EVALUATION NOTE
ROCKWOOL FOUNDATION RESEARCH UNIT

VALIDATION OF SCHOOL-BASED ANTHROPOMETRIC MEASUREMENTS AND MEASUREMENTS OF PHYSICAL FITNESS AMONG CHILDREN IN DENMARK. ATTRITION, INTRA-OBSERVER RELIABILITY AND INTER-OBSERVER RELIABILITY

Written by:
John Singhammer and
Lars Bo Andersen

JANUARY 2015

EVALUATION NOTE
Validation of school-based anthropometric measurements and measurements of physical fitness among children in Denmark. Attrition, intra-observer reliability and inter-observer reliability
Published by Rockwool Foundation Research Unit

Rockwool Foundation Research Unit
Sølvgade 10, 2. sal
DK-1307 K Copenhagen
Tlf.: 33344800
Fax: 33344899

AUTHORS:
John Singhammer
Landesamt für Gesundheit und Lebensmittelsicherheit (LGL),
Nürnberg, Germany

Lars Bo Andersen
Department of Exercise Epidemiology,
Institute of Sport Sciences and Clinical Biomechanics,
University of Southern Denmark

Sogn and Fjordane University College,
Sogndal, Norway

ABSTRACT

Aim: The purpose of the present study is to assess the intra-observer and inter-observer reliability of school-based measurements of anthropometry and physical fitness collected by schoolteachers, and to analyse attrition.

Methods: The Healthy School Network (HSN) was initiated in 2009 by the Rockwool Foundation to monitor indicators of children's health and to provide the basis for health-promoting initiatives. Teachers in the 128 schools in the Network out of 2,425 schools in Denmark in 2011 = 5%, and 114 schools out of 2,299 in 2012 =5% obtained measurements of height, weight, waist circumference (WC), vertical jump test height and cardiorespiratory fitness (CRF) among children in grades 0 to 9. Schools enrolled in the HSN project on a voluntary basis. Intra-observer reliability was assessed by comparing schools' measurement results with their own control measurements of 635 children in 50 schools collected in the school year 2010/11 (n = 700 for 35 schools in 2011/12), the original and the control measurements being taken two weeks apart. Inter-observer reliability was assessed by comparing results of tests conducted by researchers from the University of Southern Denmark (SDU) with schools' own results for a sample of 111 children, and by comparing results obtained by school nurses with schools' own results for 197 children. Reliability was assessed by calculation of the concordance correlation coefficient (ρ_c) and by visual inspection of Bland-Altman plots for separate grades.

Results: High intra-observer reliability was observed for measurements obtained in 2010/11 for height (ρ_c = 0.92), weight (ρ_c = 0.96), WC (ρ_c = 0.87), vertical jump test (ρ_c =0.86) and CRF (ρ_c = 0.83). Estimates of similar size were calculated for measurements obtained in 2011/12. Inter-observer reliability was high for measurements of height (ρ_c = 0.97) and weight (ρ_c = 0.94), moderate for WC (ρ_c = 0.81) and vertical jump test (ρ_c = 0.71), and low for CRF (ρ_c = 0.18). Inter-observer reliability for measurements obtained by school nurses and schools themselves in a subset of the HSN study in the municipality of Odense was high for height (ρ_c = 0.99), weight (ρ_c = 0.99) and WC ((ρ_c = 0.91). The participation rate for the school years 2008/09-2011/12 was 0.09% out of all schools in Denmark in 2008/09, 6.5% in 2009/10, 7.5% in 2010/11 and 11.8% in 2011/12. The percentage of children with valid measurements varied across the study period, with the most measurements being obtained in 2009/10 (50-55%) and the lowest proportion of measurements being obtained in 2011/12 (28-33%). Within participating schools, 46% of the children were measured at least once. Valid measurements of anthropometry and physical fitness varied by grade,

BMI and social factors related to the children's parents, and were highest for children in grades 0 to 6 and lowest for children in grades 7 to 9. A similar pattern was observed for the Odense sub-study.

Conclusions: Schoolteachers can provide reliable measurements of children's height, weight, and WC but need additional training in measuring vertical jump height and CRF. Training should include guided practice of the test procedures in a standardised form to ensure conformity to the test instruction sheets. Indication of selection bias suggests that the results for children in the highest grades and in the highest BMI deciles should be interpreted with caution.

CONTENTS

BACKGROUND

For the past 30 years, the proportion of obese children and adolescents in the population has increased, probably because of increased access to an energy-rich diet and a decreased level of physical activity. The importance of physical fitness of types such as cardiorespiratory fitness (CRF) to health and to the prevention of obesity is unambiguous (21). CRF and obesity are strongly related to the clustering of cardiovascular disease (CVD) risk factors. We recently found a 13-fold increase in clustered CVD risk in the least fit quartile compared to the most fit quartile of children and adolescents in the Danish population (6). We also found that low fitness levels and overweight were associated independently with clustered CVD risk. Fitness is a modifiable CVD risk factor, and a change in fitness may reflect increased physical activity level or decreased body weight. Monitoring of children's and adolescents' body composition and physical fitness of types such as muscular strength and CRF is essential for healthy development and to identify indicators of future health risks. Regular monitoring of anthropometry and/or physical fitness of school children is conducted in the USA (30 states) (9, 20) and in other countries, including Denmark. For example, all children in Denmark visit school nurses in grades 0, 1 and 9, and are measured for height and weight – but not for waist circumference or physical fitness. Thus, knowledge about anthropometry and physical fitness among Danish school children is limited to a few age groups and to selected measurements of anthropometry. Information on physical fitness is lacking, although some knowledge exists from large observational studies investigating CVD risk factors (26, 36). Thus, enabling schoolteachers to assess anthropometry and physical fitness among children in all grades on an annual basis is believed to be a time-efficient and cost-effective method of monitoring the development of health risk factors across the entire age range, and some evidence suggests that teachers can make reliable measurements of height and weight (9). Furthermore, such measurements can be used in school lessons focusing on health promotion. However, little evidence exists of teachers' ability to make reliable measurements of children's physical fitness.

The purpose of the present study is to assess whether teachers' measurements of anthropometry, vertical jump test and CRF were accurate by describing the intra-observer reliability and inter-observer reliability of the measurements using available data from all individuals participating in the Healthy School Network, as well as to analyse the extent and patterns of attrition.

METHODS

The Healthy Schools Network (HSN) was initiated by the Rockwool Foundation with the aim of engaging teachers and school staff in the monitoring of indicators of children's health (anthropometry and physical fitness). Schools were encouraged to incorporate the results of three anthropometric measurements (height, weight and waist circumference (WC)), vertical jump height (28) and cardiorespiratory fitness (CRF) as assessed by the Andersen test (5) into the curriculum. The measurements of anthropometry and physical fitness included in the present study are used in other school-based studies in Denmark and are easy to administer to a whole school class within a 90-minute PE lesson (3, 4, 36), which is the norm in Danish schools. Using these measurements also enables us to compare the results of the HSN study with those of other school-based studies recently conducted in Denmark (3, 4, 36).

The HSN covers children in grades 0 to 9. Schoolteachers measure children in the participating schools, and the resulting data are subsequently uploaded to a central database. For some tests (the vertical jump test and the Andersen test), children have assisted in the measurement procedure.

The HSN started in 2008/09 and is still operating. Schools have been recruited to the HSN throughout the period from 2008/09 to the present largely by means of recommendations from existing network members. The total number of schools in Denmark was 2,425 in 2010/11. Of these, 128 schools (5%) had volunteered to participate in the HSN (114 schools out of 2,299 possible volunteered in 2011/12 = 5%). The schools received an annual payment of €10 per child for at least one measurement (€5 per child in 2010/11). To enhance accuracy of all measurements, a set of guidelines was adopted and incorporated into detailed electronic instruction sheets (fieldwork manuals) that are available at the HSN website http://www.sundskolenettet.dk/. When they join the network, schools agree to 1) form a health committee and appoint a teacher at the school to be responsible for health, 2) measure at least 80% of the children in the school on height, weight, WC, vertical jump, and distance run within a set time, 3) adhere to the guidelines for storing and uploading the annual measurements to a central database, and 4) deliver a yearly report to the Rockwool Foundation describing the health-promoting initiatives taken at the school. In addition, schools agree to use the standard measuring equipment supplied by the Rockwool Foundation. All regulations regarding the ethical issues involved in measuring the children have to be accepted by the schools, and are incorporated into the instruction sheets. Individual children are assigned a unique identification code to ensure anonymity. Our main purpose was to evaluate schoolteachers'

ability to measure the children accurately. Consequently, we carefully selected children for our study that had been measured twice within a reasonable period of time, in order to avoid the effects of natural physical development on the consistency of the results. Measurements of anthropometry and physical fitness were obtained through the HSN for 44,126 children in the school year 2010/11 and for 43,156 children in 2011/12 (Figure 1). Control measurements (a second measurement performed by the same teacher or a colleague) of 2,348 children obtained within the same school year were conducted in 2010/11 and of 2,379 children in 2011/12 (Figure 2). Of these, valid control measurements obtained within 14 days were found for 635 children in 49 schools (700 children in 35 schools in 2011/12), and these measurements were used for the present study.

Control measurements of the vertical jump test were not collected in the school year 2011/12. Hence, we chose to focus on data from 2010/11. However, intra-observer reliability is reported for the other measurements obtained in 2011/12.

Information from Statistics Denmark enabled us to calculate the proportion of schools participating in the HSN relative to the total number of schools in Denmark. We also had information on the number of children within each grade who participated in the HSN relative to the total number of children at the particular school.

In the spring of 2012 (March to June), researchers from The Institute of Sports Science and Clinical Biomechanics (ISSCB) at the University of Southern Denmark (SDU) visited five schools. These schools were selected to reflect rural/urban differences and to ensure a geographical spread. The researchers performed the compulsory measurements of anthropometry and fitness. Children from two grades were tested at each school. It was the intention to test all the children from the selected classes, regardless of prior participation in the school's own tests. All tests were performed with similar equipment to that used by the schools and in the same settings. The testers coordinated the measurement procedures before visiting the schools, and agreed on how to address potential problems associated with the test procedures, in order to promote standardised measurement procedures across testers and to reduce error in subsequent analyses. All measurements were taken twice, on occasions two days apart. For each measurement, we calculated the arithmetic mean of the pairs of results for the children in all grades, and these scores formed the references that were compared to the schools' results obtained in 2010/11. We grouped children in different grades into broader bands to preserve statistical power (0^{th} to 3^{rd} grades, 4^{th}

to 6th grades and 7th to 9th grades). Sadly, the schools did not measure the same children in 2011/12, and this precluded a full assessment of the reliability of measurements for that year. Therefore, we assessed inter-observer reliability by comparing the measurements obtained by SDU with schools' own measurements of the same children in 2010/11, adjusted for expected growth in anthropometry and development in physical fitness.

For completeness, we evaluated the inter-observer reliability of the data by including results from a small section of the HSN initiative which were conducted by school nurses in the municipality of Odense. School nurses are responsible for the preventive care of children in all schools in Denmark, and monitoring children's growth is an integral part of their job. Thus, school nurses are trained to take measurements of height, weight and WC and have extensive experience, as all children in the 0th, 1st, 5th, and 9th grades are obliged to visit a school nurse. Included were measurements of height, weight and WC of children in grades 0, 1, 5 and 9 in 2009/10 and 2011/12 (vertical jump and CRF were not measured) (n = 6,617). We used these data in three ways in this study. First, for 2011/12 we compared the school nurses' measurements with data obtained by experienced researchers (at other schools). Second, we calculated the inter-observer reliability by comparing measurements obtained by school nurses with measurements obtained by schoolteachers 14 days later, and third, for the 14 HSN schools in Odense (which were randomly selected to participate in the HSN from among all schools in Odense (15)), we used the school nurse data to investigate whether children who were not measured by teachers in the HSN differed in terms of height, weight and WC from children with HSN measurements. For instance, HSN attrition might have been more common for heavier children. Pair-wise measurements were pooled across the study period 2009/10–2011/12.

ATTRITION

We calculated the participation rates for schools and for children within participating schools. First, we calculated the percentage of schools that participated in the HSN. Second, from within the schools that did participate, we calculated the percentage of children that participated compared with those who did not. Third, we calculated the percentage of children with complete sets of measurements compared to those with incomplete sets of measurements of anthropometry and physical fitness, and assessed whether the partially complete measurements co-varied with other measurements of anthropometry and fitness or with social factors related to the children's parents. We also investigated for possible

relationships between missing values on any measurements of anthropometry or physical fitness and children's immigrant/non-immigrant status or parental social factors by children's gender, grade and deciles of BMI.

MEASUREMENTS OF ANTHROPOMETRY

Anthropometric measurements included height, weight and WC. Height was measured in centimetres using a stadiometer attached to the wall. Children were measured in erect position and without shoes. Weight was measured in kilograms to one decimal place using a simple electronic personal scale (KORONA**™**). Children were dressed lightly and were not wearing shoes. WC was measured in centimetres using a soft non-elastic tape (Meterex DBGM). Measurements were obtained when the child was relaxed and exhaling, and were made at a point between the lowest rib and the iliac crest (approximately 2 cm above the umbilicus). Children were instructed to be dressed lightly, e.g. wearing shorts and a t-shirt.

MEASUREMENTS OF PHYSICAL FITNESS

Measurements of physical fitness included jump height in centimetres obtained with the vertical jump test, and CRF measured indirectly by the Andersen test, a 10-minute intermittent running test (5). Measurements of vertical jump height (vertical jump test) and CRF were obtained after a warm-up period. The vertical jump was measured as the maximum vertical distance in centimetres achieved in a vertical jump. Arm swinging and bending of the hips and knees were allowed. The vertical distance was measured using a special jump mat that has been developed in a collaboration between the Rockwool Foundation and the Institute of Sports and Biomechanics, University of Southern Denmark. The mat is made of 0.5-centimetre thick non-elastic rubber. The child wears a belt to which the end of a non-elastic smooth tape is attached. The tape runs through an eye in the middle of the mat, with sufficient friction to avoid shifting caused by the child's ordinary movements. Before the jump, children are instructed to stand erect with both feet on the mat. The test administrator stretches the tape, and adjusts the belt to the iliac crest, to avoid the belt moving because of inappropriate clothing. The vertical distance jumped by the child is the difference between the values on the tape before and after the jump. Each child performed the vertical jump test three times, and the greatest vertical distance achieved was recorded. A jump which caused the child to land with one foot outside the mat was discounted and the child was allowed a new trial. Teachers were assisted by older children in the measurement procedure.

CRF was measured indirectly by means of a 10-minute intermittent running test (the Andersen test (5)). In this test, two parallel lines 20 metres apart are marked on the floor of the gym. Children run from one line to the other, touching the floor behind the line before turning round and running back. The children run for 15 seconds and then take a 15-second rest, signalled by a computer sound file or the test administrator. The children are instructed to run as fast as possible during the exercise periods so as to cover the longest distance they can during the entire test period. They must stop running immediately at the sound of the signal and remain at the same position during the rest periods. The total distance covered is the test result. Children are divided into pairs, with one child running and the other counting the number of crossings between the parallel lines. The results of the CRF test are recorded in meters, representing the distance completed by the child. The teacher collects the test results for each pair of children and uploads the data on anthropometric measurements and physical fitness after completing all the tests.

Register-based information on children's date of birth, school attended each year and academic grades achieved each year was obtained for each child in the study from Statistics Denmark for the school years 2008/09-2011/12 (12) Thus, we were able to see any changes of school and/or deviance from expected academic progress from one year to another. However, the registers used do not contain information that link a particular child to a particular class within the school. Information on children's immigrant status (Western immigrant, non-Western immigrant, native Dane), parents' level of education (less than nine years of schooling versus more) and parents' employments status (employed /unemployed) was also obtained from Statistics Denmark. We used the unique identification number assigned to all persons in Denmark with permanent or temporary residency status to link the HSN database with information from Statistics Denmark.

STATISTICAL ANALYSIS

Intra-observer reliability for the HSN data was evaluated by correlating the results of the tests made by schools with the results of control tests. Intra-observer reliability is also known as test-retest reliability (8). For the 2010/11 measurements, only three children in the 9th grade were tested twice, and these results were grouped with results from children in the 8th grade. Measurements obtained in 2011/12 contained sufficient information for all grades. Evaluation was aided by the calculation of the concordance

correlation coefficient (19) (ρ_c) – a statistic equivalent to the intra-class coefficient (11)– and by visual inspection of Bland-Altman plots, both overall and separately for each grade.

We evaluated the inter-observer reliability of the measurements (8) by comparing the results for individuals obtained by schoolteachers in 2010/11 with the results for the same individuals obtained by experienced researchers from SDU in 2011/12. As the results from SDU were compared with results obtained by schools on average 345.8 (SD = 64.1) days earlier, we calculated inter-observer reliability for measurements of anthropometry and physical fitness adjusted for the expected monthly growth observed among children in the HSN. To obtain adjusted measurements, we first calculated the elapse of time in months between the schools' measurements and the measurements obtained by SDU. Second, we calculated grade- and gender-specific average changes in height, weight, WC, vertical jump height and CRF using a linear regression technique by modelling the outcome of interest as a function of grade and months, for girls and boys separately. We calculated the linear combination of estimates to obtain the predicted change in the outcome of interest for each grade level. For example, the elapsed time between schools' and SDU's measurements for boys in the 5th grade was 9.6 months. To calculate the change in height for boys from the 4th to the 5th grade, we calculated the height for boys in the 4th grade and added the change in height for boys between 4th and 5th grade but restricted to 9.6 months. The derived estimate of change in height from 4th to 5th grade was then added to the existing measurements of height obtained in 2010/11. We repeated this procedure for all measurements of anthropometry and physical fitness for girls and boys in all grades. We assessed the accuracy of the adjusted control measurements by correlating the individual measurements obtained by schools in 2011/12, individual measurements obtained by SDU in 2011/12 and individual measurements obtained by school nurses in 2011/12. After correction of the schools' measurements obtained in 2010/11 for the expected changes, we compared the schools' measurements with those taken by the researchers by calculation of the concordance correlation coefficient (ρ_c) within bands of grades and by visual inspection of Bland-Altman plots. We also calculated inter-observer reliability (ρ_c) of anthropometric measurements obtained by school nurses in 2009/10-2011/12 and by HSN schools 14 days later.

We used multiple regression analysis to calculate the differences in measurements of height, weight, WC and BMI between measurements of children made by both school nurses and schools on the one hand and similar measurements of the same children measured by the school nurses only on the other.

Significant differences were interpreted as indications of selective attrition in the schools' measurements. Logistic regression analysis was used to evaluate the relationship between missing values for all measurements by grade and for WC by deciles of BMI, for boys and girls separately. A similar analytical approach was used to evaluate the relationship between missing values for WC, vertical jump height and the Andersen test by deciles of BMI for all children in the HSN. The large sample size meant that even small differences would be statistically significant, even though they might be of little practical relevance. In all the analyses, estimates were adjusted for children's immigrant status and for their parents' level of education and employment status. Regression analyses were adjusted for possible clustering effects within schools.

For indicators of intra-observer and inter-observer reliability, we used the criteria proposed by Baumgartner(8), expecting a high degree of agreement ($\rho_c > 0.80$) for anthropometric measurements, a moderate degree of agreement ($\rho_c > 0.70$) for the vertical jump test, and a lower degree of agreement ($\rho_c > 0.50$) for the results of CRF. With respect to physical fitness, we expected a lower degree of agreement in the results for children in lower grades than in those for children in higher grades, as we assumed that younger children would have greater difficulty in understanding the instructions before each test and would be less able to apply their physical effort strategically in the Anderson test. We used Stata version 12.1 to perform all statistical analyses (31). Alpha was set to 5%.

DATA CLEANING

Initially, information assembled by the schools from 2008/09 to 2011/12 was scrutinized for erroneous values by means of univariate and bivariate analysis. We standardised all variables by subtracting the mean for each value and dividing by the standard deviation (SD). We evaluated all standardised variables separately by age, grade and year of participation, and omitted all values exceeding ± 2.57 SD. Similarly, graphical presentations such as scatterplots revealed implausible combinations of values. These were further inspected in the data matrix. Suspect values were evaluated against similar measurements for the same individual in previous or later years – if such values existed. Obviously erroneous values were corrected. For example, a value of 237 cm for height for a boy in the 5th grade was corrected to 137 cm. We deleted the suspect value if no other values for previous or later years were available or if the value could not be plausibly corrected. The same procedure was used for cleaning schools' control measurements. Details of the data cleaning procedure for the HSN database are provided in Figure 1, and

details of the cleaning procedure for the HSN control data are provided in Figure 2. Fifty schools performed the control tests in 2010/11 on 635 children. Most schools limited their control tests to children from a single grade, though they did perform all the tests. On average, 10.5 (SD = 11.2) children were tested within a grade. Some schools performed control measurements for a single child in a specific grade, while others included more than 25 children in a grade. In 2011/12, thirty-six schools performed control test on 700 children, with an average of 14.3 children per grade (SD = 10.6).

RESULTS

HSN DATA: INTRA-OBSERVER RELIABILITY

Differences in results from test to re-test in 2010/11 varied considerably between grades for the vertical jump test and CRF, but less so for measurements of anthropometry (Table 1). Across all grades, the difference in height was 1 cm (SD = 5.9) and for weight 0.7 kg (SD = 3.8). Across all grades, intra-observer reliability (ρ_c) was high (exceeding 0.80) for all measurements (Table 2) and was also high for measurements obtained among 0th, 1st, 4th and 5th grade students. However, intra-observer reliability for height was only moderate for measurements obtained for 6th grade students. Among children in the 7th grade, intra-observer reliability for height was low (ρ_c = 0.29, 95% CI 0.17-0.42, n = 96). Inspection of the data for the 7th grade students revealed that height differed by more than 3 cm in 52 measurement pairs and more than 10 cm in 26 measurement pairs. Of these, 25 measurement pairs came from the same school (Figure 3). Intra-observer reliability for WC among the 6th grade students was 0.40 (95% CI 0.25-0.55, n = 78). Visual inspection of WC revealed that 9 observations differed by 20 cm or more between test and re-test (Figure 4). These observations all came from the same school. When these measurement pairs were removed from the data, the level of agreement increased to 0.91 (95% CI 0.88-0.95, n = 69). Similarly, the intra-observer reliability for the vertical jump test among 6th grade students was low (ρ_c = 0.48, 95% CI 0.32-0.63, n = 82). Intra-observer reliability for the vertical jump test among 8-9th grade students was low (ρ_c = 0.56, 95% CI 0.38-0.74, n = 55) (Table 2). Inspection of Bland-Altman plots revealed subtle patterns of deviation from agreement between the results obtained at test and re-test, suggesting a positive correlation between differences in the vertical jump test results obtained at two points in time and the mean of the two results (heteroscedasticity) (Figure 5). Overall, intra-observer reliability was moderate for CRF (Table 2), and lowest for children in the 6th grade (ρ_c = 0.64, 95% CI 0.48-0.73, n = 72) and the 7th grade (ρ_c = 0.60, 95% CI 0.45-0.74, n = 72). On average, the distances run at re-test for children

in the 6th and 7th grades were 64.2 metres (95% CI -349.75-221.31) and 30.5 metres (95% CI -240.57-179.60) shorter than shown in the results obtained at the initial test (Table 1). The Appendix presents Bland-Altman plots of all pairs of measurements for 2010/11 (Figures 1-47) and 2011/12 (Figures 50-93).

Intra-observer reliability was high to moderate for measurements of height and weight in all grades obtained in 2011/12 (Table 3). However, measurements of weight obtained on test and re-test in the 3rd grade differed by an average of 2.6 kg (95% CI -6.1-11.2, n = 62). The intra-observer reliability for WC among children in the 3rd grade was poor, with a mean difference between test and re-test of 3.0 cm (95% CI -7.6-13.6, n = 58). Visual inspection of the BA plot revealed that the deviances between the results from the two test sessions were present across the whole range of WC measurements. Intra-observer reliability for CRF among children in the 3rd grade was acceptable (ρ_c = 0.64, 95% CI 0.47-0.82, n = 35), although measurements of CRF obtained on re-test were on average 16.7 metres longer compared with measurements obtained two weeks earlier (95% CI -140.2-173.5, n = 35).

We also evaluated the intra-observer reliability of pairs of measurements obtained by researchers from SDU (Appendix 2, Table 1). Comparison with the intra-observer reliability of teachers' measurements of height among 7th grade students and WC among 6th grade students was low, and that the researchers' ability to reproduce measurements of CRF was also generally low.

HSN DATA: INTER-OBSERVER RELIABILITY

We visually inspected the information on all children obtained from schools in 2010/11 (adjusted for expected changes until the time of measurements by SDU) and information on all children in the HSN schools in 2011/12 (Appendix 3). We calculated the correlations between grade-specific mean values for each separate measurement from schools' 2010/11 adjusted measurements, schools' measurements from 2011/12 and measurements obtained by SDU. We found high correlations between schools' adjusted measurements obtained in 2010/11 and measurements obtained by SDU for height ($r = 0.97$, $p < 0.001$), weight ($r = 0.94$, $p < 0.001$), WC ($r = 0.82$, $p < 0.001$), and vertical jump test ($r = 0.72$, $p < 0.001$), and low correlations for the Andersen test ($r = 0.2$, $p < 0.05$). The corresponding coefficients for schools' measurements obtained in 2011/12 and measurements obtained by SDU were for height, $r = 0.99$, $p < 0.001$; for weight, $r = 0.99$, $p < 0.001$; for WC, $r = 0.96$, $p < 0.001$; for the vertical jump test, $r = 0.97$, $p <$

0.001; and for the Andersen test, r = 0.83, p > 0.001. Comparison of schools' adjusted measurements with measurements obtained by school nurses in 2010/11 revealed coefficients for height of r = 0.97, p < 0.001; for weight of r = 0.98, p < 0.001; and for WC of r = 0.85, p < 0.001). The high correlations for the adjusted measurements of anthropometry and the vertical jump test suggest that our adjustment was appropriate, although the correlation between schools' measurements of the Andersen test and the measurements obtained by SDU was low. Visual inspection of the available information from the three sources revealed some deviance in the measurements obtained by SDU compared to the results from the adjusted 2010/11 measurements and from the HSN 2011/12 measurements, where the running distance for the 5th to the 9th grade students gradually decreased, while results from the other sources showed increases (Appendix 3, Figure 5).

HSN DATA: INTER-OBSERVER RELIABILITY FOR MEASUREMENTS OF INDIVIDUAL CHILDREN

For the assessment of inter-observer reliability, information from 111 children was available from five schools. Inter-observer reliability was high to moderate for height and weight (Table 4) ranging from coefficients of 0.94 (95% CI 0.91-0.97, n = 52) for weight among 4th to 6th grade students to 0.77 (95% CI 0.60-0.92, n = 20) for weight among 7th to 9th grade students. Inter-observer reliability varied considerably for measurements of WC. For example, ρ_c for 7-9th grades was 0.49 (95 % CI 0.22-0.77, n = 17) with two observations from the same school differing by more than 15 cm. Similarly, a low level of agreement between schools' and researchers' test results was observed for the vertical jump test, especially for the 4th to 6th grades (ρ_c= 0.27, 95% CI 0.1-0.53, n = 45). Eleven results differed by 10 cm or more. Inter-observer reliability for CRF was very low (Table 4). The coefficient for the 0th to 3rd grades was 0.1 (95% CI -0.02-0.42, n = 34), suggesting substantial variation between the results obtained by SDU and those obtained by the three schools who produced results. Inspection of the data showed that four children from the same school ran more than 400 metres less when tested by SDU, and that one child ran more than 300 metres further. The coefficient for the 4th to 6th grades was 0.04, (95% CI -0.23-0.31, n = 49). Visual inspection of the data revealed six observations with distances longer by 200 metres when measured by the school compared to results when measured by SDU (Figure 6). On the other hand, three observations were more than 500 metres shorter when measured by the schools. Bland-Altman plots of all pairs of measurements are presented in Appendix 1, Figures 94-113. As stated earlier, inter-observer reliability was calculated for schools' measurements adjusted for expected growth in 345 days, and this adjustment may be particularly imprecise for CRF. For completeness, we present the corresponding ρ_c for CRF before

adjustment, which was 0.2 for 0th to 3rd grade students, 0.05 for 4th to 6th grade students and 0.12 for 7th to 9th grade students. Thus, the estimates for CRF were essentially unaffected by the adjustment.

ODENSE DATA: INTER-OBSERVER RELIABILITY

Pairs of measurements from school nurses and from schoolteachers in the municipality of Odense conducted with a maximum time lag of two weeks were available for 197 children. Inter-observer reliability for measurements of height (ρ_c = 0.99, 95% CI 0.99-0.99, n = 192) and weight (ρ_c = 0.99, 95% CI 0.99-0.99, n = 197) were high (Table 5). In contrast, inter-observer reliability for measurements of WC was lower but still acceptable (ρ_c = 0.91, 95% CI 0.89-0.94, n = 166). The lowest estimate of inter-observer reliability was calculated for children in the 0th grade (ρ_c = 0.59, 95% CI 0.36-0.82, n =30). Visual inspection of the data for children in the 0th grade did not reveal any obvious pattern of systematic bias except that measurements obtained by schoolteachers tended to be higher than those taken by school nurses. Bland-Altman plots of all pairs of measurements are presented in Appendix 1, Figures 114-116.

ATTRITION: HSN DATA

In 2008/09, 15 schools participated in the HSN out of 2,441 possible in Denmark (0.6%). In 2009/10, 136 schools participated out of 2,440 possible (5.6%); in 2010/11, 128 schools participated out of 2,425 possible (5.3%); and in 2011/12, 114 schools participated out of 2,299 possible (5.0%). The number of children eligible for measurements varied during the study period 2008/09-2011/12 from 7,262 in 2008/09 to 61,556 children in 2009/10. However, the percentages of children with valid measurements varied across the study period, with the most measurements being obtained in 2011/12 (75%) and the lowest percentage of measurements being obtained in 2008/09 (43%) Participation rate varied by grade, and was highest for children in grades 0 to 6 and lowest for children in grades 7 to 9. No information was available on the number of participating children within specific classes in specific schools (since there is no information on children's allocation to specific classes, only on their grades).

MISSING VALUES FOR GIRLS, BY GRADE

In grade-specific analysis of the association between missing values for measurements of anthropometry and physical fitness obtained by schools, we observed that the odds of having missing values for height and weight gradually increased with grade (Appendix 5, Table 1). For example, girls in the 9th grade were

five times more likely to have missing values for height compared to girls in the 0th grade (95% CI 3.7-7.2, $p < 0.001$). For weight, the OR was 2.3 (95% CI 1.4-3.8, $p < 0.05$) in comparison with girls in the 0th grade. Results were adjusted for immigrant status, parental social factors and school year. The odds ratio for missing values for WC increased gradually across grades, and girls in the 6th to 9th grades had more than twice the odds of having missing values for WC than girls in lower grades. Similarly, the odds of missing values for the vertical jump test increased gradually with grade, and the OR for girls in the 9th grade was 3.6 (95% CI 2.6-5.0, $p < 0.05$) in comparison with grade 0 students. In contrast, girls in the 1st to the 5th grades were less likely to have missing values of CRF than girls in the 0th grade.

MISSING VALUES FOR BOYS, BY GRADE

The pattern of missing values by grade observed for girls was similar for boys (Appendix 5, Table 2). Boys in the 9th grade were three times more likely to have missing values for height than boys in the 0th grade (OR = 3.0, 95% CI 2.2-4.3, $p < 0.05$). Boys in the higher grades (7th to 9th) were also more likely to have missing values for weight than boys in the lower grades. The odds ratio for missing values of WC increased gradually across grades. Boys in the 9th grade were almost five times more likely to have missing values than boys in the 0th grade (OR = 3.3, 95% CI 2.3-4.7, $p < 0.05$). Missing values for the vertical jump test were significantly more likely among boys in grades 1 to 9 than for boys in the 0th grade. However, the difference compared to boys in the 0th grade did not increase linearly, as the likelihood of having missing values for the vertical jump test was similar for boys in all grades from the 2nd to the 6th (difference OR = 1.3, 95% CI 0.9-1.4, $p = 0.19$). Interestingly, boys in the 1st, 2nd and 4th grades were more likely to have measurements of CRF than boys in the 0th grade. For example, the OR for 4th grade students was 0.6 (95% CI 0.4-0.8, $p < 0.05$).

MISSING VALUES FOR GIRLS, BY BMI DECILES

For children with BMI measurements, we analysed the probability (OR) of having missing values for other measurements by BMI deciles. The odds of missing values among girls for WC were curvilinearly related to BMI (Appendix 5, Table 3). Girls in the 2nd and 4th deciles had lower odds of having missing values for WC than girls in the lowest BMI decile. However, the odds of having missing values for WC increased for girls in the 5th to the 10th deciles. In contrast, the odds of having missing values for the vertical jump test were almost equally distributed across the whole BMI distribution, except for girls in the 9th and 10th BMI deciles, where the odds ratio was statistically significantly higher than for girls in the 2nd decile. Girls in the

1st to the 6th deciles had lower odds of having missing values for the Andersen test than girls in the 1st decile. However, girls in the highest decile were more likely to have missing values for CRF compared to girls in the 1st decile (OR = 1.2, 95% CI 1.1-1.3, $p < 0.05$).

MISSING VALUES FOR BOYS, BY BMI DECILES

The odds of having missing values for WC was highest among boys in the 10th BMI decile compared to those in the 1st decile (OR = 2.9, 95% CI 2.4-3.4, $p < 0.00$) (Appendix 5, Table 4). A similar but less strong association was observed for the analysis of missing values for the vertical jump test and CRF. Boys in the 2nd to 5th deciles had lower odds of having missing values for the Andersen test than the 0th decile, but the difference was only significant for boys in the 5th decile (OR = 0.9, 95% CI 0.8-1.0, $p < 0.05$).

For completeness, we investigated the associations between missing measurements and children's gender, immigrant status, parental employment status and parental level of education. All analyses were adjusted for grade and school year (Appendix 5, Table 5). Children from immigrant families were more likely to have missing measurements of height, weight and WC. For example, the OR for having missing measurements of height was 1.2 (95% CI 1.0-1.5, $p < 0.05$) in comparison to native Danish children, and the OR for missing measurements of WC was 1.2 (95% CI 1.0-1.3, $p < 0.05$) among children of non-Western immigrants. Parental unemployment was also associated with missing values. For example, children with an unemployed father had higher odds of having missing values for weight, WC and the Andersen test (OR = 1.2; 95% CI 1.0-1.3, $p < 0.05$), while mother's unemployment was significantly associated with missing values for all measurements except for height.

ATTRITION: ODENSE DATA

School nurses made measurements of height, weight and WC in the years 2009/10 to 2011/12 for children in the 0th (n = 1,814), 1st (n = 1,628), 5th (n = 1,826) and 9th (n = 1,349) grades from 14 schools in the municipality of Odense out of 31 schools eligible to participate. A majority (68.4%, n = 4,452) of the 6,617 children included in the study were measured by both the school nurses and the HSN schools, while 31.7% (n = 2,095) had no measurements of anthropometry by the HSN schools. We compared the mean height, weight and WC obtained by school nurses for children who were measured by both the nurses and their schools with measurements for children obtained by school nurses only. The measurements of anthropometry were similar for all children except for those in the 5th grade (Appendix 6, Table 1). For

grade 5 students, average values for measurements made by both schools and school nurses of weight, WC and BMI were significantly lower than those for measurements made only by school nurses only.

MISSING VALUES BY GRADE

Girls in the 9th grade were more than seven times more likely to have missing values for WC measured by schools than girls in the 0th grade (OR = 7.5, 95% CI 3.2-17.7, $p < 0.05$), a 78% higher probability than that for girls in the 1st grade ($p > 0.05$) and 62% higher than for girls in the 5th grade ($p > 0.05$). Similarly, boys in the 9th grade had a 65% higher probability of having missing values from those obtained by schools than boys in the 0th grade (OR = 7.8, 95% CI 3.3-18.4, $p < 0.05$). No statistically significant differences were found relative to boys in the 1st and 5th grades. Estimates were adjusted for BMI, immigrant status and parental social variables.

MISSING VALUES BY BMI DECILES

Logistic regression analyses of having missing values for WC measured by schools separated by BMI deciles revealed that for girls in the 5th grade, the odds of having missing values were significantly less across the whole BMI distribution than for girls in the lowest BMI decile (Appendix 6, Table 2), although the estimates for the 2nd, 7th and the 10th deciles did not reach statistical significance. Estimates were adjusted for immigrant status, parental social variables and school year. This indicated a higher rate of attrition for girls with the lowest BMI. Boys in the 1st and 5th grades in the heaviest BMI deciles had greater odds of having missing values for WC measured by schools than boys in the lowest BMI decile (1st grade and 10th decile, OR = 3.6, 95% CI 1.4-9.5, $p < 0.05$, and 5th grade and 10th BMI decile, OR = 4.4, 95% CI 1.3-14.6, $p < 0.05$) (Appendix 6, Table 3).

Similarly, for boys in the 9th grade, there were more likely to be missing values for WC measured by schools among boys in the 7th decile (OR = 2.1, 95% CI 1.3-3.4, $p < 0.05$).

The logistic regression analysis of the association between missing values and social variables showed that few of the social factors investigated significantly predicted missing school measurements (Appendix 6, Table 4). Children in the 0th and 1st grades from non-Western immigrant families were more likely to have missing measurements of anthropometry in the HSN schools than native Danish and Western immigrant children. In fact, 25% of the children in the 1st grade from non-Western immigrant families had missing values for the HSN measurements, compared to 14% among native Danish and Western immigrant children. A difference of similar size was observed for children in the 0th grade. However, immigrant origin

was not a significant predictor of missing school measurements among children in higher grades. Missing school measurements were associated with unemployment of the father or the mother for children in grades 0 and 5, and a high level of education of the father was associated with missing school measurements among children in the 9th grade.

DISCUSSION

The purpose of this study was to evaluate the intra -observer and inter-observer reliability of measurements of anthropometric characteristics and physical fitness in a large-scale school-based programme that monitored indicators of children's health. We observed a high level of intra-observer reliability (> 0.8) for all measurements, though with some variation across grades. Inter-observer reliability was high for height and weight, moderate for WC and the vertical jump test, and low for CRF.

INTRA- OBSERVER RELIABILITY

With regard to intra -observer reliability of height and weight, our findings where in accord with those reported by Berkson et al.,(9), who also observed a high intra-observer reliability, although the estimates reported were for height and weight combined. However, España-Romero et al. (14) found only moderate intra-observer reliability in measurements of children's height in a study of 138 children aged 6 to 18 years assessed twice by six PE teachers at an interval of seven days. The average difference in height measurements between the two measuring sessions was 1.1 cm for the children aged 6-12 years, even though the PE teachers were trained prior to the assessments and supervised during them. Similarly, Berkson et al. (9) found outliers in a study of intra-observer reliability covering children in grades 1, 4 and 7 measured twice by eight teachers on the same day. Sixty percent of the erroneous measurements were associated with one teacher. Although the intra-observer reliability was higher in our study than that found by España-Romero et al. (14), our results nevertheless suggest some systematic error in the measurements of height for children in the 7th grade. Test and re-test height measurements varied substantially for children below average height, but not among children above average height. The fact that many of the erroneous results were obtained at the same school suggests failure to adhere to the standard measurement procedures; perhaps different teachers measured the same children on two different occasions, or children assisted in carrying out the measuring process, or non-standard measurement instruments were used. Such sources of error may also have affected results of WC among

6th grade students in 2010/11 and among 3rd grade students in 2011/12. However, Ruiz et al. (27) argue that it is inevitable that systematic errors will affect measurements of anthropometric characteristics, and that such error is observed in many studies. For example, España-Romero et al. (14) observed significant differences in measurements of WC obtained by PE teachers at an interval seven days.

Like the other measurements of anthropometry and physical fitness used in our study, the Vertical jump test was found to be a valid and reliable measure (22, 27). Intra-observer reliability for the vertical jump test was generally high, except for children in the 6th and 8th grades, where indications of heteroscedasticity were observed. That is, the difference between two successive measurements tended to increase as the children jumped higher. Differences between two measurements at the extremes of the distribution are not uncommon among expert anthropometrists (8, 27), and are to be expected among untrained testers. The observed heteroscedasticity may also be a result of natural variation in the ability of fit children to develop power in a vertical jump.

With regard to the CRF (Andersen test), we assume that repeated performances of the test may have increased the children's familiarity with it and thus affected their scores, for example by leading them to adjust their speed in accordance with the duration of the test. Also we must assume that the children's motivation may have influenced their results. This is supported by the change in performance between the schools test and re-test, which was perhaps most obvious among children in the 6th and 7th grades. The decrease in running distance between test and re-test may by real and indicative of lack of motivation, as the test requires the child to run as great a distance as possible. Similarly, a learning effect may have affected the partner in the test. The partner may encourage or influence the running child differently at the re-test, or cause the partner to change with respect to counting the crossings of the floor. Interestingly, Ortega et al. (24) reported a lack of systematic error influencing the result of a 20-metre shuttle run test (comparable to the Andersen test) and suggested that learning effects were not likely to have influenced the results in his study. However, the original 20-m shuttle run test developed by Léger et al. (18) was performed individually for each child, but individual assessment of children participating in a school-based study is not feasible, and some learning effect of the running child or the partner must be assumed in our study. Finally, we cannot rule out the potential influence of changes of partners between test occasions. The schoolteachers were not informed about the potential effects of changes of partners between test sessions, and we are not able to discern whether this may have influenced the intra-observer

reliability systematically. In sum, our results and the results reported by others (9, 14, 23, 27) underline the importance of ensuring the quality of the data collected by non-experts in a natural setting.

INTER-OBSERVER RELIABILITY

In comparing schoolteachers' measurements with professional measurements by SDU researchers, our inter-observer reliability findings for height and weight in the HSN data were fairly similar to those of Stoddard et al. (32), but slightly lower than those reported by Berkson et al. (9) in comparing results obtained by PE teachers with those of trained experts. In contrast, inter-observer reliability comparing schoolteacher measurements with data obtained by school nurses in the municipality of Odense were good, and quite similar to those of Berkson et al. (9). Although the adjustment for the time lag between measurements obtained by HSN schools and SDU researchers seemed appropriate, at least for the anthropometric measurements and the vertical jump test, our measurements correlated less well with the schools' measurements of the Andersen test, although the coefficient obtained was still statistically significant. A comparison of the estimates of inter-observer reliability calculated before and after adjustment revealed virtually identical results. This indicates that adjustment for the expected growth in CRF may be more difficult than anticipated, and that our estimates of inter-observer reliability should be interpreted with caution. In contrast, the inter-observer reliability estimates from the Odense subset study point to a limited influence of error in schoolteachers' and school nurses' measurements of height and weight.

The inter-observer reliability of WC measurement varied considerably, and in comparison with another field based study (33), our estimates were low. There may be several explanations for this discrepancy. Firstly, a large proportion of the errors for children in the highest grades were associated with a single school, suggesting that the school failed to comply with standard procedures for measuring WC. Secondly, obesity has been found to influence the reliability of anthropometric measurements among children and adolescents (8, 33), although this has been disputed (9), as BMI is proportional to height, and hence a higher mean BMI among older children does not reflect a higher percentage of fat in measurements of weight than among younger children. A similar variability in inter-observer reliability for measurements of WC across grades was observed in the Odense subset study.

Inter-observer reliability for vertical jump test and CRF varied across grades, and the explanations for the observed reliability of WC may apply to measurements of vertical jump as well. Although the vertical jump test is a widely-used measure of children's physical fitness, we are not aware of any other studies of inter-observer reliability in school settings for this measurement, and comparison with other studies is therefore difficult. (27, 35) Only a minority of teachers are specifically trained to teach PE, and measuring the vertical jump test and the running distance is novel to many. Thus, the moderate to low levels of agreement between teachers' results and the results of trained experts may reflect the lack of familiarity with the procedure among teachers. Another reason for the low inter-observer reliability of the Andersen test is the running child's dependency on the cooperation of the partner in counting the crossings of the gym. Lack of compliance with test procedures is a basic source of error in many tests, including tests of physical fitness among schoolchildren (8). An acceptable level of agreement depends on a variety of factors, including the age and gender of the participants, the skills of the testers, the number of days between two measurements and the complexity of the test situation and the test *per se* (8). Many of these factors could have influenced the reliability of the measurements included in the HSN initiative – in particular, the measurements of running distance. An important source of error is the involvement of children as collaborators in measuring the distance in the Andersen test. We suspect that this procedure may have influenced the results upward. As children's attention tends to drift off during the test, some may then add extra crossings to conceal their temporary loss of concentration, thereby inflating the final test result. This was particularly noticeable among children in the lowest grades during data collection for inter-observer reliability (and these data were consequently discarded), but the results obtained by schools were higher than those obtained by researchers from SDU across the whole range of grades. If this source of error is generally present when the school performs the annual measurement of children's running distance, results are inflated and the running distance is overestimated (25), resulting in a lower level of objective reliability. Lack of experience has been identified as a potential threat to the quality of the data in other school-based monitoring interventions where measurements were obtained by teachers (9), school nurses (27), or volunteers (17), and this fact emphasises the need for sufficient training of staff in measuring accurately.

ATTRITION

The number of children that participated in the HSN schools relative to the total number of children in the schools peaked in 2012. The fact that only a minority of the children in the participating schools were

measured may have introduced selection bias. Analysis of a possible association between missing values and immigrant status, parental unemployment and parental level of education supported this assumption, and such factors have been shown to be important predictors of participation in studies examining dimensions of physical activity (16, 30, 34).

In the subsample of children with measurements of weight and height, some indications were found of statistically significant associations between BMI and missing values for anthropometry and physical fitness measurements. The heavier girls and boys were less likely to participate in WC measurements compared to children in the middle and at the low end of the BMI distribution. In contrast, no such pattern was observed for missing values for the vertical jump test and the Andersen test. In this respect, it should be mentioned that even the smallest differences in odds of a positive outcome proved to be statistically significant in the logistic regression analysis, due to the large sample size. However, some selection bias is evident in all observational studies, and we advise some caution in interpreting results from future analyses.

In contrast to the complex influence of parental social factors on missing values, a clear pattern was observed with regard to missing values across grades. Children in the lowest grades had the highest share of valid measurements, and children in the highest grades had the lowest. This pattern of participation is parallel to declining participation rates in sports and physical activities among adolescents observed in numerous studies. Although the reasons for adolescents' declining interest in physical activity are not yet fully understood, factors such as peer influence, hormonal changes, adjustment of motor control due to growth spurts, fatigue and changes in self-awareness have been suggested (2, 13), and it easy to imagine that such factors may also be involved when adolescents decline to participate in the annual measurement sessions at school. Curiously, the participation rate in the Andersen test increased throughout the study period, and this could point to a complex selection process where the proportion of children in the oldest grades participating in the study as a whole decreases but the proportion of children completing the most physically demanding tests increases. This suggests that those who chose to complete the test were dedicated to the objectives of the HSN project *per se*, as opposed to those that left the study. Consequently, the pattern of attrition was not random, and suggests selection bias. Unfortunately, we do not have information on possible biological and psychological factors involved in adolescents' lower rate of participation, and we are not able to control for these factors in analyses investigating the influence of missing values. However, we must assume that the measurements available

for the highest grades (8th and 9th) are influenced by selection bias to a greater extent than measurements for children in lower grades. Thus, results for this subgroup should be interpreted with caution.

MISSING VALUES FOR THE ODENSE SUBSET STUDY

As in the analyses of missing values for measurements obtained in the full HSN study, we found some evidence that missing values were associated with various social background variables in the Odense sub-study, most notably parent's employment status and immigrant/non-immigrant status. However, the results varied across the measurements obtained. The pattern of missing values in the Odense subset study was complex. Girls with a low BMI did not participate in measurements taken by schools to the same degree as in the measurements taken by school nurses. In contrast, boys high in the BMI distribution were more likely not to be included in the schools' measurements, and this may explain why the mean weight and WC were significantly higher when measured by school nurses only – at least for the 5th grade students. The higher proportion of missing values among children in the 9th grade was similar to that observed for the full HSN study. Thus, the pattern of missing values for the 9th grade students is probably non-random, and caution should be exercised in interpreting the results for this group of children. Again in accordance with the results from the analysis of data in the full HSN study, some indications of the influence of social variables were observed; for example, children with non-Western immigrant backgrounds were less likely to participate in the measurement sessions conducted only by schools. Cultural norms may prohibit participation in such measurements among some non-Western immigrant children, although we are not aware of any research specifically investigating this phenomenon. However, the tendency for there to be a comparatively lower participation rate among subjects from immigrant families has been observed in large-scale population-based surveys (10, 29) and is associated with language barriers, poor understanding of the purpose of participation, distrust, and a general resistance to sharing information that is regarded as private. We did not investigate whether these factors were important in explaining the higher rate of attrition among children from immigrant families, and it is quite possible that other factors may have been more important. Nevertheless, a low participation rate among a subgroup of schoolchildren may limit the external validity of the estimates for this group, and further research should be conducted to discover possible initiatives to improve participation.

LIMITATIONS

It is unclear which criteria schools used to decide whether to participate in the HSN project. The selection of schools into the HSN programme was based on voluntary enrolment. Self-selection of schools into the programme is non-random, which leaves the data vulnerable to selection bias. It is likely that the schools' decision to enrol in the programme is related to the level of schools' commitment to measuring their students and to complying with the measuring instructions. If such an association does exist, it may confound the results of the present study. Similarly, less than half the children in the participating schools were measured. We do not know why children in some classes were included for measurement and why others were not, but a strong level of commitment and some perceived familiarity with the measurement procedures may have motivated some teachers to select the children in their class for measurement. Admittedly, these assumptions are speculative, as no information is available about the reasons for inclusion of children in the measurements. Another limitation of the present study was the non-random selection of schools that conducted control measurements, and the selection of schools for establishing reference values by the researchers. We assume that the non-random selection procedures may affect the generalisability of the results, but as no information on anthropometry and physical fitness is available for children in schools that did not participate in the HSN study, we are unable to assess the potential selection bias introduced. Also, some children refrained from participating in the establishment of the reference values by the researchers, and it is possible that this subsample may differ with respect to the measurements obtained. This may have influenced the reliability estimates. However, we assume that individual resistance to participation is also present when schools take their annual measurements, and hence it does not constitute a threat to the external validity of the result of the present study.

Schools selected for establishment of the reference values for the assessment of inter-observer reliability for the HSN part of the study failed to conduct measurements of the children who were measured by SDU in the same year as the SDU measurements, and this may have invalidated estimates of inter-observer reliability. Our attempt to extrapolate the HSN estimates for anthropometric measurements and physical fitness from the year before did appear to be adequate, but nevertheless we still doubt the appropriateness of this approach. Hence, estimates of inter-observer reliability based on measurements obtained by SDU should be interpreted with caution. In contrast, the estimates of inter-observer reliability

based on comparing HSN measurements with measurements obtained by school nurses, with a maximum time lag of two weeks, were high.

The large sample size strengthens this study and permits stratified analysis – here, stratification by grades. This renders the calculation of reliability estimates possible across the whole range of school grades, thereby facilitating further inspection of data for specific sub-samples. We are not aware of any other study where this is possible. Another strength is the access to register-based information on all children in Denmark. This access allowed calculation of participation rates and description of characteristics of individuals with missing values. Finally, access to register-based information provides the opportunity to address a broad spectrum of research questions that would otherwise be too costly to study.

Currently, several tests of related to the body composition and physical fitness of children and adolescents are being used in school-based settings as a sustainable method of monitoring trends in obesity among children and of providing a basis for health-promoting initiatives. A large majority of the measurements have been found to be of acceptable reliability and validity (7) when obtained by trained experts (1, 5, 7, 24, 27) or health care personnel such as school nurses (32), and these methods can easily be applied by teachers or other non-research personnel, at least in principle. However, the amount of research evaluating the reliability of the data gathered by teachers across several school based settings is limited to a single study (9). The present study extends current knowledge by adding information on the effectiveness and feasibility of engaging schoolteachers in providing reliable measurements of children's anthropometry and physical fitness. As this approach is believed to be a desirable method of monitoring the development of obesity and in children and changes in their levels of fitness, information on teachers' ability to accurately measure children at school is vital.

The practical challenges that teachers face in measuring children from a whole class within the time limit of a double PE lesson (90 minutes, which is the norm in Danish schools) may introduce measurement error. However, the measurements chosen for the HSN programme were specifically selected to enable the teacher to perform them all accurately within the time limits of a double PE lesson. As stated earlier, all the measurements have been used in several school-based studies in Denmark (3, 4, 36) and have been found to be practicable. Nevertheless, in line with the recommendations made by Berkson et al.(9) and others (14, 27, 32), we agree that teachers may benefit from assistance from external personnel in order

to carry out the task of measuring accurately. This is true for anthropometric measurements and may even be more relevant for measurements of physical fitness.

CONCLUSION

Generally, measurements of children's anthropometry and physical fitness examined in the present study are reliable. This is true for measurements obtained by school nurses and teachers in the Odense part of the HSN study and for measurements obtained by teachers in the full HSN study. Teachers can provide reliable measurements of children's height and weight, and are able to reproduce measurements of WC but need additional training in measuring vertical jump height and CRF. Training should include guided practice of the test procedures in a standardised form to ensure conformity to the test instructions. With regard to missing values, some indications of selection bias were observed, and caution is advised in interpreting results from children in the highest grades and with the highest BMI levels.

REFERENCES

1. Ahler T, Bendiksen M, Krustrup P, Wedderkopp N. Aerobic fitness testing in 6- to 9-year-old children: reliability and validity of a modified Yo–Yo IR1 test and the Andersen test. European Journal of Applied Physiology. 2012;112(3):871-76.

2. al Ue. Determinants of physical activity and sedentary behaviour in young people: a review and quality synthesis of prospective studies. Brit J Sport Med. 2011;45(11):896-905.

3. Andersen A, Froberg K. Copenhagen School Child Intervention Study (CoSCIS). Unpublished raw data. 2000.

4. Andersen A, Froberg K. European Youth Heart Study. Unpublished raw data. 2003.

5. Andersen L, Andersen T, Andersen E, Andersen S. An intermittent running test to estimate maximal oxygen uptake: the Andersen test. J Sports Med Phys Fitness. 2008;48:434-7.

6. Andersen LB, Sardinha LB, Froberg K, Riddoch CJ, Page AS, Anderssen SA. Fitness, fatness and clustering of cardiovascular risk factors in children from Denmark, Estonia and Portugal: The European Youth Heart Study. International Journal of Pediatric Obesity. 2008;3(S1):58-66.

7. Artero EG, España-Romero V, Castro-Piñero Jet al. Reliability of Field-Based Fitness Tests in Youth. Int J Sports Med. 2011;32(03):159-69. Epub 16.12.2010.

8. Baumgartner T, Jackson A, Mahar M, Rowe D. Measurement for Evaluation in Physical Education and Exercise Science: McGraw-Hill Companies,Incorporated; 2006.

9. Berkson SS, Espinola J, Corso KA, Cabral H, McGowan R, Chomitz VR. Reliability of height and weight measurements collected by physical education teachers for a school-based body mass index surveillance and screening system. J Sch Health. 2013;83(1):21-7. Epub 2012/12/21.

10. Breinholt-Larsen F, Ankersen P, Poulsen S, Søe D, Christensen S. Hvordan har du det? 2010 - Sundhedsprofil for region og kommuner - Voksne. Århus: Center for Folkesundhed, Region MIDT, 2011.

11. Carrasco JL, Jover L. Estimating the Generalized Concordance Correlation Coefficient through Variance Components. Biometrics. 2003;59(4):849-58.

12. Denmark S. Fertilitetsdatabasen. Statistic Denmark; 2013; Available from: http://www.dst.dk/da/Statistik/dokumentation/kvalitetsdeklarationer/fertilitetsdatabasen.aspx .

13. Dobbins M, Husson H, DeCorby K, LaRocca Rebecca L. School-based physical activity programs for promoting physical activity and fitness in children and adolescents aged 6 to 18. Cochrane Database of Systematic Reviews [Internet]. 2013; (2). Available from: http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD007651.pub2/abstract.

14. España-Romero V, Artero EG, Jimenez-Pavón Det al. Assessing Health-Related Fitness Tests in the School Setting: Reliability, Feasibility and Safety; The ALPHA Study. Int J Sports Med. 2010;31(07):490-97. Epub 29.04.2010.

15. Greve J, Heinesen E. Healthy School Network – a randomized experiment. Rockwool Foundation Research Unit. 2014;Study Paper.

16. Gustafson SL, Rhodes RE. Parental correlates of physical activity in children and early adolescents. Sports Medicine. 2006;36(1):79-97.

17. Ikeda JP, Crawford PB, Woodward-Lopez G. BMI screening in schools: helpful or harmful. Health Education Research. 2006;21(6):761-69.

18. Léger LA, Mercier D, Gadoury C, Lambert J. The multistage 20 metre shuttle run test for aerobic fitness. J Sport Sci. 1988;6(2):93-101.

19. Lin LIK. A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometrics. 1989;45(1):255-68.

20. Linchey J, Madsen KA. State requirements and recommendations for school-based screenings for body mass index or body composition, 2010. Preventing chronic disease. 2011;8(5):A101. Epub 2011/08/17.

21. Lubans DR, Morgan PJ, Cliff DP, Barnett LM, Okely AD. Fundamental Movement Skills in Children and Adolescents: Review of Associated Health Benefits. Sports Medicine. 2010;40(12):1019-35.

22. Markovic G, Dizdar D, Jukic I, Cardinale M. Reliability and factorial validity of squat and countermovement jump tests. J Strength Cond Res. 2004;18(3):551-5. Epub 2004/08/24.

23. Nagy E, Vicente-Rodriguez G, Manios Yet al. Harmonization process and reliability assessment of anthropometric measurements in a multicenter study in adolescents. Int J Obesity. 2008;32:S58-S65.

24. Ortega FB, Artero EG, Ruiz JRet al. Reliability of health-related physical fitness tests in European adolescents. The HELENA Study. Int J Obesity. 2008;32:S49-S57.

25. Public Act Tennesee, 194, 445,(2005) http://tennessee.gov/sos/acts/104/pub/pc0194.pdf

26. Riddoch C, Edwards D, Page Aet al. The European Youth Heart Study—Cardiovascular Disease Risk Factors in Children: Rationale, Aims, Study Design, and Validation of Methods. Journal of Physical Activity & Health. 2005;2(1):115.

27. Ruiz JR, Castro-Piñero J, España-Romero Vet al. Field-based fitness assessment in young people: the ALPHA health-related fitness test battery for children and adolescents. Brit J Sport Med. 2011;45(6):518-24.

28. Sargent DA. The physical test of man1921.

29. Singhammer J. Etniske minoriteters sundhed (Health among ethnic minorities). Århus (Aarhus): Center for folkesunhed, Region Midtjylland (Department of public health, County of Middle Jutland), 2008.

30. Stalsberg R, Pedersen AV. Effects of socioeconomic status on the physical activity in adolescents: a systematic review of the evidence. Scand J Med Sci Spor. 2010;20(3):368-83.

31. StataCorp. Stata: Release 12.1. College Station, TX: StataCorp LP.; 2012.

32. Stoddard SA, Kubik MY, Skay C. Is School-Based Height and Weight Screening of Elementary Students Private and Reliable? The Journal of School Nursing. 2008;24(1):43-48.

33. Stomfai S, Ahrens W, Bammann Ket al. Intra- and inter-observer reliability in anthropometric measurements in children. Int J Obesity. 2011;35:S45-S51.

34. Tammelin T. A review of longitudinal studies on youth predictors of adulthood physical activity. International journal of adolescent medicine and health. 2005;17(1):3-12.

35. Tomkinson GR. Global changes in anaerobic fitness test performance of children and adolescents (1958–2003). Scand J Med Sci Spor. 2007;17(5):497-507.

36. Wedderkopp N, Jespersen E, Franz Cet al. Study protocol. The Childhood Health, Activity, and Motor Performance School Study Denmark (The CHAMPS-study DK). BMC Pediatr. 2012;12:128. Epub 2012/08/22.

APPENDIX 1: FIGURES 1-6

FIGURE 1: DATA CLEANING PROCEDURE FOR THE HSN DATABASE AND SELECTION OF THE SAMPLE FOR THE PRESENT STUDY

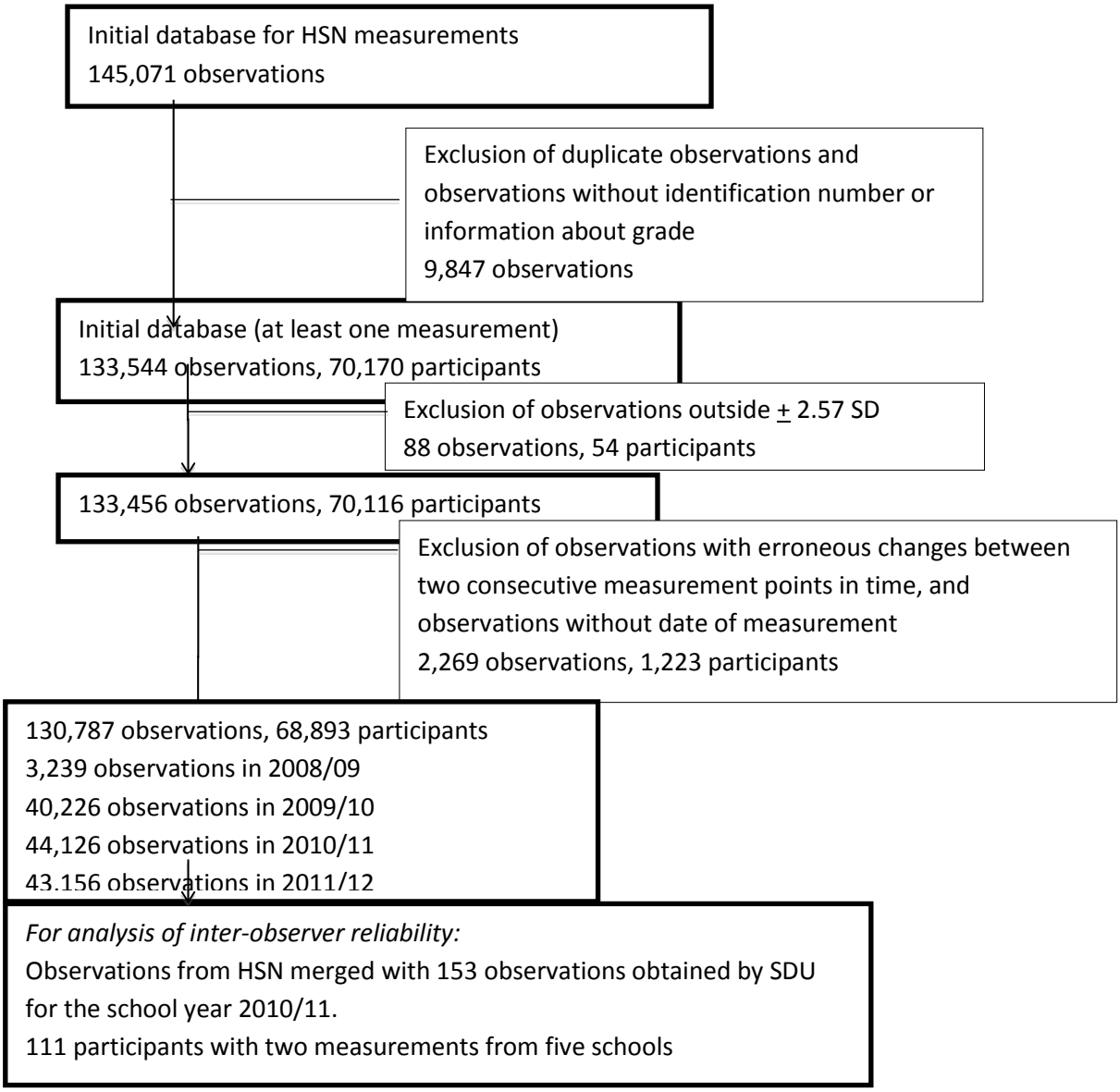


FIGURE 2: DATA CLEANING PROCEDURE FOR THE HSN CONTROL MEASUREMENTS

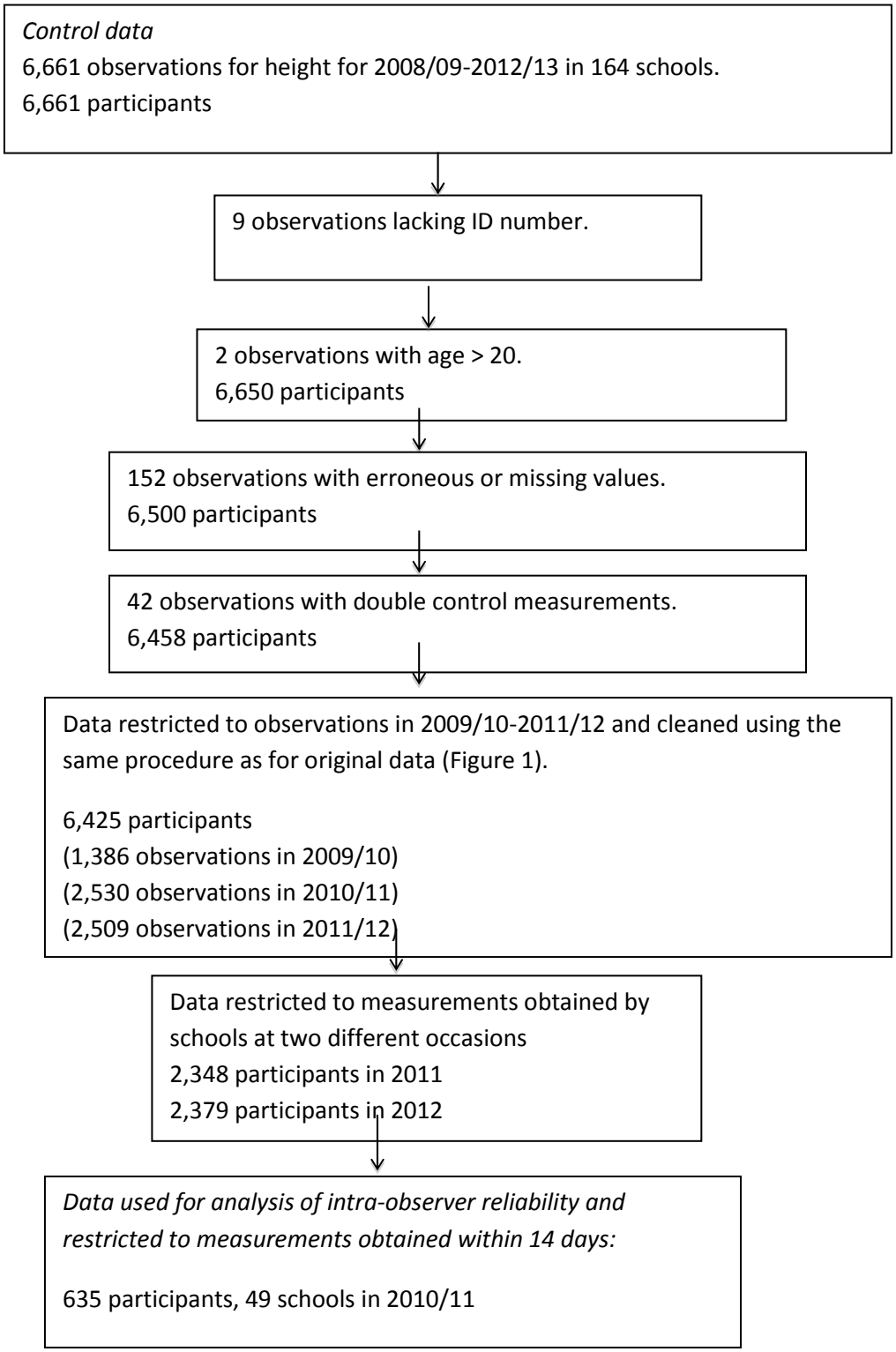
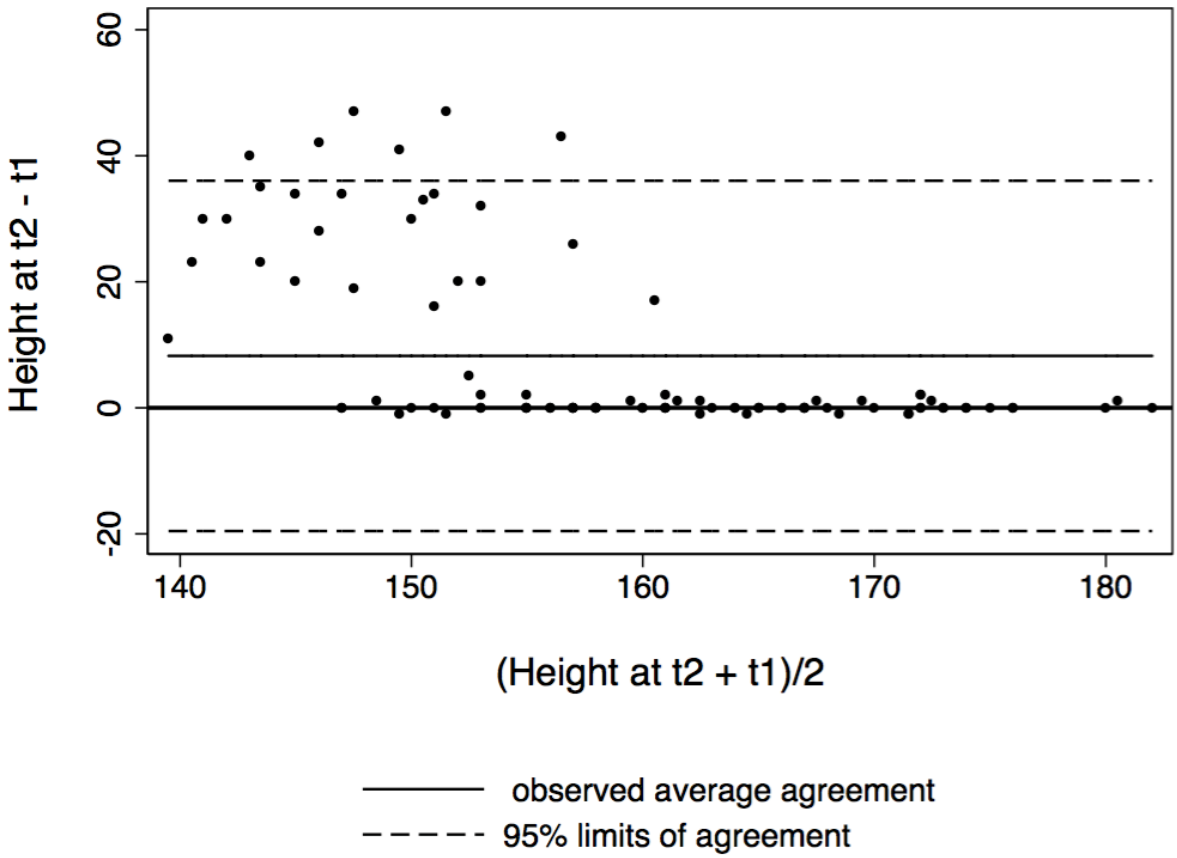
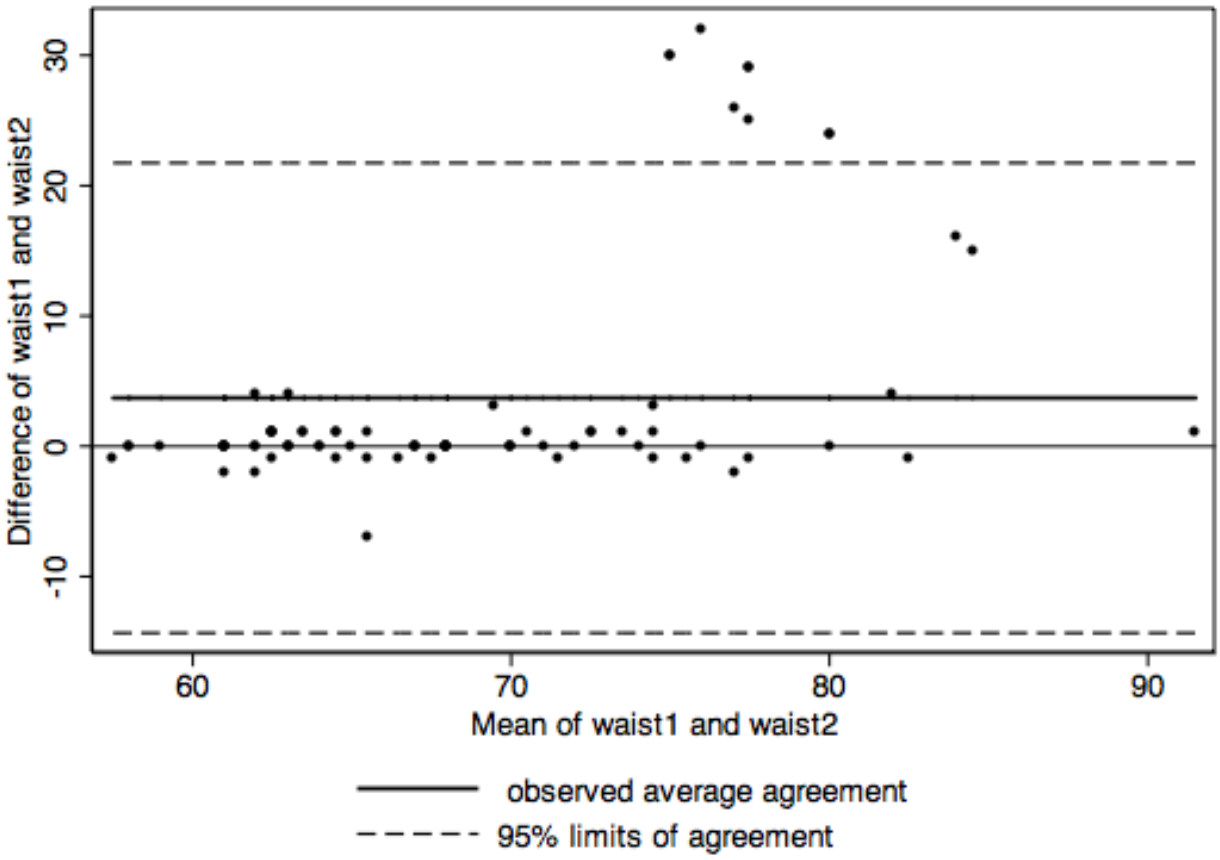


FIGURE 3: BLAND-ALTMAN PLOT OF TEACHERS' MEASUREMENTS OF HEIGHT AMONG 7TH GRADE STUDENTS



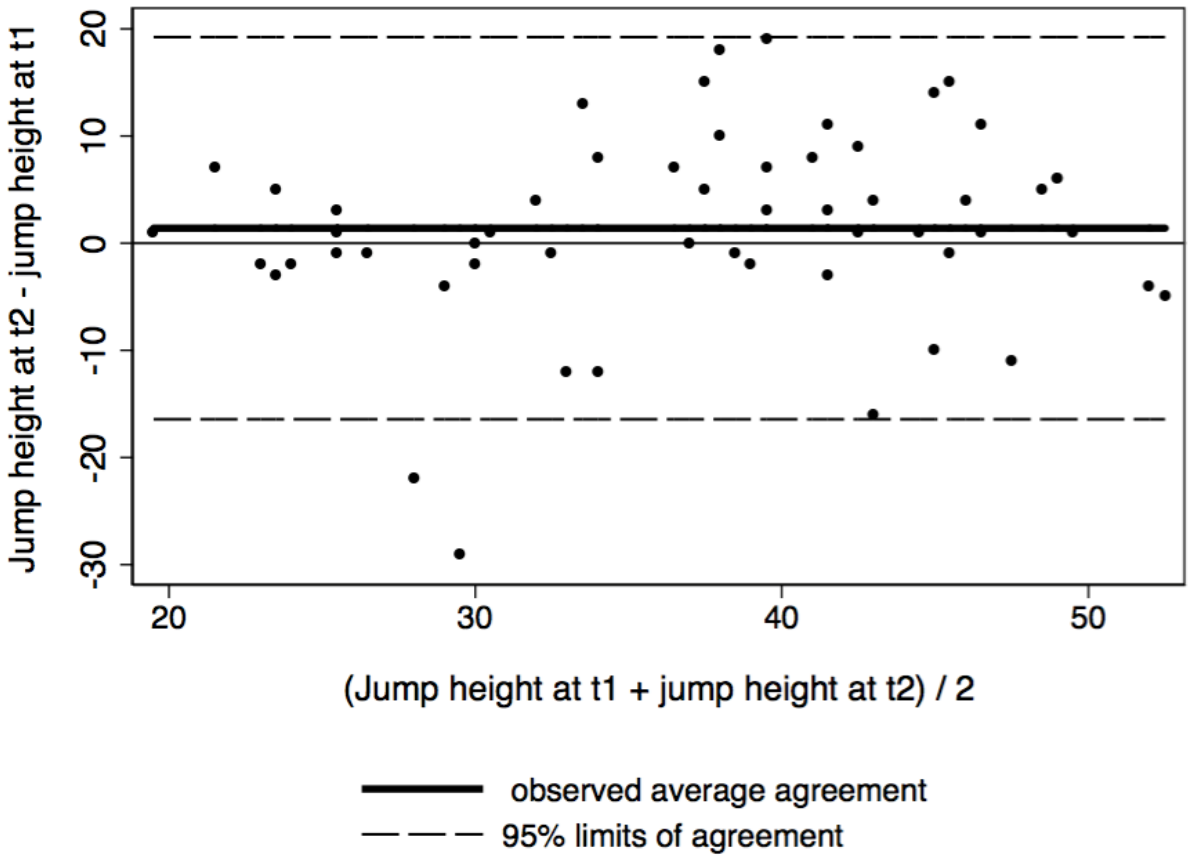
Mean absolute difference = 8.2 (SD = 14.2), ρ_c = 0.29, 95% CI 0.17-0.42, n = 96

FIGURE 4: BLAND-ALTMAN PLOT OF TEACHERS' MEASUREMENTS OF WC AMONG 6TH GRADE STUDENTS



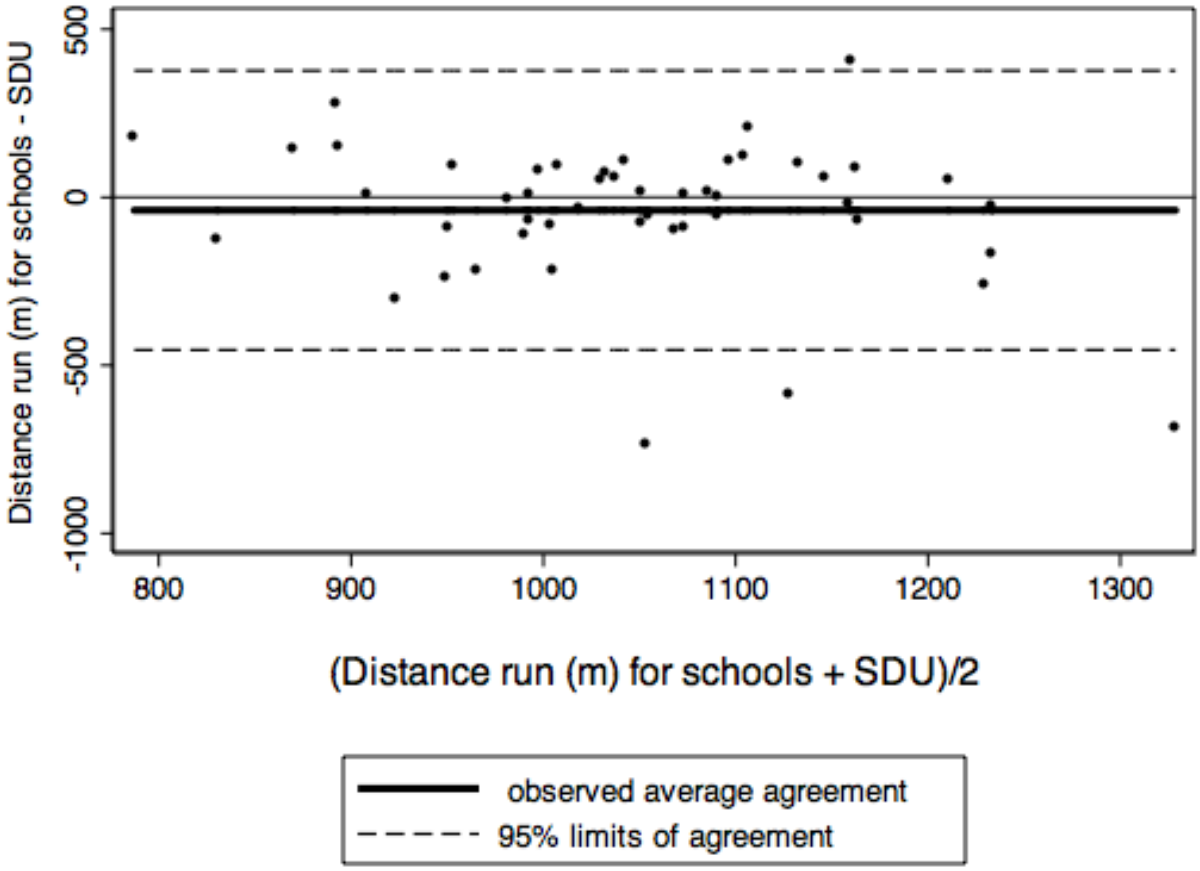
Mean absolute difference = 3.7 (SD = 9.2), ρc = 0.4, 95% CI 0.25-0.55, n = 78

FIGURE 5: BLAND-ALTMAN PLOT OF TEACHERS' MEASUREMENTS OF VERTICAL JUMP TEST FOR STUDENTS IN GRADES 8 AND 9



Mean absolute difference = 1.4 (SD = 9.1), ρ_c = 0.56, 95% CI 0.38-0.74, n = 55

FIGURE 6: BLAND-ALTMAN PLOT OF RESEARCHERS' AND SCHOOLS' MEASUREMENTS OF CRF (ANDERSEN TEST) FOR STUDENTS IN GRADES 4-6



Mean absolute difference = -38.6 (SD = 212.0), ρ_c = 0.04, 95% CI -0.23-0.31, n = 49

APPENDIX 2: TABLES 1-5

TABLE 1: SUMMARY OF AGES AND DIFFERENCES IN ANTHROPOMETRIC MEASUREMENTS AND PHYSICAL FITNESS TESTS BETWEEN ORIGINAL MEASUREMENTS AND CONTROL MEASUREMENTS TWO WEEKS LATER (MEAN AND SD) IN DANISH SCHOOLCHILDRENA

	Grade				
	0th	1st	2nd	3rd	4th
	(n = 118)	(n = 25)	(n = 40)	(n = 62)	(n = 123)
Percent female	57.5	36.0	51.2	59.0	46.0
Age (years)	7.0 (0.5)	8.1 (0.5)	9.0 (0.6)	9.8 (0.5)	11.0 (0.6)
Height (cm)	0.2 (1.6)	-0.7 (2.5)	0.8 (2.2)	0.1 (1.0)	0.2 (1.6)
Weight (kg)	0.4 (2.2)	0.2 (2.2)	0.5 (2.1)	0.1 (2.4)	0.4 (0.7)
WC (cm)	0.3 (2.5)	-0.1 (2.9)	0.4 (3.5)	-0.3 (1.6)	-0.3 (1.4)
Vertical jump (cm)	1.1 (3.3)	2.0 (4.3)	-0.3 (5.7)	0.1 (4.0)	0.8 (4.8)
CRF (m)[b]	15.8 (46.7)	-13.1 (110.3)	-7.1 (81.4)	11.8 (64.3)	15.6 (85.4)
	Grade				
	5th	6th	7th	8th-9th	All
	(n = 31)	(n = 84)	(n = 96)	(n = 56)	(n = 635)
Percent female	56.3	46.6	49.5	52.6	50.8
Age (years)	12.1 (0.5)	12.9 (0.6)	13.7 (0.5)	15.0 (0.7)	10.8 (2.6)
Height (cm)	0.3 (3.2)	1.5 (5.0)	6.5 (13.5)	-0.7 (4.5)	1.0 (5.9)
Weight (kg)	0.8 (2.1)	3.4 (8.6)	0.2 (3.7)	0.2 (4.6)	0.7 (3.8)
WC (cm)	-1.0 (2.3)	3.4 (8.8)	-0.1 (3.0)	-2.1 (5.1)	-0.1 (4.0)
Vertical jump (cm)	-0.3 (3.4)	-2.3 (6.4)	0.5 (4.0)	1.2 (9.1)	0.3 (5.2)
CRF[b]	3.4 (92.2)	-47.3 (141.1)	-46.7 (131.5)	-76.9 (221.5)	-11.8 (114.3)

a) Differences are calculated by subtracting the results of the control measurements from the original measurements. b) The Andersen test

TABLE 2. INTRA-OBSERVER RELIABILITY (ρC) OF ANTHROPOMETRIC MEASUREMENTS AND PHYSICAL FITNESS TESTS, 2010/11.

	Grade									
	0th (n=118)	1st (n=25)	2nd (n=40)	3rd (n=62)	4th (n=123)	5th (n=31)	6th (n=84)	7th (n=96)	8th-9th (n=56)	All (n=635)
	ρ_c	ρ_c	ρ_c	ρ_c	ρ_c	ρ_c	ρ_c	ρ_c	ρ_c	ρ_c
Height (cm)	0.95	0.94	0.91	0.99	0.99	0.99[b]	0.78[b]	0.29[b]	0.87	0.92[b]
Weight (kg)	0.92[b]	0.95	0.89	0.94[b]	0.99[b]	0.98[b]	0.57[b]	0.97[b]	0.92	0.96[b]
WC (cm)	0.87[b]	0.95	0.68	0.95	0.98	0.95	0.40[b]	0.98	0.75[b]	0.87[b]
Vertical jump test (cm)	0.87[b]	0.97	0.66[b]	0.87	0.92	0.91	0.48[b]	0.88	0.56	0.86
CRF (m)[a]	0.96[b]	0.95	0.73	0.77	0.81	0.95[b]	0.64[b]	0.60[b]	0.67	0.83

a) The Andersen test. Estimates are concordance correlation coefficients (ρ_c). b) Statistically significant correlations between differences and mean scores

TABLE 3: INTRA-OBSERVER RELIABILITY (ρC) OF ANTHROPOMETRIC MEASUREMENTS AND PHYSICAL FITNESS TESTS 2011/12.

	Grade										
	0th (n=118)	1th (n=25)	2th (n=40)	3rd (n=62)	4th (n=123)	5th (n=159)	6th (n=58)	7th (n=50)	8th (n=83)	9th (n=)	All (n=700)
	ρ_c	ρ_c	ρ_c	ρ_c	ρ_c	ρ_c	ρ_c	ρ_c	ρ_c	ρ_c	ρ_c
Height (cm)	0.79	0.97	0.99	0.79[b]	0.98	0.94	0.99	0.94	0.99	1.0[b]	0.99[b]
Weight (kg)	0.95	0.95[b]	0.81	0.69[b]	0.98	0.88	1.0	0.86	1.0	1.0	0.97
WC (cm)	0.81[b]	0.82	0.81	0.52[b]	0.90[b]	0.86[b]	0.98	0.76[b]	0.98	0.96	0.89
Vertical jump test (cm)[b]	-	-	-	-	-	-	-	-	-	-	-
CRF (m)[a]	0.86	0.79	0.88	0.64[b]	0.93[b]	0.81[b]	0.98	0.72	0.96	0.99	0.87

a) The Andersen test. Estimates are concordance correlation coefficients (ρ_c). b) Statistically significant correlations between differences and mean scores

b) Control measurements of the vertical jump test were not obtained in 2011/12

TABLE 4. MEANS, STANDARD DEVIATIONS (SD) AND INTER-OBSERVER RELIABILITY (ρ_C) OF ANTHROPOMETRIC MEASUREMENTS AND PHYSICAL FITNESS TESTS OBTAINED IN 2010/11 AND CORRECTED FOR GROWTH; SCHOOLS' FIGURES FROM THE LAST TIME OF MEASUREMENT; MEASUREMENTS TAKEN BY RESEARCHERS FROM THE UNIVERSITY OF SOUTHERN DENMARK (SDU) IN 2011/12

	Grades											
	0th-3rd			4th-6th			7th-9th			All		
	Schools M (SD)	SDU M (SD)	ρ_C	Schools M (SD)	SDU M (SD)	ρ_C	Schools M (SD)	SDU M (SD)	ρ_C	Schools M (SD)	SDU M (SD)	ρ_C
Height (cm)	129.9 (7.2)	131.6 (7.1)	0.92	150.6 (9.1)	148.8 (10.1)	0.92	167.0 (7.0)	165.1 (7.3)	0.88	148.1 (15.3)	146.8 (13.9)	0.97
Difference	1.7 (2.3)[c]			-1.9 (3.5)[c]			-1.9 (2.9)[c]			-0.8 (3.5)[c]		
N	25			46			16			87		
Weight (kg)	28.2 (6.3)	31.7 (8.1)	0.85	42.1 (9.8)	42.4 (11.7)	0.94	53.0 (7.8)	51.9 (6.1)	0.77	36.9 (12.0)	41.6 (12.0)	0.94
Difference	3.4 (2.2)[c]			0.3 (3.9)			-1.1 (4.6)			0.8 (4.0)		
N	25			52			20			97		
WC (cm)	62.6 (7.6)	64.4 (8.5)	0.81	65.3 (8.2)	67.5 (9.3)	0.84	71.4 (7.8)	69.5 (4.0)	0.49	66.0 (8.4)	67.2 (8.3)	0.81
Difference	1.9 (4.7)[c]			2.2 (4.4)[c]			-1.9 (6.1)[c]			1.2 (5.1)		
N	19			44			17			80		
Jump height[a] (cm)	21.4 (5.1)	23.2 (5.3)	0.57	31.5 (6.5)	29.2 (6.0)	0.27	42.7 (8.0)	41.1 (8.4)	0.60	30.6 (9.6)	29.6 (8.6)	0.71
Difference	1.8 (4.6)			-2.3 (7.5)			-1.6 (7.2)			-1.0 (6.9)		
N	24			45			15			84		
CRF (m)[b]	952.8 (200.1)	941.8 (122.9)	0.11	1067.0 (170.1)	1028.4 (134.1)	0.04	1164.5 (128.6)	1037.7 (101.4)	0.12	1049.3 (188.3)	1001.9 (130.4)	0.18
Difference	-11.0 (221.1)[c]			-38.6 (212.0)			-126.8 (147.4)[c]			-47.4 (206.4)[c]		
N	34			49			21			104		

a) Vertical jump test. b) Andersen test. c) Statistically significant correlations between differences and mean scores. Estimates are concordance correlation coefficients (ρ_C)

TABLE 5: MEANS, STANDARD DEVIATIONS (SD) AND INTER-OBSERVER RELIABILITY (ρC) OF ANTHROPOMETRIC MEASUREMENTS OBTAINED BY SCHOOL NURSES IN 2009/10–2011/12 AND BY HSN SCHOOLS 14 DAYS AFTER THE SCHOOL NURSES' MEASUREMENTS.

	Grades											
	0th			1st			5th			9th		
	Schools	School nurses		Schools	School nurses		Schools	School nurses		Schools	School nurses	
	M (SD)	M (SD)	ρ_c	M (SD)	M (SD)	ρ_c	M (SD)	M (SD)	ρ_c	M (SD)	M (SD)	ρ_c
Height (cm)	124.8 (7.0)	124.2 (7.0)	0.98	129.7 (6.8)	129.9 (6.6)	0.96	149.5 (7.4)	150.0 (7.6)	0.97	168.8 (10.5)	169.3 (10.0)	0.99
Difference	-0.6 (1.3)[a]			0.25 (1.8)			0.5 (1.7)			0.5 (1.3)		
N	34			93			57			8		
Weight (kg)	25.5 (4.4)	25.8 (4.3)	0.96	28.2 (6.4)	28.5 (6.3)	0.99	40.6 (9.1)	40.4 (9.0)	0.99	67.1 (19.4)	67.5 (20.1)	1.0
Difference	0.3 (1.2)			0.2 (1.0)			-0.2 (1.6)			0.5 (1.9)		
N	34			95			59			9		
WC (cm)	58.5 (3.8)	57.2 (4.0)	0.59	59.1 (6.2)	57.1 (6.3)	0.84	66.4 (8.2)	65.1 (9.3)	0.90	79.4 (16.3)	75.9 (14.5)	0.93
Difference	-1.3 (3.8)			-1.8 (3.1)[a]			-1.3 (3.7)[a]			-3.5 (4.6)[a]		
N	30			77			51			8		

a) Statistically significant correlation between differences and mean scores. Estimates are concordance correlation coefficients (ρ_c)