Sci Sports Exerc. 2005;37(11):S531-S43. doi:10.1249/01.mss.0000185657.86065.98.

4. Jerome GJ, Young DR, Laferriere D, Chen CH, Vollmer WM. Reliability of RT3 Accelerometers among Overweight and Obese Adults. Med Sci Sports Exerc. 2009;41(1):110-4. doi:10.1249/MSS.0b013e3181846cd8.

5. Coleman KJ, Epstein LH. Application of generalizability theory to measurement of activity in males who are not regularly active: A preliminary report. Res Q Exerc Sport. 1998;69(1):58-63.

6. Matthews CE, Ainsworth BE, Thompson RW, Bassett DR. Sources of variance in daily physical activity levels as measured by an accelerometer. Med Sci Sports Exerc. 2002;34(8):1376-81.

7. Hart TL, Swartz AM, Cashin SE, Strath SJ. How many days of monitoring predict physical activity and sedentary behaviour in older adults? Int J Behav Nutr Phys Act. 2011;8. doi:10.1186/1479-5868-8-62.

8. Basterfield L, Adamson AJ, Pearce MS, Reilly JJ. Stability of Habitual Physical Activity and Sedentary Behavior Monitoring by Accelerometry in 6-to 8-Year-Olds. J Phys Act Health. 2011;8(4):543-7.

9. Addy CL, Trilk JL, Dowda M, Byun W, Pate RR. Assessing Preschool Children's Physical Activity: How Many Days of Accelerometry Measurement. Pediatr Exerc Sci. 2014;26(1):103-9. doi:10.1123/pes.2013-0021.

10. Hinkley T, O'Connell E, Okely AD, Crawford D, Hesketh K, Salmon J. Assessing Volume of Accelerometry Data for Reliability in Preschool Children. Med Sci Sports Exerc. 2012;44(12):2436-41. doi:10.1249/MSS.0b013e3182661478.

11. Hislop J, Law J, Rush R et al. An investigation into the minimum accelerometry wear time for reliable estimates of habitual physical activity and definition of a standard measurement day in pre-school children. Physiol Meas. 2014;35(11):2213-28. doi:10.1088/0967-3334/35/11/2213.

12. Penpraze V, Reilly JJ, MacLean CM et al. Monitoring of physical activity in young children: How much is enough? Pediatr Exerc Sci 2006;18(4):483-91.

13. Ojiambo R, Cuthill R, Budd H et al. Impact of methodological decisions on accelerometer outcome variables in young children. Int J Obes. 2011;35:S98-S103. doi:10.1038/ijo.2011.40.

14. Rich C, Geraci M, Griffiths L, Sera F, Dezateux C, Cortina-Borja M. Quality Control Methods in Accelerometer Data Processing: Defining Minimum Wear Time. Plos One. 2013;8(6). doi:10.1371/journal.pone.0067206.

15. Kang M, Bassett DR, Barreira TV et al. How Many Days Are Enough? A Study of 365 Days of Pedometer Monitoring. Res Q Exerc Sport. 2009;80(3):445-53.

16. Murray DM, Catellier DJ, Hannan PJ et al. School-level intraclass correlation for physical activity in adolescent girls. Med Sci Sports Exerc. 2004;36(5):876-82. doi:10.1249/01.mss.0000126806.72453.1c.

17. Treuth MS, Sherwood NE, Butte NF et al. Validity and reliability of activity measures in African-American girls for GEMS. Med Sci Sports Exerc. 2003;35(3):532-9. doi:10.1249/01.mss.0000053702.03884.3f.

18. Trost SG, Pate RR, Freedson PS, Sallis JF, Taylor WC. Using objective physical activity measures with youth: How many days of monitoring are needed? Med Sci Sports Exerc. 2000;32(2):426-31. doi:10.1097/00005768-200002000-00025.

19. Janz KF, Witt J, Mahoney LT. The stability of childrens physical-activity as measured by accelerometry and self-report. Med Sci Sports Exerc. 1995;27(9):1326-32.

20. Bisson M, Tremblay F, Pronovost E, Julien A, Marc I. Accelerometry to measure physical activity in toddlers: Determination of wear time requirements for a reliable estimate of physical activity. J Sport Sci. 2019;37(3):298-305. doi:10.1080/02640414.2018.1499391.

21. Baranowski T, Masse LC, Ragan B, Welk G. How many days was that? We're still not sure, but we're asking the question better! Med Sci Sports Exerc. 2008;40(7):S544-S9. doi:10.1249/MSS.0b013e31817c6651.

22. Matthews CE, Hagstromer M, Pober DM, Bowles HR. Best Practices for Using Physical Activity Monitors in Population-Based Research. Med Sci Sports Exerc. 2012;44:S68-S76. doi:10.1249/MSS.0b013e3182399e5b.

23. Wickel EE, Welk GJ. Applying Generalizability Theory to Estimate Habitual Activity Levels. Med Sci Sports Exerc. 2010;42(8):1528-34. doi:10.1249/MSS.0b013e3181d107c4.

24. Aadland E, Andersen LB, Skrede T, Ekelund U, Anderssen SA, Resaland GK. Reproducibility of objectively measured physical activity and sedentary time over two seasons in children; Comparing a day-by-day and a week-by-week approach. Plos One. 2017;12(12). doi:10.1371/journal.pone.0189304.

25. 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.

26. Mattocks C, Leary S, Ness A et al. Intraindividual variation of objectively measured physical activity in children. Med Sci Sports Exerc. 2007;39(4):622-9. doi:10.1249/mss.0b013e318030631b.

27. Aadland E, Johannessen K. Agreement of objectively measured physical activity and sedentary time in preschool children. Prev Med Reports. 2015;2:635-9.

28. Aadland E, Ylvisåker E. Reliability of Objectively Measured Sedentary Time and Physical Activity in Adults. Plos One. 2015;10(7):1-13. doi:10.1371/journal.pone.0133296.

29. Atkin AJ, Sharp SJ, Harrison F, Brage S, Van Sluijs EMF. Seasonal Variation in Children's Physical Activity and Sedentary Time. Med Sci Sports Exerc. 2016;48(3):449-56. doi:10.1249/mss.0000000000000786.

30. Gracia-Marco L, Ortega FB, Ruiz JR et al. Seasonal variation in physical activity and sedentary time in different European regions. The HELENA study. J Sports Sci. 2013;31(16):1831-40. doi:10.1080/02640414.2013.803595.

31. Ridgers ND, Salmon J, Timperio A. Too hot to move? Objectively assessed seasonal changes in Australian children's physical activity. Int J Behav Nutr Phys Act. 2015;12. doi:10.1186/s12966-015-0245-x.

32. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychol Methods. 1996;1(1):30-46. doi:10.1037/1082-989x.1.4.390.

33. Nilsen AKO, Anderssen SA, Ylvisåker E, Johannessen K, Aadland E. Physical activity among Norwegian preschoolers varies by sex, age, and season. Scand J Med Sci Sports. 2019;29:862-73. doi:10.1111/sms.13405

34. John D, Freedson P. ActiGraph and Actical physical activity monitors: a peek under the hood. Med Sci Sports Exerc. 2012;44(1 Suppl 1):S86-S9.

35. Esliger DW, Copeland JL, Barnes JD, Tremblay MS. Standardizing and Optimizing the Use of Accelerometer Data for Free-Living Physical Activity Monitoring. J Phys Act Health. 2005;2(3):366-83.

36. Evenson KR, Catellier DJ, Gill K, Ondrak KS, McMurray RG. Calibration of two objective measures of physical activity for children. J Sports Sci. 2008;26(14):1557-65. doi:10.1080/02640410802334196.

37. Trost SG, Loprinzi PD, Moore R, Pfeiffer KA. Comparison of Accelerometer Cut Points for Predicting Activity Intensity in Youth. Med Sci Sports Exerc. 2011;43(7):1360-8. doi:10.1249/MSS.0b013e318206476e.

38. Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. J Strength Cond Res. 2005;19(1):231-40.

39. Chinapaw MJM, de Niet M, Verloigne M, De Bourdeaudhuij I, Brug J, Altenburg TM. From Sedentary Time to Sedentary Patterns: Accelerometer Data Reduction Decisions in Youth. Plos One. 2014;9(11). doi:10.1371/journal.pone.0111205.

40. Thompson PD, Crouse SF, Goodpaster B, Kelley D, Moyna N, Pescatello L. The acute versus the chronic response to exercise. Med Sci Sports Exerc. 2001;33(6 Suppl):S438.

41. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;1(8476):307-10.

Figure legends

Figure 1. Measured and estimated reliability for MVPA over 1–14 monitoring days across 3 seasons. The measured reliability is calculated using a period-by-period approach by accumulating and averaging MVPA over 1–14 monitoring days for each period, thus, the model is based on actual variances. The estimated reliability is calculated for 1 day (day-by-day approach), 1 week (week-by-week approach), and 1 period (period-by-period approach) and extrapolated over k days. All results are based on reliability estimates for a two-way mixed model controlling for season (i.e., a consistency definition of reliability) in addition to wear time.

Supplemental Figure 1. **Measured and estimated variance components for MVPA over 1–14 monitoring days across 3 seasons**. The between-subject variance is the part of the variance explained by subjects (“true” variation), whereas the residual variance is the unexplained variance (within-subjects variance or error).

Supplemental Table 1. Physical activity levels over 3 seasons (longitudinal sample, n = 221).

Supplemental Table 2. 95% limits of agreement and coefficients of variation for 1 week and 1 period of measurement.