Behavior Research Methods
2008, 40 (3), 873-878
doi: 10.3758/BRM.40.3.873

Actigraph data are reliable, with functional reliability increasing with aggregation

Alexis C. Wood, Jonna Kuntsi, and Philip Asherson
Kings College London, London, England

and

Kimberly J. Saudino
Boston University, Boston, Massachusetts

Motion sensor devices such as actigraphs are increasingly used in studies that seek to obtain an objective assessment of activity level. They have many advantages, and are useful additions to research in fields such as sleep assessment, drug efficacy, behavior genetics, and obesity. However, questions still remain over the reliability of data collected using actigraphic assessment. We aimed to apply generalizability theory to actigraph data collected on a large, general-population sample in middle childhood, during 8 cognitive tasks across two body loci, and to examine reliability coefficients on actigraph data aggregated across different numbers of tasks and different numbers of attachment loci. Our analyses show that aggregation greatly increases actigraph data reliability, with reliability coefficients on data collected at one body locus during 1 task (.29) being much lower than that aggregated across data collected on two body loci and during 8 tasks (.66). Further increases in reliability coefficients by aggregating across four loci and 12 tasks were estimated to be modest in prospective analyses, indicating an optimum trade-off between data collection and reliability estimates. We also examined possible instrumental effects on actigraph data and found these to be nonsignificant, further supporting the reliability and validity of actigraph data as a method of activity level assessment.

The difficulties with obtaining an objective assessment of children's activity levels have long been recognized (Teicher, Ito, Glod, & Barber, 1996). Parent and teacher reports of children's activity levels are commonly used as activity level measures in both clinics and research, but there is awareness of possible biases and errors (Saudino, Ronald, & Plomin, 2005; Thapar, Harrington, Ross, & McGuffin, 2000; Verhulst, Achenbach, Althaus, & Akkerhuis, 1988) such as contrast effects (Eaves et al., 2000; Saudino, Cherny, & Plomin, 2000; Simonoff et al., 1998) or halo effects (Abikoff, Courtney, Pelham, & Koplewicz, 1993; Schachar, Sandberg, & Rutter, 1986; Stevens, Quittner, & Abikoff, 1998). Although observer-rated data may be considered more objective than parent or teacher reports, the expense and practicalities of obtaining such data have made them difficult to use in large-scale studies and unsuitable for long-term field studies of activity level. Motion sensor devices, such as actigraphs, provide a potential way around these pitfalls, by providing objective, technologically simple activity level data (Eaton, McKeen, & Saudino, 1996), in a method that can be used over large periods of time and has been shown to be readily accepted by the majority of young people (Van Coevering et al., 2005). Despite useful additions to research in varied fields such as sleep assessment, drug efficacy, behavior genetics, and obesity, the validity and reliability of actigraphic assessment has been queried (Jean-Louis et al., 1997). Empirical research has largely validated the use of actigraphy as a method of activity level assessment, showing significant correlations with activity level assessments by methods such as room respiration calometry (Puyau, Adolph, Vohra, & Butte, 2002), oxygen consumption over activities of varying intensity (Treuth et al., 2004), spinners and precision pendulums (Tryon, 2005), and questionnaire data (Saudino, Wertz, Gagne, & Chawla, 2004). A comparison study found higher correlations for three actigraphy monitors and treadmill pace, but found that the CSA Monitor (Computer Science and Applications, Inc.) was the most accurate at predicting energy expenditure as measured by indirect calorimetry, compared with the Tritac and Biotrainer systems (Welk, Blair, Wood, Jones, & Thompson, 2000).

However, there are still methodological issues to be addressed regarding the use of motion sensor data. Although the validity of actigraphs as a method of activity level data has been addressed in the above studies, the reliability of the actigraph method is still open to question (Sadeh, Sharkey, & Carskadon, 1994). Reliability studies have largely focused on sleep–wake identification (see Cole, Kripke, Gruen, Mullaney, & Gillin, 1992, for a review), with little data on the reliability of actigraphy assessment

A. C. Wood, alexis.wood@iop.kcl.ac.uk

 Copyright 2008 Psychonomic Society, Inc.

in ascertaining an assessment of overall activity level, although a study used generalizability theory to show that CSA accelerometers have the lowest error variance when compared with Biotraner Pro, Titrac, and Actical machines (Welk, Schaben, & Morrow, 2004). However, few studies have focused on how to quantify and maximize the reliability of actigraph data.

In his seminal paper "Aggregation and Beyond," Epstein (1983) discussed issues surrounding the reliability of all behavioral data, highlighting the need to minimize situation-specific influences on behavior measurement, and the role of aggregation across measurement occasions in achieving this. He defined aggregation as "a basic procedure for reducing error of measurement and for enhancing and establishing the range of generality of a phenomenon by averaging over many measurements" (p. 367), in particular, by canceling out—or at least reducing—instances of unrepresentative behavior within the measurement of a phenotype. For data aggregation to be appropriate, the measurements must measure the same concept and have a common variance. Aggregation over measurement occasions is an important concept for increasing the reliability of data, and aggregated summed components are generally expected to have better reliability than single variables are (Rousson, Gasser, & Seifert, 2002); combining scores across theoretically related cognitive variables has been shown to increase data reliability and decrease error variance (Kuntsi et al., 2006).

One of the advantages of motion sensor data is that they facilitate aggregation over a large sample (Eaton et al., 1996). However, given that children's activity levels can be highly variable, further aggregation—such as across measurements and collecting data across several periods—can be useful to minimize momentary or unrepresentative influences. This may be particularly important for fields such as behavior genetics, where it is not the mean score over the sample that is important but the correlation between individuals. Earlier work has highlighted the importance of aggregating actometer data across instruments (Eaton, 1983) and limbs (Eaton et al., 1996), to reduce behavioral variability and produce the most reliable score of overall activity level, and these studies have also speculated that aggregation over longer data collection periods such as 24–48 h, would increase functional reliability over shorter periods of 15 min (Eaton et al., 1996). This idea was supported in one actigraph study, which reported an increase in the reliability of actigraph data collected over three 5-min bouts of treadmill walking over that collected during one (Welk et al., 2004). However, there is a lack of data on this latter issue for actigraph data not collected during bouts of regulated physical activity, when the demands of different tasks may elicit differing activity levels (Dane, Schachar, & Tannock, 2000). Nor is there available data on the suitability of aggregation across limbs for actigraphs themselves, which record different information than that of the actometers used by Eaton and colleagues.

Classical test theory, item response theory, and generalizability theory are three major ways to assess reliability. However, the former two approaches consider only one source of measurement error at a time; nor do they provide an overall estimate of reliability, or explicit information about how many extra measurements would be needed to obtain a specific reliability coefficient in future studies (Mushquash & O'Connor, 2006). Generalizability, sometimes called *G theory*, allows the estimation of variance due to person, facet, and residual variance components (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). Using available data, a G study estimates reliability coefficients for different numbers of measurements. Further, prospective analyses can estimate the number of measurements needed in future studies to achieve a chosen reliability coefficient.

We aimed to examine whether there is justification for aggregation across body loci and tasks for actigraph data, and, if so, whether aggregation results in an increase in reliability by using G theory to calculate reliability coefficients for different levels of aggregation. Furthermore, we aimed to examine whether effects of remaining battery life, intermachine differences, and research assistant differences have a significant effect on the data collected.

METHOD

Sample

Participants were members of the Study of Activity and Impulsivity Levels in Children (SAIL), a study of a general population sample of twins between the ages of 7 and 9. The sample was recruited from a birth cohort study, the Twins Early Development Study (TEDS; Trouton, Spinath, & Plomin, 2002), which had invited parents of all twins born in England and Wales between 1994 and 1996 to enroll. Despite attrition, the TEDS families continue to be representative of the U.K. population with respect to parental occupation, education, and ethnicity (Spinath & O'Connor, 2003).

Families on the TEDS register were invited to take part if they fulfilled the following SAIL project inclusion criteria: their twins' birthdates were between September 1, 1995, and December 31, 1996; they lived within feasible traveling distance of the research center (return day trip); their ethnic origin was white European (to reduce population heterogeneity for molecular genetic studies); they had recently participated in TEDS, as indicated by return of questionnaires at either the 4- or the 7-year data collection point; they had no extreme pregnancy or perinatal difficulties (15 pairs excluded), specific medical syndromes, chromosomal anomalies (2 pairs excluded), or epilepsy (1 pair excluded); they were not participating in other current TEDS substudies (45 pairs excluded); and they were not on stimulants or other neuropsychiatric medications (2 pairs excluded).

The present analyses focus on data obtained following contact with the first 693 suitable families on the register. Of these, 400 families agreed to participate, reflecting a participation rate of 58%. Of the 800 participants, data from 9 individual children were excluded from analyses on the SAIL dataset (5 children with IQs below 70, and 1 child because of each of the following: neurofibromatosis, epilepsy, hypothyroidism, and ADHD with scioptic tendencies). A further 3 children were excluded because of difficulties during test sessions that inappropriately affected the data (e.g., playing with the actigraph). Actigraph data were therefore available for 797 children; but of these, 109 had incomplete data and were therefore excluded from the present analyses. This was a necessary step for the generalizability analyses and arose either as a mechanical failure or as a testing-related issue (such as a specific computer failure leading to nonadministration of a task). This left for the present analyses a final sample of 598 children, mean age 8.36 years ($SD = 0.26$). All participants gave informed consent, and the study was approved by the Institute of Psychiatry ethical committee (approval no. 286/01).

Measures

Actigraph measurements. The families visited the research center for the actigraph assessments (for further details, see Wood,

Saudino, Rogers, Asherson, & Kuntsi, 2007). The actigraph readings used in the present analyses are taken from a laboratory-based test session, when the twins were apart completing a short-form IQ test and several cognitive–experimental tasks, under the supervision of separate experimenters who administered standardized instructions. The total length of the testing session was approximately 2 h, excluding a 25-min unstructured break approximately halfway through the session. The children completed 4 separate tasks, with differing task conditions. The children completed all tasks while seated; no task required any movement other than arm movement, usually computer mouse control. However, for the block design subtest in the Wechsler Intelligence Scale for Children (Wechsler, 1991), the children manipulated objects on the desk. Each task condition was separated by a 3- to 4-min break, during which instructions for the next task were given and prizes awarded. For the purposes of simplicity, therefore, as far as these analyses are concerned, each condition is treated and referred to as a separate "task." The children completed the following four tasks:

1. Vocabulary, similarities, picture completion, and block design tests were completed from Wechsler Intelligence Scales for Children (Wechsler, 1991);

2. The go/no-go task (Börger & van der Meere, 2000; Kuntsi, Andreou, Ma, Börger, & van der Meere, 2005; van der Meere, Stemerdink, & Gunning, 1995), which instructed children to respond by pressing a computer key to an onscreen "go" stimulus but not respond to an onscreen "stop" stimulus under three conditions distinguished by different interstimuli intervals (see Kuntsi et al., 2006, for further details);

3. The Fast Task (Kuntsi et al., 2005), a standard four-choice reaction time (RT) task under two conditions: a baseline and a fast incentive, where the latter condition was distinguished by a faster event rate and the addition of smiley-face incentives;

4. The Maudsley index of childhood delay aversion (Kuntsi, Stevenson, Oosterlaan, & Sonuga-Barke, 2001), a computer-based task which asked children to choose between a small immediate reward and a large delayed reward, conducted under two conditions: no postreward delay where the smaller reward led immediately to the next trial, and a postreward delay when choosing the small reward rather than the larger one led to a decrease in delay time. All tasks are discussed in more detail in Kuntsi et al. (2006).

The children wore two actigraphs, each slightly larger than a watch (MTI Health Services, version 323, Health One Technology), one on the dominant leg (established by asking with which leg they would kick a ball and start upstairs on), and one on the waist. These attachment loci were chosen to minimize the relationship between actigraph data and task performance. MTI actigraphs have been shown to have the least variability across overall G coefficients and the highest reliability compared with other personal motion sensors (Welk et al., 2004). These devices contain accelerometer technology, which records the number of movements as well as the cumulative magnitude. The actigraph data output was set to readings per minute, measured in gravitational acceleration (G) units, a standard measure of acceleration. This acceleration is then sampled and digitized by a 12-bit analog-to-digital converter and passed through a digital filter, which band-limits the accelerometer to 0.25–2.5 Hz. This was selected to detect normal human activity, while rejecting motion from other sources.

In all analyses, we obtained an average reading by dividing the cumulative magnitude by the number of minute readings; this removed the effect of time, since some children spent longer in some conditions than did others. Actigraph data collected during the eight "tasks" were used in the generalizability analyses. There was an average task length of 10.60 min, with a range of 8–14 min.

Analyses

All analyses were conducted using Stata statistical software release 9.2 (Stata Corporation, College Station, TX), with the exception of the generalizability analyses, which were conducted with SPSS version 14.0 (SPSS, Inc.) using an adaptation of a syntax script provided by Mushquash and O'Connor (2006). Log transformations were applied to data (optimized minimal skew through the lnskew0 command in Stata version 9.1) to normalize skewed distributions. Since the data were collected on a twin sample, analyses were conducted using the "cluster" command in Stata to control for the genetic relationship between members of the sample (Armitage, Berry, & Matthews, 2001). Where this was not possible—for example, in the generalizability analyses and the principal components analyses—analyses were conducted separately on Twin 1 and Twin 2, where the assignation to Twin 1 or Twin 2 status was random, to control for the nonindependence of the data. For these analyses, analyses for Twin 1 and Twin 2 yielded a similar pattern of results, so data from Twin 1 only is presented here (Twin 2's data are available from the first author on request).

Principal components analyses. To investigate whether there is shared variance underlying individual task data, a principal components analysis was run on task level data to see how many latent components were extracted from measurements, and to see whether all measurements loaded onto a shared latent component(s). This should indicate whether, as required for aggregation, all measurements do in fact share a common variance, and are, therefore, likely to measure the same concept (Epstein, 1983).

Generalizability analyses. Generalizability analyses reveal and compare the sources of variance in a common metric (Mushquash & O'Connor, 2006), decomposing the variance into person-specific variance, facet-specific variance, and error variance. This makes G theory preferable to true-score theory, but following the advice of O'Brien (1995), G coefficients are here called *reliability coefficients*, a more familiar term, given that reliability and G coefficients are analogous. In these analyses, the common metric is actigraph data and the number of tasks and number of body loci are the two facets. A fully crossed, two-facet design was used that estimated person, task, body loci, and residual (error) variance. Using these component variances, the reliabilities across different levels of facet can be estimated—that is, the reliability of actigraph data collected from, and averaged across, both different numbers of tasks and different numbers of body loci. We use Cicchetti's (1994) differentiation for interjudge reliability coefficients of a clinical significance where $<.40$ indicates "poor" reliability; .4 to $<.6$, "fair" reliability; .6 to $<.74$, "good" reliability; and $>.75$, "excellent" reliability.

The analyses are presented from two applications of G theory. The first set is a "G study" which presents reliability coefficients for the data collected. The second set of analyses relate to a "D study" where, using G theory, prospective analyses can indicate what reliability coefficients would be in future studies for numbers and combinations of facets not collected as part of the G study. In this case, we present reliability coefficients for a D study in which the number of body loci is extended to four and the number of tasks is extended to 12.

Assessing situational effects on actigraph data. The effect of individual differences between the two experimenters administering the tasks and intermachine effects on the data were assessed through a regression model on which the assumption of nonindependence data was relaxed.

RESULTS

Increasing Reliability Through Task Aggregation

Shared-variance analyses. For the leg data, a principal components analysis showed that one major factor accounted for 58% of the variance between the actigraph data collected during the 8 "tasks" (eigenvalue, 4.62). No other clear factors emerged. Between 40% and 69% of the variance in each task measurement was explained by this one factor. As such, there was shared common variance between the actigraph data for all 8 tasks, and none of the tasks contributed significantly to a unique aspect of actigraph data.

For the waist data, once again, only one clear factor emerged (eigenvalue, 4.42), explaining 55% of the variance between task level actigraph data. Between 41% and

Table 1
Reliability Correlations for Differing Levels of Aggregation Across Different Numbers of Tasks, and Data Collected From Different Numbers of Body Loci

Body Loci in Aggregation	Tasks in Aggregation											
	1	2	3	4	5	6	7	8	9	10	11	12
1	.29	.39	.43	.46	.48	.50	.51	.52	*.52*	*.53*	*.53*	*.54*
2	.39	.51	.57	.60	.62	.64	.65	.66	*.67*	*.67*	*.68*	*.68*
3	*.43*	*.57*	*.63*	*.66*	*.70*	*.71*	*.72*	*.73*	*.74*	*.74*	*.75*	*.75*
4	*.47*	*.60*	*.67*	*.70*	*.73*	*.75*	*.76*	*.77*	*.77*	*.78*	*.79*	*.79*

Note—Data collected on Twin 1 only. Data in normal typeface are calculated from available data. Data in italic typeface are calculated using prospective analyses based on G theory.

68% of the variance in each task level actigraph set of data were explained by the extracted factor. Similarly, for data averaged across all eight tasks, a principal components analysis extracted one component that explained 74% of the variance shared between leg and waist data for Twin 1 (eigenvalue, 1.48). This suggests that both body loci share a common variance, and that no single locus contributes significantly to a unique aspect of actigraph data.

Generalizability analyses. Reliability coefficients for aggregated data are presented in Table 1. Using data collected from one locus, the G study showed that reliability coefficients increased from .29 to .52 as data were aggregated across 1 and 8 tasks, respectively (Table 1; Figure 1). Prospective analyses for the D study indicate that the reliability coefficient would increase to .54 if data were aggregated across 12 tasks (Table 1; Figure 1), which indicates that that the increase in reliability is more modest as one aggregates across subsequent numbers of tasks.

When data were aggregated across two body loci, the reliability coefficients increased in comparison with using data collected from one locus. The G study showed that for data collected during one task, but aggregated across two loci, the reliability coefficient increased by .1 to .39 (Table 1; Figure 1). When data are aggregated across eight tasks, the reliability coefficient increases from .52 for data collected on one locus to .66 when aggregated across two loci (Figure 1). Prospective analyses for the D study indicated that using data aggregated across four body loci would increase this latter reliability coefficient to .77 (Table 1; Figure 1).

There was more variance across person (.32) than across the facets (.05 across task and .18 across body loci), confirming the shared variance in actigraph data across individual tasks and body loci. However, the interaction between persons and tasks (.23) and between persons and body loci (.23) indicates that the rank ordering of persons changes across tasks and loci, so, as might be expected, to some extent different tasks elicit different activity levels.

Instrumentation and Rater Effects on Actigraph Data

Situational effects on actigraph data were assessed using actigraph data aggregated across all eight tasks and both body loci, since the analyses above indicate increased reliability for data aggregated in this way. To minimize data disruption, since the data were on the same scale the raw data were first summed, then log transformed (optimized minimal skew through the lnskew0 command in Stata version 9.1) to normality to create an aggregate actigraph score per person, across tasks and limbs.

Intermachine differences did not have a significant effect on actigraph data collected across the sample [$t(322) = -0.32$, $p = .75$]. Remaining battery life of the actigraph did not have a significant effect on actigraph data collected across the sample [$t(226) = -1.26$, $p = .21$]. There were no significant differences in the actigraph data collected across the sample between the two research assistants [$t(327) = 0.77$, $p = .44$].

DISCUSSION

With evidence rapidly accumulating that actigraphs are a valid measure of activity level (Puyau et al., 2002; Saudino et al., 2004; Treuth et al., 2004; Tryon, 2005), questions remained about the reliability of activity level as measured by these devices over short measurement occasions, given the inherent variability of human activity. We showed that actigraph data, collected during a single visit to the research center, can have good reliability (Cicchetti, 1994), with many instrumentation variables not having a significant effect on the data collected. However, we also showed that this reliability is dependent on aggregation across actigraph measurements.

The task level and attachment loci level (leg and waist) data for both leg and waist measurements measure the same construct of activity level and have a common variance, fulfilling the main criteria to justify aggregation (Epstein, 1983). The G study indicated that, once aggregated across tasks and loci, the data showed higher reliability coefficients, with an average coefficient of .52, when one body locus was averaged across 8 tasks, or .39 when 1 task was averaged across the waist and leg data. When averaged across two body loci and 8 tasks, the reliability coefficient rose to .66, compared with a reliability coefficient of .29 when just 1 task and one body locus were used. On the basis of G theory, D study analyses indicated that, had data been further aggregated (e.g., across 12 tasks and four body loci), the reliability coefficient was estimated to become .79. This suggests that using longer chunks of actigraph data to measure children's activity levels, here by aggregating across tasks, would produce a more reliable measure of overall activity level over shorter chunks by canceling out momentary or unrepresentative influences. It is likely that this finding will extrapolate to data collected over a longer period, since the tasks were

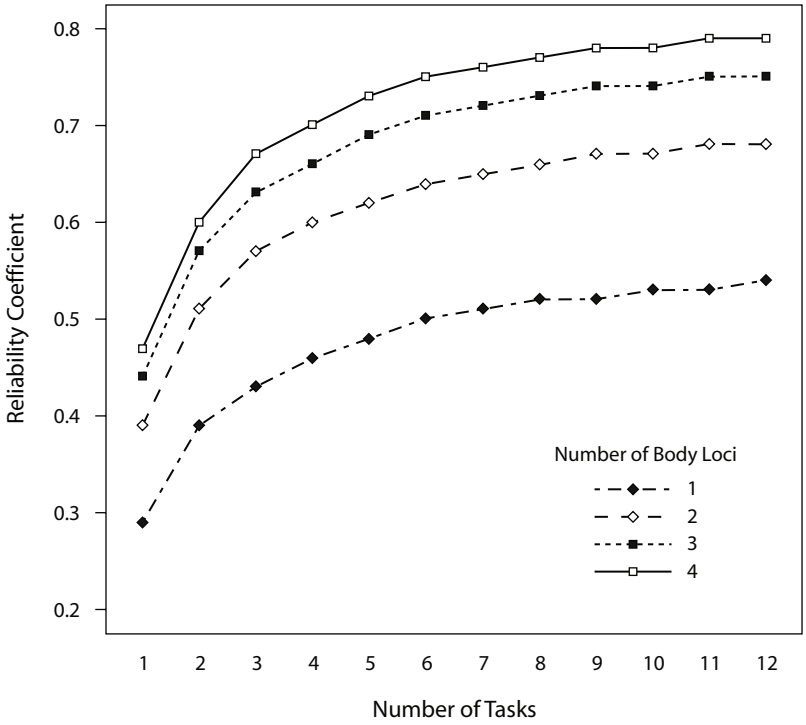

Figure 1. Reliability correlations for differing levels of aggregation across different numbers of tasks, and data collected from different numbers of body loci. Data were collected across two body loci and during eight tasks. Reliability coefficients are presented for this data, as part of the G study, and for prospective data across up to twelve tasks and four body loci as a result of the D study.

administered consecutively, with only a small break between each. This finding is supported by personal communication with the researchers who administered the cognitive tasks, and with informal viewing of video tapes of the session. As increasing numbers of researchers exploit the advantages of actigraphs to collect activity level data (Morgenthaler et al., 2007), this is an important point for study design: Testing task level data, when tasks are fairly short, may not be a suitable method for examining activity level differences, and conclusions drawn from small chunks of data should be treated with caution (e.g., Inoue et al., 1998). Reducing error variance in actigraph data through aggregation has many potential applications, but previous analyses on these data used aggregation across limb scores as an important step in maximizing sensitivity to the underlying genotype (Wood et al., 2007).

Encouragingly, of three possible instrumental and rater effects on actigraph data (intermachine differences, interrater individual differences, and the effect of battery life), none had a significant effect on the data. This finding reinforces both the reliability and validity of actigraph data. Nonetheless, we would still encourage researchers to counterbalance devices across individuals and across body loci, and to be aware of potential maintenance issues with the actigraphs.

The analyses presented relate to actigraph data collected in a laboratory setup; however, the applicability of the findings to a more naturalistic setting requires further study. A benefit of actigraph measurement in a laboratory setup is that it more closely resembles a classroom setting, in which increased activity level may be most problematic. Finally, it may be task level or loci differences that researchers are interested in. These analyses highlight the potential importance of such differences through highlighting task-specific variance, and the present aggregation methods relate more to those seeking to assess overall activity level. When investigating task-specific differences, our data suggest that researchers should make across-data comparisons aggregated across theoretically related tasks of interest, and/or aggregate across several body loci, especially if tasks are fairly short, which is likely to be the case in child research.

Our results show that actigraph data are reliable. We would recommend that actigraph data be aggregated across theoretically related tasks and body loci. Data aggregated over an average of 80 min and over two body loci show higher reliability than data collected from one locus for an average of 10 min.

AUTHOR NOTE

The Study of Activity and Impulsivity Levels in Children (SAIL) is funded by a project grant from the Wellcome Trust (GR070345MF). K.J.S. is supported by Grant MH062375 from the National Institute of Mental Health. A.C.W. is supported by the Economic and Social Research Council. We thank the TEDS-SAIL families, Eda Salih, Hannah Rogers, Rebecca Gibbs, Greer Swinard, Kate Lievesley, Kayley O'Flynn, Suzi Marquis and Rebecca Whittemore, Vlad Mereuța, Desmond Campbell, and everyone on the TEDS team. Correspondence concerning this

article should be addressed to A. C. Wood, Institute of Psychiatry, Kings College London, London WC2R 2LS, England (e-mail: alexis.wood@iop.kcl.ac.uk).

REFERENCES

ABIKOFF, H., COURTNEY, M., PELHAM, W. E., JR., & KOPLEWICZ, H. S. (1993). Teachers' ratings of disruptive behaviors: The influence of halo effects. *Journal of Abnormal Child Psychology*, **21**, 519-533.

ARMITAGE, P., BERRY, G., & MATTHEWS, J. N. S. (2001). *Statistical methods in medical research* (4th ed.). Oxford: Blackwell.

BÖRGER, N., & VAN DER MEERE, J. (2000). Motor control and state regulation in children with ADHD: A cardiac response study. *Biological Psychology*, **51**, 247-267.

CICCHETTI, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, **6**, 284-290.

COLE, R. J., KRIPKE, D. F., GRUEN, W., MULLANEY, D. J., & GILLIN, J. C. (1992). Automatic sleep/wake identification from wrist activity. *Sleep*, **15**, 461-469.

CRONBACH, L. J., GLESER, G. C., NANDA, H., & RAJARATNAM, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

DANE, A. V., SCHACHAR, R. J., & TANNOCK, R. (2000). Does actigraphy differentiate ADHD subtypes in a clinical research setting? *Journal of the American Academy of Child & Adolescent Psychiatry*, **39**, 752-760.

EATON, W. O. (1983). Measuring activity level with actometers: Reliability, validity, and arm length. *Child Development*, **54**, 720-726.

EATON, W. O., MCKEEN, N. A., & SAUDINO, K. J. (1996). Measuring human individual differences in general motor activity with actometers. In K. Ossenkopp, M. Kavaliers, & P. R. Sanberg (Eds.), *Measuring movement and locomotion: From invertebrates to humans* (pp. 79-92). Austin, TX: Landes.

EAVES, L., RUTTER, M., SILBERG, J. L., SHILLADY, L., MAES, H., & PICKLES, A. (2000). Genetic and environmental causes of covariation in interview assessments of disruptive behavior in child and adolescent twins. *Behavior Genetics*, **30**, 321-334.

EPSTEIN, S. (1983). Aggregation and beyond: Some basic issues on the prediction of behavior. *Journal of Personality*, **51**, 360-392.

INOUE, K., NADAOKA, T., OIJI, A., MORIOKA, Y., TOTSUKA, S., KANBAYASHI, Y., & HUKUI, T. (1998). Clinical evaluation of attention-deficit hyperactivity disorder by objective quantitative measures. *Child Psychiatry & Human Development*, **28**, 179-188.

JEAN-LOUIS, G., VON GIZYCKI, H., ZIZI, F., SPIELMAN, A., HAURI, P., & TAUB, H. (1997). The actigraph data analysis software: II. A novel approach to scoring and interpreting sleep–wake activity. *Perceptual & Motor Skills*, **85**, 219-226.

KUNTSI, J., ANDREOU, P., MA, J., BÖRGER, N. A., & VAN DER MEERE, J. (2005). Testing assumptions for endophenotype studies in ADHD: Reliability and validity of tasks in a general population sample. *BMC Psychiatry*, **5**, 40.

KUNTSI, J., ROGERS, H., SWINARD, G., BÖRGER, N., VAN DER MEERE, J., RIJSDIJK, F., & ASHERSON, P. (2006). Reaction time, inhibition, working memory and "delay aversion" performance: Genetic influences and their interpretation. *Psychological Medicine*, **36**, 1613-1624.

KUNTSI, J., STEVENSON, J., OOSTERLAAN, J., & SONUGA-BARKE, E. J. S. (2001). Test–retest reliability of a new delay aversion task and executive function measures. *British Journal of Developmental Psychology*, **19**, 339-348.

MORGENTHALER, T., ALESSI, C., FRIEDMAN, L., OWENS, J., KAPUR, V., BOEHLECKE, B., ET AL. (2007). Practice parameters for the use of actigraphy in the assessment of sleep and sleep disorders: An update for 2007. *Sleep*, **30**, 519-529.

MUSHQUASH, C., & O'CONNOR, B. P. (2006). SPSS and SAS programs for generalizability theory analyses. *Behavior Research Methods*, **38**, 542-547.

O'BRIEN, R. M. (1995). Generalizability coefficients are reliability coefficients. *Quality & Quantity*, **29**, 421-428.

PUYAU, M. R., ADOLPH, A. L., VOHRA, F. A., & BUTTE, N. F. (2002). Validation and calibration of physical activity monitors in children. *Obesity Research*, **10**, 150-157.

ROUSSON, V., GASSER, T., & SEIFERT, B. (2002). Assessing intrarater, interrater and test–retest reliability of continuous measurements. *Statistical Medicine*, **21**, 3431-3446.

SADEH, A., SHARKEY, K. M., & CARSKADON, M. A. (1994). Activity-based sleep–wake identification: An empirical test of methodological issues. *Sleep*, **17**, 201-207.

SAUDINO, K. J., CHERNY, S. S., & PLOMIN, R. (2000). Parent ratings of temperament in twins: Explaining the "too low" DZ correlations. *Twin Research*, **3**, 224-233.

SAUDINO, K. J., RONALD, A., & PLOMIN, R. (2005). The etiology of behavior problems in 7-year-old twins: Substantial genetic influence and negligible shared environmental influence for parent ratings and ratings by same and different teachers. *Journal of Abnormal Child Psychology*, **33**, 113-130.

SAUDINO, K. J., WERTZ, A. E., GAGNE, J. R., & CHAWLA, S. (2004). Night and day: Are siblings as different in temperament as parents say they are? *Journal of Personality & Social Psychology*, **87**, 698-706.

SCHACHAR, R., SANDBERG, S., & RUTTER, M. (1986). Agreement between teachers' ratings and observations of hyperactivity, inattentiveness, and defiance. *Journal of Abnormal Child Psychology*, **14**, 331-345.

SIMONOFF, E., PICKLES, A., HERVAS, A., SILBERG, J. L., RUTTER, M., & EAVES, L. (1998). Genetic influences on childhood hyperactivity: Contrast effects imply parental rating bias, not sibling interaction. *Psychological Medicine*, **28**, 825-837.

SPINATH, F. M., & O'CONNOR, T. G. (2003). A behavioral genetic study of the overlap between personality and parenting. *Journal of Personality*, **71**, 785-808.

STEVENS, J., QUITTNER, A. L., & ABIKOFF, H. (1998). Factors influencing elementary school teachers' ratings of ADHD and ODD behaviors. *Journal of Clinical Child Psychology*, **27**, 406-414.

TEICHER, M. H., ITO, Y., GLOD, C. A., & BARBER, N. I. (1996). Objective measurement of hyperactivity and attentional problems in ADHD. *Journal of the American Academy of Child & Adolescent Psychiatry*, **35**, 334-342.

THAPAR, A., HARRINGTON, R., ROSS, K., & MCGUFFIN, P. (2000). Does the definition of ADHD affect heritability? *Journal of the American Academy of Child & Adolescent Psychiatry*, **39**, 1528-1536.

TREUTH, M. S., SCHMITZ, K., CATELLIER, D. J., MCMURRAY, R. G., MURRAY, D. M., ALMEIDA, M. J., ET AL. (2004). Defining accelerometer thresholds for activity intensities in adolescent girls. *Medicine & Science in Sports & Exercise*, **36**, 1259-1266.

TROUTON, A., SPINATH, F. M., & PLOMIN, R. (2002). Twins early development study (TEDS): A multivariate, longitudinal genetic investigation of language, cognition and behavior problems in childhood. *Twin Research*, **5**, 444-448.

TRYON, W. W. (2005). The reliability and validity of two ambulatory monitoring actigraphs. *Behavior Research Methods*, **37**, 492-497.

VAN COEVERING, P., HARNACK, L., SCHMITZ, K., FULTON, J. E., GALUSKA, D. A., & GAO, S. (2005). Feasibility of using accelerometers to measure physical activity in young adolescents. *Medicine & Science in Sports & Exercise*, **37**, 867-871.

VAN DER MEERE, J., STEMERDINK, N., & GUNNING, B. (1995). Effects of presentation rate of stimuli on response inhibition in ADHD children with and without tics. *Perceptual & Motor Skills*, **81**, 259-262.

VERHULST, F. C., ACHENBACH, T. M., ALTHAUS, M., & AKKERHUIS, G. W. (1988). A comparison of syndromes derived from the child behavior checklist for American and Dutch girls aged 6–11 and 12–16. *Journal of Child Psychology & Psychiatry*, **29**, 879-895.

WECHSLER, D. (1991). *Wechsler Intelligence Scale for Children* (3rd ed.). London: Psychological Corporation.

WELK, G. J., BLAIR, S. N., WOOD, K., JONES, S., & THOMPSON, R. W. (2000). A comparative evaluation of three accelerometry-based physical activity monitors. *Medicine & Science in Sports & Exercise*, **32**, S489-S497.

WELK, G. J., SCHABEN, J. A., & MORROW, J. R., JR. (2004). Reliability of accelerometry-based activity monitors: A generalizability study. *Medicine & Science in Sports & Exercise*, **36**, 1637-1645.

WOOD, A. C., SAUDINO, K. J., ROGERS, H., ASHERSON, P., & KUNTSI, J. (2007). Genetic influences on mechanically-assessed activity level in children. *Journal of Child Psychology & Psychiatry*, **48**, 695-702.

(Manuscript received December 19, 2007;
revision accepted for publication March 11, 2008.)