

REVIEW

Open Access



A systematic review of the validity, reliability, and feasibility of measurement tools used to assess the physical activity and sedentary behaviour of pre-school aged children

Sophie M. Phillips^{1,2*} , Carolyn Summerbell^{1,2}, Matthew Hobbs³, Kathryn R. Hesketh⁴, Sonia Saxena⁵, Cassey Muir^{2,6} and Frances C. Hillier-Brown^{2,6,7,8}

Abstract

Physical activity (PA) and sedentary behaviour (SB) of pre-school aged children are associated with important health and developmental outcomes. Accurate measurement of these behaviours in young children is critical for research and practice in this area. The aim of this review was to examine the validity, reliability, and feasibility of measurement tools used to assess PA and SB of pre-school aged children.

Searches of electronic databases, and manual searching, were conducted to identify articles that examined the measurement properties (validity, reliability or feasibility) of measurement tools used to examine PA and/or SB of pre-school aged children (3–7 years old). Following screening, data were extracted and risk of bias assessment completed on all included articles.

A total of 69 articles, describing 75 individual studies were included. Studies assessed measurement tools for PA ($n = 27$), SB ($n = 5$), and both PA and SB ($n = 43$). Outcome measures of PA and SB differed between studies (e.g. moderate to vigorous activity, step count, posture allocation). Most studies examined the measurement properties of one measurement tool only ($n = 65$). Measurement tools examined included: calorimetry, direct observation, combined heart rate and accelerometry, heart rate monitors, accelerometers, pedometers, and proxy report (parent, carer or teacher reported) measures (questionnaires or diaries). Studies most frequently assessed the validity (criterion and convergent) ($n = 65$), face and content validity ($n = 2$), test-retest reliability ($n = 10$) and intra-instrument reliability ($n = 1$) of the measurement tools. Feasibility data was abstracted from 41 studies.

Multiple measurement tools used to measure PA and SB in pre-school aged children showed some degree of validity, reliability and feasibility, but often for different purposes. Accelerometers, including the Actigraph

* Correspondence: sophie.m.phillips@durham.ac.uk

¹Department of Sport and Exercise Sciences, Durham University, Durham City, UK

²The Centre for Translational Research in Public Health (Fuse), Newcastle upon Tyne, UK

Full list of author information is available at the end of the article



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

(in particular GT3X versions), Actical, ActivPAL and Fitbit (Flex and Zip), and proxy reported measurement tools used in combination may be useful for a range of outcome measures, to measure intensity alongside contextual information.

Keywords: Physical activity, Sedentary behaviour, Pre-school, Validity, Reliability, Feasibility, Measurement

Background

Physical activity (PA) and sedentary behaviour (SB) in children are associated with numerous health and developmental outcomes [1–3]. Evidence of the importance of these associations in pre-school aged children has been a relatively recent area of research enquiry and is still emerging [4–7]. A pre-school aged child refers to any child who has not yet reached the age of formal schooling, usually aged between 3 and 5 years old but varies internationally [8]. The World Health Organization (WHO) [9] recommend that pre-school aged children should spend at least 180 minutes per day in a variety of physical activities, of which 60 minutes should include moderate to vigorous PA (MVPA). Recommendations for SB suggest that children should not be sedentary for extended periods of time, should not be restrained (such as in a pram) for more than 60 minutes at a time, and should engage in no more than 60 minutes of sedentary screen time per day. The WHO guidelines also suggest that when sedentary, pre-school aged children should engage in activities such as reading and storytelling [9]. Although estimates of guideline adherence vary in the literature, there is evidence to suggest that high proportions of pre-school aged children meet the 180 minutes PA guideline [10–12], but do not always engage in 60 minutes of MVPA [12]. Additionally, pre-school aged children are thought to spend extensive periods of their day sedentary [13, 14], and often do not meet the 60 minutes screen time guideline [10, 11, 13].

It is crucial to monitor PA and SB in pre-school aged children in response to changes in national and local policy; to survey guideline adherence; to develop appropriate policies and programmes; and to establish the efficacy of interventions and initiatives aimed at changing these behaviours [15, 16]. The measurement of PA and SB using quality tools which have optimal measurement properties, including validity, reliability and feasibility, are fundamental as they underpin research and practice in this area [15, 16]. However, there are no clear and up-to-date recommendations, or guidance, on the best tools to measure PA and SB in pre-school aged children.

PA and SB can be measured using various tools (or methods) including proxy report measures (questionnaires/diaries), device-based measurement tools (e.g. accelerometers, pedometers, heart rate monitors, combined

heart rate and accelerometry), direct observation and measures of energy expenditure (e.g. doubly labelled water (DLW) and whole room calorimetry). Selecting the best quality tool to use for a particular purpose in any age group can be difficult [17] and there are additional and specific issues to consider for pre-school aged children. These include the more sporadic and intermittent nature of their movement [18], reduced cognitive capabilities which limit the ability to recall their own behaviour [19], and the increased likelihood that they will tamper with device based measurement tools [20].

Existing reviews have examined the measurement properties of selected measurement tools used to assess PA and SB in children [16, 21–25], including those which focused specifically on questionnaire based measures [26–29]. However, these reviews did not examine the full range of measurement tools used to assess PA and SB, and did not focus specifically on pre-school aged children.

In 2007, Oliver and colleagues, conducted a review examining prevalence and measurement issues in assessing the PA of pre-school aged children [19]. The authors summarised studies that had examined the validity and reliability of a range of measurement tools used to assess PA of young children; however, a rapid scoping search of relevant studies that we conducted prior to the review presented here, identified a large number of potential studies for inclusion that were published after Oliver et al's review. Further, this review did not examine SB: the most likely reason for this is that the important associations of SB with health and developmental outcomes, independent of PA, have only started to emerge in the last 10–15 years [6, 7, 30].

Therefore, the aim of the present review was to examine the validity, reliability, and feasibility, of measurement tools used to assess PA and SB in pre-school aged children.

Methods

This systematic review is reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) criteria [31] (see Additional file 1). The protocol for this systematic review was registered at the International Prospective Register for Systematic Reviews (PROSPERO), registration number CRD42019133613.

Search strategy

Systematic searches of seven major electronic databases (Scopus, Web of Science, PsycARTICLES, PsycINFO, MEDLINE, CINAHL and SPORTDiscus); topic specific journals: Journal for the Measurement of Physical Behaviour and Pediatric Exercise Science; and the grey literature (opengrey.eu and Research Gate), were conducted to identify relevant studies. Searches were conducted in March 2019, and updated in March 2020. No restrictions were placed on language, year of study or publication status.

The search strategy included combinations of the: construct (physical activity, sedentary, sitting), population (pre-school, early years, early childhood, young children and kindergarten) and measurement properties (assessment, measurement, method, valid, reliable, feasible). Searches were adapted to each database, alongside the use of appropriate boolean operators and database specific filters (see Additional file 2). The lead review author (SMP) worked with the Durham University information science team, to refine the search strategies. References and citation searches of included studies, as well as checking the reference lists of selected existing reviews [16, 19, 21–29] were conducted for completeness.

Eligibility criteria for included studies

Studies were eligible for inclusion if their aim was to examine the measurement properties (validity and/or reliability and/or feasibility) of a tool used to measure PA and/or SB of a general population sample of pre-school aged children between the ages of 3 and 7 years old. Table 1 provides an overview of the definitions of measurement properties that were examined in this review.

There remains some debate over the terminology used for validity in the field of PA and SB measurement, particularly in relation to criterion validity [32]. For the purpose of this review, we developed a level of evidence scheme to distinguish between validity studies that used different comparison measurement tools. Level 1 includes criterion validity studies; the only methods considered ‘criterion’ were calorimetry based methods (e.g. whole room calorimetry and DLW) when compared against a measurement tool aiming to measure energy expenditure. Levels 2–4 include convergent validity, separated by the quality of the comparison measurement tool used. Level 2 includes studies where a tool has been compared against a measure which is considered to have relatively high validity (but not criterion), such as direct observation. Level 3 includes studies where a tool has been compared against a measure which is considered to have relatively low validity, such as device based measurement tools. Level 4 includes the comparison of two (or more) of the same type of measurement tool where neither tool is considered to have known higher validity, such as the outcomes of two makes of accelerometer being compared. Table 2 provides a full explanation on what constitutes each level of evidence.

Studies were excluded if:

- a) The aim of the study *was not* to examine the measurement properties of the tool. For example, studies were excluded if the aim was to examine the reproducibility or tracking of behaviours over time, inter-observer reliability, epoch lengths, wear time, the calibration of cut points or prediction equations (with no separate validation study);

Table 1 Definition of each of the measurement properties included in this review

Measurement property	Definition
Validity	<i>Ability for a measure to accurately reflect the construct it is designed to measure.</i>
Criterion validity	Output of a measure produces similar results to a ‘criterion’ measure. This includes studies that have examined a tool for determining energy expenditure with calorimetry (including doubly labelled water) used as the criterion measure.
Convergent validity	Output of a measure produces similar results to a reference measure not considered a criterion.
Face validity	Appearance of a measure, in that it appears to measure what it claims to measure
Content validity	Extent to which a measure covers all aspects of the intended domains or dimensions that it claims to measure
Reliability	<i>Extent to which a tool gives measurements that are consistent, stable and repeatable.</i>
Test-retest reliability	The extent to which a measure can obtain similar results in repeated trials, keeping as many conditions stable as possible
Inter/Intra Instrument Reliability	The extent to which scores are consistent when measurements are taken by different versions of the same instrument (inter-instrument) or by the same version of an instrument repeatedly (intra-instrument)
Feasibility	The extent to which a measurement tool: is suitable for the target population; can be successfully delivered in the target population/context; shows promise of being successful within the intended population. This can include: participant acceptability, researcher acceptability and cost, which can be assessed for all measurement tools through qualitative feedback of participants and through missing or lost data occurred from the measurement tool (with the exception of proxy or self-reported tools that can only be determined through qualitative means including the comprehensiveness and relevance of items).

Table 2 Level of evidence for validity studies included in this review

Level of evidence	Explanation	Criteria of comparisons (examples)	
		Measurement tool under study	Comparison tool/measure
1	Criterion validity	Any measurement tool to determine energy expenditure	Calorimetry, including doubly labelled water
2	Convergent validity: measurement tool compared with a measure which is considered to have <i>relatively high validity</i> (but not considered criterion)	Direct observation protocols (newly devised) Device based measurement tools: combined heart rate and accelerometer, heart rate monitor, accelerometer, pedometer Proxy reported measurement tool: questionnaire, diaries	Direct observation protocol (pre-existing) Direct observation Electrodiagram (for heart rate monitors) Direct observation
3	Convergent validity: measurement tool compared with a measure which is considered to have <i>relatively low validity</i>	Direct observation protocols (newly devised)	Device based measurement tools: combined heart rate and accelerometer, heart rate monitor, accelerometer or pedometer
4	Convergent validity: two (or more) of the same type of measurement tools being compared, where <i>neither tool is considered to have known higher validity</i>	Device based measurement tools: combined heart rate and accelerometer, heart rate monitor, accelerometer, pedometer Proxy reported measurement tool: questionnaire, diaries Device based measurement tools: combined heart rate and accelerometer, heart rate monitor, accelerometer Proxy reported measurement tool: questionnaire, diaries	Device based measurement tools with known higher validity than tool under study: combined heart rate and accelerometer, heart rate monitor, accelerometer Device based measurement tools: combined heart rate and accelerometer, heart rate monitor, accelerometer or pedometer Device based measurement tools: combined heart rate and accelerometer, heart rate monitor, accelerometer Device based measurement tools: combined heart rate and accelerometer, heart rate monitor, accelerometer

- b) The aim of the study was *not* to examine a tool measuring PA and/or SB. For example, studies were excluded if their aim was to examine physical fitness, motor skills, PA environment, correlates of PA or SB, or the impact of interventions;
- c) The study included children under 3 years or over 7 years of age, or if the population sample included children with chronic conditions;
- d) The study was a study protocol or a review. Higher degree theses and conference abstracts were included, however, where the relevant information from the theses were also provided in published peer-reviewed journal articles, the journal article was included and the thesis excluded.

Screening for relevant studies to include in the review

Following the searches, all identified articles were imported into a referencing manager software (Endnote X9.1) and duplicates were removed. Titles and abstracts of included articles were screened by the lead review author (SMP) for inclusion, with a further 10% screened by a second reviewer (MH). There was very high agreement between the two reviewers for title and abstract screening, with a Cohen's kappa statistic [37] of $k = 0.94$. Any disagreement on the eligibility of particular studies was resolved through discussion, without the need for escalation to a third reviewer (FCHB). Full texts of potentially relevant studies were then double screened by two reviewers (SMP, MH, and/or FCHB) for inclusion.

Data extraction of individual studies included in the review

Data from all relevant studies were extracted independently by two data extractors (SMP, CM, and/or FCHB) using a pre-piloted data extraction form. Any discrepancies were resolved by discussion. Extracted information included: study characteristics, participant characteristics, the measurement tool explored (e.g. accelerometry), the measurement tool(s) used as a comparison, data interpretation choices for device based measurement tools (e.g. cut points, epoch, placement), statistical method used to compare measurement tools, behaviour category (PA and/or SB), and the details of the units of measure (e.g. moderate-to-vigorous PA), measurement properties assessed (e.g. criterion validity), the results of the study, and the sources of funding for the study.

Risk of bias assessment of individual studies included in the review

Risk of bias assessment was conducted independently by two reviewers (SMP, FCHB), with any discrepancies resolved through discussion. The risk of bias of all individual studies included in this review was assessed using a modified version of the Downs and Black Checklist, a

method suitable for appraising non-randomised studies [38]. This modified checklist has been successfully used in previous systematic reviews examining PA assessment measures [21, 39]. The tool includes seventeen questions: eight focus on the quality of reported criteria, three on the external validity and five the internal validity. The maximum quality score a study could receive was 17, with study quality rated as good (13-17), fair (9-< 13) or poor (< 9), based on the protocols used in previous reviews [40-42].

An additional risk of bias assessment was conducted on studies examining proxy reported measurement tools, using the COnsensus-based Standards for the selection of health Measurement INSTRUMENTS (COSMIN) risk of bias checklist [43, 44]. This checklist was devised specifically for assessing the risk of bias of participant reported measurement property studies. Based on the studies included in our review we conducted the assessment using the sub-sections relating to reliability and construct (convergent) validity. It was not possible to assess the quality of the content validity of the studies due to such minimal information available. Each item was scored using a 4 point scale (very good, adequate, doubtful, inadequate). The overall methodological rating of a study was determined using the COSMIN protocol of 'the worst score counts' principle [45] (e.g. if the lowest rating of all items was 'doubtful', the overall methodological quality of the measurement property in that study would be rated as 'doubtful').

Interpretation of validity, reliability, and feasibility

Studies commonly use a number of different statistical analyses to define absolute (agreement between the two measurement tools) or relative (the degree to which the two measurement tools rank individuals in the same order) validity and reliability [32, 35]. These types of statistical analyses include correlations (Pearson's; Spearman's; Kendall's; Intraclass), linear regressions, root mean square error (RMSE), Bland Altman, kappa statistics and area under the receiver operating curve (AUC-ROC) [35, 46-51]. Additionally, studies use different methods of analysing and reporting the feasibility of measurement tools. In order to demonstrate consistency in the interpretation of the results across studies, and also to compare and rank these results, we scoped the relevant literature to search for guidance.

We found no consensus in the literature as to which statistical test results indicate weak, moderate, or good validity or reliability. However, in line with a number of previous reviews of this type [23, 26, 28, 29, 36, 52], we provide an overview of what constitutes a 'weak', 'moderate', or 'good' statistical result for validity or reliability, to rank individual studies in this way.

For proxy reported based measurement tools, feasibility can only be determined using qualitative methods, including to determine if questions are relevant, comprehensible and understandable [34]. However, there is no standardised way of determining feasibility of other measurement tools. Therefore, feasibility in the present review was based on qualitative acceptability or feasibility of the measurement tools where data was available. Additionally, an indication of feasibility was also based on compliance and usable data information using values from the Effective Public Health Practice Project (EPHPP) quality assessment tool for withdrawal and drop out [53]. The EPHPP is primarily used for assessing the quality of quantitative intervention based studies in systematic reviews and rates studies as ‘strong’, ‘moderate’, or ‘weak’ based on the percentage of participants completing the study. We applied the criteria to provide an indication on the feasibility of the measurement tools in our review, to indicate a ‘weak’, ‘moderate’, or ‘good’ level of feasibility based on the percentage of usable data from the measurement tools (as a result of missing data and/or drop out) [53]. This was used to provide an indication on feasibility only, as the true feasibility of measurement tools should be determined through the use of qualitative research methods [36, 54–56]. When interpreting summary scores for feasibility, more weight was given to qualitative findings; the scores of these studies were based on the qualitative data provided in the original study. Information on the interpretation of the studies can be found in Table 3.

Combining the results of individual studies for overall interpretation

We summarised the results of studies where they aimed to compare a particular measurement property

(separated by level of evidence for validity) of a particular measurement tool (e.g. Actigraph GT3X accelerometers). We have included summaries of this information in: Table 4 (level 1 validity evidence), Table 5 (level 2 validity evidence), Table 6 (level 3 validity evidence), Table 7 (level 4 validity evidence), and Table 8 (reliability and feasibility evidence). The evidence outlined in the tables was interpreted as follows:

- Where the specific measurement property for a specific measurement tool was deemed ‘good’ in over half of these studies, the summary assessment was deemed to be ‘good’.
- Where the specific measurement property for a specific measurement tool was deemed ‘moderate’ in over half of these studies, the overall assessment was deemed to be ‘moderate’.
- Where the specific measurement property for a specific measurement tool was deemed ‘weak’ in over half of these studies, the overall assessment was deemed to be ‘weak’.
- In instances where the specific measurement property of a measurement tool had mixed evidence in the studies, such as studies with outcomes of ‘weak’ and ‘moderate’, or ‘moderate’ and ‘good’, the overall assessment was deemed to be the most positive of the two outcomes.

For completeness, we have included information about all tools of reasonable quality in the summary tables (Tables 4, 5, 6, 7 and 8) where there was any available evidence, including where there was only one or two studies that reported results. The information is greyed out to highlight where the evidence is based on ≥ 3 studies.

Table 3 Main statistical analyses and interpretation of statistics

	Relative or Absolute Validity/reliability?	Weak	Moderate	Good
Correlations (r)				
Pearson’s	Relative	< 0.60	0.60–0.79	≥ 0.80
Spearman’s	Relative	< 0.60	0.60–0.79	≥ 0.80
Kendall’s	Relative	< 0.60	0.60–0.79	≥ 0.80
Intraclass correlation coefficient	Absolute	< 0.60	0.60–0.69	≥ 0.70
Linear Regression (% variance explained by the measurement tool)	Relative	< 60%	60–79%	$\geq 80\%$
Root mean squared error	Absolute	*	*	*
Bland Altman (mean difference, limits of agreement, bias)	Absolute	*	*	*
Kappa (r)	Absolute	< 0.60	0.60–0.69	≥ 0.70
Area under the receiver operating curve (AUC-ROC)	Relative	< 0.70	0.70–0.79	≥ 0.80
Feasibility (% of usable data)		< 60%	60–79%	80–100%

(References: ([23, 26, 28, 29, 35, 36, 46–53]))

*Depends on the units of measurement

Table 4 Summary table of level 1 validity evidence of the measurement tools

Measurement tools under study	Outcome measures		References
	Energy expenditure	VO ₂	
Combined heart rate and accelerometer			
Actiheart			[57]
Accelerometers			
Actigraph (MTI)	1-2	3	[58, 59]
Actigraph (GT3X)	2-3	3	[60, 61]
Actical	4-6	5	[57, 62, 63]
Actigraph (wGT3X-BT)			[64]
ActivPAL			[60, 65]
GENEActiv			[60]
Triaxial Research Tracker 3 (RT3)			[57]
Actiwatch (AW16)			[66]
TracmorD			[67]
Proxy reported measurement tools			
Children's physical activity questionnaire (CPAQ)			[68]

This table shows a summary of the results of studies where they aimed to compare a particular measurement tool (eg, Actigraph GT3X accelerometer) against calorimetry (including doubly labelled water). The summary ratings were based on the quality of the tools for this specific measurement property. Where the measurement tool was deemed 'good' in the majority of the studies, the summary assessment was deemed 'good'. Where the measurement tool was deemed 'moderate' in the majority of the studies, the summary assessment was deemed 'moderate'. Where the measurement tool was deemed 'weak' in the majority of the studies, the summary assessment was deemed 'weak'. In instances where the measurement tool had mixed evidence in the studies, such as studies with outcomes of 'weak' and 'moderate', or 'moderate' and 'good', the overall assessment was deemed to be the most positive of the two outcomes. All tools of reasonable quality where any evidence was available are included in this table, including where only one or two studies reported that result.

Cut points: ¹Ekelund et al. 2001 [69]; ²Puyau et al. 2002 [70]; ³Pate et al. 2006 [58]; ⁴Evenson et al. 2008 [71]; ⁵Pfeiffer et al. 2006 [63]; ⁶Adolph et al. 2012 [57]

*Methodology used to assess the ability of the tool is detailed in the methods above and is indicated in the summary table as:

Good = Moderate = Weak =

Key for colour of boxes:

= evidence from ≥3 studies

= evidence from <3 studies

Table 5 Summary table of level 2 validity evidence of the measurement tools (Continued)

Method	Outcome measures						References
	SB	Posture allocation	LPA	MVPA	TPA	Levels of activity	
Actiwatch (Spectrum)	● 16		● 16	● 16			[76]
Actiwatch (MiniMitter)						●	[93]
Actiwatch (AW16)						●	[83]
Actiwatch L						●	[94]
Pedometers							
Yamax Digiwalker (SW-200)						●	[82, 95–97]
Yamasa AM-5 Pedometer				●			[98]
MVP 4 Walk4Life Digital				●		●	[99]
Proxy reported measurement tools							
Teacher/mother reported habitual PA	●					●	[98, 100]

Published cut points used:¹ Evenson et al., 2008 [71]; ² Johansson et al., 2015 [101]; ³Pate et al., 2006 [58]; ⁴ Sitarid et al., 2005 [85]; ⁵Van Cauwenberghe et al., 2011 [102]; ⁶Puyau et al., 2002 [70]; ⁷Reilly et al., 2003 [84]; ⁸Frederick et al., 2005 [103]; ⁹Adolph et al., 2012 [57]; ¹⁰ Pfeiffer et al., 2006 [63]; ¹¹ Schaefer et al., 2014 [104]; ¹²Vanhelst et al., 2000 [105]; ¹³Rowlands et al., 2004 [106]; ¹⁴Sun et al., 2008 [107]; ¹⁵Chu et al., 2007 [108]; ¹⁶Ekblom et al., 2012 [109]

***Methodology used to assess the ability of the tool is detailed in the methods above and is indicated in the summary table as:**

Good = ● Moderate= ● Weak= ●

Key for colour of boxes:

■ = evidence from ≥3 studies

□ = evidence from <3 studies

This table shows a summary of the results of studies where they aimed to compare a particular measurement tool (e.g. Actigraph GT3X accelerometer) against direct observation (or electrodiagram for the heart rate monitor). The summary ratings were based on the quality of the tools for this specific measurement property. Where the measurement tool was deemed 'good' in the majority of the studies, the summary assessment was deemed 'good'. Where the measurement tool was deemed 'moderate' in the majority of the studies, the summary assessment was deemed 'moderate'. Where the measurement tool was deemed 'weak' in the majority of the studies, the summary assessment was deemed 'weak'. In instances where the measurement tool had mixed evidence in the studies, such as studies with outcomes of 'weak' and 'moderate', or 'moderate' and 'good', the overall assessment was deemed to be the most positive of the two outcomes. All tools of reasonable quality where any evidence was available are included in this table, including where only one or two studies reported that result

Table 6 Summary table of level 3 validity evidence of the measurement tools

	Outcome measures							References
	SB	Posture allocation	LPA	MVPA	TPA	Levels of activity	Step count	
Direct observation								
OSRAC-P (Observation System for Recording Physical Activity in Children- Preschool)						●		[110]
SOFIT-P (System for Observing Fitness Instruction Time for Preschoolers)	●			●				[111]
Accelerometers								
Fitbit (Flex)	●			●	●			[112]
ActivPAL	●							[113]
Best Fit Friend						●		[114]
Pedometers								
Yamax Digi-Walker (SW-200)							●	[115, 116]
Yamax Digi-Walker (SW-700)							●	[117]
Omron Walking Style Pro (HJ-720IT-E2)							●	[118, 119]
Proxy reported measurement tools								
Nursery teacher's report (based on Toyama Cohort Study survey questions)							●	[120]
Leisure time report				●				[121]
Pre School Physical Activity Questionnaire (PRE-PAQ)	●		●	●				[122]
Netherland's Physical Activity Questionnaire (NPAQ)					●			[123]
Questionnaire developed for parents of pre-schoolers in Mexico	●			●				[124]
Teacher activity rating				●				[121]
7 day activity diary (adapted from CLASS)	●						●	[125]
Habitual Activity Estimation Scale (HAES)				●				[126]
Children's physical activity questionnaire (CPAQ)				●				[68]

Table 6 Summary table of level 3 validity evidence of the measurement tools (Continued)

	Outcome measures						References
	SB	Posture allocation	LPA	MVPA	TPA	Levels of activity	
Children's Leisure Activities Study Survey (CLASS)				●	●		[127]
TV Diary	●						[128]
Teacher/mother reported habitual PA					●		[86, 98]

This table shows a summary of the results of studies where they aimed to compare a particular measurement tool (e.g. Pre School Physical Activity Questionnaire (PRE-PAQ)) against a measure which is considered to have relatively low validity such as a device based measurement tool (see Table 2 for a full explanation). The summary ratings were based on the quality of the tools for this specific measurement property. Where the measurement tool was deemed 'good' in the majority of the studies, the summary assessment was deemed 'good'. Where the measurement tool was deemed 'moderate' in the majority of the studies, the summary assessment was deemed 'moderate'. Where the measurement tool was deemed 'weak' in the majority of the studies, the summary assessment was deemed 'weak'. In instances where the measurement tool had mixed evidence in the studies, such as studies with outcomes of 'weak' and 'moderate', or 'moderate' and 'good', the overall assessment was deemed to be the most positive of the two outcomes. All tools of reasonable quality where any evidence was available are included in this table, including where only one or two studies reported that result.

*Methodology used to assess the ability of the tool is detailed in the methods above and is indicated in the summary table as:

Good = ● Moderate= ● Weak= ●

Key for colour of boxes:

■ = evidence from ≥3 studies

□ = evidence from <3 studies

Table 7 Summary table of level 4 validity evidence of the measurement tools

Measurement tools under study	Outcome measures					Reference
	SB	MVPA	TPA	Activity counts	Habitual physical activity reports	
Combined heart rate and accelerometer and accelerometers						
Actiheart, Actical and Triaxial Research Tracker 3 (RT3)						[57]
Accelerometers						
Actigraph (GT1M) ¹⁻⁶ and RT3 ⁷⁻¹⁰						[80]
Actigraph (GT1M) ⁴ and ActivPAL						[78]
Actigraph (GT3X+) ³ and Actical ¹¹						[129]
Actigraph (GT3X) ^{1,3} and Actiwatch (Spectrum) ¹²						[76]
Actigraph (CSA/MTI) and Actiwatch (AW16)						[83]
Actical ⁴ and ActivPAL						[130]
Proxy reported measurement tools						
Parental vs teacher reports						[98, 100]

This table shows a summary of the results of studies where they aimed to compare two (or more) of the same type of measurement tools where *neither tool is considered to have known higher validity* (e.g. comparison between Actical and ActivPAL). The summary ratings were based on the quality of the tools for this specific measurement property. Where the measurement tool was deemed 'good' in the majority of the studies, the summary assessment was deemed 'good'. Where the measurement tool was deemed 'moderate' in the majority of the studies, the summary assessment was deemed 'moderate'. Where the measurement tool was deemed 'weak' in the majority of the studies, the summary assessment was deemed 'weak'. In instances where the measurement tool had mixed evidence in the studies, such as studies with outcomes of 'weak' and 'moderate', or 'moderate' and 'good', the overall assessment was deemed to be the most positive of the two outcomes. All tools of reasonable quality where any evidence was available are included in this table, including where only one or two studies reported that result.

Cut points: ¹Sirard et al. 2005 [85]; ²Freedson et al. 2005 [103]; ³Pate et al., 2006 [58]; ⁴Evenson et al. 2008 [71]; ⁵Van Cauwenberghe et al. 2011 [102]; ⁶Puyau et al. 2002 [70]; ⁷Vanhelst et al. 2000 [105]; ⁸Rowlands et al. 2004 [106]; ⁹Sun et al. 2008 [107]; ¹⁰Chu et al. 2007 [108]; ¹¹Pfeiffer et al. 2006 [63]; ¹²Eklom et al. 2012 [109]

*Methodology used to assess the ability of the tool is detailed in the methods above and is indicated in the summary table as:

Key for colour of boxes:

- = Good
- = Moderate
- = Weak

= evidence from ≥3 studies

= evidence from <3 studies

Table 8 Summary table of the reliability and feasibility of the measurement tools























Method	Reliability	Feasibility	Reference
Calorimetry			[60–62, 65–67, 131, 132]
Direct observation			[72, 76, 87, 91, 111, 133]
Heart rate monitors			
Polar Vantage XL Monitor			[73]
Accelerometers			
Fitbit (Zip)			[91]
Triaxial Research Tracker 3 (RT3)			[117]
Actigraph (GT3X/+)			[75, 76, 113, 114, 129]
Actigraph (CSA/MTI)			[68, 82, 83, 85, 115, 122]
Actigraph (wGT3X-BT)			[64]
Actigraph (GT1M)			[78, 113, 116, 118, 124, 125, 128]
ActivPAL			[65, 78, 87–89, 113, 130]
Actical			[86, 129, 130]
Actiwatch (AW16)			[66, 83]
Actiwatch (Spectrum)			[76]
Actiwatch-L			[120]
Caloriecounter			[120]
Tracmor _D			[67]
Caltrac			[134]
Best Fit Friend			[114]
Pedometers			
Yamax (SW-700)			[117, 126]
Yamax Digiwalker (SW-200)			[95, 115, 116]

Table 8 Summary table of the reliability and feasibility of the measurement tools (Continued)

Method	Reliability	Feasibility	Reference
MVP 4 Walk4Life Digital			[99]
Omron Walking Style Pro (HJ-720IT-E2)			[118]
Proxy reported measurement tools			
Children’s Leisure Activities Study Survey (CLASS)			[127]
Questionnaire developed for parents of pre-schoolers in Mexico			[124]
Teacher activity rating			[121]
TV Diary			[128]
Leisure time report			[121]
Pre School Physical Activity Questionnaire (PRE-PAQ)			[122]
Netherlands’s Physical Activity Questionnaire (NPAQ)			[123]
‘Toybox’ Primary Caregivers Questionnaire			[135]
Children’s physical activity questionnaire (CPAQ)			[68]

***Methodology used to assess the ability of the tool is detailed in the methods above and is indicated in the summary table as:**

Good = Moderate= Weak=

Key for colour of boxes: = evidence from ≥3 studies

= evidence from <3 studies

This table shows a summary of the results of studies where they tested the reliability or feasibility of the measurement tool. The summary ratings were based on the quality of the tools for the specific measurement property. Where the measurement tool was deemed ‘good’ in the majority of the studies, the summary assessment was deemed ‘good’. Where the measurement tool was deemed ‘moderate’ in the majority of the studies, the summary assessment was deemed ‘moderate’. Where the measurement tool was deemed ‘weak’ in the majority of the studies, the summary assessment was deemed ‘weak’. In instances where the measurement tool had mixed evidence in the studies, such as studies with outcomes of ‘weak’ and ‘moderate’, or ‘moderate’ and ‘good’, the overall assessment was deemed to be the most positive of the two outcomes. All tools of reasonable quality where any evidence was available are included in this table, including where only one or two studies reported that result

Results

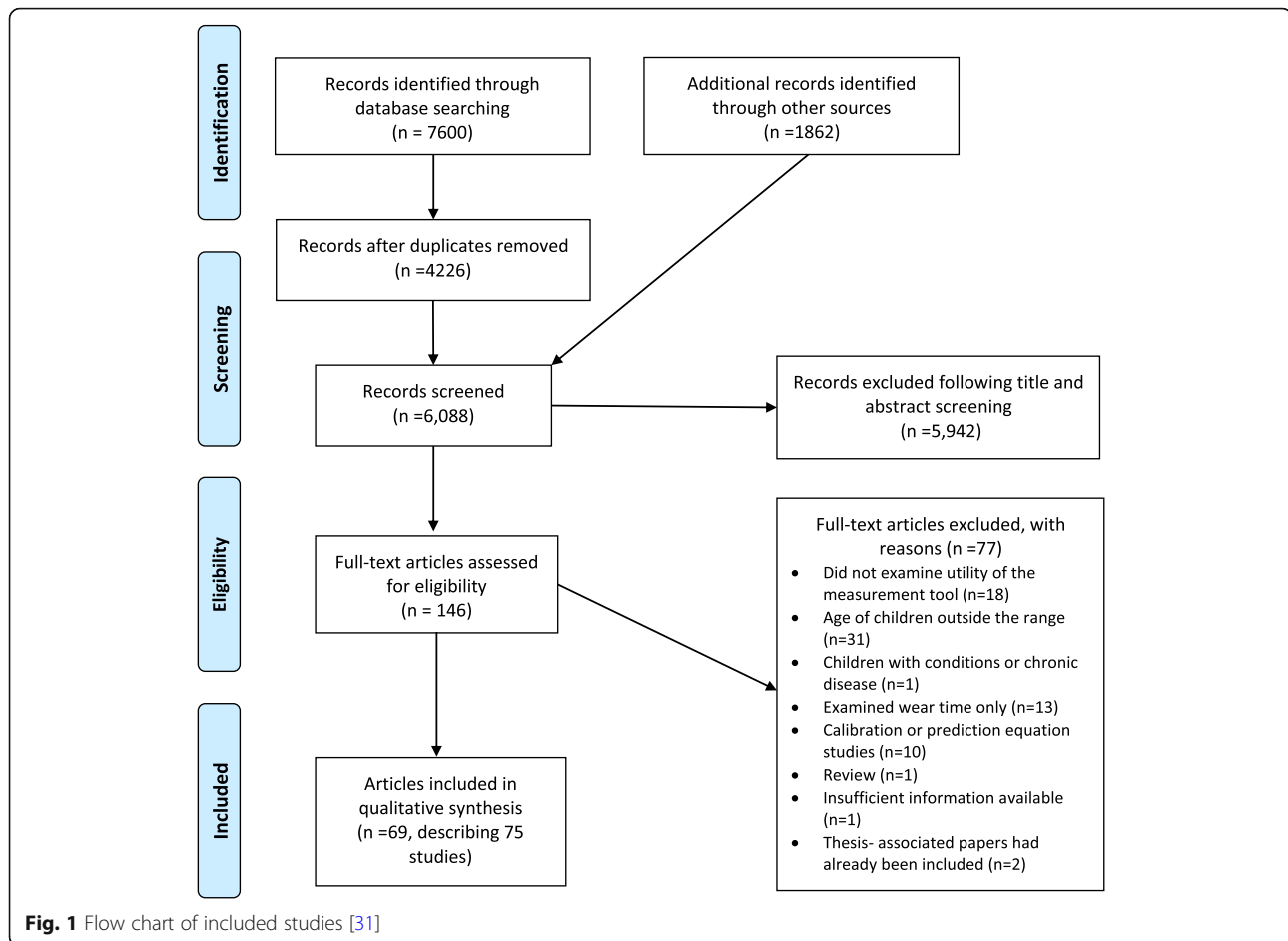
Study selection

A total of 6088 records were screened for inclusion, of which 146 were included for full text screening and 69 articles were included in the review, describing 75 individual studies (See Fig. 1). Sixty seven articles were retrieved following the initial searches, and a further two from the updated searches [60, 75]. All included articles were in English and, in most cases, were peer reviewed journal articles (n = 66) [57–68, 72–98, 100, 110–113, 115–118, 120–137]. We included one abstract [119] and two Masters theses [99, 114] due to the articles meeting the inclusion criteria and providing sufficient information. A list of excluded studies with reasons can be found in Additional file 3.

Description of studies

Detailed information of all included studies can be found in Additional files: 4 (level 1 validity evidence), 5 (level 2

validity evidence), 6 (level 3 validity evidence), 7 (level 4 validity evidence), 8 (reliability) and 9 (feasibility). The majority of the studies were conducted in high income countries: USA (n = 24), UK (n = 17), Australia (n = 10), Belgium (n = 3), Canada (n = 3), Hong Kong (n = 3), New Zealand (n = 3), Greece (n = 2), Japan (n = 2), Netherlands (n = 2), Germany (n = 1), Korea (n = 1), Sweden (n = 1), and both the USA and Sweden (n = 1). One study was conducted in Mexico, an upper to middle income country [124] and one study was a collaboration between high income and an upper to middle income country: Belgium, Bulgaria, Germany, Greece, Poland and Spain [135]. No studies were conducted in low income countries. Study sample sizes ranged from 4 [110] to 269 [85] (median n = 34). Based on criteria outlined in the COSMIN measurement property checklist [43], only 7 of the included studies had excellent (n = > 100), 15 had good (n = 50–99), 26 had moderate (n = 30–49), and 27 had small (n = < 30) sample sizes. All studies



reporting child sex ($n = 69$) included both male and female participants. The median age of children included in the studies was 4.5 years. The studies commonly explored the measurement properties of the measurement tools in free-living conditions ($n = 54$), used structured protocols that were reflective of habitual behaviour of pre-school aged children ($n = 11$) or were conducted in laboratory settings ($n = 10$). A large proportion of the studies did not report on any funding received for the research ($n = 31$). Many of the studies that reported funding were supported by more than one funding source (see Additional file 10 for funding sources of studies).

Studies assessed the measurement tools for PA only ($n = 27$), SB only ($n = 5$), and both PA and SB ($n = 43$). Units of measure varied across studies, and included: activity preferences, activity levels, activity counts, activity energy expenditure, energy expenditure, frequency of activity, heart rate, metabolic equivalent of task (MET) values, posture allocation, step count, time spent in different intensities of activity and VO_2 .

The majority of the studies examined the measurement properties of one measurement tool only ($n =$

65). Several studies examined the measurement properties of two ($n = 8$) or infrequently three ($n = 2$) measurement tools simultaneously, in comparison with the reference methods. Articles examined the measurement properties of: calorimetry ($n = 2$), direct observation ($n = 4$), combined heart rate and accelerometry ($n = 1$), heart rate monitors ($n = 1$), accelerometers ($n = 44$), pedometers ($n = 13$), and proxy report measures (questionnaires or diaries) reported by parent, carer or teacher ($n = 13$).

Validity of the measurement tools was the most frequently examined measurement property; level 1 validity ($n = 12$), level 2 validity ($n = 36$), level 3 validity ($n = 23$) and level 4 validity ($n = 9$). Only two studies examined the face and content validity of the measurement tools. Ten studies described the test-retest reliability of the measurement tools and 1 study the intra-instrument reliability. Feasibility data was abstracted from 41 studies; 13 of these studies had a primary aim of determining the feasibility of the measurement tool, the remaining 28 studies commented on reasons for drop out or exclusion of data, which also classified as assessing feasibility. Table 9 provides an overview of the measurement

Table 9 Overview of all measurement tools examined and measurement properties explored

Measurement tools examined	Validity ^a					Reliability		Feasibility
	Criterion: Level 1	Convergent: Level 2	Convergent: Level 3	Convergent: Level 4	Face/Content	Test-retest-	Intra-instrument	
Calorimetry								.
Direct observation		.	.					.
Combined heart rate and accelerometer								
Actiheart	.	.			.			
Heart rate monitor								
Polar Vantage XL Monitor		.						.
Accelerometers								
Actical
Actigraph (CSA/MTI)
Actigraph (GT1M)		.			.			.
Actigraph (GT3X)
Actigraph (GT3X+)		.			.			.
Actigraph (wGT3X-BT)	.							.
Actometer		.						.
ActivPAL
Actiwatch (AW16)
Actiwatch (MiniMitter)		.						.
Actiwatch (Spectrum)		.			.			.
Actiwatch-L		.						.
Best Fit Friend			.					.
Caloriecounter								.
Caltrac								.
Fitbit (Flex)		.	.					.
Fitbit (Zip)		.					.	.
GENEActiv	.							.
New Lifestyles NL-1000		.						.
Tracmor _D	.							.
Triaxial Research Tracker 3 (RT3)
Pedometers								
MVP 4 Walk4Life Digital		.				.		.
Omron Walking Style Pro (HJ-720IT-E2)		.	.					.
Yamax Digi-Walker (SW-200)		.	.					.
Yamax Digi-Walker (SW-700)			.			.		.
Yamasa AM-5 Pedometer		.						.
Pedometer (type not specified)		.	.					.
Proxy reported measurement tools								
Children's Leisure Activities Study Survey (CLASS)			.			.		.
Children's physical activity questionnaire (CPAQ)
Habitual Activity Estimation Scale (HAES)			.					.

Table 9 Overview of all measurement tools examined and measurement properties explored (*Continued*)

Measurement tools examined	Validity ^a					Reliability		Feasibility
	Criterion: Level 1	Convergent: Level 2	Convergent: Level 3	Convergent: Level 4	Face/Content	Test-retest-	Intra-instrument	
Leisure time report			•				•	
Netherland's Physical Activity Questionnaire (NPAQ)			•				•	
Nursery teacher's report			•					
Pre School Physical Activity Questionnaire (PRE-PAQ)			•		•		•	
Questionnaire developed for parents of pre-schoolers in Mexico			•		•		•	
Teacher activity rating			•				•	
Teacher/mother reported habitual PA		•	•		•			
'Toybox' Primary Caregivers Questionnaire							•	
TV Diary			•				•	
7 day activity diary (adapted from CLASS)			•					

^aValidity separated per level of evidence depending on the quality of measurement tool used (see Table 2 for full explanation)

properties that were examined for each tool, to help determine which type of evaluation was conducted on each of the measurement tools. This table does not state the quality of the tools; please refer to Tables 4, 5, 6, 7 and 8 that outline summaries of the quality of the tools based on the available evidence.

Risk of bias assessment

Risk of bias was assessed for all included studies using the modified Downs and Black checklist [38]. Studies consistently described the main aims of the research, the main outcomes to be measured, the exposures of interest and the main findings. In most cases, the staff, places and facilities were representative of what would usually be the case for the children under study, as testing often took place in nursery settings, at home, or was assessing habitual activity behaviours consisting of children's usual behaviours and routines. However, a proportion of the studies were not reflective of usual activity for children due to them being laboratory based or involving structured protocols.

The main concern regarding potential bias of the studies was related to the lack of reporting of key information in the studies. This included lack of clarity on the representativeness of the sample population compared with the population from which they were recruited (*n* = 63). Also, many studies failed to consistently report reasons for drop out (e.g. non-completion, or missing or incomplete data) (*n* = 34). Many of the studies only reported the number of children included in analysis, but did not include the

number who started the study sample, and so it was unclear as to whether drop out was an issue in these studies.

The majority of the studies demonstrated fair to good methodological quality ratings. Only two studies received a low methodological quality score [136, 137] and were removed from the overall summary analysis, however this did not affect the overall results of the review. The full Downs and Black risk of bias assessment for all included studies can be found in Additional file 11.

An additional risk of bias assessment was conducted on studies reporting proxy reported measurement tools, using the COSMIN risk of bias [43, 44]. The checklist demonstrated that the majority of tools examining convergent validity were of low quality; five studies were rated doubtful [100, 122, 125, 127, 128] and seven studies inadequate [68, 98, 120, 121, 123, 124, 126]. In the majority of the studies it was clear what the comparator tool measured and the statistical methods used were generally appropriate. Main concerns with the methodological quality of the studies related to the measurement properties of the comparator tools, with inadequate information provided. Proxy reported tools examined for test-retest reliability showed a range of quality, with some studies rated as adequate [122, 127, 128], two as doubtful [68, 121] and some inadequate [123, 124, 135]. The full COSMIN risk of bias assessment can be found in Additional file 12. This risk of bias assessment highlighted that the studies on proxy reported measurement tools were generally of low quality. However no studies were removed from the overall summary analysis

based on this additional assessment, due to this evidence being the best available evidence for proxy reported measurement tools. The quality of the evidence is regarded in the interpretation of the studies throughout the review.

Summary of measurement properties of measurement tools, separated by measurement property type

Results presented here are in line with the level of evidence scheme displayed in Table 2. The results will therefore be discussed as follows: Level 1 validity evidence where the tool under study is compared with calorimetry; level 2 validity, where the tool under study is compared with a measurement tool with relatively high validity e.g. direct observation; level 3 validity, where the tool under study is compared with a measurement tool with relatively low validity e.g. device based method such as accelerometry; level 4 validity, where two of the same type of measurement tool are compared where neither tool is considered to have known higher validity e.g. two makes of accelerometer; reliability; followed by feasibility. Detailed study tables of all included studies can be found in Additional files 4, 5, 6, 7, 8 and 9 presented by measurement property examined and separated by level of evidence. Where reported, these tables also include details on the interpretation choices used for device based measurement tools, including: cut points, epoch length, placement, wear time, non-wear time and valid number of days. It is critical that when using this review to help with measurement tool choice decisions, researchers should replicate the procedures in which the tool was validated (e.g. using the same cut points, epoch length and placement that the tool has shown to be valid for), which can all be found in the Additional files.

Level 1 validity

The criterion method of calorimetry (including DLW) for the outcome of energy expenditure was used in 12 studies. Multiple accelerometers were shown to have reasonable ability to determine energy expenditure, but often based on very limited evidence (one study only). There was however stronger evidence to suggest that the Actigraph (in particular the MTI and GT3X) and Actical were both able to determine EE and VO_2 max [58–61], and the ActivPAL to determine EE [60, 65].

Table 4 provides a summary table of included studies that examined level 1 validity evidence of measurement tools (detailed information can be found in Additional file 4).

Level 2 validity

The most commonly used comparison methods for convergent validity of device based measurement tools was direct observation. Studies demonstrated that a number

of device-based measurement tools were reasonably accurate at determining different PA and SB outcomes [81–84, 86, 93, 96]. The Actigraph accelerometer was the most frequently evaluated tool. Overall, these studies showed the Actigraph devices (in particular GT3X versions) had a good ability to determine SB, vigorous PA (VPA) and moderate to vigorous PA (MVPA) [61, 74–77, 79, 80]. Fewer studies evaluated the Actical accelerometer but these showed similar results [57, 62]. The ActivPAL accelerometer was shown to be suitable at assessing SB, MVPA and posture allocation [65, 87–89]; a unique quality that is not identified by other measurement tools. However, there is space for this to be developed further, as the accuracy of this measurement tool for identifying posture allocation is lower than in other population samples due to the amount of time that preschool aged children will spend in ‘other’ postures, such as kneeling and crawling [138]. Fitbits also show some promising results for the measurement of SB, MVPA, total PA (TPA), and step counts [90, 91]; however, these conclusions are based on a very limited number of studies.

For pedometers, when compared against direct observation, study results were mixed but there is limited evidence to suggest that the Yamax Digiwalker SW-200 is able to determine step counts with reasonable accuracy [95, 96].

Table 5 provides a summary table of included studies that examined level 2 validity evidence of measurement tools (detailed information can be found in Additional file 5).

Level 3 validity

Level 3 validity evidence mainly consisted of proxy reported measurement tools and they were most frequently compared to accelerometry. Proxy reported measurement tools were generally poor at determining PA and SB outcomes. However, the Pre-PAQ was shown to be moderately accurate at determining stationary behaviour, light PA and VPA [122] and the *leisure time report* was able to determine MVPA [121]. The Netherland’s physical activity questionnaire and *nursery teacher’s report (based on Toyama Cohort Study survey questions)* could distinguish between different levels of activity [120, 123]. Although relatively few proxy reported tools demonstrated reasonable criterion or convergent validity, this could be due to a lack of face and content validity testing during the development of these tools [34, 139]. It is also worth noting that this evidence is based on very few, low quality studies. An advantage of the proxy reported tools over the other measurement tools is that they were able to capture the context and type of the behaviours, such as screen time, rather than just determine movement.

The direct observation protocol Observation System for Recording Physical Activity in Children- Preschool (OSRAC-P) showed promising agreement with a heart rate monitor and pedometer for determining different levels of activity [110]. Whilst the System for Observing Fitness Instruction Time for Preschoolers (SOFIT-P) did not demonstrate strong correlations with the output of the Actigraph (GT3X) [111].

The Fitbit (Flex) showed excellent ability to determine SB and TPA, but not MVPA, when compared with the Actigraph (GT3X+) [112]. For pedometers, activity counts from the Actigraph (CSA/MTI/GT1M) accelerometer were moderately correlated with step counts of the pedometer Yamax Digiwalker SW-200 [115, 116] and the Omron Walking Style Pro Pedometer (HJ-720IT-E2) [118] (values ranging from $r = 0.64$ to 0.92); suggesting that pedometers may be a plausible cheaper alternative to accelerometers in some instances [82].

Table 6 provides a summary table of included studies that examined level 3 validity evidence of measurement tools (detailed information can be found in Additional file 6).

Level 4 validity

A selection of studies included a comparison between two different makes of accelerometer each with unknown validity. These studies demonstrated that the combined heart rate with accelerometer, Actiheart, was shown to be similar in the activity count outcome to the accelerometers, Actical and RT3 ($r = 0.80$ to 0.95) [57]. Similarly, the Actigraph (GT1M) and RT3 showed reasonable similarity in activity count outcome [80] ($r = 0.72$). However, the majority of the studies did not show reasonable convergence, demonstrating that the various types of accelerometer can produce different outcomes [76, 78, 83, 129, 130]. The increased availability of accelerometers and differences in the outcomes of these studies demonstrates the importance of assessing validity of different devices simultaneously, alongside comparison measures [19]. There was weak comparisons between parental and teacher reported habitual physical activity [98, 100].

Table 7 provides a summary table of included studies that examined level 4 validity evidence of measurement tools (detailed information can be found in Additional file 7).

Face and content validity

Only two studies commented on the face and content validity of the proxy reported measurement tools [122, 124]. Face and content validity was determined by focus groups with parents and pre-school staff and consulting experts during questionnaire development [122]; and a pilot study with 21 parents, to determine the comprehension and reproducibility of the measure [124]. The Pre-PAQ included individual response options for both weekend days, due to parents indicating that children's

PA varied more on a weekend than during the week [122]. No further information was reported on the level of face and content validity within these studies; however, no major comprehension concerns were reported. There was minimal information about the procedures in these included studies and so it was not possible to assess the quality or provide firm conclusions on the content validity.

Test-retest reliability

Several of the proxy reported tools showed reasonable test-retest reliability (values ranging from $r = 0.76$ to 0.94) [121, 124, 127, 128] or variable test-retest results, whereby good test-retest for some items on the questionnaire but poor for other items [121–123, 135]. There was very limited evidence showing the test-retest of accelerometry and pedometry, with good test-retest for some activities but poor for others (ICC range from 0.34 to 0.87) [117]. Table 8 provides a summary table of included studies that examined reliability of measurement tools (detailed information can be found in Additional file 8).

Intra-instrument reliability

One accelerometer, the Fitbit (Zip) was the only tool to be examined for intra-instrument reliability in the included studies; showing excellent intra-instrument reliability (ICC = 0.91) when two devices were worn simultaneously on the right hip during a 5 minutes structured walking task in the nursery [91]. Table 8 provides a summary table of included studies that examined reliability of measurement tools (detailed information can be found in Additional file 8).

Feasibility

Whole room calorimetry was shown to be accepted by pre-school aged children as a way of measuring PA and SB [131, 132]. However, this method is expensive, cannot examine free living activity, may only be feasible for smaller scale projects and is highly burdensome on researchers due to the required training and expertise; thereby not being viable for surveillance and the majority of research projects [140]. Similarly, there were promising results for the feasibility of direct observation protocols [111, 133]. However, due to the intensive and demanding nature of direct observation, there are limits on the practicality of this method. Observations usually take place in just one location and for a short period of time, impacting the viability of this method in large samples and to identify habitual activity [140].

Feasibility and acceptability of device based measurement was generally high [73, 78, 115, 130, 134], even when more than one device was worn simultaneously

[115, 130]. Although high acceptability of the ActivPAL accelerometer was reported [130], there was evidence of concerns of irritability of the ActivPAL accelerometer based devices due to these being attached directly to the skin [78, 88]. This may also be a concern when using other devices that attach in a similar way [54, 114]. The proportions of missing and excluded data, for reasons such as device malfunction or children not wearing the device for a sufficient period of time, should be considered when calculating the sample size for studies.

There were no studies that determined the feasibility of proxy reported measurement tools. Table 8 provides a summary table of included studies that examined feasibility of measurement tools (detailed information can be found in Additional file 9).

Generalisability of results

The majority of the studies reported the age ($n = 74$) and sex ($n = 69$) of the included children. However, only ten studies [66, 86, 99, 100, 111, 122, 123, 128, 134] reported other key attributes that help determine the generalisability of the results to the wider population, including: ethnic origin and socioeconomic profile (SEP). Therefore, there is insufficient evidence to suggest that results from the individual studies are generalisable across other populations.

Ethnicity

Eighteen studies described the ethnicity of the sample in which the measurement property of the tool was examined. The majority of the studies that reported on ethnicity had samples with children whom were primarily white or Caucasian [57, 60, 66, 77, 85, 93, 94, 100, 122, 123, 134]. Three of the studies reported primarily Hispanic populations samples [99, 111, 128], followed by primarily African American samples [58, 63]. There was no indication of lower measurement properties of the tools in any specific ethnic group in these studies, however, this was not examined directly. Details of these studies are outlined in Additional files 4, 5, 6, 7, 8 and 9.

Socioeconomic profile

Fourteen studies reported on the socioeconomic profile (SEP) of their sample. Some studies reported that at least some of the participants were recruited from pre-schools characterised to have individuals with lower SEP, such as Head Start Centres, which require proof of income to demonstrate that families are at or below poverty level [66, 86, 99, 111, 125, 128, 135]. The remaining studies that reported on SEP were based on individual level demographics. These studies reported that the samples were primarily made up of individuals of higher SEP [100, 122, 123, 134]. Only one study reported that the

children in their sample were from lower to lower-middle SEP families [126]. Whilst one study also reported an equal amount of participants from both high and low SEP families [115]. None of the studies directly examined whether SEP affected validity or reliability, or whether there was reduced feasibility of the tools in different SEP groups. There was no evidence to suggest that SEP was affecting the validity or reliability of the measurement tools being evaluated. The majority of the studies found no indication of reduced feasibility amongst the different SEP, however, two studies conducted with participants of lower SEP reported a lack of feasibility when using pedometer-based measurement tools [99, 126]. Details of these studies are outlined in Additional files 4, 5, 6, 7, 8 and 9.

Discussion

This systematic review identified 69 articles, describing 75 studies that were examining the measurement properties of measurement tools used to assess PA and/or SB in pre-school aged children. In this review, we provide an overview on what measurement tools have been examined for what outcome measures, with an indication on whether these have been shown to be valid, reliable or feasible.

The heterogeneity of the studies included in this review emphasises the complexity of measurement of PA and SB behaviours, identified previously by others [141, 142], alongside the additional challenges associated with measurement in a pre-school aged population [20, 143]. We show that different measurement tools often examine different dimensions of PA and SB (e.g. time spent in different intensities of activity, posture allocation, step count, energy expenditure) in line with previous literature [32, 144]. Measurement tools may all have a specific use depending on their derived purpose, desired measurement outcomes and the context in which the tool is being used [145]. However, when selecting an appropriate and useful measurement tool to assess PA and SB amongst children, there is often a trade-off between the three main utilities (validity, reliability, feasibility), alongside further considerations, such as the sample size of the study, budget, and availability of resources [144]. As such, it is not a question of which measurement tool is 'best' for assessing the PA and SB of pre-school aged children, but rather, what measure or combination of measures, are most appropriate in the given context for the desired outcome measures [55, 141, 144, 146, 147]. In line with this, the measurement properties of the tools are only reflective of the context in which they have been tested and cannot be generalised to other contexts, for example, if a tool showed good validity, reliability or feasibility but was examined in a laboratory based setting for use over a short period of time then

the tool can only be said to be valid, reliable or feasible in this context, and the assessment does not apply to free living longer term measurement. We will, however, make recommendations based on our findings that we believe would be of interest to those involved in research based on the frequently used PA and SB outcomes.

Overall, based on the current evidence, which included a limited number of studies of varying quality, of the measurement properties of measurement tools used to examine PA and SB of pre-school aged children, multiple accelerometers, including the Actigraph (in particular GT3X versions), Actical, ActivPAL and Fitbits (Flex and Zip), can provide valid measures, with some evidence of feasibility, of movement-related behaviours that would be of interest in a range of research where resources and capacity allow. However, disadvantages include the need for expertise in the analysis of data, device malfunctions and continuous technological advances and development of new and improved activity monitoring devices, which increases difficulty for standardisation within and between studies [148, 149]. Alongside this, there are also differences on subjective decisions when using accelerometry, including: epochs, cut points, placement, wear time, non-wear time and valid number of days, which determines whether the data is valid or not [19, 20, 138, 150, 151]. Proxy report based measurement tools, in particular the PRE-PAQ and *leisure time report*, also show some promise, for some dimensions of PA and SB. Although, the evidence quality is weak, therefore, much more evaluation of the measurement properties of these types of tools is needed. These measurement tools have advantage in terms of identifying contextual behaviour, cost, accessibility and can be consistently used in a standardised way, as they are not reliant on upgrades in technology [152]. Additionally, whole room calorimetry and direct observation were shown to be accepted and feasible methods for measuring PA and SB of pre-school aged children. However, this was based on a very small number of studies and small sample sizes. As these methods are expensive, highly burdensome and intensive, require extensive training, and can only capture a short period of time, they may only be feasible for smaller scale projects and for only some dimensions of PA and SB.

Face and content validity have been suggested to be a crucial first step in determining the appropriateness of a measure, to determine if the measurement tool is assessing what it intends to measure and to identify whether the measure is understandable to the target population [34, 153, 154]. However, only two of the included studies reported on an element of face and content validity of their measures [122, 124].

A limited number of monitors were examined for reliability; the Fitbit Zip showed excellent intra-instrument

reliability, however, this was based on one study only [91]. More research studies examining the intra-instrument reliability of measurement tools in pre-school aged children are warranted. Similarly, very few studies explicitly examined the feasibility, including acceptability, of the measurement tools [73, 78, 88, 99, 111, 114, 115, 128, 130–134], although, a proportion of the studies did report exclusion of participants and missing data. The measurement properties of measurement tools may be comprised if they are not feasible in the context and with the population in which they are to be used. The ability to identify reasons for missing data, reasons for non-compliance and overall acceptability and feasibility of the measurement tools, to determine end user experience, will help with understanding which tools are most applicable for use [36].

The review revealed a discrepancy in the amount of studies examining each measurement tool; with an emphasis placed on examining the measurement properties of device based measurement tools, primarily accelerometers, making up 70% of all studies, with multiple studies examining the same type of accelerometer. Accelerometers were also frequently used as a comparison measurement tool. A major limitation of the included studies utilising accelerometers was the considerable variation in the interpretation of data, due to differences in the subjective decisions on epochs, cut points, placement, wear time, non-wear time and valid number of days [19, 20, 138, 150, 151]. At least sixteen different published cut points were used in the included studies [57, 58, 63, 69–71, 84, 85, 101, 103–109], with some studies applying their own. Such methodological inconsistencies can be problematic, as applying different cut points to the same data can result in statistically and biologically significant differences in the outcomes of PA [102, 155, 156]. Although we provide reference to the published cut points used for each study in the summary tables, ultimately, the overall aims of the studies included in this review were not to show which data interpretation choices were most valid, and so we cannot determine which data interpretation choices are best to use. Therefore, it is important to note that the devices shown to be valid can only be said to be valid with the data interpretation choices in which they were tested. If researchers decide that a device based measurement tool is best for their research study, it is important to ensure use of the respective epochs, cut points and monitor placement outlined in the validation studies. We provide information on the interpretation choices used in each study in the tables outlined in Additional files 4, 5, 6 and 7.

Additionally, tools in the included studies were often evaluated over a short space of time, usually around 1 h, due to use of an intensive reference method (such as whole room calorimetry or direct observation). Consequently, wear time, non-wear time and information on

the amount of valid number of days required for longer term data collection were not required. However, for examining PA, a full 7 days of measurement is preferred where possible [157, 158], with a minimum of 3 days required for reasonable estimates [159–161]. For examining SB, research suggests that ≥ 4 days of monitoring is required for reasonable estimates [162]. In addition to this, it is suggested that data include at least one week-end day [163]. Similarly, there are differences in the recommended number of hours wear time of the device, ranging from 6 to 10 hour wear time per day required for the most reliable estimates of PA [158–160, 164]. Collecting data using a 7 day wear protocol is recommended to increase the chances of there being enough data. Studies should ensure that they include the minimum amount of data, including both number of days and hours per day, to meet a reliability score of at least 0.7, and should report this in their study [163, 165]. Similarly, there is no consensus on defining non-wear time of accelerometry data [20, 138]. Non-wear time is often determined by consecutive zero counts in the data sets, however, this varies between studies and becomes more challenging when also examining SB using accelerometry [138]. A common and useful way in which non-wear time can be determined is by completion of a log to state when the accelerometer was removed and reasons for this; which can then be cross-validated with numbers of consecutive zeros within the data sets [20, 166].

The majority of studies included in this review were conducted in a free living (habitual) context, including at the pre-school and/or at home [32]. However, the results of the studies were often lacking in potential for generalisability and translation. As is the case for much research, identified studies have traditionally been conducted in high income countries. There was no evidence to suggest differences in the measurement properties of the measurement tools across different ethnicities or SEP, however a very small proportion of the studies explored these factors, and none directly. Similarly, there was minimal evidence of the measurement properties of the tools being evaluated with large sample sizes, so the capability of the tools at scale are unknown.

A major limitation in the field of PA and SB measurement is the lack of criterion methods for the majority outcomes, and indeed a lack of consensus on what might be considered a criterion. We used the level of evidence scheme to distinguish between differing levels of validity of the comparison measurement tools, but caution should always be applied when validity is measured against non-criterion methods, especially when the validity of these methods themselves are not well established.

Strengths and weaknesses of the review

A primary strength of this review is that we assessed international literature for evidence on this topic, which included extensive searching of multiple platforms; published academic research articles through database searching, searches of the grey literature, and manual searches. Another strength of this review is that we examined various measurement properties of measurement tools; including the feasibility of the measures.

A potential weakness of this review was the risk of bias assessment, due to the tool that was used. Although we believe that we used the best available tool, it was not ideal because it was not devised specifically for studies examining the measurement properties of measurement tools. However, we did use an additional risk of bias assessment on proxy reported measurement tools due to this being available for these types of studies. Additionally, although two reviewers screened a 10% sample of the title and abstracts with very high agreement, only one reviewer screened the remaining studies and so it is plausible that studies were missed through human error.

A limitation of the research is that we did not explore the measurement property of ‘responsiveness’ in our search terms. Responsiveness, defined as ‘*The ability of a PROM to detect change over time in the construct to be measured*’ [145], is a measurement property that is not well established in the field of PA and SB measurement. This is apparent from the literature [28, 29] and when conducting initial scoping searches for this review, no studies exploring responsiveness were identified. No studies were identified, also, through our additional hand searches. We are, therefore, currently unable to determine whether any proxy reported tools are able to detect changes in behaviour over time (e.g. in response to an intervention) and it is essential that this is the focus of future research.

Implications for research

The lack of multiple studies examining the same measurement tool makes it difficult to produce firm conclusions on the measurement properties of some tools [28]. Differences in the number of studies examining each measurement tool, the ways in which the studies were conducted, the differences in comparison measurement tools used, the included samples, and reported outcome measures can make direct comparisons between the results of the studies difficult.

Future research should focus on the measurement properties of measurement tools being examined in different populations to ensure external validity of the measures and to extend the generalisability of the findings from these types of studies. This should include using large sample sizes, individuals with varying SEP and with different ethnicities, and conducted in lower income countries; this would help to evaluate the measurement

properties, including feasibility and acceptability, of the measures in different contexts [16]. However, only after a measurement tool is shown to be valid, reliable, and feasible in the population for which it was originally developed [145].

In addition to this, research projects examining the PA or SB of pre-school aged children should look to identify the context in which the measurement properties of tools have been examined prior to choosing which tool may be most appropriate for their purpose. If the measurement properties of the selected tool are unknown in the context in which they are to be used, researchers should aim to conduct a validation check even if just on a sub-sample of children included in the study. This would be useful to identify whether the measurement tools are working as intended and to ensure greater confidence in the results of the study.

Our review highlights the importance of reporting on internal validity, such as missing data and non-completion within studies. Studies that reported such data revealed various important implications of using measurement tools, including that malfunctions of measurement tools, primarily with device based measurement tools and calorimetry, can have sufficient impact on the included final sample [76, 118, 122]. We highlight the need to examine and report missing data, as these considerations can substantially impact on the utility of a tool [56].

Further qualitative work in this area is needed, with two main purposes: 1) to determine the feasibility and acceptability of measurement tools and 2) to examine face and content validity. In some instances in this review, an indication of feasibility of the tools was provided based on numerical scores due to this being the only available feasibility data within the studies. However, we wish to highlight that the true feasibility of a measure cannot be expressed in numbers only, and much more qualitative work is needed in this area. This has been recognised in previous research, highlighting the importance of conducting qualitative research to examine the feasibility of measurement tools in the target population prior to use [54, 55, 139]; with such work being important to understand recruitment bias and reasons for missing data or non-completion [36, 56]. Additionally, qualitative work to determine validity is rarely conducted, demonstrated by only two studies included in the current review that commented on these aspects [122, 124]. However, face and content validity conducted using qualitative methods have been highlighted as a crucial first step in examining the validity of measurement tools for PA and SB, to determine whether tools are valid for their intended purpose [28, 32].

Future research should focus on further development and evaluation of proxy report methods together with

the target population and ensure representativeness and feasibility of the measurement tool in the context in which it is intended to be used.

Conclusions

The measurement tools used to measure PA and SB in pre-school aged children show mixed measurement properties, and were generally based on minimal studies providing variable quality of evidence. There is a clear need for further and more in-depth evaluation work. Based on currently available evidence, we conclude that the Actigraph (in particular GT3X versions), Actical and ActivPAL, have the greatest measurement properties for assessing common movement related outcomes (e.g. SB, MVPA, TPA) for free living activity of pre-school aged children, and should be the tool of choice where resources allow and where logistically possible. The Fitbit (Flex and Zip) also shows very promising results; however, these were based on a very limited sample of studies. Where measurement of a large sample is required and where budgets are limited, proxy measures can provide some valid data, alongside useful contextual information not captured by device-based measurement tools. A combination of accelerometers and proxy reported measurement tools (based on parent or carer reports) may be most useful for a range of PA and SB outcome measures.

Abbreviations

AUC-ROC: Area under the receiver operating curve; COSMIN: Consensus-based Standards for the selection of health status Measurement Instruments; DLW: Doubly labelled water; EPHPP: Effective Public Health Practice Project; LPA: Light physical activity; MVPA: Moderate to vigorous physical activity; PA: Physical activity; PRE-PAQ: Pre School Physical Activity Questionnaire; PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses; PROSPERO: Prospective Register for Systematic Reviews; SB: Sedentary behaviour; SEP: Socioeconomic profile; TPA: Total physical activity; UK: United Kingdom; USA: United States of America

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12966-021-01132-9>.

Additional file 1. PRISMA Checklist.

Additional file 2. Search strategy and outcomes.

Additional file 3. Excluded studies with reasons.

Additional file 4. Study details of level 1 validity evidence.

Additional file 5. Study details of level 2 validity evidence.

Additional file 6. Study details of level 3 validity evidence.

Additional file 7. Study details of level 4 validity evidence.

Additional file 8. Study details of reliability evidence.

Additional file 9. Study details of feasibility evidence.

Additional file 10. Source of funding for each study.

Additional file 11. Downs and Black risk of bias assessment.

Additional file 12. COSMIN risk of bias assessment.

Acknowledgements

The work described in this publication was supported and funded by the National Institute for Health Research School for Public Health Research (NIHR SPHR). We thank James Bissett and Benjamin Taylorson at Durham University for their advice in developing the search strategies. We would also like to thank the anonymous reviewers for their useful feedback and suggested changes which we feel improved this paper.

Authors' contributions

SMP was involved in the conception, design, data screening, data extraction, risk of bias, data synthesis, interpretation and write up of the manuscript. FCHB was involved in the conception, design, data screening, data extraction, risk of bias, interpretation and write up of the manuscript. CS was involved in the conception, design, interpretation and write up of the manuscript. MH was involved in the data screening, interpretation and write up of the manuscript. KRH and SS were involved in the interpretation and write up of the manuscript. CM was involved in the data extraction and review of the manuscript. SMP initially drafted the article and all authors contributed to subsequent drafts and approved the final manuscript. All authors have approved the submitted version and have agreed to be personally accountable for their own contributions and to ensure that questions related to the accuracy or integrity of any part of the work, are appropriately investigated, resolved, and the resolution documented in the literature.

Funding

This study was funded and supported by the National Institute for Health Research (NIHR) School for Public Health Research (SPHR), Grant Reference Number PD-SPH-2015. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. The NIHR School for Public Health Research is a partnership between the Universities of Sheffield; Bristol; Cambridge; Imperial; University College London; The London School for Hygiene and Tropical Medicine (LSHTM); LiLaC—a collaboration between the Universities of Liverpool and Lancaster; and Fuse - The Centre for Translational Research in Public Health a collaboration between Newcastle, Durham, Northumbria, Sunderland and Teesside Universities. The funders had no role in study design, data analysis and interpretation, or preparation of the manuscript. KRH is funded by the Wellcome Trust (107337/Z/15/Z).

Availability of data and materials

All data generated or analysed during this study are included in this published article and its supplementary information files.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Department of Sport and Exercise Sciences, Durham University, Durham City, UK. ²The Centre for Translational Research in Public Health (Fuse), Newcastle upon Tyne, UK. ³School of Health Sciences, University of Canterbury, Christchurch, New Zealand. ⁴Population Policy & Practice Research and Teaching Department, UCL Great Ormond Street Institute of Child Health, London, UK. ⁵Faculty of Medicine, School of Public Health, Imperial College London, London, UK. ⁶Population Health Sciences Institute, Newcastle University, Newcastle upon Tyne, UK. ⁷Human Nutrition Research Centre, Newcastle University, Newcastle upon Tyne, UK. ⁸Newcastle University Centre of Research Excellence in Healthier Lives Newcastle University, Newcastle upon Tyne, UK.

Received: 9 September 2020 Accepted: 1 May 2021

Published online: 04 November 2021

References

- Janssen I, LeBlanc AG. Systematic review of the health benefits of physical activity and fitness in school-aged children and youth. *Int J Behav Nutr Phys Act*. 2010. <https://doi.org/10.1186/1479-5868-7-40>.
- Poitras VJ, Gray CE, Borghese MM, Carson V, Chaput JP, Janssen I, et al. Systematic review of the relationships between objectively measured physical activity and health indicators in school-aged children and youth. *Appl Physiol Nutr Metab*. 2016. <https://doi.org/10.1139/apnm-2015-0663>.
- Carson V, Hunter S, Kuzik N, Gray CE, Poitras VJ, Chaput JP, et al. Systematic review of sedentary behaviour and health indicators in school-aged children and youth: an update. *Applied Physiology, Nutrition, and Metabolism*. 2016. <https://doi.org/10.1139/apnm-2015-0630>.
- Carson V, Hunter S, Kuzik N, Wiebe SA, Spence JC, Friedman A, et al. Systematic review of physical activity and cognitive development in early childhood. *J Sci Med Sport*. 2016. <https://doi.org/10.1016/j.jsams.2015.07.011>.
- Carson V, Lee EY, Hewitt L, Jennings C, Hunter S, Kuzik N, et al. Systematic review of the relationships between physical activity and health indicators in the early years (0–4 years). *BMC Public Health*. 2017. <https://doi.org/10.1186/s12889-017-4860-0>.
- LeBlanc AG, Spence JC, Carson V, Connor Gorber S, Dillman C, Janssen I, et al. Systematic review of sedentary behaviour and health indicators in the early years (aged 0–4 years). *Appl Physiol Nutr Metab*. 2012. <https://doi.org/10.1139/h2012-063>.
- Poitras VJ, Gray CE, Janssen X, Aubert S, Carson V, Faulkner G, et al. Systematic review of the relationships between sedentary behaviour and health indicators in the early years (0–4 years). *BMC Public Health*. 2017. <https://doi.org/10.1186/s12889-017-4849-8>.
- World Bank Data. Primary School Starting age. 2020. <https://data.worldbank.org/indicator/se.prm.ages?view=map>. Accessed 12 Nov 2020.
- World Health Organization. Guidelines on physical activity, sedentary behaviour and sleep for children under 5 years of age. 2019. <https://www.who.int/publications/i/item/9789241550536>. Accessed 12 Nov 2020.
- Chaput JP, Colley RC, Aubert S, Carson V, Janssen I, Roberts KC, et al. Proportion of preschool-aged children meeting the Canadian 24-hour movement guidelines and associations with adiposity: results from the Canadian health measures survey. *BMC Public Health*. 2017;17(5):147–54. <https://doi.org/10.1186/s12889-017-4854-y>.
- Cliff DP, McNeill J, Vella SA, Howard SJ, Santos R, Batterham M, et al. Adherence to 24-hour movement guidelines for the early years and associations with social-cognitive development among Australian preschool children. *BMC Public Health*. 2017;17(5):207–15. <https://doi.org/10.1186/s12889-017-4858-7>.
- Hesketh KR, McMinn AM, Ekelund U, Sharp SJ, Collings PJ, Harvey NC, et al. Objectively measured physical activity in four-year-old British children: a cross-sectional analysis of activity patterns segmented across the day. *Int J Behav Nutr Phys Act*. 2014;11(1):1–9. <https://doi.org/10.1186/1479-5868-11-1>.
- Downing KL, Hinkley T, Salmon J, Hnatiuk JA, Hesketh KD. Do the correlates of screen time and sedentary time differ in preschool children? *BMC Public Health*. 2017;17(1):1–2. <https://doi.org/10.1186/s12889-017-4195-x>.
- Pereira JR, Cliff DP, Sousa-Sá E, Zhang Z, Santos R. Prevalence of objectively measured sedentary behavior in early years: systematic review and meta-analysis. *Scand J Med Sci Sports*. 2019;29(3):308–28. <https://doi.org/10.1111/sms.13339>.
- Bauman A, Phongsavan P, Schoeppe S, Owen N. Physical activity measurement—a primer for health promotion. *Promot Educ*. 2006. <https://doi.org/10.1177/10253823060130020103>.
- Kohl HW III, Fulton JE, Caspersen CJ. Assessment of physical activity among children and adolescents: a review and synthesis. *Prev Med*. 2000. <https://doi.org/10.1006/pmed.1999.0542>.
- McClain JJ, Tudor-Locke C. Objective monitoring of physical activity in children: considerations for instrument selection. *J Sci Med Sport*. 2009. <https://doi.org/10.1016/j.jsams.2008.09.012>.
- Ruiz RM, Sommer EC, Tracy D, Banda JA, Economos CD, JaKa MM, et al. Novel patterns of physical activity in a large sample of preschool-aged children. *BMC Public Health*. 2018. <https://doi.org/10.1186/s12889-018-5135-0>.
- Oliver M, Schofield GM, Kolt GS. Physical activity in preschoolers. *Sports Med*. 2007. <https://doi.org/10.2165/00007256-200737120-00004>.

20. Cliff DP, Reilly JJ, Okely AD. Methodological considerations in using accelerometers to assess habitual physical activity in children aged 0–5 years. *J Sci Med Sport*. 2009. <https://doi.org/10.1016/j.jsams.2008.10.008>.
21. Adamo KB, Prince SA, Tricco AC, Connor-Gorber SA, Tremblay M. A comparison of indirect versus direct measures for assessing physical activity in the pediatric population: a systematic review. *Int J Pediatr Obes*. 2009. <https://doi.org/10.1080/17477160802315010>.
22. Loprinzi PD, Cardinal BJ. Measuring children's physical activity and sedentary behaviors. *J Exerc Sci Fit*. 2011. [https://doi.org/10.1016/S1728-869X\(11\)60002-6](https://doi.org/10.1016/S1728-869X(11)60002-6).
23. Lubans DR, Hesketh K, Cliff DP, Barnett LM, Salmon J, Dollman J, et al. A systematic review of the validity and reliability of sedentary behaviour measures used with children and adolescents. *Obes Rev*. 2011. <https://doi.org/10.1111/j.1467-789X.2011.00896.x>.
24. Pate RR, O'Neill JR, Mitchell J. Measurement of physical activity in preschool children. *Med Sci Sports Exerc*. 2010;42(3):508–12. <https://doi.org/10.1249/mss.0b013e3181cea116>.
25. Sirard JR, Pate RR. Physical activity assessment in children and adolescents. *Sports Med*. 2001. <https://doi.org/10.2165/00007256-200131060-00004>.
26. Chinapaw MJ, Mokkink LB, van Poppel MN, van Mechelen W, Terwee CB. Physical activity questionnaires for youth. *Sports Med*. 2010. <https://doi.org/10.2165/11530770-000000000-00000>.
27. Helmerhorst HH, Brage S, Warren J, Besson H, Ekelund U. A systematic review of reliability and objective criterion-related validity of physical activity questionnaires. *Int J Behav Nutr Phys Act*. 2012. <https://doi.org/10.1186/1479-5868-9-103>.
28. Hidding LM, Altenburg TM, Mokkink LB, Terwee CB, Chinapaw MJ. Systematic review of childhood sedentary behavior questionnaires: what do we know and what is next? *Sports Med*. 2017;47(4):677–99. <https://doi.org/10.1007/s40279-016-0610-1>.
29. Hidding LM, Chinapaw MJ, van Poppel MN, Mokkink LB, Altenburg TM. An updated systematic review of childhood physical activity questionnaires. *Sports Med*. 2018. <https://doi.org/10.1007/s40279-018-0987-0>.
30. Vale S, Silva P, Santos R, Soares-Miranda L, Mota J. Compliance with physical activity guidelines in preschool children. *J Sports Sci*. 2010. <https://doi.org/10.1080/02640411003702694>.
31. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Int J Surg*. 2009;8(5):336–41.
32. Kelly P, Fitzsimons C, Baker G. Should we reframe how we think about physical activity and sedentary behaviour measurement? Validity and reliability reconsidered. *Int J Behav Nutr Phys Act*. 2016. <https://doi.org/10.1186/s12966-016-0351-4>.
33. Bowen DJ, Kreuter M, Spring B, Cofta-Woerpel L, Linnan L, Weiner D, et al. How we design feasibility studies. *Am J Prev Med*. 2009. <https://doi.org/10.1016/j.amepre.2009.02.002>.
34. Terwee CB, Prinsen CA, Chiarotto A, Westerman MJ, Patrick DL, Alonso J, et al. COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. *Qual Life Res*. 2018. <https://doi.org/10.1007/s11136-018-1829-0>.
35. Forouhi N et al. Diet, anthropometry and physical activity (DAPA) measurement toolkit. 2017. <https://dapa-toolkit.mrc.ac.uk/>. Accessed 6 Jan 2020.
36. Evenson KR, Goto MM, Furberg RD. Systematic review of the validity and reliability of consumer-wearable activity trackers. *Int J Behav Nutr Phys Act*. 2015. <https://doi.org/10.1186/s12966-015-0314-1>.
37. McHugh ML. Interrater reliability: the kappa statistic. *Biochemia Med*. 2012; 22(3):276–82.
38. Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *J Epidemiol Community Health*. 1998. <https://doi.org/10.1136/jech.52.6.377>.
39. Prince SA, Adamo KB, Hamel ME, Hardt J, Gorber SC, Tremblay M. A comparison of direct versus self-report measures for assessing physical activity in adults: a systematic review. *Int J Behav Nutr Phys Act*. 2008. <https://doi.org/10.1186/1479-5868-5-56>.
40. Jutai JW, Strong JG, Russell-Minda E. Effectiveness of assistive technologies for low vision rehabilitation: a systematic review. *J Vis Impairment Blindness*. 2009;103(4):210–22. <https://doi.org/10.1177/0145482X0910300404>.
41. Korakakis V, Whiteley R, Tzavara A, Malliaropoulos N. The effectiveness of extracorporeal shockwave therapy in common lower limb conditions: a systematic review including quantification of patient-rated pain reduction. *Br J Sports Med*. 2018;52(6):387–407. <https://doi.org/10.1136/bjsports-2016-097347>.
42. Maidment DW, Barker AB, Xia J, Ferguson MA. A systematic review and meta-analysis assessing the effectiveness of alternative listening devices to conventional hearing aids in adults with hearing loss. *Int J Audiol*. 2018; 57(10):721–9. <https://doi.org/10.1080/14992027.2018.1493546>.
43. Mokkink LB, De Vet HC, Prinsen CA, Patrick DL, Alonso J, Bouter LM, et al. COSMIN risk of bias checklist for systematic reviews of patient-reported outcome measures. *Qual Life Res*. 2018;27(5):1171–9. <https://doi.org/10.1007/s11136-017-1765-4>.
44. Prinsen CA, Mokkink LB, Bouter LM, Alonso J, Patrick DL, De Vet HC, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. *Qual Life Res*. 2018;27(5):1147–57. <https://doi.org/10.1007/s11136-018-1798-3>.
45. Terwee CB, Mokkink LB, Knol DL, Ostelo RW, Bouter LM, de Vet HC. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Qual Life Res*. 2012 May;21(4):651–7. <https://doi.org/10.1007/s11136-011-9960-1>.
46. De Vet HC, Terwee CB, Mokkink LB, Knol DL. Measurement in medicine: a practical guide. United Kingdom: Cambridge University Press; 2011. <https://doi.org/10.1017/CBO9780511996214>.
47. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016. <https://doi.org/10.1016/j.jcm.2016.02.012>.
48. Schmidt ME, Steindorf K. Statistical methods for the validation of questionnaires. *Methods Inf Med*. 2006;45(04):409–13. <https://doi.org/10.1055/s-0038-1634096>.
49. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol*. 2007. <https://doi.org/10.1016/j.jclinepi.2006.03.012>.
50. Twomey PJ, Kroll MH. How to use linear regression and correlation in quantitative method comparison studies. *Int J Clin Pract*. 2008. <https://doi.org/10.1111/j.1742-1241.2008.01709.x>.
51. Lohr KN. Assessing health status and quality-of-life instruments: attributes and review criteria. *Qual Life Res*. 2002. <https://doi.org/10.1023/A:1015291021312>.
52. van Poppel MN, Chinapaw MJ, Mokkink LB, Van Mechelen W, Terwee CB. Physical activity questionnaires for adults. *Sports Med*. 2010. <https://doi.org/10.2165/11531930-000000000-00000>.
53. Effective Public Health Practice Project. Quality assessment tool for quantitative studies. 2009. https://merst.ca/wp-content/uploads/2018/02/quality-assessment-tool_2010.pdf. Accessed 24 July 2020.
54. Costa S, Barber SE, Griffiths PL, Cameron N, Clemes SA. Qualitative feasibility of using three accelerometers with 2–3-year-old children and both parents. *Res Q Exerc Sport*. 2013. <https://doi.org/10.1080/02701367.2013.812002>.
55. Kowalski K, Rhodes R, Naylor PJ, Tuokko H, MacDonald S. Direct and indirect measurement of physical activity in older adults: a systematic review of the literature. *Int J Behav Nutr Phys Act*. 2012. <https://doi.org/10.1186/1479-5868-9-148>.
56. Dössegger A, Ruch N, Jimmy G, Braun-Fahrlander C, Mäder U, Hänggi J, et al. Reactivity to accelerometer measurement of children and adolescents. *Med Sci Sports Exerc*. 2014;46(6):1140–6. <https://doi.org/10.1249/MSS.0000000000000215>.
57. Adolph AL, Puyau MR, Vohra FA, Nicklas TA, Zakeri IF, Butte NF. Validation of uniaxial and triaxial accelerometers for the assessment of physical activity in preschool children. *J Phys Act Health*. 2012. <https://doi.org/10.1123/jpah.9.7.944>.
58. Pate RR, Almeida MJ, McIver KL, Pfeiffer KA, Dowda M. Validation and calibration of an accelerometer in preschool children. *Obesity*. 2006. <https://doi.org/10.1038/oby.2006.234>.
59. Reilly JJ, Kelly LA, Montgomery C, Jackson DM, Slater C, Grant S, et al. Validation of Actigraph accelerometer estimates of total energy expenditure in young children. *Int J Pediatr Obes*. 2006;1(3):161–7. <https://doi.org/10.1080/17477160600845051>.
60. Steenbeck B, Wright MN, Wirsik N, Brandes M. Accelerometry-based prediction of energy expenditure in preschoolers. *J Meas Phys Behav*. 2019. <https://doi.org/10.1123/jmpb.2018-0032>.
61. Janssen X, Cliff DP, Reilly JJ, Hinkley T, Jones RA, Batterham M, et al. Predictive validity and classification accuracy of ActiGraph energy

- expenditure equations and cut-points in young children. *PLoS One*. 2013a. <https://doi.org/10.1371/journal.pone.0079124>.
62. Janssen X, Cliff D, Reilly J, Hinkley T, Jones R, Batterham M, et al. Evaluation of Actical equations and thresholds to predict physical activity intensity in young children. *J Sports Sci*. 2015. <https://doi.org/10.1080/02640414.2014.949826>.
63. Pfeiffer KA, Mciver KL, Dowda M, Almeida MJ, Pate RR. Validation and calibration of the Actical accelerometer in preschool children. *Med Sci Sports Exerc*. 2006;38(1):152–7. <https://doi.org/10.1249/01.mss.0000183219.44127.e7>.
64. Nyström CD, Pomeroy J, Henriksson P, Forsum E, Ortega FB, Maddison R, et al. Evaluation of the wrist-worn ActiGraph wGT3x-BT for estimating activity energy expenditure in preschool children. *Eur J Clin Nutr*. 2017. <https://doi.org/10.1038/ejcn.2017.114>.
65. Janssen X, Cliff DP, Reilly JJ, Hinkley T, Jones RA, Batterham M, et al. Validation and calibration of the activPAL™ for estimating METs and physical activity in 4–6 year olds. *J Sci Med Sport*. 2014. <https://doi.org/10.1016/j.jsams.2013.10.252>.
66. Lopez-Alarcon M, Merrifield J, Fields DA, Hilario-Hailey T, Franklin FA, Shewchuk RM, et al. Ability of the activPAL accelerometer to predict free-living energy expenditure in young children. *Obes Res*. 2004. <https://doi.org/10.1038/oby.2004.231>.
67. Sijtsma A, Schierbeek H, Goris AH, Joosten KF, van Kessel I, Corpeleijn E, et al. Validation of the TracmorD triaxial accelerometer to assess physical activity in preschool children. *Obesity*. 2013. <https://doi.org/10.1002/oby.20401>.
68. Corder K, van Sluijs EM, Wright A, Whincup P, Wareham NJ, Ekelund U. Is it possible to assess free-living physical activity and energy expenditure in young people by self-report? *Am J Clin Nutr*. 2009. <https://doi.org/10.3945/ajcn.2008.26739>.
69. Ekelund UL, Sjöström M, Yngve A, Poortvliet E, Nilsson A, Froberg KA, et al. Physical activity assessed by activity monitor and doubly labeled water in children. *Med Sci Sports Exerc*. 2001;33(2):275–81. <https://doi.org/10.1097/00005768-200102000-00017>.
70. Puyau MR, Adolph AL, Vohra FA, Butte NF. Validation and calibration of physical activity monitors in children. *Obes Res*. 2002. <https://doi.org/10.1038/oby.2002.24>.
71. Evenson KR, Catellier DJ, Gill K, Ondrak KS, McMurray RG. Calibration of two objective measures of physical activity for children. *J Sports Sci*. 2008. <https://doi.org/10.1080/02640410802334196>.
72. Klesges RC, Woolfrey J, Vollmer J. An evaluation of the reliability of time sampling versus continuous observation data collection. *J Behav Ther Exp Psychiatry*. 1985. [https://doi.org/10.1016/0005-7916\(85\)90004-7](https://doi.org/10.1016/0005-7916(85)90004-7).
73. Bar-Or T, Bar-Or O, Waters H, Hirji A, Russell S. Validity and social acceptability of the Polar Vantage XL for measuring heart rate in preschoolers. *Pediatr Exerc Sci*. 1996. <https://doi.org/10.1123/pes.8.2.115>.
74. Hislop J, Palmer N, Anand P, Aldin T. Validity of wrist worn accelerometers and comparability between hip and wrist placement sites in estimating physical activity behaviour in preschool children. *Physiol Meas*. 2016;37(10):1701, 1714. <https://doi.org/10.1088/0967-3334/37/10/1701>.
75. Dobell AP, Eyre EL, Tallis J, Chinapaw MJ, Altenburg TM, Duncan MJ. Examining accelerometer validity for estimating physical activity in preschoolers during free-living activity. *Scand J Med Sci Sports*. 2019. <https://doi.org/10.1111/sms.13496>.
76. Alhassan S, Sirard JR, Kurdziel LB, Merrigan S, Greever C, Spencer RM. Cross-validation of two accelerometers for assessment of physical activity and sedentary time in preschool children. *Pediatr Exerc Sci*. 2017. <https://doi.org/10.1123/pes.2016-0074>.
77. Kahan D, Nicaise V, Reuben K. Convergent validity of four accelerometer cutpoints with direct observation of preschool children's outdoor physical activity. *Res Q Exerc Sport*. 2013. <https://doi.org/10.1080/02701367.2013.762294>.
78. De Decker E, De Craemer M, Santos-Lozano A, Van Cauwenbergh E, De Bourdeaudhuij I, Cardon G. Validity of the ActivPAL™ and the ActiGraph monitors in preschoolers. *Med Sci Sports Exerc*. 2013;45(10):2002–11. <https://doi.org/10.1249/mss.0b013e318292c575>.
79. Hislop JF, Bulley C, Mercer TH, Reilly JJ. Comparison of accelerometer cut points for physical activity and sedentary behavior in preschool children: a validation study. *Pediatr Exerc Sci*. 2012a. <https://doi.org/10.1123/pes.24.4.563>.
80. Hislop JF, Bulley C, Mercer TH, Reilly JJ. Comparison of epoch and uniaxial versus triaxial accelerometers in the measurement of physical activity in preschool children: a validation study. *Pediatr Exerc Sci*. 2012b. <https://doi.org/10.1123/pes.24.3.450>.
81. Fairweather SC, Reilly JJ, Grant S, Whittaker A, Paton JY. Using the Computer Science and Applications (CSA) activity monitor in preschool children. *Pediatr Exerc Sci*. 1999. <https://doi.org/10.1123/pes.11.4.413>.
82. Hands B, Larkin D. Physical activity measurement methods for young children: a comparative study. *Meas Phys Educ Exerc Sci*. 2006;10(3):203–14. https://doi.org/10.1207/s15327841mpee1003_5.
83. Kelly LA, Reilly JJ, Fairweather SC, Barrie S, Grant S, Paton JY. Comparison of two accelerometers for assessment of physical activity in preschool children. *Pediatr Exerc Sci*. 2004. <https://doi.org/10.1123/pes.16.4.324>.
84. Reilly JJ, Coyle J, Kelly L, Burke G, Grant S, Paton JY. An objective method for measurement of sedentary behavior in 3-to 4-year olds. *Obes Res*. 2003. <https://doi.org/10.1038/oby.2003.158>.
85. Sirard JR, Trost SG, Pfeiffer KA, Dowda M, Pate RR. Calibration and evaluation of an objective measure of physical activity in preschool children. *J Phys Act Health*. 2005. <https://doi.org/10.1123/jpah.2.3.345>.
86. Ettienne R, Nigg CR, Li F, Su Y, McGlone K, Luick B, et al. Validation of the Actical accelerometer in multiethnic preschoolers: the Children's healthy living (CHL) program. *Hawaii J Med Public Health*. 2016;75(4):95–100.
87. Alghaeed Z, Reilly JJ, Chastin SF, Martin A, Davies G, Paton JY. The influence of minimum sitting period of the ActivPAL™ on the measurement of breaks in sitting in young children. *PLoS One*. 2013. <https://doi.org/10.1371/journal.pone.0071854>.
88. Davies G, Reilly JO, McGowan A, Dall P, Granat M, Paton J. Validity, practical utility, and reliability of the activPAL in preschool children. *Med Sci Sports Exerc*. 2012;44(4):761–8. <https://doi.org/10.1249/MSS.0b013e31823b1dc7>.
89. Janssen X, Cliff DP, Reilly JJ, Hinkley T, Jones RA, Batterham M, et al. Validation of activPAL defined sedentary time and breaks in sedentary time in 4-to 6-year-olds. *Pediatr Exerc Sci*. 2013b. <https://doi.org/10.1123/pes.2013-0106>.
90. Byun W, Lee JM, Kim Y, Brusseau TA. Classification Accuracy of a Wearable Activity Tracker for Assessing Sedentary Behavior and Physical Activity in 3–5-Year-Old Children. *Int J Environ Res Public Health*. 2018a. <https://doi.org/10.3390/ijerph15040594>.
91. Sharp CA, Mackintosh KA, Erjavec M, Pascoe DM, Horne PJ. Validity and reliability of the Fitbit zip as a measure of preschool children's step count. *BMJ Open Sport Exerc Med*. 2017;3(1):e000272. <https://doi.org/10.1136/bmjsem-2017-000272>.
92. Liggett L, Gray A, Parnell W, McGee R, McKenzie Y. Validation and reliability of the New Lifestyles NL-1000 accelerometer in New Zealand preschoolers. *J Phys Act Health*. 2012. <https://doi.org/10.1123/jpah.9.2.295>.
93. Finn KJ, Specker BO. Comparison of Actiwatch® activity monitor and Children's activity rating scale in children. *Med Sci Sports Exerc*. 2000;32(10):1794–7. <https://doi.org/10.1097/00005768-200010000-00021>.
94. Djfarian K, Speakman JR, Stewart J, Jackson DM. Comparison of activity levels measured by a wrist worn accelerometer and direct observation in young children. *Open J Pediatr*. 2013. <https://doi.org/10.4236/ojped.2013.34076>.
95. Louie L, Chan L. The use of pedometry to evaluate the physical activity levels among preschool children in Hong Kong. *Early Child Dev Care*. 2003. <https://doi.org/10.1080/0300443022000022459>.
96. McKee DP, Boreham CA, Murphy MH, Nevill AM. Validation of the Digiwalker™ pedometer for measuring physical activity in young children. *Pediatr Exerc Sci*. 2005. <https://doi.org/10.1123/pes.17.4.345>.
97. Oliver M, Schofield GM, Kolt GS, Schluter PJ. Pedometer accuracy in physical activity assessment of preschool children. *J Sci Med Sport*. 2007. <https://doi.org/10.1016/j.jsams.2006.07.004>.
98. Nishikido N, Kashiwazaki H, Suzuki T. Preschool children's daily activities: direct observation, pedometry or questionnaire. *J Hum Ergol*. 1982. <https://doi.org/10.11183/jhe1972.11.214>.
99. Murray ME. Validity and reliability of using MVP 4 Function Walk4Life digital pedometers to assess physical activity levels among preschool-aged Head Start children. 2009. (Doctoral dissertation, The University of Texas School of Public Health).
100. Noland M, Danner F, Dewalt K, McFadden M, Kotchen JM. The measurement of physical activity in young children. *Res Q Exerc Sport*. 1990. <https://doi.org/10.1080/02701367.1990.10608668>.

101. Johansson E, Ekelund U, Nero H, Marcus C, Hagströmer M. Calibration and cross-validation of a wrist-worn Actigraph in young preschoolers. *Pediatr Obes*. 2015. <https://doi.org/10.1111/j.2047-6310.2013.00213.x>.
102. Van Cauwenberghe E, Labarque V, Trost SG, De Bourdeaudhuij I, Cardon G. Calibration and comparison of accelerometer cut points in preschool children. *Int J Pediatr Obes*. 2011. <https://doi.org/10.3109/17477166.2010.526223>.
103. Freedson P, Pober D, Janz KF. Calibration of accelerometer output for children. *Med Sci Sports Exerc*. 2005;37(11):523–30.
104. Schaefer CA, Nace H, Browning R. Establishing wrist-based cutpoints for the actical accelerometer in elementary school-aged children. *J Phys Act Health*. 2014. <https://doi.org/10.1123/jpah.2011-0411>.
105. Vanhelst J, Béghin L, Turck D, Gottrand F. New validated thresholds for various intensities of physical activity in adolescents using the Actigraph accelerometer. *Int J Rehabil Res*. 2011;34(2):175–7. <https://doi.org/10.1097/MRR.0b013e328340129e>.
106. Rowlands AV, Thomas PW, Eston RG, Topping R. Validation of the RT3 triaxial accelerometer for the assessment of physical activity. *Med Sci Sports Exerc*. 2004;36(3):518–24. <https://doi.org/10.1249/01.MSS.0000111715.14542.E7>.
107. Sun DX, Schmidt G, Teo-Koh SM. Validation of the RT3 accelerometer for measuring physical activity of children in simulated free-living conditions. *Pediatr Exerc Sci*. 2008. <https://doi.org/10.1123/pes.20.2.181>.
108. Chu EY, McManus AM, Yu CC. Calibration of the RT3 accelerometer for ambulation and nonambulation in children. *Med Sci Sports Exerc*. 2007;39(11):2085–91. <https://doi.org/10.1249/mss.0b013e318148436c>.
109. Ekblom O, Nyberg G, Bak EE, Ekelund U, Marcus C. Validity and comparability of a wrist-worn accelerometer in children. *J Phys Act Health*. 2012. <https://doi.org/10.1123/jpah.9.3.389>.
110. Larson TA, Normand MP, Hustyi KM. Preliminary evaluation of an observation system for recording physical activity in children. *Behav Interv*. 2011. <https://doi.org/10.1002/bin.332>.
111. Sharma SV, Chuang RJ, Skala K, Atteberry H. Measuring physical activity in preschoolers: reliability and validity of the System for Observing Fitness Instruction Time for Preschoolers (SOFIT-P). *Meas Phys Educ Exerc Sci*. 2011. <https://doi.org/10.1080/1091367X.2011.594361>.
112. Byun W, Kim Y, Brusseau TA. The Use of a Fitbit Device for Assessing Physical Activity and Sedentary Behavior in Preschoolers. *J Pediatr*. 2018b. <https://doi.org/10.1016/j.jpeds.2018.03.057>.
113. Martin A, McNeill M, Penpraze V, Dall P, Granat M, Paton JY, et al. Objective measurement of habitual sedentary behavior in pre-school children: comparison of activPAL With Actigraph monitors. *Pediatr Exerc Sci*. 2011. <https://doi.org/10.1123/pes.23.4.468>.
114. Shin JC. Calibration of an accelerometer to measure physical activity in preschool children: a feasibility study. 2015. (Doctoral dissertation).
115. Cardon G, De Bourdeaudhuij I. Comparison of pedometer and accelerometer measures of physical activity in preschool children. *Pediatr Exerc Sci*. 2007. <https://doi.org/10.1123/pes.19.2.205>.
116. Pagels P, Boldemann C, Raustorp A. Comparison of pedometer and accelerometer measures of physical activity during preschool time on 3-to 5-year-old children. *Acta Paediatr*. 2011. <https://doi.org/10.1111/j.1651-2227.2010.01962.x>.
117. Bikchu CH. Convergent validity of the electronic pedometer with the TriTrac RT3 accelerometer for measuring structured play activities in preschool children. *J Bio Educ*. 2014;11:10.
118. De Craemer M, De Decker E, Santos-Lozano A, Verloigne M, De Bourdeaudhuij I, Deforche B, et al. Validity of the Omron pedometer and the actigraph step count function in preschoolers. *J Sci Med Sport*. 2015. <https://doi.org/10.1016/j.jsams.2014.06.001>.
119. Lee M, Cho J, Lee H, Park CH, Oh J, Choi MC. Validity evidence of objective physical activity measures for early childhood. *Res Q Exerc Sport*. 2014;85(S1):77.
120. Chen X, Sekine M, Hamanishi S, Wang H, Hayashikawa Y, Yamagami T, et al. The validity of nursery teachers' report on the physical activity of young children. *J Epidemiol*. 2002. <https://doi.org/10.2188/jea.12.367>.
121. Manios Y, Kafatos A, Markakis G. Physical activity of 6-year-old children: Validation of two proxy reports. *Pediatr Exerc Sci*. 1998. <https://doi.org/10.1123/pes.10.2.176>.
122. Dwyer GM, Hardy LL, Peat JK, Baur LA. The validity and reliability of a home environment preschool-age physical activity questionnaire (Pre-PAQ). *Int J Behav Nutr Phys Act*. 2011. <https://doi.org/10.1186/1479-5868-8-86>.
123. Janz KF, Broffitt B, Levy SM. Validation evidence for the Netherlands physical activity questionnaire for young children: the Iowa bone development study. *Res Q Exerc Sport*. 2005. <https://doi.org/10.1080/02701367.2005.10599308>.
124. Bacardi-Gascón M, Reveles-Rojas C, Woodward-Lopez G, Crawford P, Jiménez-Cruz A. Assessing the validity of a physical activity questionnaire developed for parents of preschool children in Mexico. *J Health Popul Nutr*. 2012;30(4). <https://doi.org/10.3329/jhpn.v30i4.13327>.
125. Wen LM, Van der Ploeg HP, Kite J, Cashmore A, Rissel C. A validation study of assessing physical activity and sedentary behavior in children aged 3 to 5 years. *Pediatr Exerc Sci*. 2010. <https://doi.org/10.1123/pes.22.3.408>.
126. Chow BC, Au YC. Validation of parent proxy report with Pedometry on preschool Children's physical activity. *Asian J Phys Educ Recreation*. 2009;15(2):65–74. <https://doi.org/10.24112/ajper.151781>.
127. Telford A, Salmon J, Jolley D, Crawford D. Reliability and validity of physical activity questionnaires for children: The Children's Leisure Activities Study Survey (CLASS). *Pediatr Exerc Sci*. 2004. <https://doi.org/10.1123/pes.16.1.64>.
128. Mendoza JA, McLeod J, Chen TA, Nicklas TA, Baranowski T. Convergent validity of preschool children's television viewing measures among low-income Latino families: a cross-sectional study. *Child Obes*. 2013. <https://doi.org/10.1089/chi.2012.0116>.
129. Vanderloo LM, Di Cristofano NA, Proudfoot NA, Tucker P, Timmons BW. Comparing the Actical and ActiGraph approach to measuring young children's physical activity levels and sedentary time. *Pediatr Exerc Sci*. 2016. <https://doi.org/10.1123/pes.2014-0218>.
130. Van Cauwenberghe E, Wooller L, Mackay L, Cardon G, Oliver M. Comparison of Actical and activPAL measures of sedentary behaviour in preschool children. *J Sci Med Sport*. 2012. <https://doi.org/10.1016/j.jsams.2012.03.014>.
131. Janssen X, Cliff D, Okely AD, Jones RA, Batterham M, Ekelund U, et al. Practical utility and reliability of whole-room calorimetry in young children. *Br J Nutr*. 2013c. <https://doi.org/10.1017/S0007114512003820>.
132. Oortwijn AW, Plasqui G, Reilly JJ, Okely AD. Feasibility of an activity protocol for young children in a whole room indirect calorimeter: a proof-of-concept study. *J Phys Act Health*. 2009. <https://doi.org/10.1123/jpah.6.5.633>.
133. Puhl J, Greaves K, Hoyt M, Baranowski T. Children's Activity Rating Scale (CARS): description and calibration. *Res Q Exerc Sport*. 1990. <https://doi.org/10.1080/02701367.1990.10607475>.
134. Ellison RC, Freedson PS, Zavallos JC, White MJ, Marmor JK, Garrahe EJ, et al. Feasibility and costs of monitoring physical activity in young children using the Caltrac accelerometer. *Pediatr Exerc Sci*. 1992. <https://doi.org/10.1123/pes.4.2.136>.
135. González-Gil EM, Mouratidou T, Cardon G, Androustos O, De Bourdeaudhuij I, Gózdź M, et al. ToyBox-study group. Reliability of primary caregivers reports on lifestyle behaviours of European pre-school children: the Toy Box-study. *Obes Rev*. 2014. <https://doi.org/10.1111/obr.12184>.
136. Fotini V, Antonis K, Dimitra GM. The validity of two Omron pedometers in preschool children under different conditions. *Sylvan*. 2015;159:60–89.
137. Saris WH, Binkhorst RA. The use of pedometer and actometer in studying daily physical activity in man. Part II: validity of pedometer and actometer measuring the daily physical activity. *Eur J Appl Physiol Occup Physiol*. 1977;37(3):229–35. <https://doi.org/10.1007/BF00421778>.
138. Janssen X, Cliff DP. Issues related to measuring and interpreting objectively measured sedentary behavior data. *Meas Phys Educ Exerc Sci*. 2015. <https://doi.org/10.1080/1091367X.2015.1045908>.
139. Hidding LM, Chinapaw MJ, Belmon LS, Altenburg TM. Co-creating a 24-hour movement behavior tool together with 9–12-year-old children using mixed-methods: MyDailyMoves. *Int J Behav Nutr Phys Act*. 2020. <https://doi.org/10.1186/s12966-020-00965-0>.
140. Fulton JE, Burgeson CR, Perry GR, Sherry B, Galuska DA, Alexander MP, et al. Assessment of physical activity and sedentary behavior in preschool-age children: priorities for research. *Pediatr Exerc Sci*. 2001. <https://doi.org/10.1123/pes.13.2.113>.
141. Rowe DA. Back to the future? Algorithms and equipment vs. simplicity and common sense in physical activity measurement. *Int J Hum Mov Sci*. 2011;5(2):25–45.
142. Thompson D, Peacock O, Western M, Batterham AM. Multidimensional physical activity: an opportunity not a problem. *Exerc Sport Sci Rev*. 2015;43(2):67–74. <https://doi.org/10.1249/JES.0000000000000039>.

143. Welk GJ, Corbin CB, Dale D. Measurement issues in the assessment of physical activity in children. *Res Q Exerc Sport*. 2000. <https://doi.org/10.1080/02701367.2000.11082788>.
144. Trost SG. State of the art reviews: measurement of physical activity in children and adolescents. *Am J Lifestyle Med*. 2007. <https://doi.org/10.1177/1559827607301686>.
145. Terwee CB, Mokkink LB, Hidding LM, Altenburg TM, van Poppel MN, Chinapaw MJ. Comment on "Should we reframe how we think about physical activity and sedentary behavior measurement? Validity and reliability reconsidered". *Int J Behav Nutr Phys Act*. 2016;13(1):66. <https://doi.org/10.1186/s12966-016-0392-8>.
146. Rennie KL, Wareham NJ. The validation of physical activity instruments for measuring energy expenditure: problems and pitfalls. *Public Health Nutr*. 1998. <https://doi.org/10.1079/PHN19980043>.
147. Troiano RP. Can there be a single best measure of reported physical activity? *Ther Am J Clin Nutr*. 2009. <https://doi.org/10.3945/ajcn.2008.27461>.
148. Bauman A, Pedišić Ž, Bragg K. Objective measurement in physical activity surveillance: present role and future potential. In: *The objective monitoring of physical activity: contributions of Accelerometry to epidemiology, exercise science and rehabilitation*. Springer: Cham; 2016. p. 347–67. https://doi.org/10.1007/978-3-319-29577-0_13.
149. Lee IM, Shiroma EJ. Using accelerometers to measure physical activity in large-scale epidemiological studies: issues and challenges. *Br J Sports Med*. 2014. <https://doi.org/10.1136/bjsports-2013-093154>.
150. Cain KL, Sallis JF, Conway TL, Van Dyck D, Calhoun L. Using accelerometers in youth physical activity studies: a review of methods. *J Phys Act Health*. 2013. <https://doi.org/10.1123/jpah.10.3.437>.
151. Migueles JH, Cadenas-Sanchez C, Ekelund U, Nyström CD, Mora-Gonzalez J, Löf M, et al. Accelerometer data collection and processing criteria to assess physical activity and other outcomes: a systematic review and practical considerations. *Sports Med*. 2017. <https://doi.org/10.1007/s40279-017-0716-0>.
152. Hallal PC, Andersen LB, Bull FC, Guthold R, Haskell W, Ekelund U, et al. Global physical activity levels: surveillance progress, pitfalls, and prospects. *Lancet*. 2012. [https://doi.org/10.1016/S0140-6736\(12\)60646-1](https://doi.org/10.1016/S0140-6736(12)60646-1).
153. Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. *BMC Med Res Methodol*. 2010. <https://doi.org/10.1186/1471-2288-10-22>.
154. Terwee CB, Mokkink LB, van Poppel MN, Chinapaw MJ, van Mechelen W, de Vet HC. Qualitative attributes and measurement properties of physical activity questionnaires. *Sports Med*. 2010. <https://doi.org/10.2165/11531370-000000000-00000>.
155. Banda JA, Haydel KF, Davila T, Desai M, Bryson S, Haskell WL, et al. Effects of varying epoch lengths, wear time algorithms, and activity cut-points on estimates of child sedentary behavior and physical activity from accelerometer data. *PLoS One*. 2016. <https://doi.org/10.1371/journal.pone.0150534>.
156. Ojiambo R, Cuthill R, Budd H, Konstabel K, Casajus JA, González-Agüero A, et al. Impact of methodological decisions on accelerometer outcome variables in young children. *Int J Obes*. 2011;35(S1):S98–S103. <https://doi.org/10.1038/ijo.2011.40>.
157. Addy CL, Trilk JL, Dowda M, Byun W, Pate RR. Assessing preschool children's physical activity: how many days of accelerometry measurement. *Pediatr Exerc Sci*. 2014;26(1):103–9. <https://doi.org/10.1123/pes.2013-0021>.
158. Penpraze V, Reilly JJ, MacLean CM, Montgomery C, Kelly LA, Paton JY, et al. Monitoring of physical activity in young children: how much is enough? *Pediatr Exerc Sci*. 2006;18(4):483–91. <https://doi.org/10.1123/pes.18.4.483>.
159. Bingham DD, Costa S, Clemes SA, Routen AC, Moore HJ, Barber SE. Accelerometer data requirements for reliable estimation of habitual physical activity and sedentary time of children during the early years—a worked example following a stepped approach. *J Sports Sci*. 2016;34(20):2005–10. <https://doi.org/10.1080/02640414.2016.1149605>.
160. Hislop J, Law J, Rush R, Grainger A, Bulley C, Reilly JJ, et al. An investigation into the minimum accelerometry wear time for reliable estimates of habitual physical activity and definition of a standard measurement day in pre-school children. *Physiol Meas*. 2014;35(11):2213–28.
161. Ricardo LI, Wendt A, Galliano LM, de Andrade Muller W, Niño Cruz GI, Wehrmeister F, et al. Number of days required to estimate physical activity constructs objectively measured in different age groups: findings from three Brazilian (Pelotas) population-based birth cohorts. *PLoS One*. 2020;15(1):e0216017. <https://doi.org/10.1371/journal.pone.0216017>.
162. Byun W, Beets MW, Pate RR. Sedentary behavior in preschoolers: how many days of accelerometer monitoring is needed? *Int J Environ Res Public Health*. 2015;12(10):13148–61. <https://doi.org/10.3390/ijerph121013148>.
163. Hinkley T, O'Connell E, Okely AD, Crawford D, Hesketh K, Salmon J. Assessing volume of accelerometry data for reliability in preschool children. *Med Sci Sports Exerc*. 2012;44(12):2436–41. <https://doi.org/10.1249/MSS.0b013e3182661478>.
164. Lima RA, Barros SS, Cardoso Júnior CG, Silva G, Farias Júnior JC, Andersen LB, et al. Influence of number of days and valid hours using accelerometry on the estimates of physical activity level in preschool children from Recife, Pernambuco, Brazil. *Rev Bras Cineantropometria Desempenho Humano*. 2014, 16;(2). <https://doi.org/10.5007/1980-0037.2014v16n2p171>.
165. Barber SE, Bingham DD, Akhtar S, Jackson C, Ainsworth H, Hewitt C, et al. 'Pre-schoolers in the Playground' (PIP)—a pilot cluster randomised controlled trial of a physical activity intervention for children aged 18 months to 4 years old. *Trials*. 2013. <https://doi.org/10.1186/1745-6215-14-326>.
166. Vanhelst J, Vidal F, Drumez E, Béghin L, Baudalet JB, Coopman S, et al. Comparison and validation of accelerometer wear time and non-wear time algorithms for assessing physical activity levels in children and adolescents. *BMC Med Res Methodol*. 2019;19(1):72. <https://doi.org/10.1186/s12874-019-0712-1>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

