

## Practice of Epidemiology

### The Role of Measurement Error in Estimating Levels of Physical Activity

Pietro Ferrari<sup>1</sup>, Christine Friedenreich<sup>2</sup>, and Charles E. Matthews<sup>3</sup>

<sup>1</sup> Nutrition and Hormones Group, International Agency for Research on Cancer, Lyon, France.

<sup>2</sup> Division of Population Health and Information, Alberta Cancer Board, Calgary, Alberta, Canada.

<sup>3</sup> Division of General Internal Medicine, Department of Medicine, Vanderbilt University School of Medicine, Nashville, TN.

Received for publication November 15, 2006; accepted for publication April 9, 2007.

Epidemiologic studies have demonstrated that physical inactivity is an important determinant of numerous chronic diseases. However, self-reported estimates of physical activity contain measurement errors responsible for attenuating relative risk estimates. A validation study conducted in 2002–2003 at the Alberta Cancer Board (Canada) included a physical activity questionnaire, four 7-day physical activity logs, and four sets of accelerometer data from 154 study subjects (51% women) aged 35–65 years. The authors used a measurement error model to evaluate validity of the different types of physical activity assessment, and the attenuation factors, after taking into account error correlations between self-reported measurements. The validity coefficients, which express the correlation between measured and true exposure, were higher for accelerometers (0.81, 95% confidence interval (CI): 0.76, 0.85) compared with the physical activity log (0.57, 95% CI: 0.47, 0.66) and questionnaire measurements (0.26, 95% CI: 0.12, 0.40). The estimate of the attenuation factor for questionnaires was 0.13 (95% CI: 0.05, 0.23). Accuracy of physical activity questionnaire measurements was higher for men than for women, for younger individuals, and for those with a lower body mass index. Because the degree of attenuation in relative risk estimates is substantial, after the role of error correlations was considered, validation studies quantifying the impact of measurement errors on physical activity estimates are essential to evaluate the impact of physical inactivity on health.

bias (epidemiology); measurement error; questionnaires

Abbreviations: CI, confidence interval; PA LOG, physical activity log; PAQ, physical activity questionnaire.

It is recognized that physical activity measurements used in large epidemiologic investigations contain a sizable level of measurement error (1–3). Under certain conditions, when measurement errors are nondifferentially expressed, the association between physical activity and an outcome of interest is attenuated toward the null hypothesis of no association (4–7), thus making it difficult to evaluate the likely beneficial effects of physical activity on the full range of health outcomes.

It is possible to estimate the magnitude of measurement errors and the impact on the estimation of associations when assessing individuals' physical activity exposure. In so do-

ing, the effect of measurement errors can be quantified. Numerous statistical methods to correct for bias attributable to measurement error when evaluating a statistical association between an exposure and an outcome of interest have been proposed (8–10).

No "gold standard" measurements of physical activity levels among free-living individuals are available to directly measure true levels of exposure. Physical activity assessment methods that make use of recall methods, such as questionnaires, are a subjective means of estimating individual exposure because they rely on individuals' ability to remember levels of exposure. It has been observed that

Correspondence to Dr. Pietro Ferrari, Nutrition and Hormones Group, International Agency for Research on Cancer, 150 Cours Albert Thomas 69372 Lyon, France (e-mail: ferrari@iarc.fr; www.iarc.fr).

many factors influence the accuracy of exposure estimates, all based primarily on personal perception of the exposure being evaluated and quantified.

Design and analysis of validation studies in terms of statistical latent variable models (11, 12) have shown that the validity coefficient can be estimated by comparison with a third objective measurement. It is assumed that errors in the objective measurements are independent from self-reported measurements.

Currently, the methods that provide the most feasible and accurate measurements of activity in large validation studies are physical activity logs (PA LOGs) and objective measures obtained from accelerometers. PA LOGs are thought to minimize errors associated with long-term recall of activity and estimation of activity frequency, whereas objective measures do not rely on self-report at all. When assessing the validity of physical activity measurements, it is therefore important to use statistical models that account for random and systematic errors, as well as error correlation between questionnaires and PA LOG measurements and between replicates of the same instrument.

In a recent study (13), validity and reliability of a self-administered past-year physical activity questionnaire (PAQ) developed for use in research studies conducted at the Alberta Cancer Board (Canada) were assessed by comparison with 7-day PA LOGs and 7-day accelerometer data. In the absence of individual estimates of true physical activity level, 7-day PA LOG and accelerometer measurements were used as the reference methods. In this paper, we propose to acknowledge and take into account measurement errors in all observed measurements and to estimate the validity of PAQs, PA LOGs, and accelerometer data. This assessment is accomplished by means of structural equation models, where inclusion of objective measures of physical activity allows estimation of correlation between errors of self-report instruments (questionnaire and PA LOGs). In addition, we estimate the attenuation factor that quantifies the effect of measurement error on the association with an outcome. Numerous aspects of measurement error in individual estimates of physical activity level are evaluated and discussed.

## MATERIALS AND METHODS

Data were available on 154 men and women aged 35–65 years who completed four PAQ measurements (PAQ1–PAQ4), four 7-day PA LOGs, and four sets of accelerometer data over a 1-year period. The design and methods for this initial validation study have been described previously (13).

### Assessment of physical activity

In brief, two measurements of physical activity in the previous 12 months were taken at baseline in a 1-year follow-up study by using the self-administered PAQ known as the Past Year Total Physical Activity Questionnaire. These repeat measurements were taken, on average, 9 weeks apart. At the end of the 1-year follow-up validation study, the questionnaire was completed again, first as a self-administered (PAQ3) and then as a telephone-administered

(PAQ4) questionnaire. During the year-long follow-up, study participants wore the accelerometers for four 1-week periods at intervals approximately 12 weeks apart. Immediately after wearing the accelerometers for 1 week, they completed the 7-day PA LOGs by coding their activity for each 15-minute interval of every hour of the day for 7 days consecutively. These data were transformed to express metabolic equivalent task-hours per week, as determined from the Compendium of Physical Activities (14, 15).

Hence, during the year, four sets of accelerometer data and PA LOGs were collected. Accelerometer data estimate the intensity and duration of motion in the sampling interval (i.e., counts/minute). The basic activity count data were summarized in terms of hours per week and, using total energy expenditure (metabolic equivalent task-hours/week), were derived from the equation of Swartz et al. (16), after censoring recorded values below 150 counts/minute. In addition, physical activity that was done when the accelerometers were not worn was recorded, with a separate activity monitor log that the respondents completed.

### Statistical model

In the present study, the validity of PAQ3, accelerometer, and PA LOG measurements was evaluated. This evaluation was achieved by estimating the validity coefficient that reflects the correlation between the observed measurement of exposure, be it questionnaire, accelerometer data, or PA LOG, and the unknown true level ( $T$ ). A measurement error model was therefore defined that assumes relations between questionnaire ( $Q$ ), PA LOG ( $R$ ), and accelerometer ( $A$ ) measurements and true level of activity, for subject  $i = 1, \dots, I$ , and measurements  $j = 1, \dots, J_X$  ( $X = R, A$ ), as

$$Q_i = \alpha_Q + \beta_Q T_i + \varepsilon_{Qi} \quad (1a)$$

$$R_{ij} = T_i + \varepsilon_{Rij} \quad (1b)$$

$$A_{ij} = \alpha_A + \beta_A T_i + \varepsilon_{Aij} \quad (1c)$$

The following relations are assumed for random measurement errors  $\varepsilon_{Qi}$  and  $\varepsilon_{Xij}$ ,  $E(\varepsilon_{Qi}|T_i) = 0$ ,  $E(\varepsilon_{Rij}|T_i) = 0$ ,  $E(\varepsilon_{Aij}|T_i) = 0$ , and  $\text{Var}(\varepsilon_{Qi}) = \sigma_{\varepsilon_Q}^2$ ,  $\text{Var}(\varepsilon_{Rij}) = \sigma_{\varepsilon_R}^2$ , and  $\text{Var}(\varepsilon_{Aij}) = \sigma_{\varepsilon_A}^2$ . For the questionnaire measurement  $Q$ , the coefficients  $\alpha_Q$  and  $\beta_Q$  in model 1a express constant and proportional scaling biases (17), while the residual terms  $\varepsilon_{Qi}$  models the random part of measurement errors in  $Q$  and are assumed to be uncorrelated with true physical activity level  $T_i$ , after  $\alpha_Q$  and  $\beta_Q$  have captured the systematic component of measurement error (12).

In this analysis, the 7-day PA LOG was chosen to be the “reference measurement,” with  $\alpha_R = 0$  and  $\beta_R = 1$ . The relation expressed in model 1b assumes that errors are strictly random and that variation around individuals’ unknown true physical activity is entirely attributable to within-person random variability or to random measurement errors in reporting physical activity levels (i.e.,  $\text{cov}(\varepsilon_{Rij}, T_i) = 0$ , for  $\forall j \in J_R$ ).

In addition, throughout this analysis, it is assumed that random errors in questionnaire and 7-day PA LOG measurements are correlated ( $\text{cov}(\varepsilon_{Qi}, \varepsilon_{Rij}) \neq 0$ , for  $\forall j \in J_R$ ), as a result of individuals’ tendency to consistently misreport

their activity level. Furthermore, correlation between errors in questionnaire-based assessments may arise because these measurements may share some common source of variability. Similar assumptions were made for the replicates of PA LOG measurements ( $\text{cov}(\varepsilon_{Rij}, \varepsilon_{Rik}) \neq 0; j \neq k$ ).

Although measurements collected by using an accelerometer have been reported to be reliable for estimating dynamic physical activities (18–22), some concerns have been raised about the capacity of accelerometers to provide a valid estimate of absolute physical activity levels (20). For this reason, in expression 1c, the terms  $\alpha_A$  and  $\beta_A$  were introduced for accelerometer data to take into account systematic error, while  $\varepsilon_A$  captures the random measurement error component.

Accelerometer measurements provide objective estimates of physical activity levels. It was assumed that random errors in questionnaire-based and accelerometer measurements are independent, that is,  $\text{cov}(\varepsilon_{Qij}, \varepsilon_{Aij}) = \text{cov}(\varepsilon_{Rij}, \varepsilon_{Aij}) = 0$  for  $\forall j \in J_A$  and  $\forall j \in J_R$ , under the hypothesis that errors in the latter are not influenced by study subjects' ability to recall past activity levels. Measurement errors in model 1 could also be expressed in terms of person-specific bias and random variation (23, 24). However, the two notations are equivalent, as has been shown previously (25).

Model parameter estimates can be obtained through structural equation models. In this context, given the linear associations in expression 1 and the related assumptions regarding error variances and covariances, the population covariance matrix ( $\Sigma$ ) of a set of observed variables ( $X_1, \dots, X_p$ ) is a function of a set of basic parameters ( $\mathbf{q} = (q_1, \dots, q_r)$ ), as  $\hat{\Sigma} = \Sigma(\hat{\mathbf{q}})$  (26, 27). For the model to be identifiable, a situation that arises when a unique set of parameter estimates corresponds to the observed data, it is necessary that the components of  $\mathbf{q}$  are not more than the elements of the sample covariance matrix  $S$ , that is,  $t \leq p(p+1)/2$ . Classically, the maximum likelihood method is used for such model fitting, which usually requires the assumption that variables are normally distributed (27). Notably, in this study, the components of  $\mathbf{q}$  are the regression coefficients in equation 1 as well as the variance of the independent variable ( $\sigma_T^2$ ), and the error variances and covariances.

Having estimated the parameter vector  $\mathbf{q}$ , the validity coefficient for questionnaire  $Q$  can be calculated, as detailed in the Appendix:

$$\hat{\rho}_{QT} = \sqrt{\frac{1}{1 + \frac{\hat{\sigma}_{\varepsilon_Q}^2}{\hat{\beta}_Q^2 \hat{\sigma}_T^2}}}. \quad (2)$$

The validity coefficients for the accelerometer measurements,  $\rho_{AT}$ , and PA LOG,  $\rho_{RT}$ , are estimated in the same manner.

When measurement errors in exposure variables are present, the estimated association describing the exposure-disease relation (e.g., relative risk estimates) may be biased, typically toward the null. A number of statistical methods have been proposed for correcting the effects of covariate measurement error (7–9). One simple adjustment for measurement error is the regression calibration method (8), where the attenuation factor,  $\lambda$ , is estimated by the slope

of true on observed exposure in a validation study. The log-relative risk in the disease model is corrected (deattenuated) by dividing by  $\hat{\lambda}$  (8). For this reason, values of the attenuation factor close to 0 lead to more serious underestimations of risk. In this work, the attenuation factor is calculated as

$$\hat{\lambda}_{QT} = \frac{\hat{\beta}_Q \hat{\sigma}_T^2}{\hat{\beta}_Q^2 \hat{\sigma}_T^2 + \hat{\sigma}_{\varepsilon_Q}^2}. \quad (3)$$

The regression calibration method corrects for random and systematic within-person error in observed measurements of exposure, provided that a reference measure is available (24). By estimating the validity coefficient and the attenuation factor using models 1 and 2, the possible error correlation between questionnaire and PA LOG measurements is quantified and the effect of taking into account this error correlation is evaluated.

### A model with an instrumental variable

It is conceivable that replicates of accelerometers contain components of intraindividual systematic errors, because accelerometer data for subjects who engage in very dynamic kinds of pursuits (e.g., walking, running) for a larger proportion of their overall activity would be predicted to have less systematic error, whereas the actigraph data for participants who engage in more static pursuits for a larger proportion of their activity, which may expend considerable energy but cause less acceleration of the body (e.g., household chores, gardening), might contain systematic errors. For this reason, two strategies have been used to examine these possible errors.

Estimation of error correlation in replicates of accelerometer data in models 1a–1c leads to identifiability issues. To address this, two complementing strategies have been followed. First, a sensitivity analysis was conducted in model 1 by imputing values of the correlation between errors in replicates of accelerometer data, thus pursuing estimation of the validity coefficients. Second, an additional model was considered by introducing an instrumental variable (24, 28, 29), as follows:

$$\begin{aligned} Q_i &= \alpha_Q + \beta_Q T_i + \varepsilon_{Qi} \\ R_{ij} &= T_i + \varepsilon_{Rij} \\ A_{ij} &= \alpha_A + \beta_A T_i + \varepsilon_{Aij} \\ B_i &= \alpha_B + \beta_B T_i + \varepsilon_{Bi}, \end{aligned} \quad (4)$$

where  $T_i$ ,  $\alpha_Q$ ,  $\alpha_A$ ,  $\beta_Q$ ,  $\beta_A$ ,  $\varepsilon_{Qi}$ ,  $\varepsilon_{Rij}$ , and  $\varepsilon_{Aij}$  have been defined in model 1. It is further assumed that the errors in  $B_i$  measurements are  $E(\varepsilon_{Bi}|T_i) = 0$ ,  $\text{Var}(\varepsilon_{Bi}) = \sigma_{\varepsilon_B}^2$  and are uncorrelated with any other residual error terms in model 4. Furthermore, it is assumed that  $\text{cov}(\varepsilon_{Aij}, \varepsilon_{Aik}) \neq 0; \forall j \in J_A$ , with  $j \neq k$ . Parameter estimation was conducted by using structural equation models, similarly to model 1. Details are provided in the Appendix.

Structural equation models assume that the random variables approximate a multivariate normal distribution. For

**TABLE 1. Overall and gender-specific sample size, mean, and standard deviation of log-transformed PAQ\* and 7-day PA LOG,\* ACC\* measurements (log-MET\*-hours/week), and weight (kg)**

Measurement	Variable name	Overall		Men		Women	
		No.	Mean (SD*)	No.	Mean (SD)	No.	Mean (SD)
PAQ	<i>Q</i>	154	4.78 (0.51)	75	4.84 (0.45)	79	4.73 (0.56)
PA LOG1	<i>R1</i>	153	4.82 (0.42)	74	4.85 (0.42)	79	4.78 (0.41)
PA LOG2	<i>R2</i>	153	4.83 (0.46)	75	4.81 (0.53)	78	4.85 (0.39)
PA LOG3	<i>R3</i>	153	4.72 (0.50)	74	4.72 (0.55)	79	4.72 (0.45)
PA LOG4	<i>R4</i>	151	4.78 (0.44)	74	4.79 (0.48)	77	4.78 (0.41)
ACC1	<i>A1</i>	143	4.91 (0.23)	73	4.91 (0.24)	70	4.92 (0.22)
ACC2	<i>A2</i>	144	4.92 (0.24)	72	4.90 (0.26)	72	4.93 (0.22)
ACC3	<i>A3</i>	142	4.82 (0.25)	67	4.81 (0.28)	75	4.83 (0.22)
ACC4	<i>A4</i>	137	4.85 (0.26)	67	4.86 (0.29)	70	4.85 (0.24)
Weight	<i>B</i>	154	78.1 (15.2)	75	84.2 (12.4)	79	72.3 (15.4)

\* PAQ, physical activity questionnaire; PA LOG, physical activity log; ACC, accelerometer; MET, metabolic-equivalent task; SD, standard deviation.

this reason, the data have been log transformed. To deal with missing values ( $n = 56$  over the nine different physical activity variables), an expectation-maximization algorithm was used to estimate the maximum likelihood covariance matrix that was used as input data to estimate model parameters in the structural equation model. This approach assumes that data are missing at random (30). Parameter estimates were obtained by using the CALIS procedure in SAS software (31), as well as associated approximate standard errors. For validity coefficient and attenuation factor estimates, 95 percent confidence intervals were obtained by

computing the 2.5th and 97.5th percentiles of a distribution determined with bootstrap sampling (32). A total of 1,000 repetitions gave sufficiently stable intervals.

Models for men and women combined were systematically adjusted for gender by regressing physical activity measurements on gender and using residuals. To evaluate the degree of correlated errors after controlling for study subjects' characteristics, residuals of physical activity measurements on, in turn, age, body mass index, and level of physical fitness at baseline were computed. However, results were very similar, and only unadjusted results are presented here.

**TABLE 2. Overall ( $n = 154$ ) and sex-specific (men,  $n = 75$ ; women,  $n = 79$ ) estimates of validity coefficient ( $\hat{\rho}_{XT}$ ) and attenuation factors ( $\hat{\lambda}$ ), estimates of true variance ( $\hat{\sigma}_T^2$ ), error correlations ( $\hat{\rho}_{eijk}$ ), and associated 95% confidence intervals under model 1 for measurements  $X$ , in turn, PAQ,\* 7-day PA LOG,\* and ACC\* measurements**

Measurement	Variable name	$\hat{\rho}_{XT}$		$\hat{\lambda}$		$\hat{\sigma}_T^2$		$\hat{\rho}_{e j;c;k}$	
		Estimate	95% CI*	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI
Overall†									
PAQ	<i>Q</i>	0.258	0.115, 0.398	0.132	0.052, 0.233	0.067	0.037, 0.098	0.225‡	0.066, 0.379
PA LOG	<i>R</i>	0.568	0.467, 0.662					0.553§	0.445, 0.638
ACC	<i>A</i>	0.813	0.762, 0.853						
Men									
PAQ	<i>Q</i>	0.386	0.197, 0.576	0.228	0.092, 0.414	0.070	0.020, 0.120	0.273‡	0.102, 0.415
PA LOG	<i>R</i>	0.534	0.384, 0.650					0.545§	0.390, 0.661
ACC	<i>A</i>	0.786	0.689, 0.847						
Women									
PAQ	<i>Q</i>	0.146	0.010, 0.330	0.066	−0.030, 0.183	0.065	0.027, 0.104	0.211‡	−0.050, 0.439
PA LOG	<i>R</i>	0.613	0.452, 0.736					0.558§	0.392, 0.664
ACC	<i>A</i>	0.863	0.760, 0.877						

\* PAQ, physical activity questionnaire; PA LOG, physical activity log; ACC, accelerometer; CI, confidence interval.

† Gender-adjusted values.

‡ Correlation between errors in *Q* and *R* measurements.

§ Correlation between errors among replicates of *R* measurements.

In addition, in an attempt to evaluate the effect of measurement errors on the validity of physical activity measurements, models were run according to study subjects' age, body mass index, and level of physical fitness at baseline. Separately for each variable, two groups were created by taking, respectively, cutoffs equal to 50 years for age, the gender-specific body mass index median values, equal to 27.5 and 25.7 (kg/m<sup>2</sup>) in men and women, respectively, and the gender-specific level of fitness, based on predicted maximum oxygen consumption at baseline, equal to 31.3 and 25.8 (ml/kg per minute) in men and women, respectively.

Log-transformed body weight was used to estimate the validity of questionnaire measurements using an instrumental variable in model 4. Using level of fitness instead of weight provided very similar results.

## RESULTS

Although average values in log-transformed questionnaire, PA LOG, and accelerometer physical activity measurements were not very different, greater heterogeneity was observed for variability in the different types of data (table 1). Mean values were very similar for men and women, but variability of different measurements was higher in men than in women.

The values for estimates of the validity coefficients were higher for accelerometer measurements (0.81, 95 percent confidence interval (CI): 0.76, 0.85) compared with PA LOG (0.57, 95 percent CI: 0.47, 0.66) and questionnaire (0.26, 95 percent CI: 0.12, 0.40) measurements (table 2). Higher validity coefficients for questionnaire measurements were observed for men (0.39, 95 percent CI: 0.20, 0.58) than for women (0.15, 95 percent CI: 0.01, 0.33), while similar coefficients were observed for PA LOG and accelerometer data for the two genders. The 95 percent confidence intervals showed that these parameters were estimated with precision. Attenuation factor estimates for questionnaires were equal to 0.13 (95 percent CI: 0.05, 0.23), with a lower value for women (0.07, 95 percent CI: -0.03, 0.18) than for men (0.23, 95 percent CI: 0.09, 0.41). True variability was equal to 0.07 (95 percent CI: 0.04, 0.10).

Higher error correlations were observed between replicates of PA LOG measurements (0.55, 95 percent CI: 0.45, 0.64) than between questionnaire and PA LOG (0.23, 95 percent CI: 0.07, 0.38). Similar values were observed for men and women.

The results of assessment of measurement error by study subjects' specific characteristics (table 3) show that validity of questionnaire measurements varied by age, with lower coefficients observed for subjects older than age 50 years, and also by body mass index, with lower validity for participants with a higher body mass index. Although adjustment for age did not produce noticeable changes, some minor confounding by age cannot be ruled out. Similarly, higher coefficients were observed for individuals with a better level of fitness, although the lower body mass index values and, to a lesser extent, the younger age of participants with higher predicted maximum oxygen consumption might have partially driven these findings.

**TABLE 3. Age, BMI,\* and level-of-fitness group-specific means and estimates of gender-adjusted validity coefficients ( $\hat{\rho}_{XT}$ ), and associated 95% confidence interval under model 1, for measurements  $X$ , in turn, PAQ,\* 7-day PA LOG,\* and ACC\* measurements**

Measurement	Variable name	Younger age (n = 75)†			Older age (n = 79)			Lower BMI (n = 76)‡			Higher BMI (n = 78)			Lower fitness level (n = 77)§			Higher fitness level (n = 75)		
		Mean	95% CI¶	¶	Mean	95% CI¶	¶	Mean	95% CI¶	¶	Mean	95% CI¶	¶	Mean	95% CI¶	¶	Mean	95% CI¶	¶
Age (years)		42.1	35.0, 49.0		55.0	50.0, 63.0		47.1	35.0, 63.0		50.2	37.0, 63.0		50.3	37.0, 63.0		47.0	35.0, 62.0	
BMI (kg/m <sup>2</sup> )		26.4	18.5, 36.3		27.9	20.7, 42.5		23.7	18.5, 27.4		30.6	25.8, 42.5		29.2	21.8, 42.5		25.2	18.5, 32.9	
Level of fitness		30.6	20.8, 41.3		28.7	13.5, 49.9		31.9	21.6, 49.9		27.4	13.5, 41.3		25.1	13.5, 31.1		34.2	26.0, 49.9	
		$\hat{\rho}_{XT}$	95% CI#		$\hat{\rho}_{XT}$	95% CI#		$\hat{\rho}_{XT}$	95% CI#		$\hat{\rho}_{XT}$	95% CI#		$\hat{\rho}_{XT}$	95% CI#		$\hat{\rho}_{XT}$	95% CI#	
PAQ	Q	0.420	0.255, 0.582		0.056	0.004, 0.267		0.326	0.139, 0.496		0.189	0.015, 0.407		0.069	0.004, 0.281		0.382	0.205, 0.560	
PA LOG	R	0.654	0.523, 0.762		0.454	0.290, 0.586		0.572	0.441, 0.696		0.524	0.356, 0.665		0.493	0.330, 0.644		0.652	0.525, 0.745	
ACC	A	0.825	0.727, 0.877		0.799	0.726, 0.852		0.822	0.748, 0.877		0.790	0.697, 0.845		0.750	0.651, 0.819		0.861	0.800, 0.900	

\* BMI, body mass index; PAQ, physical activity questionnaire; PA LOG, physical activity log; ACC, accelerometer; CI, confidence interval.

† Study subjects aged <50 years (51.3% of women and 48.7% of men).

‡ Study subjects below gender-specific median values of BMI (27.5 kg/m<sup>2</sup> for men and 25.7 kg/m<sup>2</sup> for women).

§ Study subjects below gender-specific median values of fitness level based on predicted maximum oxygen consumption at baseline (31.3 ml/kg per minute in men and 25.8 ml/kg per minute in women).

¶ Based on the 2.5th and 97.5th percentiles.

# Calculated with bootstrap sampling for attenuation factors for Q, R, and A measurements.

**TABLE 4.** Values of the validity coefficients for questionnaire measurements ( $\hat{\rho}_{QT}$ ), attenuation factor ( $\hat{\lambda}$ ), correlation between error in questionnaire and PA LOG ( $\hat{\rho}_{e_{Rij}, e_{Qik}}$ ), and correlation between errors in replicates of PA LOG ( $\hat{\rho}_{e_{Rij}, e_{Rik}}$ ) for different values of the correlation between error in replicates of accelerometer data ( $\rho_{e_{Aij}, e_{Aik}}$ ), as estimated in a sensitivity analysis

$\rho_{e_{Aij}, e_{Aik}} (j \neq k)$	$\hat{\rho}_{QT}$	$\hat{\lambda}$	$\hat{\rho}_{e_{Rij}, e_{Qik}}$	$\hat{\rho}_{e_{Rij}, e_{Rik}}$
−0.400	0.244	0.116	0.238	0.579
−0.200	0.251	0.122	0.233	0.569
0*	0.258*	0.132*	0.225*	0.553*
0.200	0.281	0.153	0.203	0.520
0.400	0.323	0.202	0.148	0.408
0.500	0.374	0.273	0.039	0.113

\* No correlation between errors of accelerometer measurements is assumed.

The sensitivity analysis showed that the validity coefficients and the attenuation factors increased slightly with increasing values of the correlation between errors in replicates of accelerometer measurements, although only for values of accelerometer error correlations higher than 0.4 was the change (bias) appreciable (table 4).

When an instrumental variable was used, the questionnaire validity coefficient was equal to 0.27 (95 percent CI: 0.13, 0.49), a value very similar to the coefficient estimated in model 1 (table 5). The correlation between error in replicates of accelerometer measurements was equal to 0.17 (95 percent CI: −0.75, 0.54). The confidence intervals for correlation between errors in model 4 were wide and asymmetrical.

## DISCUSSION

We evaluated the validity of PAQ, PA LOG, and accelerometer measurements by using an integrated structural equation model. It was assumed that none of these measurements are themselves true gold standards but rather contain components of random and systematic measurement error.

Although earlier studies investigated the role of measurement error in physical activity assessments (33, 34), to our knowledge we evaluated and quantified for the first time in this work the effect of error correlations in self-reported physical activity measurements.

The models used in this study relied on several statistical assumptions. First, it was assumed that PA LOG measurements can be used as the reference and that errors are strictly random and unrelated to true long-term physical activity level (3, 35, 36). It should be noted that PA LOGs are burdensome for participants to complete, and they measure activities specified a priori, thus having the potential to introduce systematic errors. However, if the assumption of random error in PA LOG measurements does not hold (i.e., model 1b), underestimation of the validity coefficients can result. Therefore, our estimates are likely to be conservative. Second, to estimate all the parameters, thus making the model identifiable, a third objective measurement is needed whose errors are assumed to be independent of the self-reported physical activity measurements (11, 12, 24, 37).

For this purpose, accelerometer data were used, which were assumed to contain random and systematic measurement errors. Accelerometers optimally capture more dynamic activities (e.g., walking and running) but provide limited or no estimates of cycling or swimming, which in this study were systematically recorded with monitor logs. Although this could introduce an aspect of self-report to the actigraph, the contribution of these activities to total activity is likely very low. We found the validity coefficients of PAQ measurements to be threefold lower for women, and of similar magnitude for men, compared with values previously observed (13). This result is mainly attributable to the fact that the methods used in this study allow the error correlation between replicates of the same types of instruments, as well as between different types, to be evaluated and taken into account. Estimating the validity coefficients by comparison of measurements that share the same source of errors can lead to overestimating the quantity of interest (23).

It is realistic to assume that the sources of errors in self-reported estimates of physical activity are similar, thus leading to error correlations between the measurements of two types of instruments—in our study, PAQ and PA LOGs—and between replicates of the same instrument. The results

**TABLE 5.** Estimates of validity coefficient ( $\hat{\rho}_{XT}$ ) and attenuation factors ( $\hat{\lambda}$ ) for measurements *X*, in turn, PAQ,\* 7-day PA LOG,\* and ACC\* measurements, estimates of true variance ( $\hat{\sigma}_T^2$ ), and error correlations ( $\hat{\rho}_{e_{ijk}}$ ) and associated 95% confidence interval under model 4, assuming correlation between errors in replicates of accelerometer measurements

Measurement	Variable name	$\hat{\rho}_{XT}$		$\hat{\lambda}$		$\hat{\sigma}_T^2$		$\hat{\rho}_{e_{ijk}}$	
		Estimate	95% CI*	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI
PAQ	<i>Q</i>	0.269	0.129, 0.486	0.145	0.059, 0.381	0.076	0.015, 0.134	0.213†	−0.170, 0.375
PA LOG	<i>R</i>	0.603	0.479, 0.879					0.523‡	−0.400, 0.626
ACC	<i>A</i>	0.762	0.507, 0.897					0.177§	−0.750, 0.537

\* PAQ, physical activity questionnaire; PA LOG, physical activity log; ACC, accelerometer; CI, confidence interval.

† Correlation between errors in *Q* and *R* measurements.

‡ Correlation between errors among replicates of *R* measurements.

§ Correlation between errors among replicates of *A* measurements.

of this study confirm that, after the role of error correlations has been taken into account, the validity coefficients of physical activity levels from questionnaire measurements can be relatively low. It is therefore of primary interest to set up validation studies in which physical activity levels are estimated through different assessment instruments.

In this study, the assumption that errors in replicates of accelerometer measurements are uncorrelated was challenged through a sensitivity analysis and by considering a model with an instrumental variable (24, 28, 29). Our results provide some evidence that the impact of error correlation in replicates of accelerometers is minor, as the validity coefficient is not sizably affected for values of accelerometer error correlations ranging from  $-0.4$  to  $0.4$ . To estimate the validity coefficient, it was assumed that individual weights were linearly related to the true level of the log of physical activity and that errors in this relation were uncorrelated with errors in physical activity measurements ( $Q$ ,  $R$ , and  $M$ ). Concerns regarding the appropriateness of these assumptions may exist, and our results need to be replicated in models with objective measurements with independent errors over replicates (e.g., doubly labeled water).

We also evaluated the effects of measurement errors on the validity of questionnaire measurements according to study subjects' characteristics, notably age, body mass index, and level of fitness. Higher validity coefficients were observed for younger subjects (below age 50 years) and for subjects below the gender-specific body mass index medians, thus suggesting that these factors are associated with accuracy of estimates of physical activity level. These results provide evidence that measurement errors in the assessment of past year physical activity patterns have a strong systematic component, which can be the result of within-person systematic error not randomly distributed between subjects (38). Kipnis et al. (23) refer to this bias as person- and group-specific bias. Although the effect of body mass index on the validity coefficients of dietary measurements has been repeatedly suggested in the dietary assessment literature (39–41), the evidence on the role of study subjects' characteristics in the accuracy of physical activity measurements has been more mixed, with some studies demonstrating an effect of body mass index on reporting errors (42, 43), while others have not (3, 44). However, concerns could be raised about whether these results could be generalized to populations different from the one under study, that is, with a different age structure or ethnic composition.

We have shown that the degree of attenuation of risk estimates derived from instruments such as the PAQ is likely substantial. Thus, the magnitude of the associations between physical activity and health outcomes, such as cancer or cardiovascular disease, may be severely attenuated by measurement errors. We estimated an attenuation factor equal to 0.13 for the PAQ, and, given this level of attenuation, a true relative risk of 2 would be estimated as a relative risk of 1.10. Therefore, to estimate the magnitude of the association between physical activity and health more accurately, carefully conducted calibration studies may be required, in which different types of physical activity assessment are used, possibly on a subsample of a large-scale epidemiologic study.

Our findings are based on log-transformed variables and therefore apply to risk models in which log of physical activity measurements is related to a specific disease outcome. However, results on untransformed variables provided similar results (data not shown). Results from this study also have implications for the design of calibration studies. The correlation between errors in self-reported measurements was substantial and must be properly accounted for in the measurement error model; otherwise, the model might only partially correct for the effect of measurement error. In our study, estimates of attenuation factors using a measurement error model were generally two- to threefold lower than attenuation estimated in a "standard" regression calibration model, in which questionnaire measurements were directly regressed on PA LOG values, without using accelerometer data. The attenuation factor estimated in this way was equal to 0.30 (data not shown), thus leading to underestimation of the deattenuated risk, which would be equal to 1.23 for a true relative risk of 2. These results suggest that the overall level of attenuation may be greater than previously expected, which confirms the importance of estimating the attenuation factors accurately because quantification of the impact of physical activity on public health outcomes will be directly affected by such attenuation.

In the absence of objective measurements that provide accurate estimates of absolute level of physical activity, questionnaires should be combined with accelerometer data and PA LOG measurements. Accelerometers ensure independence of errors with self-reported measurements, thus fulfilling fundamental requirements for the identifiability of statistical models for complex validation study designs. In this way, it is possible to estimate the various quantities that compose the error structure and to better understand the associations between physical activity and health.

The design of validation studies therefore requires great care, as well as thoughtful considerations concerning the sample size of the study, to enable the potential sources of errors, attributable to study subjects' individual characteristics and/or particular types of physical activity, to be investigated in depth.

## ACKNOWLEDGMENTS

The authors thank Heather Neilson for reviewing drafts of the manuscript. They also thank Bernard Rosner for his helpful suggestions regarding formalization of the statistical model.

Conflict of interest: none declared.

## REFERENCES

1. Philippaerts RM, Westerterp KR, Lefevre J. Comparison of two questionnaires with a tri-axial accelerometer to assess physical activity patterns. *Int J Sports Med* 2001;22:34–9.

2. Schuit AJ, Schouten EG, Westerterp KR, et al. Validity of the Physical Activity Scale for the Elderly (PASE): according to energy expenditure assessed by the doubly labeled water method. *J Clin Epidemiol* 1997;50:541–6.
3. Adams SA, Matthews CE, Ebbeling CB, et al. The effect of social desirability and social approval on self-reports of physical activity. *Am J Epidemiol* 2005;161:389–98.
4. de Klerk NH, English DR, Armstrong BK. A review of the effects of random measurement error on relative risk estimates in epidemiological studies. *Int J Epidemiol* 1989;18:705–12.
5. Armstrong BG. The effects of measurement errors on relative risk regressions. *Am J Epidemiol* 1990;132:1176–84.
6. Fuller WA. Measurement error model. New York, NY: Wiley, 1987.
7. Carroll R, Ruppert D, Stefansky LA. Measurement error in nonlinear models. London, United Kingdom: Chapman & Hall, 1995.
8. Rosner B, Willett WC, Spiegelman D. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Stat Med* 1989;8:1051–69.
9. Armstrong BG, Whittemore AS, Howe GR. Analysis of case-control data with covariate measurement error: application to diet and colon cancer. *Stat Med* 1989;8:1151–63; discussion 1165–6.
10. Prentice RL, Willett WC, Greenwald P, et al. Nutrition and physical activity and chronic disease prevention: research strategies and recommendations. *J Natl Cancer Inst* 2004;96:1276–87.
11. Plummer M, Clayton D. Measurement error in dietary assessment: an investigation using covariance structure models. Part I. *Stat Med* 1993;12:925–35.
12. Kaaks R, Ferrari P, Ciampi A, et al. Uses and limitations of statistical accounting for random error correlations in the validation of dietary questionnaire assessments. *Public Health Nutr* 2002;5:969–76.
13. Friedenreich CM, Courneya KS, Neilson HK, et al. Reliability and validity of the Past Year Total Physical Activity Questionnaire. *Am J Epidemiol* 2006;163:959–70.
14. Ainsworth BE, Haskell WL, Leon AS, et al. Compendium of physical activities: classification of energy costs of human physical activities. *Med Sci Sports Exerc* 1993;25:71–80.
15. Ainsworth BE, Haskell WL, Whitt MC, et al. Compendium of physical activities: an update of activity codes and MET intensities. *Med Sci Sports Exerc* 2000;32:S498–504.
16. Swartz AM, Strath SJ, Bassett DR Jr, et al. Estimation of energy expenditure using CSA accelerometers at hip and wrist sites. *Med Sci Sports Exerc* 2000;32:S450–6.
17. Kaaks R, Riboli E, Estève J. Estimating the accuracy of dietary questionnaire assessments: validation in terms of structural equation models. *Stat Med* 1994;13:127–42.
18. Freedson PS, Melanson E, Sirard J. Calibration of the Computer Science and Applications, Inc. accelerometer. *Med Sci Sports Exerc* 1998;30:777–81.
19. Janz KF. Validation of the CSA accelerometer for assessing children's physical activity. *Med Sci Sports Exerc* 1994;26:369–75.
20. Matthews CE. Calibration of accelerometer output for adults. *Med Sci Sports Exerc* 2005;37(11 suppl):S512–22.
21. Nichols JF, Morgan CG, Chabot LE, et al. Assessment of physical activity with the Computer Science and Applications, Inc. accelerometer: laboratory versus field validation. *Res Q Exerc Sport* 2000;71:36–43.
22. Welk GJ, Schaben JA, Morrow JR Jr. Reliability of accelerometry-based activity monitors: generalizability study. *Med Sci Sports Exerc* 2004;36:1637–45.
23. Kipnis V, Carroll RJ, Freedman LS, et al. Implications of a new dietary measurement error model for estimation of relative risk: application to four calibration studies. *Am J Epidemiol* 1999;150:642–51.
24. Spiegelman D, Zhao B, Kim J. Correlated errors in biased surrogates: study designs and methods for measurement error correction. *Stat Med* 2005;24:1657–82.
25. Michels KB, Day NE. Re: "Implications of a new dietary measurement error model for estimation of relative risk: application to four calibration studies." (Letter). *Am J Epidemiol* 2000;152:494–5.
26. Bentler PM, Weeks DG. Linear structural equations with latent variables. *Psychometrika* 1980;45:289–308.
27. Bollen KA. A new incremental fit index for general structural equation models. *Sociol Methods Res* 1989;17:303–16.
28. Greenland S. An introduction to instrumental variables for epidemiologists. *Int J Epidemiol* 2000;29:722–9.
29. Fraser GE, Butler TL, Shavlik D. Correlations between estimated and true dietary intakes: using two instrumental variables. *Ann Epidemiol* 2005;15:509–18.
30. Little RJA, Rubin DB. Statistical analysis with missing data. New York, NY: Wiley, 1987.
31. SAS Institute, Inc. SAS Online Doc, version 8. Cary, NC: SAS Institute, Inc, 1999.
32. Efron B, Tibshirani RJ. An introduction to the bootstrap. London, United Kingdom: Chapman & Hall, 1998.
33. Wong MY, Day NE, Wareham NJ. Measurement error in epidemiology: the design of validation studies II: bivariate situation. *Stat Med* 1999;18:2831–45.
34. Wareham NJ, Wong MY, Hennings S, et al. Quantifying the association between habitual energy expenditure and blood pressure. *Int J Epidemiol* 2000;29:655–60.
35. Haggarty P, McNeill G, Manneh MK, et al. The influence of exercise on the energy requirements of adult males in the UK. *Br J Nutr* 1994;72:799–813.
36. Racette SB, Schoeller DA, Kushner RF. Comparison of heart rate and physical activity recall with doubly labeled water in obese women. *Med Sci Sports Exerc* 1995;27:126–33.
37. Kipnis V, Subar AF, Midthune D, et al. Structure of dietary measurement error: results of the OPEN biomarker study. *Am J Epidemiol* 2003;158:14–21.
38. Willett W. An overview of issues related to the correction of non-differential exposure measurement error in epidemiologic studies. *Stat Med* 1989;8:1031–40.
39. Bandini LG, Schoeller DA, Cyr HN, et al. Validity of reported energy intake in obese and nonobese adolescents. *Am J Clin Nutr* 1990;52:421–5.
40. Heitmann BL. The influence of fatness, weight change, slimming history and other lifestyle variables on diet reporting in Danish men and women aged 35–65 years. *Int J Obes* 1993;17:329–36.
41. Heitmann BL, Lissner L. Dietary underreporting by obese individuals—is it specific or non-specific? *BMJ* 1995;311:986–9.
42. Irwin ML, Ainsworth BE, Conway JM. Estimation of energy expenditure from physical activity measures: determinants of accuracy. *Obes Res* 2001;9:517–25.
43. Lichtman SW, Pisarska K, Berman ER, et al. Discrepancy between self-reported and actual caloric intake and exercise in obese subjects. *N Engl J Med* 1992;327:1893–8.
44. Klesges RC, Eck LH, Mellon MW, et al. The accuracy of self-reports of physical activity. *Med Sci Sports Exerc* 1990;22:690–7.



## APPENDIX

In this work, parameter estimates were obtained through the CALIS procedure (31) that uses the maximum likelihood estimation algorithm. As described in the text, one questionnaire, four PA LOGs, and four accelerometer measurements were modeled in expression 1. In this study, six linear constraints were used for error correlation among replicates of PA LOG measurements, as  $\sigma_{\varepsilon R_{ij}, \varepsilon R_{ik}} = \sigma_{\varepsilon R, \varepsilon R}$  for  $\forall j, k \in J_R$  with  $j \neq k$ , as well as four constraints for the error correlation between questionnaire and PA LOG, as  $\sigma_{\varepsilon Q_i, \varepsilon R_{ij}} = \sigma_{\varepsilon Q, \varepsilon R}$  for  $\forall j \in J_R$ . Parameter estimates can be alternatively obtained with the method-of-moments estimators, as shown by Spiegelman et al. (24). In model 1, there are eight second moments that allow eight model parameters to be uniquely determined, as

$$\begin{aligned} \theta_1 &= \text{Var}(Q_i) = \beta_Q^2 \sigma_T^2 + \sigma_{\varepsilon Q}^2 & \beta_A &= \theta_8 / \theta_7 \\ \theta_2 &= \text{Var}(R_{ij}) = \sigma_T^2 + \sigma_{\varepsilon R}^2 & \sigma_T^2 &= \theta_7^2 / \theta_8 \\ \theta_3 &= \text{Cov}(R_{ij}, R_{ik}) = \sigma_T^2 + \sigma_{\varepsilon R, \varepsilon R}, \quad j \neq k & \beta_Q &= \theta_6 / \theta_7 \\ \theta_4 &= \text{Cov}(Q_i, R_{ij}) = \beta_Q \sigma_T^2 + \sigma_{\varepsilon Q, \varepsilon R} & \sigma_{\varepsilon Q}^2 &= \theta_1 - \theta_6^2 / \theta_8 \\ \theta_5 &= \text{Var}(A_{ij}) = \beta_A^2 \sigma_T^2 + \sigma_{\varepsilon A}^2 & \sigma_{\varepsilon R}^2 &= \theta_2 - \theta_7^2 / \theta_8 \\ \theta_6 &= \text{Cov}(A_{ij}, Q_i) = \beta_A \beta_Q \sigma_T^2 & \sigma_{\varepsilon R, \varepsilon R} &= \theta_3 - \theta_7^2 / \theta_8 \\ \theta_7 &= \text{Cov}(A_{ij}, R_{ik}) = \beta_A \sigma_T^2, \quad j \neq k & \sigma_{\varepsilon A}^2 &= \theta_5 - \theta_8 \\ \theta_8 &= \text{Cov}(A_{ij}, A_{ik}) = \beta_A^2 \sigma_T^2, \quad j \neq k & \sigma_{\varepsilon Q, \varepsilon R} &= \theta_4 - \theta_6 \theta_7 / \theta_8 \end{aligned}$$

Following the above associations, the validity coefficient for questionnaire measurements that estimates the correlation between questionnaire  $Q$  and true physical activity level,  $T$ , is equal to

$$\rho_{QT} = \frac{\text{Cov}(Q, T)}{\sqrt{\text{var}(Q)} \sqrt{\text{var}(T)}} = \sqrt{\frac{1}{1 + \frac{\hat{\sigma}_{\varepsilon Q}^2}{\hat{\beta}_Q^2 \hat{\sigma}_T^2}}} = \theta_6 / (\theta_8 \theta_1)^{1/2} = \text{Cov}(A_{ij}, Q_i) / [\text{Cov}(A_{ij}, A_{ik}) \text{Var}(Q_i)]^{1/2}.$$

Similarly, for the attenuation factor,

$$\lambda_{QT} = \frac{\text{Cov}(Q, T)}{\text{var}(Q)} = \frac{\hat{\beta}_Q \hat{\sigma}_T^2}{\hat{\beta}_Q^2 \hat{\sigma}_T^2 + \hat{\sigma}_{\varepsilon Q}^2} = \theta_6 \theta_7 / (\theta_8 \theta_1) = \text{Cov}(A_{ij}, Q_i) \text{Cov}(A_{ij}, R_{ij}) / [\text{Var}(Q_i) \text{Cov}(A_{ij}, A_{ik})].$$

In model 4, an instrumental variable has been added to the model. It is assumed that  $\sigma_{\varepsilon A_{ij}, \varepsilon A_{ik}} = \sigma_{\varepsilon A, \varepsilon A}$  for  $\forall j, k \in J_A$  with  $j \neq k$ . Similarly to model 1,

$$\begin{aligned} \theta_1 &= \text{Var}(Q_i) = \beta_Q^2 \sigma_T^2 + \sigma_{\varepsilon Q}^2 & \beta_A &= \theta_{12} / \theta_{11} \\ \theta_2 &= \text{Var}(R_{ij}) = \sigma_T^2 + \sigma_{\varepsilon R}^2 & \sigma_T^2 &= \theta_{11} \theta_7 / \theta_{12} \\ \theta_3 &= \text{Cov}(R_{ij}, R_{ik}) = \sigma_T^2 + \sigma_{\varepsilon R, \varepsilon R}, \quad j \neq k & \sigma_{\varepsilon R}^2 &= \theta_2 - \theta_{11} \theta_7 / \theta_{12} \\ \theta_4 &= \text{Cov}(Q_i, R_{ij}) = \beta_Q \sigma_T^2 + \sigma_{\varepsilon Q, \varepsilon R}, & \sigma_{\varepsilon R, \varepsilon R} &= \theta_3 - \theta_{11} \theta_7 / \theta_{12} \\ \theta_5 &= \text{Var}(A_{ij}) = \beta_A^2 \sigma_T^2 + \sigma_{\varepsilon A}^2 & \beta_B &= \theta_{12} / \theta_7 \\ \theta_6 &= \text{Cov}(A_{ij}, Q_i) = \beta_A \beta_Q \sigma_T^2 & \sigma_{\varepsilon A}^2 &= \theta_5 - \theta_{12} \theta_7 / \theta_{11} \\ \theta_7 &= \text{Cov}(A_{ij}, R_{ij}) = \beta_A \sigma_T^2, \quad j \neq k & \sigma_{\varepsilon B}^2 &= \theta_9 - \theta_{12} \theta_{11} / \theta_7 \\ \theta_8 &= \text{Cov}(A_{ij}, A_{ik}) = \beta_A^2 \sigma_T^2 + \sigma_{\varepsilon A, \varepsilon A}, \quad j \neq k & \sigma_{\varepsilon A, \varepsilon A} &= \theta_8 [1 - \theta_8 \theta_{11} / (\theta_7 \theta_{12})] \\ \theta_9 &= \text{Var}(B_i) = \beta_B^2 \sigma_T^2 + \sigma_{\varepsilon B}^2 & \beta_Q &= \theta_6 / \theta_7 = \theta_{10} / \theta_{11} \\ \theta_{10} &= \text{Cov}(B_i, Q_i) = \beta_B \beta_Q \sigma_T^2 & \sigma_{\varepsilon Q, \varepsilon R} &= \theta_4 - \theta_6 \theta_{11} / \theta_{12} = \theta_4 - \theta_{10} \theta_7 / \theta_{12} \\ \theta_{11} &= \text{Cov}(B_{ij}, R_{ik}) = \beta_B \sigma_T^2, \quad j \neq k & \sigma_{\varepsilon Q}^2 &= \theta_1 - \theta_7 \theta_{10}^2 / (\theta_{12} \theta_{11}) \\ \theta_{12} &= \text{Cov}(B_i, A_{ij}) = \beta_B \beta_A \sigma_T^2 \end{aligned}$$

In this model, the number of moments ( $n = 12$ ) exceeds the number of parameters ( $n = 11$ ). Therefore, more than one estimator is available for the parameters  $\beta_Q$ ,  $\sigma_{\varepsilon Q, \varepsilon R}$ , and  $\sigma_{\varepsilon Q}^2$ . In this work, model parameters are estimated iteratively by a nonlinear optimization algorithm in the CALIS procedure (31), which uses maximum likelihood theory.