

Systematic review to identify and appraise outcome measures used to evaluate childhood obesity treatment interventions (CoOR): evidence of purpose, application, validity, reliability and sensitivity

Maria Bryant, Lee Ashton, Julia Brown, Susan Jebb, Judy Wright, Katharine Roberts and Jane Nixon



***National Institute for
Health Research***

Systematic review to identify and appraise outcome measures used to evaluate childhood obesity treatment interventions (CoOR): evidence of purpose, application, validity, reliability and sensitivity

Maria Bryant,^{1*} Lee Ashton,¹ Julia Brown,¹
Susan Jebb,² Judy Wright,³ Katharine Roberts⁴
and Jane Nixon¹

¹Clinical Trials Research Unit, University of Leeds, Leeds, UK

²Medical Research Council (MRC) Human Nutrition Research, Cambridge, UK

³Institute of Health Sciences, University of Leeds, Leeds, UK

⁴National Obesity Observatory (NOO), Oxford, UK

*Corresponding author

Declared competing interests of authors: SJ is a current member of the Tanita Medical Advisory Board; past member of Nestlé Advisory Board (ceased 2010), Coca-Cola Advisory Board (ceased 2011) and Heinz Advisory Board (ceased 2011), and contributor to the Rosemary Conley *Diet & Fitness* magazine.

Published August 2014

DOI: 10.3310/hta18510

This report should be referenced as follows:

Bryant M, Ashton L, Brown J, Jebb S, Wright J, Roberts K, *et al.* Systematic review to identify and appraise outcome measures used to evaluate childhood obesity treatment interventions (CoOR): evidence of purpose, application, validity, reliability and sensitivity. *Health Technol Assess* 2014;**18**(51).

Health Technology Assessment is indexed and abstracted in *Index Medicus/MEDLINE*, *Excerpta Medica/EMBASE*, *Science Citation Index Expanded (SciSearch®)* and *Current Contents®/Clinical Medicine*.

ISSN 1366-5278 (Print)

ISSN 2046-4924 (Online)

Impact factor: 5.116

Health Technology Assessment is indexed in MEDLINE, CINAHL, EMBASE, The Cochrane Library and the ISI Science Citation Index and is assessed for inclusion in the Database of Abstracts of Reviews of Effects.

This journal is a member of and subscribes to the principles of the Committee on Publication Ethics (COPE) (www.publicationethics.org/).

Editorial contact: nihredit@southampton.ac.uk

The full HTA archive is freely available to view online at www.journalslibrary.nihr.ac.uk/hta. Print-on-demand copies can be purchased from the report pages of the NIHR Journals Library website: www.journalslibrary.nihr.ac.uk

Criteria for inclusion in the *Health Technology Assessment* journal

Reports are published in *Health Technology Assessment* (HTA) if (1) they have resulted from work for the HTA programme, and (2) they are of a sufficiently high scientific quality as assessed by the reviewers and editors.

Reviews in *Health Technology Assessment* are termed 'systematic' when the account of the search appraisal and synthesis methods (to minimise biases and random errors) would, in theory, permit the replication of the review by others.

HTA programme

The HTA programme, part of the National Institute for Health Research (NIHR), was set up in 1993. It produces high-quality research information on the effectiveness, costs and broader impact of health technologies for those who use, manage and provide care in the NHS. 'Health technologies' are broadly defined as all interventions used to promote health, prevent and treat disease, and improve rehabilitation and long-term care.

The journal is indexed in NHS Evidence via its abstracts included in MEDLINE and its Technology Assessment Reports inform National Institute for Health and Care Excellence (NICE) guidance. HTA research is also an important source of evidence for National Screening Committee (NSC) policy decisions.

For more information about the HTA programme please visit the website: www.hta.ac.uk/

This report

The research reported in this issue of the journal was funded by the HTA programme as project number 09/127/07. The contractual start date was in September 2011. The draft report began editorial review in February 2013 and was accepted for publication in July 2013. The authors have been wholly responsible for all data collection, analysis and interpretation, and for writing up their work. The HTA editors and publisher have tried to ensure the accuracy of the authors' report and would like to thank the reviewers for their constructive comments on the draft document. However, they do not accept liability for damages or losses arising from material published in this report.

This report presents independent research funded by the National Institute for Health Research (NIHR). The views and opinions expressed by authors in this publication are those of the authors and do not necessarily reflect those of the NHS, the NIHR, NETSCC, the HTA programme or the Department of Health. If there are verbatim quotations included in this publication the views and opinions expressed by the interviewees are those of the interviewees and do not necessarily reflect those of the authors, those of the NHS, the NIHR, NETSCC, the HTA programme or the Department of Health.

© Queen's Printer and Controller of HMSO 2014. This work was produced by Bryant *et al.* under the terms of a commissioning contract issued by the Secretary of State for Health. This issue may be freely reproduced for the purposes of private research and study and extracts (or indeed, the full report) may be included in professional journals provided that suitable acknowledgement is made and the reproduction is not associated with any form of advertising. Applications for commercial reproduction should be addressed to: NIHR Journals Library, National Institute for Health Research, Evaluation, Trials and Studies Coordinating Centre, Alpha House, University of Southampton Science Park, Southampton SO16 7NS, UK.

Published by the NIHR Journals Library (www.journalslibrary.nihr.ac.uk), produced by Prepress Projects Ltd, Perth, Scotland (www.prepress-projects.co.uk).

Editor-in-Chief of *Health Technology Assessment* and NIHR Journals Library

Professor Tom Walley Director, NIHR Evaluation, Trials and Studies and Director of the HTA Programme, UK

NIHR Journals Library Editors

Professor Ken Stein Chair of HTA Editorial Board and Professor of Public Health, University of Exeter Medical School, UK

Professor Andree Le May Chair of NIHR Journals Library Editorial Group (EME, HS&DR, PGfAR, PHR journals)

Dr Martin Ashton-Key Consultant in Public Health Medicine/Consultant Advisor, NETSCC, UK

Professor Matthias Beck Chair in Public Sector Management and Subject Leader (Management Group), Queen's University Management School, Queen's University Belfast, UK

Professor Aileen Clarke Professor of Public Health and Health Services Research, Warwick Medical School, University of Warwick, UK

Dr Tessa Crilly Director, Crystal Blue Consulting Ltd, UK

Dr Peter Davidson Director of NETSCC, HTA, UK

Ms Tara Lamont Scientific Advisor, NETSCC, UK

Professor Elaine McColl Director, Newcastle Clinical Trials Unit, Institute of Health and Society, Newcastle University, UK

Professor William McGuire Professor of Child Health, Hull York Medical School, University of York, UK

Professor Geoffrey Meads Professor of Health Sciences Research, Faculty of Education, University of Winchester, UK

Professor Jane Norman Professor of Maternal and Fetal Health, University of Edinburgh, UK

Professor John Powell Consultant Clinical Adviser, National Institute for Health and Care Excellence (NICE), UK

Professor James Raftery Professor of Health Technology Assessment, Wessex Institute, Faculty of Medicine, University of Southampton, UK

Dr Rob Riemsma Reviews Manager, Kleijnen Systematic Reviews Ltd, UK

Professor Helen Roberts Professor of Child Health Research, University College London, UK

Professor Helen Snooks Professor of Health Services Research, Institute of Life Science, College of Medicine, Swansea University, UK

Please visit the website for a list of members of the NIHR Journals Library Board:
www.journalslibrary.nihr.ac.uk/about/editors

Editorial contact: nihredit@southampton.ac.uk

Abstract

Systematic review to identify and appraise outcome measures used to evaluate childhood obesity treatment interventions (CoOR): evidence of purpose, application, validity, reliability and sensitivity

Maria Bryant,^{1*} Lee Ashton,¹ Julia Brown,¹ Susan Jebb,² Judy Wright,³ Katharine Roberts⁴ and Jane Nixon¹

¹Clinical Trials Research Unit, University of Leeds, Leeds, UK

²Medical Research Council (MRC) Human Nutrition Research, Cambridge, UK

³Institute of Health Sciences, University of Leeds, Leeds, UK

⁴National Obesity Observatory (NOO), Oxford, UK

*Corresponding author

Background: Lack of uniformity in outcome measures used in evaluations of childhood obesity treatment interventions can impede the ability to assess effectiveness and limits comparisons across trials.

Objective: To identify and appraise outcome measures to produce a framework of recommended measures for use in evaluations of childhood obesity treatment interventions.

Data sources: Eleven electronic databases were searched between August and December 2011, including MEDLINE; MEDLINE In-Process and Other Non-Indexed Citations; EMBASE; PsycINFO; Health Management Information Consortium (HMIC); Allied and Complementary Medicine Database (AMED); Global Health, Maternity and Infant Care (all Ovid); Cumulative Index to Nursing and Allied Health Literature (CINAHL) (EBSCOhost); Science Citation Index (SCI) [Web of Science (WoS)]; and The Cochrane Library (Wiley) – from the date of inception, with no language restrictions. This was supported by review of relevant grey literature and trial databases.

Review methods: Two searches were conducted to identify (1) outcome measures and corresponding citations used in published childhood obesity treatment evaluations and (2) manuscripts describing the development and/or evaluation of the outcome measures used in the childhood intervention obesity evaluations. Search 1 search strategy (review of trials) was modelled on elements of a review by Luttikhuis *et al.* (Oude Luttikhuis H, Baur L, Jansen H, Shrewsbury VA, O'Malley C, Stolk RP, *et al.* Interventions for treating obesity in children. *Cochrane Database Syst Rev* 2009;1:CD001872). Search 2 strategy (methodology papers) was built on Terwee *et al.*'s search filter (Terwee CB, Jansma EP, Riphagen II, de Vet HCW. Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. *Qual Life Res* 2009;18:1115–23). Eligible papers were appraised for quality initially by the internal project team. This was followed by an external appraisal by expert collaborators in order to agree which outcome measures should be recommended for the Childhood obesity Outcomes Review (CoOR) outcome measures framework.

Results: Three hundred and seventy-nine manuscripts describing 180 outcome measures met eligibility criteria. Appraisal of these resulted in the recommendation of 36 measures for the CoOR outcome measures framework. Recommended primary outcome measures were body mass index (BMI) and dual-energy X-ray absorptiometry (DXA). Experts did not advocate any self-reported measures where

objective measurement was possible (e.g. physical activity). Physiological outcomes hold potential to be primary outcomes, as they are indicators of cardiovascular health, but without evidence of what constitutes a minimally importance difference they have remained as secondary outcomes (although the corresponding lack of evidence for BMI and DXA is acknowledged). No preference-based quality-of-life measures were identified that would enable economic evaluation via calculation of quality-adjusted life-years. Few measures reported evaluating responsiveness.

Limitations: Proposed recommended measures are fit for use as outcome measures within studies that evaluate childhood obesity treatment evaluations specifically. These may or may not be suitable for other study designs, and some excluded measures may be more suitable in other study designs.

Conclusions: The CoOR outcome measures framework provides clear guidance of recommended primary and secondary outcome measures. This will enhance comparability between treatment evaluations and ensure that appropriate measures are being used. Where possible, future work should focus on modification and evaluation of existing measures rather than development of tools de nova. In addition, it is recommended that a similar outcome measures framework is produced to support evaluation of adult obesity programmes.

Funding: The National Institute for Health Research Health Technology Assessment programme.

Contents

List of tables	xi
List of figures	xiii
Glossary	xv
List of abbreviations	xix
Scientific summary	xxi
Chapter 1 Aims and objectives	1
Chapter 2 Background	3
Chapter 3 Methods	5
Protocol and registration	5
Design	5
<i>Evidence synthesis</i>	5
Search strategy	5
<i>Search 1 terms: identification of childhood obesity treatment intervention evaluations</i>	5
<i>Search 2 terms: identification of studies describing the development or evaluation of relevant outcome measures</i>	5
<i>Grey literature and evidence from clinical trials databases</i>	6
<i>Data management</i>	7
Eligibility criteria	7
<i>Childhood obesity treatment evaluation studies</i>	7
<i>Outcome development/evaluation methodology studies</i>	10
Data extraction process	11
<i>Phase I: Trial description extraction</i>	11
<i>Phase II: Outcome measure methodology extraction</i>	11
Appraisal of quality of outcome measures	12
<i>Internal appraisal</i>	12
<i>Expert appraisal</i>	15
Chapter 4 Results	19
Number and type of studies identified	19
Number and type of studies excluded, with reasons	19
Study characteristics	21
<i>Manuscripts describing childhood obesity treatment trials</i>	21
<i>Manuscripts describing the development/evaluation of outcome measures</i>	22
Findings of the systematic review	22
<i>Anthropometry</i>	22
<i>Diet</i>	24
<i>Eating behaviour</i>	26
<i>Physical activity</i>	27
<i>Sedentary behaviour/time</i>	28
<i>Fitness</i>	29
<i>Physiology</i>	30

<i>Economic evaluation</i>	31
<i>Health-related quality of life</i>	31
<i>Psychological well-being</i>	32
<i>Environment</i>	32
Results of internal appraisal	34
<i>Anthropometry (1 certainty = '1'; 2 certainty = '2'; 35 certainty = '3')</i>	34
<i>Diet (3 certainty = '1'; 9 certainty = '2'; 19 certainty = '3')</i>	34
<i>Eating behaviours (5 certainty = '1'; 6 certainty = '2'; 11 certainty = '3')</i>	35
<i>Physical activity (4 certainty = '1'; 9 certainty = '2'; 11 certainty = '3')</i>	35
<i>Sedentary behaviour/time (0 certainty = '1'; 0 certainty = '2'; 6 certainty = '3')</i>	35
<i>Fitness (1 certainty = '1'; 5 certainty = '2'; 7 certainty = '3')</i>	36
<i>Physiology (2 certainty = '1'; 0 certainty = '2'; 10 certainty = '3')</i>	36
<i>Health-related quality of life (4 certainty = '1'; 2 certainty = '2'; 6 certainty = '3')</i>	36
<i>Psychological well-being (4 certainty = '1'; 1 certainty = '2'; 12 certainty = '3')</i>	36
<i>Environment (5 certainty = '1'; 1 certainty = '2'; 4 certainty = '3')</i>	37
Results of expert appraisal	37
<i>Anthropometry</i>	37
<i>Diet</i>	38
<i>Eating behaviours</i>	39
<i>Physical activity</i>	39
<i>Sedentary time</i>	40
<i>Fitness</i>	40
<i>Physiology</i>	40
<i>Health-related quality of life</i>	41
Psychological well-being	41
<i>Environment</i>	41
Summary of key findings	42
Final included studies: results from appraisal	43
Chapter 5 Discussion	57
Summary of evidence	57
How to use the Childhood obesity Outcomes Review outcome measures framework	60
Limitations of the research	61
Future recommendations	62
Chapter 6 Conclusions	63
Acknowledgements	65
References	67
Appendix 1 Search 1 search strategy	95
Appendix 2 Search 2 search strategy	97
Appendix 3 Search 1 references (included childhood obesity treatment trials)	101
Appendix 4 Data extraction form for search 1	115
Appendix 5 Data extraction form for search 2: dietary assessment	123
Appendix 6 Anthropometry studies: summary table	141

Appendix 7 Dietary assessment evaluation studies: summary table	197
Appendix 8 Eating behaviour studies: summary table	215
Appendix 9 Physical activity measurement studies: summary table	235
Appendix 10 Sedentary time/behaviour measurement studies: summary table	249
Appendix 11 Fitness measurement studies: summary table	253
Appendix 12 Physiology measures studies: summary table	261
Appendix 13 Health-related quality-of-life studies: summary table	275
Appendix 14 Psychological well-being measures: summary table	287
Appendix 15 Environment measures: summary	295
Appendix 16 Additional scoping searches for quality-adjusted life-years and clinical cut-offs in physiological measures	303
Appendix 17 Appraisal decision forms: anthropometry	305
Appendix 18 Diet methodology studies: development and evaluation scores	319
Appendix 19 Eating behaviour methodology studies: development and evaluation scores	325
Appendix 20 Physical activity methodology studies: development and evaluation scores	329
Appendix 21 Sedentary time/behaviour methodology studies: development and evaluation scores	335
Appendix 22 Fitness methodology studies: development and evaluation scores	337
Appendix 23 Physiology methodology studies: development and evaluation scores	339
Appendix 24 Health-related quality-of-life studies: development and evaluation scores	343
Appendix 25 Psychological well-being studies: development and evaluation scores	347
Appendix 26 Environment studies: development and evaluation scores	351
Appendix 27 Non-English manuscripts of search 1 trials (data not extracted)	355
Appendix 28 Childhood obesity Outcomes Review appraisal decision form: secondary outcomes	359

List of tables

TABLE 1 Included secondary outcomes with examples of outcome measures	8
TABLE 2 Criteria used to allocate robustness scores for evaluation of quality	13
TABLE 3 Number of eligible manuscripts with corresponding measures by outcome domain	23
TABLE 4 The CoOR outcome measures framework	43

List of figures

FIGURE 1 Summary of study design	6
FIGURE 2 Expert appraisal process	16
FIGURE 3 Summary of study selection and exclusion	20
FIGURE 4 Frequency (%) of primary outcome measures used in search 1 trials	21
FIGURE 5 Frequency (%) of trials using each type of secondary outcome	22
FIGURE 6 Diet methodologies included in appraisal	24
FIGURE 7 Using the CoOR outcome measures framework	60

Glossary

Some of the following glossary/definitions may be presented alternatively elsewhere. These, however, were specifically chosen to support the Childhood obesity Outcomes Review study.

Design

Eligibility criteria The requirements that a subject must fulfil to be allowed to enter a study. These are usually devised to ensure that the subject has the appropriate disease and that he or she is the type of subject that the researchers wish to study. Inclusion criteria should not simply be the opposites of the exclusion criteria.

End point A variable that is one of the primary interests in a study. The variable may relate to efficacy, effectiveness or safety.

Feasibility study Pieces of research done before a main study in order to answer the question 'Can this study be done?'. Feasibility studies for randomised controlled trials may not themselves be randomised. Crucially, feasibility studies do not evaluate the outcome of interest – that is left to the main study.

Outcome measure Measure used to evaluate the primary or secondary end points of an intervention evaluation; the standard against which the end result of the intervention is assessed.

Pilot study A version of the main study that is run in miniature to test whether the components of the main study can all work together. In some cases this will be the first phase of the substantive study, and data from the pilot phase may contribute to the final analysis – an internal pilot.

Primary end point The principal end point in a study, providing the primary data.

Secondary end point One of (possibly many) less important end points in a study than the primary end point.

Sample

Age Age categories have been assigned in extraction of data for the Childhood obesity Outcomes Review study as follows: infants = < 36 months; child = 36 months to 12 years; adolescents = 13–18 years.

Ethnicity Information has been extracted for the Childhood obesity Outcomes Review study on ethnicities for all ethnic groups that contribute to at least 5% of the sample within each study.

Weight status Weight status of participants was assigned using the predetermined status reported within each paper. Data extraction pertaining to weight status includes (1) 'All obese'; (2) 'All overweight'; (3) 'All overweight or obese'; (4) 'Mixed stratified' (includes all weight status groups, with results stratified by weight status); or (5) 'Mixed non-stratified' (includes all weight status groups, but results not stratified by weight status).

Evaluation

Reliability

Inter-rater Consistency between raters (people taking/reporting measures).

Internal reliability The extent to which the questions or tasks on a test are similar. It is the correlation between different items on the same test or scale, and measures whether several items that propose to measure the same general construct produce similar scores; commonly presented as an alpha result (with values of > 0.7 generally considered to be acceptable).

Test–retest A measure of the ability of a tool to produce the same result for two different test periods. The distance/duration between tests should be considered so that any variation detected reflects reliability of the instrument rather than changes in the behaviour/construct being measured.

Validity

Confirmatory factor analysis A multivariate statistical procedure that is used to test how well the measured variables represent the number of constructs. Researchers can specify the number of factors required in the data and which measured variable is related to which latent variable, and confirmatory factor analysis is a tool that is used to confirm or reject the measurement theory.

Construct Ability of a measurement tool to actually measure the concept being studied (e.g. association between a food environment tool and dietary intake).

Content How items and domains appear to relate to construct being measured. Often done by experts to help in the selection of items.

Convergent/concurrent Degree to which the tool relates to other similar tools measuring the same or similar constructs.

Criterion How closely the tool relates to the criterion measure or the actual truth, which is a reflection of the success of the tool to predict or estimate something. For example, a tool to measure blood pressure should closely relate to actual blood pressure taken from a gold standard. Gold standard measure within each outcome domain have been predefined for the Childhood obesity Outcomes Review study (i.e. they are not based on what is reported by authors).

Face Whether the test 'looks valid', i.e. whether a test looks as if it measures what it is supposed to measure. Face validity is often established with people who are administering the tool and/or with the participants for whom the tool is developed.

[Content and face validity have been extracted as part of the Childhood obesity Outcomes Review but were not dependently scored as part of the internal appraisal (although were considered relevant as part of appraisal of tool developments within 'participant involvement').]

Internal: Exploratory factor analysis Test to explore the possible underlying factor structure of a set of observed variables without imposing a preconceived structure on the outcome to identify the number of constructs and the underlying factor structure.

Responsiveness A measure of change over time (similar to Sensitivity to Change); which is often associated with treatment. Responsiveness can also be measured without an intervention if the construct is expected to change and refers to the ability of a tool to measure clinically important change (and should also remain stable when no change has occurred).

Statistical analysis

%variance Percentage of total variance among the variables accounted for by each factor (factor analysis).

Cronbach's alpha A coefficient of reliability often used to measure internal consistency. Cronbach's alpha values reflect how closely related a set of items are as a group. A 'high' value of alpha is often used as evidence that the items measure an underlying (or latent) construct (with a reliability coefficient of ≥ 0.70 or being considered 'acceptable' in most social science research situations).

Eigenvalue (within factor analysis) The eigenvalue for a given factor reflects the variance in all of the variables, which is accounted for by that factor. The first factor will always account for the most variance (and hence have the highest eigenvalue), and the next factor will account for as much of the left over variance as it can, and so on. Hence, each successive factor will account for less and less variance. A factor that has a low eigenvalue is contributing little to the explanation of variances in the variables and may be excluded (with eigenvalues of ≥ 1 considered for inclusion).

Factor loading Correlation coefficients between the variables (rows) and factors (columns) in factor analysis. Cut-offs are arbitrary and vary considerably, but values of between ≥ 0.4 and ≥ 0.7 are often used to confirm that independent variables are represented by a particular factor. The Childhood obesity Outcomes Review considered values of ≥ 0.4 to demonstrate sufficient loading.

Intraclass correlation coefficient An index of concordance for dimensional measurements ranging between 0 and 1, where 0.75 is considered excellent reliability. The Childhood obesity Outcomes Review considered that intraclass correlation coefficients of ≥ 0.4 demonstrated sufficient correlation.

Kappa coefficients Reliability defined for nominal variables. Kappa is analogous to a correlation coefficient and has the same range of values (-1 to $+1$).

Limits of agreement Descriptive measure of agreement and the mean difference between the two tests ± 2 standard deviations, in which 95% of the differences between the two tests lie within this interval.

Pearson's r (Pearson product-moment correlation coefficient) A measure of the linear relationship between two variables. Results are presented generally as 'r values' and range from $+1$ to -1 . A correlation of $+1$ means that there is a perfect positive linear relationship between variables.

Receiver operating characteristic curve (area under the curve) A measure of a diagnostic test's discriminatory power, with an area under the curve value of 1.0 theoretically representing a perfect test (i.e. 100% sensitive and 100% specific) and a value of 0.5 indicating no discriminative value (i.e. 50% sensitive and 50% specific). The latter is represented graphically as a diagonal line extending from the lower left corner to the upper right. There are several scales for area under the curve value interpretation but, in general, receiver operating characteristic curves with an area under the curve value of < 0.75 are not clinically useful, and an area under the curve value of 0.97 has a very high clinical value, correlating with likelihood ratios of approximately 10 and 0.1.

Regression Assessment of the relationship between several independent or predictor variables and a dependent or criterion variable.

Spearman's rho (Spearman's rank correlation coefficient) Non-parametric equivalent to Pearson's correlation.

List of abbreviations

3C model	three-compartmental model	DEBQ-C	Dutch Eating Behaviour Questionnaire (child reported)
4C model	four-compartmental model		
5D FFQ	5-day food frequency questionnaire	DEBQ-P	Dutch Eating Behaviour Questionnaire parent-reported
ADP	air displacement plethysmography	df	degree of freedom
AUC	area under the curve	DLW	doubly labelled water
BIA	bioelectrical impedance analysis	DXA	dual-energy X-ray absorptiometry
BMI	body mass index	EAH-C	Eating in the Absence of Hunger-Children
BMI-SDS	body mass index standard deviation score	EES-C	Emotional Eating Scale for Children
BMR	basal metabolic rate	EHC	euglycaemic–hyperinsulinaemic clamp
C-BEDS	Children’s Binge Eating Disorder Scale	EI	energy intake
C-PSPP	Children’s Physical Self-Perception Profile	EMA	Electronic Momentary Assessment
CEBQ	Child Eating Behaviour Questionnaire	EPAO	Environment and Policy Assessment and Observation
CEHQ-FFQ	Children’s Eating Habits Questionnaire food frequency questionnaire	EQ-5D	European Quality of Life-5 Dimensions
CFQ	Child Feeding Questionnaire	EQ-5D-Y	European Quality of Life-5 Dimensions (youth version)
ChEAT	Children’s Eating Attitudes Test	ES	effect size
ChEDE-Q	Child Eating Disorder Examination Questionnaire	FA	factor analysis
CI	chief investigator	FBQ	Food Behaviour Questionnaire
CLASS	Children’s Leisure Activities Study Survey	FDA	Food and Drug Administration
CoOR	Childhood obesity Outcomes Review	FEAHQ	Family Eating and Activity Habits Questionnaire
CPSS	Children’s Physical Self-Concept Scale	FFQ	food frequency questionnaire
CSAPPA	Children’s Self-Perceptions of Adequacy in and Predilection for Physical Activity	HbA _{1c}	glycated haemoglobin
DEBQ	Dutch Eating Behaviour Questionnaire	HES	Home Environment Survey
		HHS	Healthy Home Survey
		HMIC	Health Management Information Consortium
		HR	heart rate
		HRQoL	health-related quality of life

HSFFQ	Harvard Service Food Frequency Questionnaire	PEAS	Parenting Strategies for Eating and Activity Scale
IC	internal consistency	PFQ	Preschool Feeding Questionnaire
IFIS	International Fitness Scale	PRO	patient-reported outcome
IFQ	Infant Feeding Questionnaire	QALY	quality-adjusted life-year
IFSQ	Infant Feeding Style Questionnaire	QEWP	Questionnaire of Eating and Weight Patterns
IGF	insulin-like growth factor	RCT	randomised controlled trial
IGF-1	insulin-like growth factor 1	SAC	Scientific Advisory Committee
IGFBP-1	insulin-like growth factor binding protein 1	SCRS	Self-Control Rating Scale
IGFBP-3	insulin-like growth factor binding protein 3	SES	socioeconomic status
IOTF	International Obesity Task Force	SFT	skinfold thickness
IWQoL	Impact of Weight on Quality of Life	Short YAQ	Short-list Youth/Adolescent Questionnaire
LBM	lean body mass	SOCARP	System for Observing Children's Activity and Relationships during Play
LOA	limits of agreement	SPPC	Self-Perception Profile for Children
LOC	loss of control	SRM	standardised response mean
MET	metabolic equivalent	TBW	total body water
MID	minimally important difference	TEE	total energy expenditure
MRC	Medical Research Council	TOBEC	total body electrical conductivity
MRFS-III	McKnight Risk Factor Survey-III	TRT	test-retest
NAPSACC	Nutrition and Physical Activity Self-Assessment for Child Care	TSFFQ	Toddler Snack Food Feeding Questionnaire
NICE	National Institute for Health and Care Excellence	WC	waist circumference
NIR	near-infrared interactance	WHR	waist-to-hip ratio
NOO	National Obesity Observatory	YAQ	Youth Adolescent Questionnaire
NOO SEF	National Obesity Observatory Standard Evaluation Framework	YEDE-Q	Youth Eating Disorder Examination-Questionnaire
OSRAC-P	Observational System for Recording Physical Activity in Children-Preschool version	YQOL-W	Youth Quality-of-Life Instrument-Weight Module
PA	physical activity	YRBS	Youth Risk Behaviour Survey
PAQ	Physical Activity Questionnaire		
PAQ-C	Physical Activity Questionnaire for Older Children		

Scientific summary

Background

The lack of uniformity in the outcome measures used in the evaluation of childhood obesity treatment interventions often impedes the ability to truly assess effectiveness and limits comparisons across trials. In part, this arises because of the lack of consensus on what outcomes are required and the most appropriate outcome measures to use within outcome domains.

Objective

This study aimed to systematically review the literature in order to produce a database of outcome measures that have been used (or developed for use) in childhood obesity treatment interventions and to use expert appraisal to develop a framework of recommended outcome measures for use as a resource to guide researchers when designing childhood obesity treatment evaluations. Secondary objectives include (1) a summary of the description and measurement properties of all outcome measures identified and (2) a methodology to determine the quality of outcome measures and/or aid in the development of new outcome measures in this area.

Methods

Search strategy

Two searches were performed with the aim to identify (1) outcome measures that had already been used in existing evaluations by searching trials of childhood obesity treatment interventions and (2) methodology studies that developed and/or evaluated the outcome measures for childhood obesity research. Both searches were conducted in 11 databases and were supported with literature obtained from relevant citations (including reviews of measurement tools), conference proceedings and information from registered clinical trials in progress.

Search strategies were developed by the Information Specialist (JW), with contributions of search terms from the project team. Searches were agreed by the project team and conducted from the date of inception, with no language restrictions, from August to October 2011. Terms and keywords were selected for search 1 to identify manuscripts detailing randomised controlled trials (or pilot/feasibility studies) aimed at evaluating childhood obesity treatment interventions. Search 2 included keywords/terms pertaining to the development and/or evaluation of outcome measures.

Process of study selection

Assessment of titles and abstracts was performed independently by two reviewers (MB, LA). Agreement between reviewers was tested after review of the first 130 search 1 papers and the first 50 search 2 papers. For search 1, 98% agreement was reached; for search 2, 96% agreement was reached. Disagreements were discussed to refine eligibility clarification. Papers were retained at title and abstract review if there was any degree of uncertainty by either reviewer. Full papers were then assessed against eligibility criteria, and disagreements were resolved by discussion. Measures had to have been developed specifically for childhood obesity research or evaluated in a paediatric obese population (or present results stratified by obesity) and included those in the following domains: anthropometry (primary outcome), diet, eating behaviours, physical activity (PA), sedentary time/behaviour, fitness, physiology, environment, psychological well-being and quality of life.

Data extraction

Data were extracted from relevant search 1 papers (i.e. trials), including information concerning all included outcome measures used and corresponding citations of measurement development/evaluation papers. These cited papers were then located and added to search 2 (i.e. methodology) papers. Data pertaining to the sample, design, development, evaluation and feasibility of each outcome measurement development/evaluation paper from search 2 were then extracted on prespecified extraction forms. Disagreements were resolved by discussion.

Quality assessment

Search 1 trials were not judged on quality/bias, as the study outcome information (i.e. intervention efficacy/effectiveness) was not relevant to the aims of this study. Quality assessment for measurement papers in search 2 was based on internal and external appraisal of the rigour in development and evaluation of each outcome measure. For internal appraisal, members of the internal project team (MB, LA) appraised each outcome measure related to evidence of development, reliability testing and validity testing using international guidelines for the development of patient-reported outcomes (e.g. Food and Drug Administration) and previous work already conducted by the chief investigator. This resulted in a database of outcome measures with a detailed description of each measure, in addition to a parallel assessment of quality (based on a scoring system). The internal project team then considered whether each measure was (1) fit for purpose (i.e. recommended for inclusion to the outcome measures framework); (2) not fit for purpose (i.e. not recommended for inclusion); or (3) uncertain (i.e. requires further consideration). This decision was based on existing evidence gathered and was reached by consensus. External appraisal was then conducted via an expert appraisal meeting, which was held (in person) with 10 national collaborators (plus five applicants). Collaborators were invited based on their experience and expertise within evaluation of childhood obesity interventions and/or measurement. Prior to the meeting they were provided with (1) a list of all included outcome measure development/evaluation papers alongside access to all papers; (2) tables describing each paper (summarised from the data extraction forms); and (3) internal appraisal documents, including scores for quality (e.g. for development, reliability and validity) and degree of certainty from the internal appraisal for each measure related to whether it should be included in the final framework. They were asked to review all measures but to focus on the outcome domain that was most closely aligned to their area of expertise (defined by the project team). The purpose of the meeting was then to agree on whether or not each measure was suitable for inclusion in the final outcome measures framework based on the evidence provided and any relevant personal experience/knowledge in using the measures.

Methods of analysis/synthesis

This report provides a narrative summary of outcome measures, which are grouped according to outcome domain. Analysis of reliability and validity testing was considered for appraisal, but results were not pooled.

Results

Results of search strategy

A total of 25,486 papers were identified from both searches. Eligible search 1 papers (of existing evaluations) cited 417 additional papers linked to included outcome measures, of which only 56 were eligible methodology papers. A further 323 outcome development/evaluation methodology papers from search 2 met eligibility criteria. Combined, these 379 papers described 180 outcome measures.

Results of quality assessment

Based on the reliability and validity evidence, eligible measures were appraised by the internal team, resulting in 29 outcomes that were considered to be fit for inclusion in the framework as a recommended tool (i.e. degree of certainty = 1); 35 outcomes deemed unfit for inclusion (i.e. degree of certainty = 2); and 121 requiring further consideration (i.e. degree of certainty = 3). External appraisal considered these findings alongside their experience and expertise, and concluded that 52 outcomes were fit for inclusion

across the 10 outcome domains (remaining 128 tools deemed unfit (degree of certainty = 2). Of these, two [body mass index (BMI) and dual-energy X-ray absorptiometry DXA] out of the 38 anthropometry measures were recommended. In secondary outcomes, recommended tools included 6 (out of 22) diet measures (all food frequency questionnaires); 12 (out of 22) eating behaviour measures; 4 (out of 24) PA measures (with no self-reported measures); 1 (out of 6) sedentary time measure; 1 (out of 13) fitness measure; 1 (out of 12) physiological measure; 10 (out of 12) health-related quality of life questionnaires; 10 (out of 17) psychological well-being measures; and 5 (out of 10) environmental measures.

The childhood obesity outcome measures framework

Recommended outcome measures are presented by outcome domain alongside details relating to feasibility of implementation (e.g. number of items, costs, licensing, etc.). This framework is a tool to guide researchers but the final decision for inclusion of measures must be based on those that are (1) aligned with the targets of the intervention and (2) appropriate for use in a given population (e.g. age/ethnicity specific). This framework is recommended as an initial guide outcome measure selection. In exceptional cases when no measures meet the needs of a particular study, a detailed description of all measures meeting the eligibility criteria is provided so that researchers are able to self-select the most appropriate measure given the information available on its validity.

Conclusions

The key findings of this study are:

1. Only 13% of trials correctly cited outcome measures used.
2. Approximately 20% of eligible primary and secondary outcome measures were recommended by experts.
3. Primary outcome measures recommended are BMI or DXA.
4. Objective measurement must be applied if available (i.e. use of activity monitors instead of self-reported PA).
5. Physiological outcomes have the potential to be primary outcomes (as they are measured with high precision and are related to adverse health outcomes) but, at present, there is insufficient evidence on what constitutes a clinically meaningful change (although it is recognised that this is also the case in existing primary outcomes).
6. Evidence of ability of measures to detect change was poor or lacking.
7. While new tools are pending, there remains no published preference-based measures for use in economic evaluations in this population. Cost-effectiveness should therefore include measures most pertinent to the targets of the intervention [e.g. costs per reduction in body mass index standard deviation score (BMI-SDS)].
8. The proposed recommended outcome measures are fit for use specifically within studies that evaluate childhood obesity treatment evaluations. They may or may not be suitable for other study designs.

Implications for clinical practice

The results of the expert appraisal provide clear guidance to researchers about appropriate outcomes domains and recommended measures in each of these domains to encourage greater adoption of well-validated tools. This will make it easier to judge clinical effectiveness and enhance the comparability between different studies or treatment interventions. The review also provides details of other measures that may be appropriate for other settings with details of the extent of methodological testing already conducted to inform decision-making. Researchers wishing to use novel tools are recommended to adopt these alongside the recommended tools, wherever possible, to encourage evolution and the development of new knowledge. Details of the validity of each of the recommended outcome tools provide a knowledge trail to encourage more accurate reporting of these measures in future studies.

Implications for future research

In the case of economic evaluations, primary research is urgently needed as this review did not identify a single measure that was able to calculate quality-adjusted life-years (although we are aware that some work in this area is in progress). In all other domains, a large number of outcome measures have been proposed, but in many cases robust evidence of validity is scant. There may be opportunities to make rapid progress with further testing and modifications, where necessary, of existing measures. Many outcome measures rely on self-report and more objective measures would add value, especially for dietary outcomes. There are also opportunities to consider the use of new technologies to replace pen-and-paper retrospective questionnaires to collect information on some outcomes measures. Given that a number of different types of outcome measures were identified within many outcome domains, findings from this study suggest that future research should invest in the modification (if appropriate) and evaluation of existing measures (not the development of new measures when others are available).

Research is needed to determine the ability of measures to detect change. For some (more historical) measures, such as BMI, evidence demonstrating a level of precision over multiple assessment periods may be sufficient. However, there is a lack of testing of responsiveness in many of the recommended questionnaire outcomes. Lastly, the lack of data describing the clinically meaningful change and/or appropriate cut-offs was noted as part of the expert appraisal, specifically for anthropometry, physiology and fitness outcomes.

Funding

Funding for this study was provided by the Health Technology Assessment programme of the National Institute for Health Research.

Chapter 1 Aims and objectives

This study aimed to perform a systematic review to identify and appraise existing outcome measures for use in the evaluation of childhood obesity treatment interventions. This aim was met via the following objectives:

1. Systematic review of the literature in order to produce a database of outcome measures that have been used (or developed for use) in childhood obesity treatment interventions.
2. Appraisal of outcome measures to identify and highlight those that have been developed and evaluated using high-quality, fully rigorous methods.
3. Creation of a childhood obesity outcome measures framework, categorised by (1) anthropometry/weight status; (2) diet; (3) eating behaviours; (4) physical activity (PA); (5) sedentary time; (6) fitness; (7) psychological well-being; (8) quality of life; (9) environmental measures; and (10) physiological outcomes. This framework was intended to guide researchers as to the best tool to use in their evaluation of childhood obesity treatment interventions and aimed to include:
 - outcome measure description (name, purpose, number of items and mode of administration)
 - outcome-specific issues (population intended for, theoretical orientation)
 - content (any evidence given for an underlying conceptual model; list of domains/scales covered)
 - measurement evaluation properties (development method, item reduction, validity, reliability, feasibility and responsiveness)
 - cost and practical considerations (details of licensing fees, duration of administration).

Chapter 2 Background

Many interventions to treat obesity are aimed at children but there remains a lack of high-quality evidence on effective childhood obesity interventions in the literature.¹ Existing systematic reviews aimed at comparing effectiveness of intervention programmes (particularly those conducting meta-analysis) are hampered by a lack of quality in the conduct and reporting of trials in this area. There has been some attenuation in the rising rates of childhood obesity in recent years, and it is therefore probable that many attempts to prevent and treat obesity in children have been of some success.² The problem, therefore, may lie in the methods used to evaluate and report interventions.

The degree to which weight management leads to improvements in a child's health is reflected by measuring change in outcomes in clinical trials. Outcomes either directly measure a definitive clinical change (i.e. primary outcome of weight loss) or assess proximal/secondary outcomes (e.g. change in diet) that impact on the primary outcome. In the design phase of a trial, choosing the appropriate outcomes is essential. Use of inappropriate outcomes will result in data that are inaccurate or biased and that do not indicate the effectiveness of an intervention. Moreover, collection of data using poorly chosen outcomes is a waste of resources, both for the researchers and participants involved in the trial.³ Inappropriate selection of outcomes in childhood obesity research is probably due to the uncertainty about which outcome domains are most relevant to children and their families.⁴ Furthermore, there is a lack of knowledge on which can be most reliably measured.

Guidance tools are available to facilitate the design of high-quality research, including the Medical Research Council (MRC) guidance for the evaluation of complex interventions and, more specifically, the National Obesity Observatory Standard Evaluation Framework (NOO SEF) for childhood obesity evaluation (www.noo.org.uk/core/SEF).⁵ The latter (commissioned by the Department of Health) was produced via consensus of prominent obesity researchers to aid clinicians in their evaluation of childhood obesity programmes. It now stands as a grounded tool to enable consistency with research design. The primary audience for the NOO SEF is those evaluating public health obesity programmes. However, much of the advice is of relevance to researchers conducting trial evaluations. For example, recommended outcomes are listed and described as 'essential' or 'desirable'. This resembles the output of a core outcome set, although the inclusion of each outcome has not been based on formal consensus methodologies, such as those described by 'COMET' (Core Outcome Measures in Effectiveness Trials). Core outcome sets are a minimum set of outcomes that should be measured and reported within trials or other forms of research for a specific condition (www.comet-initiative.org/). The use of core outcome sets permits comparisons between trials that are agreed on by experts within each disease area. At present, there is not a core outcome set for obesity research – partly because of the complexity and variability in intervention targets (requiring potentially different outcomes). The NOO SEF therefore stands as a guide, rather than a minimum set of outcomes. Importantly, the NOO SEF does not provide advice or details of outcome measures that should be used within each outcome domain that it recommends. Although there have been reviews published on some individual measures and their general application (e.g. measurement of television exposure⁶), there has not been a review that has focused specifically on outcome measures for used in childhood obesity treatment intervention evaluations.

The lack of consensus in determining appropriate outcome measures for the reliable and valid assessment of childhood obesity interventions means that comparisons between interventions are consequently difficult, partly because of a shortage of validated outcome measures available, but also because the selected outcome measures differ between studies. Consequently, it is a challenge to identify which interventions are most effective. Such a lack of consistency and inadequacy impedes the progress of childhood obesity research.

Chapter 3 Methods

Protocol and registration

A current review protocol exists and can be accessed via the Health Technology Assessment (HTA) website or direct correspondence with the chief investigator [(CI): MB]. Registration of this study was not required, as there is no process for doing this at present for systematic reviews of outcome measures.

Design

Evidence synthesis

A systematic review that will guide the production of a childhood obesity outcome measures framework that will be crucial in guiding researchers aiming to assess the impact of obesity treatment interventions in children. Resulting outcome measures that were identified by this review were appraised by a two-stage process of internal and external appraisal. A summary of the design is shown in *Figure 1*.

Search strategy

Two searches were performed to identify outcome measures. Search 1 identified randomised controlled trials (RCTs), pilot and feasibility studies of childhood obesity treatment evaluation studies [with the intent of identifying outcome measures (and corresponding citations) already used in trials]. Search 2 aimed to identify manuscripts describing the development and/or evaluation of outcome measures intended for use in childhood obesity intervention evaluations.

Both searches were conducted from August 2011 to October 2011 in 11 databases, including MEDLINE; MEDLINE In-Process and Other Non-Indexed Citations; EMBASE; PsycINFO; Health Management Information Consortium (HMIC); Allied and Complementary Medicine Database (AMED); Global Health, Maternity and Infant Care (all Ovid); Cumulative Index to Nursing and Allied Health Literature (CINAHL) (EBSCOhost); Science Citation Index (SCI) [Web of Science (WoS)]; and The Cochrane Library (Wiley) – from the date of inception, with no language restrictions.

Search 1 terms: identification of childhood obesity treatment intervention evaluations

Search concepts included obesity terms *and* child terms *and* evaluative studies terms. The evaluative studies search consisted of focused 'text-word' and subject heading searches (MeSH: exp clinical trial/, or evaluation studies/or meta-analysis/or validation studies/, Randomised Controlled Trials as Topic/). Child obesity terms identified in the Cochrane Review¹ were also incorporated where appropriate. In addition to full RCTs, pilot and feasibility trials were searched. Differences in the configuration of databases in particular for the subject heading searches, led to slight adaptations of the terms used.

Search 2 terms: identification of studies describing the development or evaluation of relevant outcome measures

Search concepts included obesity terms *and* child terms *and* outcome measure properties terms. The search terms for 'obesity' and 'child' searches replicated those in search 1. Studies evaluating outcome measures are recognised as difficult to identify owing to a lack of appropriate indexing terms and highly inconsistent indexing (and text) terms used across database records. The 'outcome measures properties' search was adapted from the validated sensitive search filter developed by Terwee *et al.*⁷ Terwee's filter⁷ offers a 97.1% sensitivity of retrieving all relevant documents and a precision of 4.4% (references that

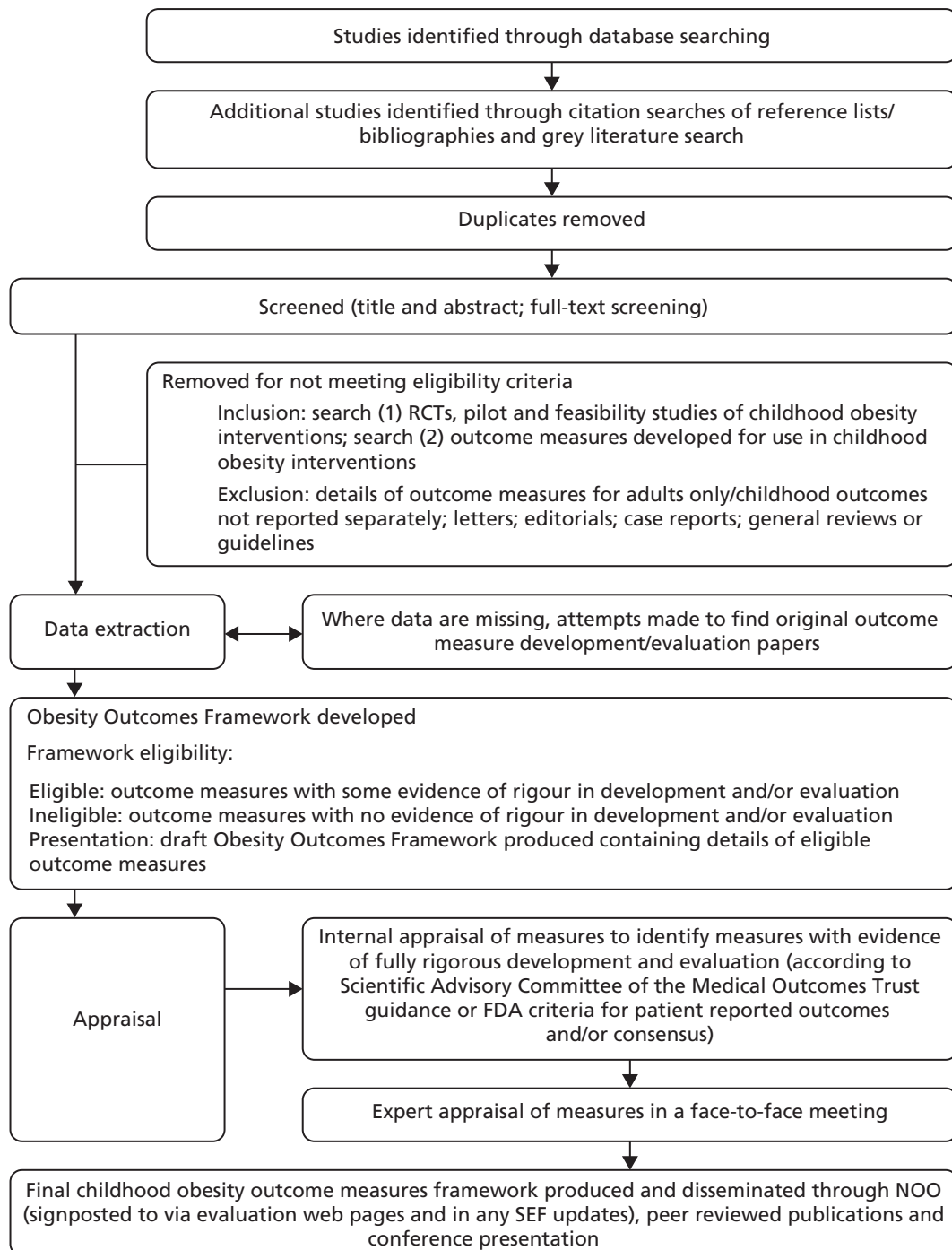


FIGURE 1 Summary of study design. FDA, Food and Drug Administration.

pass the screening stage). Again slight adaptations were applied for specific search terms to meet requirements of each database.

MEDLINE search strategies used are provided in *Appendix 1* (search 1) and *Appendix 2* (search 2) as examples of typical strategies used.

Grey literature and evidence from clinical trials databases

Additional searching was conducted via citation searches of studies that satisfied the inclusion criteria for search 1 – in particular, references on outcome measures used and cited in the relevant treatment interventions from identified childhood obesity treatment evaluations. Relevant reviews picked up in either

search were examined to identify any other additional relevant articles. Unpublished literature was obtained by grey literature, which was sought by searching a range of relevant databases including Inside Conferences, Systems for Information in Grey Literature (SIGLE), Web of Science Conference Proceedings Citation Index-Science (Thomson) and ClinicalTrials.gov. The same eligibility criteria were applied for each of these additional sources.

Data management

Search results were combined and stored in an EndNote® library (Thomson Reuters, CA, USA), and duplicates were identified and removed. Results of the abstracts and full-text screenings were recorded in the EndNote Library and appropriately filed (i.e. by inclusion/exclusion according to outcome domain).

Eligibility criteria

Childhood obesity treatment evaluation studies

Study design

Primary research of obesity *treatment* intervention evaluation studies including: RCTs, pilot studies and feasibility studies (with the intention of carrying out RCT). Although a quality assessment was not made on search 1 papers, the decision to focus on only these designs was based on the capacity of the study to deliver the results in a timely fashion. However, identified papers with pre–post study designs were retained and are available on request.

Sample

Any childhood study population (≤ 18 years at baseline). Studies with special populations (i.e. those with a cause of obesity such as Prader–Willi syndrome) were included.

Type of interventions used

Any intervention to treat obesity, including drug and surgery interventions. These are defined according to categories of strategies set by a Cochrane Review of childhood obesity treatment trials:¹

- lifestyle (dietary, PA and/or behavioural therapy interventions)
- drug (orlistat, metformin, sibutramine, rimonabant)
- surgical interventions.

Types of outcome measures used

All studies had to have obesity reduction as a primary outcome, as measured by any of the following methods:

- body mass index (BMI)/(also known as Quetelet index)
- waist circumference (WC)
- waist-to-hip ratio (WHR)
- skinfold thickness (SFT; multiple sites or one site – measured with calipers)
- mid-arm circumference
- dual-energy X-ray absorptiometry (DXA)
- bioelectrical impedance analysis (BIA)
- hydrodensitometry weighing
- near-infrared interactance (NIR)
- BOD POD (air displacement)
- total body electrical conductivity (TOBEC)
- magnetic resonance imaging (MRI)
- computed tomography (CT).

Included secondary outcomes are shown in *Table 1* below. Against each outcome is a list of all potential types of outcome measures, which was provided as a means to support identification and categorisation of outcome measures. It was not exclusive, therefore citations from trials describing the use of additional outcome measures within prespecified domains could have been included provided all other eligibility criteria were met.

TABLE 1 Included secondary outcomes with examples of outcome measures

Outcome measure domain	Example outcome measures
Diet	Weighed food diary/record
	Estimated food diary/record
	FFQ
	Semiquantitative FFQ
	Multiple-pass dietary recall
	24-hour dietary recall
	Food intake checklist [i.e. specific food/groups (e.g. F&V intake checklist)]
	Diet history
	Diet observation (DVD or direct observation)
	DLW
Eating behaviour	Dietary nitrogen
	Eating behaviour checklists
	Eating disorders questionnaires/observations
PA	Feeding styles questionnaires
	Activity monitor/movement sensors
	Activity diaries
	Retrospective questionnaires
	Activity recalls
	Screen time questionnaires
Sedentary behaviour/time	Direct observation (recorded or researcher conducted)
	Television questionnaires
	Screen time questionnaires
	Activity monitor/movement sensors
	Direct observation

TABLE 1 Included secondary outcomes with examples of outcome measures (*continued*)

Outcome measure domain	Example outcome measures
Fitness	HR (resting and/or recovery)
	Aerobic capacity/agility (step test, shuttle runs, sprints, timed/endurance runs/walk/bike)
	DLW
	Respiratory exchange ratio
	Packed cell volume
	Muscular strength
	Muscular endurance
	Flexibility
Psychological well-being	Self-esteem
	Self-perception
	Depression
	Anxiety
	Behaviour
	Psychiatric dysfunction
	Perceived competence
	Body image
HRQoL	Quality-of-life scales
Environment	Geospatial (food/retail outlets)
	Built environment (e.g. neighbourhood layout)
	Home environment [physical (e.g. food availability) and social (e.g. rules and policies)]
Physiological	Blood pressure
	Metabolic markers (e.g. lipids, glucose, insulin, leptin, adipocytokines)
	Room calorimetry (CO ₂ /VO ₂ , energy expenditure)
	Indirect calorimetry (CO ₂ /VO ₂ , energy expenditure)

DLW, double labelled water; F&V, fruit and vegetable; FFQ, food frequency questionnaire; HR, heart rate; HRQoL, health-related quality of life; VO₂, oxygen consumption.

Childhood obesity treatment evaluation studies: exclusions

1. Studies without a primary outcome of obesity reduction, such as weight loss, BMI or adiposity reduction.
2. Those with a secondary aim of obesity reduction (e.g. those with a primary aim to control diabetes).
3. Those providing details of outcome measures for adults (or childhood outcomes are not reported separately).
4. Obesity prevention studies (or designs other than those listed in the inclusion criteria, including letters, editorials, commentaries, dissertations, books, errata, notes, introductory, conference proceedings, meeting abstracts* and case reports).
5. General reviews or guidelines [unless specifically about the evaluation of childhood obesity treatment interventions (e.g. Luttikhuis *et al.*¹)].
6. Papers without sufficient information to determine eligibility (where author cannot provide missing information).
7. Those not specifically focusing on all obese subjects for intervention. Sample must all be obese and not just a proportion (e.g. obesity prevention studies with a subsample of obese).
8. Maintenance studies that are retrospective to studies previously carried out.
9. School-based interventions considered only if the sample is obese and/or stratified, i.e. treatment.
10. Phase I testing for drug trials (i.e. safety, tolerance, effect).

[*Conference proceedings and meeting abstracts were considered for specific conferences only as part of the grey literature search in search 2 (see below).]

Outcome development/evaluation methodology studies

Study design

Methodological studies describing the development (e.g. conceptual framework) and evaluation of outcome methods, including quantitative measurement, qualitative assessment, feasibility and psychometrics.

Sample

Participants must be obese or results have been stratified by weight status (presenting results separately in obese), or measures had to be developed, modified or utilised for children (≤ 18 years at baseline). Studies with special populations (i.e. those with a cause of obesity such as Prader–Willi syndrome) were included.

Type of outcome measures

In line with study aims, outcome measures were eligible if they had been (1) previously used as outcomes in a trial (i.e. cited in search 1 trials) or (2) developed for childhood obesity research. The latter was defined by demonstration of the following: (1) the underlying concept for development was based on measurement within childhood obesity; (2) the development/evaluation was conducted in overweight or obese children; or (3) the results were stratified by weight status categories.

The exception was with primary outcome measures, in which manuscripts were not included purely on the basis that they had been used previously in a childhood obesity treatment trial. Given the wealth of literature describing these methodologies, Childhood obesity Outcomes Review (CoOR) eligibility for those identified in search 2 (methodology papers) were applied. As they were unlikely to be developed specifically for childhood obesity research, manuscripts describing primary outcome measures were eligible only if they conducted evaluation in an overweight or obese sample (or stratified results by weight status category).

Outcome development/evaluation methodology studies

1. Not primary research (letters, editorials, case reports, general reviews).
2. Papers with no data relating to children unless there is evidence that they have been modified or utilised for children.
3. General reviews [unless specific to outcomes in childhood obesity research (e.g. Bryant *et al.*⁶).
4. Papers without sufficient information to determine eligibility (where the missing information cannot be sourced from the manuscript authors).
5. Comparisons of different cut-off points or population equations [e.g. World Health Organization (WHO), International Obesity Task Force (IOTF) and Must *et al.*⁸].
6. Standards of population-based criteria.

Data extraction process

Data from studies fulfilling the systematic review eligibility criteria were extracted on to prepared standardised data extraction forms. Where data were missing, attempts were made to find the original outcome measures papers with data pertaining to the development and evaluation.

There were two phases of data extraction:

- *Phase I* Trial description extraction (search 1)
- *Phase II* Outcome measure methodology extraction (search 2 and citations from search 1).

Phase I: Trial description extraction

A description of papers fulfilling the eligibility criteria for search 1 was entered on to a trial specific data extraction form (see *Appendix 4*). Three versions of paper-based forms were initially piloted until a final form was created and incorporated into the 'Bristol Online Survey' (BOS: www.survey.bris.ac.uk). This enabled relocation of all data into an Microsoft Excel 2010 database (Microsoft Corporation, Redmond, WA, USA). Two modes of extraction (electronic and paper based) were conducted for all manuscripts.

Phase II: Outcome measure methodology extraction

This phase of data extraction included papers that were identified through search 2 (methodology papers) and papers that were located following a citation search of search 1 (intervention studies) (i.e. sourcing methodology papers that were cited for each of the measures provided by the evaluation studies). Separate data extraction forms were developed for the extraction of each outcome domain, as the methodology to develop and evaluate measures differs. For example, whereas it is common to conduct internal consistency (IC) on questionnaire measures, this is not appropriate for non-survey/questionnaire measures. Similarly, gold standard comparators are dependent on the type of measure. As an example, *Appendix 5* provides the data extraction form for the diet domain. Extraction forms for other domains are available on request.

Each data extraction form began with gathering detailed information on the characteristics of the manuscript (authors, year of publication), study (e.g. country of origin) and sample (e.g. age, ethnicity). Where possible, predefined categorical responses were developed to avoid the need to code open response data. Extraction forms then went on to gather information related to outcome measurement development (e.g. conceptual framework, involvement of users), reliability, validity, responsiveness and feasibility. Again, predefined categorical responses were developed as appropriate. Specific sections within reliability included internal reliability (e.g. IC), test–retest (TRT) reliability and inter-rater reliability. Validity sections included internal validity [e.g. factor analysis (FA)], criterion validity (with prespecified 'permitted' gold standard/criterion measures), convergent validity [described here as the association with another measure, aimed at assessing the same or similar construct(s)], and construct validity (i.e. ability of a measurement tool to measure the concept being studied). Data describing face and content validity were also extracted but were considered to be part of the outcome measurement development.

Sample size was recorded for each type of evaluation. Validity and reliability evidence was extracted for each questionnaire scale or category where available. Overall means and ranges were also extracted if provided by authors. Otherwise, these were derived from data provided in manuscripts. Mean (and ranges) were then entered into domain-specific tables for each study.

The Bristol Online Survey was not used in extraction of methodology because of difficulties in amending the on-line form once it had gone live. Given the volume of data to collect across 10 domains (resulting in several rounds of piloting of the forms), the team decided to extract data using paper forms, which were then entered directly in Excel.

Unlike all other outcome domains, evaluation of anthropometric tools is generally limited to assessment of 'criterion' validity. As this domain also had multiple papers describing the evaluation of the same measures, it was not necessary to repeatedly extract full information on the method itself. Instead, key findings related to the population and the validation were extracted. Often, this information was available within the article abstract, although reviewers extracted information from other parts of the manuscripts as appropriate.

Appraisal of quality of outcome measures

Each outcome measure was appraised for quality in order to identify those that demonstrate rigorous methods in both development and evaluation procedures. Appraisal involved two stages: (1) internal appraisal and (2) external appraisal.

Internal appraisal

Principles of international guidelines^{9,10} were drawn on (where appropriate) to appraise rigour (i.e. development and measurement properties) of outcome measures meeting eligibility criteria. Measures within outcome domains were specifically appraised according to its construct and/or clinical context, as strict adherence to any individual guideline is not always appropriate. For example, many anthropometric and physiological outcomes are derived from standard clinical tests, and it would therefore be unlikely to find published data on measurement development in relation to childhood obesity; thus anthropometric outcome measures were not expected to have involved obese children in the development stage.

Specific international guidelines that were used in developing the data extraction and scoring systems were the Scientific Advisory Committee (SAC) of the Medical Outcomes Trust guidelines¹¹ and Food and Drug Administration (FDA) guidelines on the development of patient-reported outcomes (PROs).¹⁰ The SAC defines key attributes that should form part of the development and evaluation of instruments. With this, there are clear rules on what the committee considered to be important in the reporting of a reproducibility or validation study (e.g. a clear description of the methods of data collection and reporting of specific estimates and standard errors). In addition, standards for evaluation are provided, such as for assessment of reliability and some criteria for good measurement properties, including cut-off points for intraclass correlations. These criteria were used as a guide rather than explicitly regulating which measures were and were not considered as rigorously developed and evaluated. FDA guidance describes the best practice in the review, and evaluation of existing, modified or newly created PRO instruments. The criteria helped to guide appraisal procedures related to the conceptual framework (definition of the concepts being measured with description of relationships between items/domains and scores) and measurement properties (reliability, validity, ability to detect change) of each measure. Specific characteristics that were included in the CoOR appraisal method include concepts being measured, number of items, conceptual framework, intended use, population for intended use, data collection method, administration mode, response options, recall period, scoring, weighting, format and response burden.

A scoring system was also applied to the development and evaluation of each secondary outcome measure. Scores were based on quality in the conduct and results of evaluation where appropriate and ranged from '1' to '4' (with '1' being the lowest). These were developed from criteria set by the

international guidelines,^{9,10} in addition to previous research conducted by the lead applicant (MB). For example, in reporting the study sample, a maximum score of '4' was assigned to manuscripts reporting a minimum of the four characteristics: age, gender, ethnicity and socioeconomic status (SES). Those describing three of these were assigned a score of '3' and so on. *Appendix 5* provides the data extraction form for the diet outcome domain in which the scoring system is fully detailed. In addition, *Table 2* provides criteria that were applied in assigning scores.

TABLE 2 Criteria used to allocate robustness scores for evaluation of quality

Measurement development and reporting			
The concept to be measured was clearly stated (rationale and description)		4 = strongly agree (concepts are named and clearly defined)	
		3 = agree (concepts are named and general described)	
		2 = disagree (concepts named only but not defined)	
		1 = strongly disagree (concepts are not clearly named or defined)	
Was a theoretical or conceptual framework used or referenced?		4 = strongly agree (theory/framework used as a basis for development)	
		3 = agree (theory/framework named and incorporated)	
		2 = disagree (theory/framework named but not used)	
		1 = strongly disagree (no theory/framework described)	
		0 = N/A = (biochemical/anthropometry, direct measures/observations)	
Populations that the measure was intended for were adequately described		4 = strongly agree (describes at least four characteristics, including age, gender, race/ethnicity and SES)	
		3 = agree (three characteristics reported)	
		2 = disagree (two characteristics reported)	
		1 = strongly disagree (no characteristics reported)	
Were the populations for which the measure was intended involved in measurement development?		4 = strongly agree (at least three methods of involvement, including part of study team, steering committee, pilot testing, cognitive interviews/focus groups)	
		3 = agree (involved using at least two methods)	
		2 = disagree (populations minimally involved in one method)	
		1 = strongly disagree (populations not involved)	
		0 = N/A (biochemical/anthropometry)	
Measurement evaluation			
	Sample size	Appropriate statistics ^a	Results/findings
IC	Five or more participants per item	Cronbach's alpha KR-20 (Kuder–Richardson coefficient) Split half	$\alpha = 0.7$
TRT reliability	≥ 50	Spearman Pearson Kappa Agreement	$r = 0.4$ $\kappa = 0.4$ Agreement (not used to score but reported for comparisons)

continued

TABLE 2 Criteria used to allocate robustness scores for evaluation of quality (*continued*)

Measurement development and reporting			
Inter-rater reliability	Study specific (depending on design)	Pearson/ICC/rho = kappa K = Krippendorff's alpha	$r = 0.4$ $\kappa = 0.40$
FA	Five or more participants per item	Eigenvalue Factor loading %variance	Eigenvalue ≥ 1 Factor loading = high > 0.6 , low < 0.4 CFA RNSEA < 0.06 , RNI close to 1
Criterion validity	≥ 50 [less for objective such as DLW (≥ 20)]	Pearson Spearman Regression Agreement <i>t</i> -test (not in isolation) ANOVA	Pearson's/Spearman ≥ 0.4 Regression coefficient = $p > 0.5$ or $r \geq 0.50$ Agreement <i>t</i> -test $p > 0.05$, <i>t</i> -value > 1
Convergent validity	≥ 100	Sensitivity/specificity Pearson Spearman Regression Agreement <i>t</i> -test (not in isolation) ANOVA	AUC > 0.7 Pearson/Spearman ≥ 0.4 Regression coefficient = $p > 0.5$ or $r \geq 0.50$ Agreement <i>t</i> -test $p > 0.05$, <i>t</i> -value > 1
Construct validity	≥ 100	Sensitivity/specificity Pearson Spearman Regression Agreement <i>t</i> -test (not in isolation) ANOVA	AUC > 0.7 Pearson's/Spearman ≥ 0.4 Regression coefficient = $p > 0.5$ or $r \geq 0.50$ Agreement <i>t</i> -test $p > 0.05$, <i>t</i> -value > 1
Responsiveness	≥ 100	Sensitivity/specificity MCID SRM ROC AUC ES <i>t</i> -test	AUC > 0.7 MCID/SRM > 0.5 ROC AUC > 0.7 ES > 0.5 <i>t</i> -test $p < 0.05$

ANOVA, analysis of variance; AUC, area under the curve; CFA, confirmatory factor analysis; ES, effect size; ICC, intraclass correlation coefficient; MCID, minimal clinically important difference; N/A, not applicable; RNI, reference nutrient intake; RNSEA, root-mean-square error of approximation; ROC, receiver operating characteristic; SRM, standardised response mean.

a The protocol for consideration of statistical tests that were not listed included consideration by the team statistician (JB).

It was not feasible to assign scores to all of the anthropometry (primary) outcome measure studies. The majority of manuscripts meeting criteria for eligibility evaluated multiple measures, which would mean that scores would have to be provided for an amount of studies that was beyond the capacity of this study (estimated to be > 300 studies). This was also deemed inappropriate, as multiple studies evaluated the same measures (generating multiple scores for the same measures). Instead, CoOR members grouped all manuscripts evaluating the same measure and reported the overall conclusions (reported by authors) of each paper as: (1) yes, authors advocate its use; (2) no, authors do not advocate its use; and (3) conclusions drawn by authors are unclear (?). The form used to record this information is provided in *Appendix 17* for clarity. (Note: This also provides the findings.)

Internal recommendation of measures to include or exclude (degree of certainty)

Two members of the CoOR internal team (MB and LA) classified each of the primary and secondary measures into one of three categories (by discussion and consensus) in relation to their confidence of whether or not each measure should be recommended for inclusion into the final CoOR outcome measures framework: (1) 'certain, good evidence, fit for purpose' (i.e. confident that the measure is robust and should be recommended for use); (2) 'certain, poor evidence, not fit for purpose'; and (3) 'uncertain, requiring further consideration'. Assignment of certainty considered the data extracted from each study alongside the scoring system. For example, a measure that was assigned a score of 3 out of 4 for quality of reliability testing was further investigated to determine why one point was lost. If lost because of poor reporting methodology, the team may have been more likely to deem a measure 'uncertain' rather than 'unfit' than lost points due to poor results or inadequate sample size. This was conducted separately for each domain in order to facilitate comparisons between measures (i.e. questionnaire-style outcome measures would be expected to include a measure of IC, which was not applicable in objective measures. Similarly, historical physiological measures, such as blood pressure, would not be expected to have included obese children in their development). Tools were placed into Category 1 or 2 only, providing that mutual agreement had been established. Category 1 was assigned only when the tool was clearly highly robust in terms of development and evaluation. Similarly, Category 2 was assigned only when the tools was very poorly developed and evaluated. Any disagreements were placed into Category 3 to be further discussed at the expert appraisal meeting.

Expert appraisal

Results of the systematic review and corresponding files from the internal appraisal were reviewed by experts with specific proficiency in each outcome, in addition to methodological experts. Each expert was asked to review all of the included outcome measures that met eligibility criteria of CoOR, as well as considering the internal appraisal decisions. *Figure 2* shows the process in which external appraisal was conducted.

In Phase I, experts were provided with all materials (via a web-based file share facility: Dropbox). Provided documents included:

1. A list of all included manuscripts (with information on the pathway in which each was included). This included manuscripts that did not fully meet eligibility criteria but which the internal team felt had potential for inclusion.
2. PDFs of all manuscripts meeting eligibility criteria (with copies of measurement questionnaire if available).
3. Summary tables providing details of all data that were extracted for each measure according to domain (see *Appendices 6–15*).
4. Tables providing internal scoring for development and evaluation of each measure (see *Appendices 18–26*).
5. Appraisal decision of certainty for each measure [see *Appendix 17* (primary) and *Appendix 28* (secondary)].

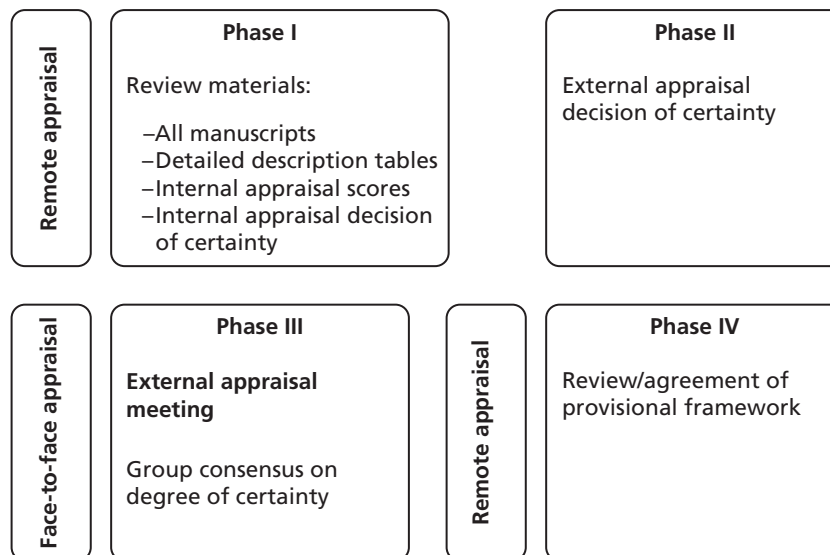


FIGURE 2 Expert appraisal process.

Experts were asked to look at material for all 10 domains. As part of Phase II, they were then asked to more closely examine documents within areas of their expertise (predefined by the CoOR team), so that they could lead discussion in these domains at a future face-to-face meeting. Experts involved were Susan Jebb and Carolyn Summerbell (diet and eating behaviours), John Reilly (anthropometry/weight status), Ashley Cooper and Ulf Ekelund (PA and sedentary time/behaviour), Lucy Griffiths and Andrew Hill (psychological well-being), Maria Bryant and Steven Cummins (environmental outcomes), Paul Kind (economics/quality of life), and Julian Hamilton-Shield (physiological outcomes). Two further consultants with expertise in outcome evaluation and clinical trial methodology reviewed the framework (Claudia Gorecki and Julia Brown, respectively). In addition, a specialist in public health evaluation from the NOO (Katharine Roberts) facilitated in consideration of measure applicability for public health interventions.

Experts were provided with instruction asking them to consider factors such as appropriateness of categorisation (i.e. ensure within correct outcome domain); obvious omissions not identified by search strategy (including knowledge of modified versions of outcomes); and personal and theoretical experience of use of outcome measures related to feasibility.

Phase III of the external appraisal involved a face-to-face meeting with all experts. A physical (rather than a remote) meeting was chosen because it was more likely to create a richer, in-depth discussion of the inclusion (or exclusion) of all outcomes. Experts were provided with a short presentation by the CI (MB) describing the study aims and methodology. They were then divided into two groups. Group A included experts for the domains: diet, eating behaviour, psychological well-being, economics/health-related quality of life (HRQoL) and environment. Group B included experts from the domains of anthropometry, PA, sedentary behaviour/time, fitness and physiology. Discussions began by determining expert agreement on the internal appraisal decisions '1' (certain, fit for purpose) and '2' (certain, unfit for purpose). Disagreements were resolved by discussion. Outcome measures that had been given an internal appraisal decision of '3' (uncertain, requiring further consideration) were then more fully discussed. Justifications for decisions were provided at the meeting and final rulings of the tools were made based on consensus. This was recorded directly on to a predefined pro forma that permitted the recording of internal and external decisions (see *Appendices 16* and *27*), alongside any relevant discussion. In addition, discussions from both groups were recorded and transcribed.

After each group had made decisions regarding certainty, a final discussion was held by both groups together to review key decisions. All final decisions contributed towards the development of a provisional framework, which was then forwarded to each expert to secure their final agreement (Phase IV).

Note: At the time of the expert appraisal meeting, data from some of manuscripts had not been extracted. These included those that had to be ordered by The British Library and which had not yet been delivered to the team. The exact same methodology was later applied to these manuscripts; however, experts were asked to review them remotely. Outcome measures that were appraised using this approach are highlighted within *Appendix 16*. The exception to this was with manuscripts written in languages other than English. Where possible, data were extracted via translation of methodology papers. However, these were not appraised for quality.

Chapter 4 Results

Number and type of studies identified

Combined, searches 1 and 2, conducted in 11 databases, identified 25,486 manuscripts (after removal of 8674 duplicates). A further 25 were identified through hand-searching [grey literature, citations and references from relevant reviews (including manuscripts cited in 48 reviews)]. Of these, 14,419 were search 1 trial manuscripts and 11,092 were search 2 methodology manuscripts. Screening for eligibility at both the title and abstract stage and the full paper review resulted in the inclusion of 200 trial manuscripts from search 1. After data were extracted from these papers, 417 further manuscripts were identified that were citations linked to the outcome measures used by the trials. However, only 56 cited methodology manuscripts met eligibility criteria for inclusion as methodology papers. The majority of other citations were linked to a previous study using the outcome measure (i.e. not papers describing development or evaluation) or were completely incorrect citations (28 were duplicates, already found in search 2). Screening of search 2 methodology papers resulted in the inclusion of 320 manuscripts meeting eligibility criteria. Combined with search 1, a total of 376 manuscripts were identified that described 180 outcome measures (*Figure 3*).

Note, although this study did not exclude manuscripts that were not written in English, there was no formal protocol for translation or extraction of papers. Eligible manuscripts written in languages other than English ($n = 53$) that were identified via search 1 are listed in *Appendix 27* but data have not been extracted from them. Manuscripts written in languages other than English ($n = 23$) that were identified via search 2 (i.e. pertaining directly to development/evaluation of outcome measures) were included for data extraction. These are listed within study findings and the language is indicated in the detailed summary tables (see *Appendices 5–14*). However, as the level of extraction was not as detailed as with English papers, measures described by these papers were not considered in appraisal unless already included within another study manuscript written in English.

Number and type of studies excluded, with reasons

In search 1 (of trials evaluating obesity treatment interventions), a large number of identified studies (almost 13,000) were not eligible for inclusion when screened by title and abstract. Description of the reasons for exclusion for each of these has been noted and is available on request, but is not feasible to provide here (non-eligible manuscripts are also listed in supplementary on-line material). Details are provided for the 1175 manuscripts from search 1 that were excluded at full-text screening. Of these, 200 papers did not have a primary outcome of obesity reduction, 30 had a secondary aim of obesity reduction and 85 papers focused on the prevention of childhood obesity. The sample in 465 of the papers was reported in adults or was not reported by children separately. Three hundred and fifteen manuscripts reported a non-eligible study design and one paper was a Phase I trial for drug testing. In 20 papers a pilot study was implemented but failed to express any intentions of producing a future RCT. Twenty-eight did not specifically focus on all obese children for the intervention (i.e. school-based interventions with a subsample of obese). Twenty-one papers were weight maintenance evaluations, with most investigating the long-term success of interventions that had already been identified. Eight manuscripts described studies that had already been published (i.e. several publications coming from the same trial). A further two papers were without sufficient information to determine eligibility. Two reviewers independently screened manuscripts for eligibility (MB and LA). To ensure consistency, the first 132 articles were reviewed by both people, which resulted in an agreement of 98% (two disagreements). Issues related to these disagreements were discussed and the protocol was amended as appropriate.

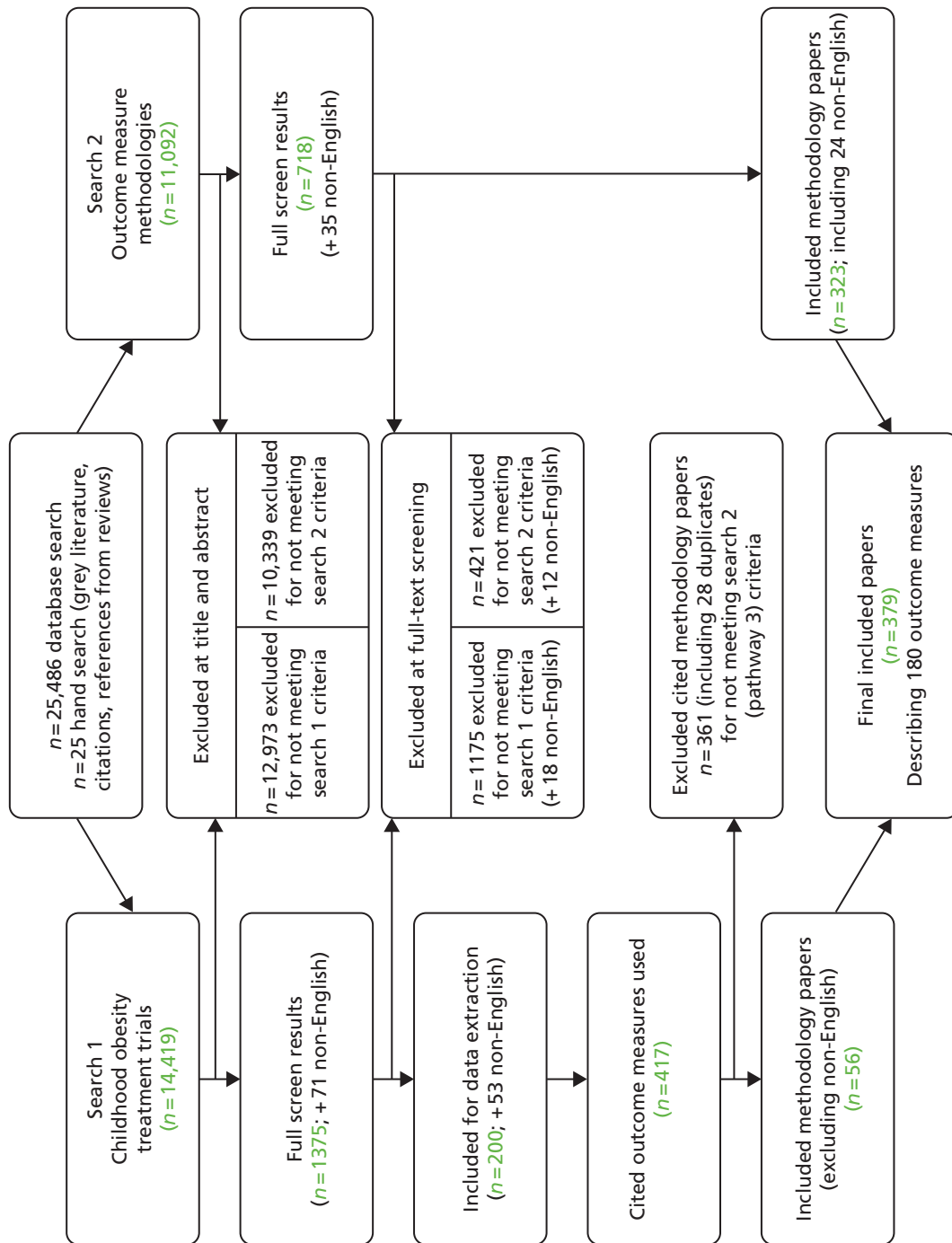


FIGURE 3 Summary of study selection and exclusion.

In search 2 (methodology papers), 421 manuscripts failed to meet the inclusion criteria at full-text screening and were excluded from the review. Of these, 107 papers had an ineligible design (with no assessment of development and/or evaluation of outcome methods for childhood obesity treatment intervention evaluations), and seven papers conducted minimal psychometric testing and the development/evaluation was not the main aim of the paper. One hundred and seventy-one papers did not include an obese sample or results were not stratified by obese. In 95 papers, outcome measures were developed for adults and had not been modified for children. Six manuscripts described studies that were not primary research (i.e. reviews, editorials, case reports, etc.), 19 manuscripts compared cut-off thresholds or population equations (e.g. WHO vs. IOTF cut-offs) and 11 considered reference standards for population databases. Two further papers assessed the evaluation tools that were not outcome measures. One study included psychometric testing but results were also available in another publication, and one study included an outcome measure within a domain not specified in *Table 1*. Finally, one paper was without sufficient information to determine eligibility. Agreement between reviewers for search 2 papers was 96% (48 out of 50 agreed). Similar to search 1, issues with disagreement were resolved by discussion and the protocol was amended to clarify these issues.

Study characteristics

Manuscripts describing childhood obesity treatment trials

Data were extracted from 200 manuscripts describing the evaluation of a childhood obesity treatment intervention (see *Appendix 3*). The majority (156 manuscripts) described a phase III evaluation of a childhood obesity treatment. Nine manuscripts described a feasibility study, 30 manuscripts described a pilot study and nine manuscripts were protocol papers for future RCTs. Publication dates ranged from 1960 to 2012, and included sample sizes ranging from 8¹¹ to 2112.¹² Most studies evaluated a lifestyle intervention, but there were also evaluations of cognitive interventions, drug and surgical interventions, drug/surgical interventions combined with lifestyle change and those that focused on reducing sedentary behaviours. *Figure 4* shows the different types of primary outcome measures used by identified trials. The most common primary outcome was BMI [including those deriving body mass index standard deviation score (BMI-SDS) or %BMI]. However, measurement of weight was also popular, with 37 evaluations assessing absolute weight or percentage weight change as the primary outcome.

Eighty-two (41%) of trials included a measure of diet as a secondary outcome. Sixty-eight (34%) studies included a measure of PA, with the most popular measures being activity recalls and objective measures (e.g. accelerometers or pedometers). Seventy (35%) of the trials included an evaluation of psychological

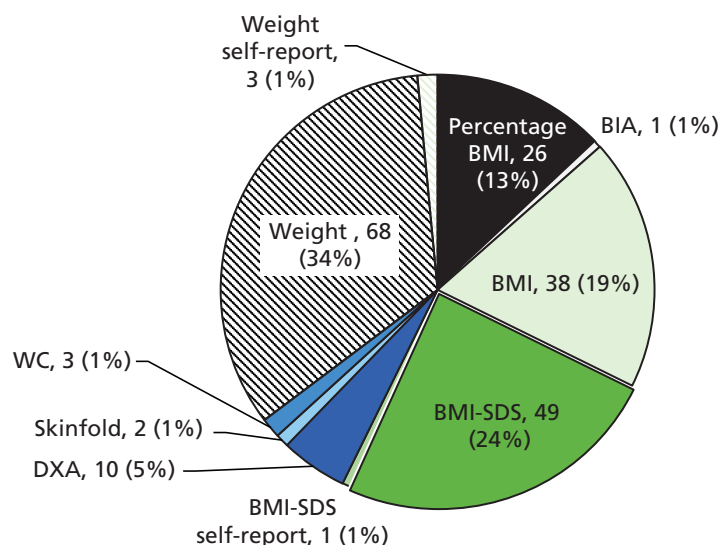


FIGURE 4 Frequency (%) of primary outcome measures used in search 1 trials.

well-being, measuring a variety of concepts, including self-esteem, depression and body image. Physiological measurement was also popular, with 94 (47%) trials measuring outcomes such as blood pressure, insulin or blood lipids. Other secondary outcomes were used less frequently (Figure 5).

Four hundred and seventeen citations that were linked to primary and secondary outcome measures within all of the 200 included manuscripts were located. However, only 56 of these referred to manuscripts that described the development and/or evaluation of outcome measures. Incorrect citations were linked to the majority of outcome measures, most commonly linking to a previous study that had used the same measure.

Manuscripts describing the development/evaluation of outcome measures

A total of 379 manuscripts that describe the development or evaluation of 180 measures met inclusion criteria to CoOR. Fifty-six of the included manuscripts were derived from searching citations of the trials (from search 1) and the remaining 323 were identified directly from search 2. Of these, 24 were written in a language other than English. Efforts were made to translate these (and gain information from English abstracts), resulting in the inclusion of all except for three studies.^{13–15} A further paper that was not translated describes a measure that has already been included within the eating behaviour domain.¹⁴ It has been included in the summary table (see Appendix 8), but no data have been extracted from this paper. Table 3 provides detail on the number of manuscripts and corresponding measures (excluding the three written in non-English that could not be translated). Some manuscripts evaluated more than one measure (hence there is a discrepancy between the number of manuscripts and the number of studies). In addition, some measures have multiple manuscripts describing their evaluation, thus the number of manuscripts and number of measures are not equal.

Findings of the systematic review

The following text summarises data extraction of measures pertaining to evaluation of reliability and validity within outcome domains. Key findings are provided with 'in-text' citations for some manuscripts. However, given the volume of included manuscripts, not all are cited within the text. However, full details of data extracted from every manuscript are provided in the corresponding Appendices 5–14 and within the reference list.

Anthropometry

Data from a total of 162 papers with 38 tools were extracted (see Appendix 6). Of these 162 manuscripts, 15 were written in a language other than English. Data were extracted only from abstracts (which were

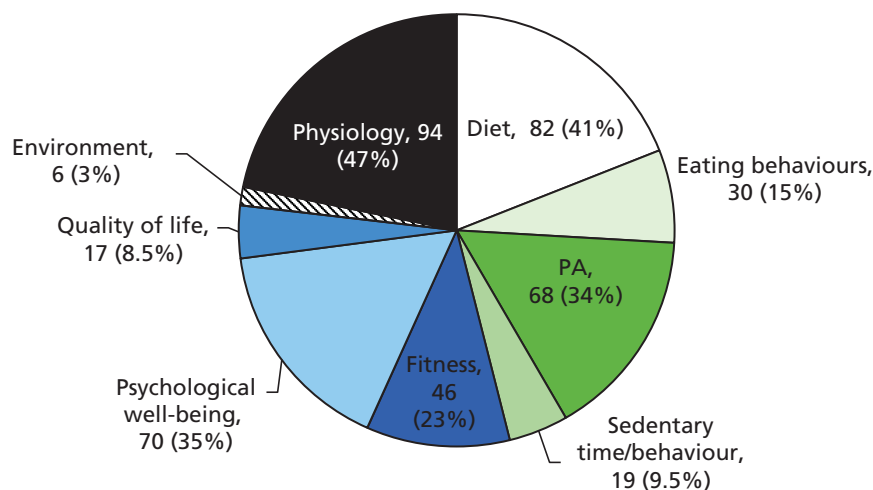


FIGURE 5 Frequency (%) of trials using each type of secondary outcome.

TABLE 3 Number of eligible manuscripts with corresponding measures by outcome domain

Outcome domain	No. of manuscripts	No. of studies	No. of measures
Anthropometry	162 (including 15 non-English)	258	38 (none exclusively non-English)
Diet	40	44	22
Eating behaviour	39 (including one non-English)	40	22 (none exclusively non-English)
PA	35	45	24
Sedentary time/behaviour	5	6	6
Fitness	14	14	13
Physiology	28 (including two non-English)	28	12 (none exclusively non-English)
HRQoL	25 (including three non-English)	25	16 (including three non-English)
Psychological well-being	19	20	17
Environment	9	10	10
Total	376	490	180

available in English); however, all non-English papers described further evaluation of outcome measures that were also described in multiple other papers written in English. Appraisal decisions to 'recommend' or 'not recommend' tools were therefore not based on non-English papers.

Of the 162 papers, only eight evaluated the validity of primary outcomes against a gold standard measure of body composition using either the four-compartmental (4C) model^{16–19} or total body water (TBW) by deuterium dilution.^{20–24} Each of the four papers using the 4C model as a gold standard describes the validation of DXA [with Gately *et al.*¹⁷ also validating air displacement plethysmography (ADP) and total body water]. Wells *et al.*¹⁶ and Gately *et al.*¹⁷ validated DXA in 174 and 30 overweight and obese adolescents, respectively. Findings were similar, with Gately *et al.*¹⁷ finding that the total error and mean difference [\pm 95% limits of agreement (LOA)] compared with the 4C model were 2.74 kg and 1.9 kg (\pm 4.0 kg), respectively, and Wells *et al.*¹⁶ finding similar LOA at \pm 4.2 kg, with overestimations of fat mass by DXA of 0.9 kg. However, interpretation of these results differs by authors, with Wells *et al.*¹⁶ applying more caution to the validity of DXA. Additionally, Wells *et al.*¹⁶ showed that the bias in fat mass was significantly related to the magnitude of fat mass (so that greater inaccuracies were seen with increasing fat mass). Further longitudinal analysis was conducted by Wells *et al.*¹⁶ in a subsample of 66 children. Although average bias was not found to differ significantly from zero for 'change' in both lean mass and fat mass, the LOA in individuals were described as 'large' (\pm 3 kg) compared with an average weight change of 1.7 kg (lean mass) or 0.6 kg (fat mass). Combined with problems encountered in actually using the equipment in very obese children, authors conclude that further work (including investment by companies manufacturing DXA machines to develop technology capable of measuring obese participants) may be required to enhance measurement accuracy. Variability in accuracy in DXA according to other factors was also found by Williams *et al.*¹⁸ in a study that compared groups of obese children, 'normal' weight children and children with cystic fibrosis. Bias in measurement was found according to the sex, size, degree of adiposity and disease state of the subjects, indicating that DXA is unreliable for studies of persons who undergo significant changes in nutritional status between measurements (comparisons with obese children were based on 28 children). The final paper identified by CoOR in which DXA was validated against the 4C model also highlighted limitations of the method, although concluded that it remains of use in longitudinal population comparisons.¹⁹ Comparisons were made in per cent body fat in a sample of children and adolescents and show a mean difference between DXA and 4C of -3.5% ($p = 0.171$), with LOA at $+5\%$ to -12% .

Further comparisons by Gately *et al.*¹⁷ were made between the 4C model and other anthropometric measures of ADP and TBW, finding strong correlations for all measures ($r \geq 0.95$, $p < 0.001$; standard error ≤ 2.14). The anthropometric measurement demonstrating the highest validity in this study was ADP (total error 2.5, mean difference 1.8 kg, 95% limit of agreement ± 3.5 kg for ADP with Siri equations, and total error 1.82, mean difference 0.04 kg, ± 3.6 for ADP with Loh equations).

Many studies evaluated BIA, but only two reported comparisons against the gold standard methodology of TBW by deuterium dilution.^{21,24} Wabitsch *et al.*²⁴ was also one of the few studies to measure the ability of a measure to detect change following an intervention. In comparisons between BIA and TBW, cross-sectional comparisons showed good agreement between BIA and TBW. However, correlations were poor ($r = 0.21$) with change, where BIA was not accurate at predicting small changes in TBW. Rush *et al.*²¹ also used the deuterium dilution method to compare BIA and BMI in their study of 172 children and adolescents, although the focus of the paper was actually to develop prediction equations in three ethnic groups. A further study made comparisons with a three-compartmental (3C) model²⁵ and found that BIA (using Tanita equations) overestimated fat-free mass by 2.7 kg ($p < 0.001$), although new equations by the authors improved correlations.

Fifty-five papers tested the use of BMI as a valid measure of change in body fat by deuterium dilution.^{22,23} Findings from these suggest that fat mass (from TBW) is well correlated with BMI across ethnic groups (Caucasians $r = 0.81$, $p < 0.001$; Sri Lankans $r = 0.92$, $p < 0.001$)²² and genders (girl $r = 0.82$, $p < 0.001$; boy $r = 0.87$, $p < 0.001$), but that BMI cut-offs often fail to detect obesity as defined by the gold standard methods. Use of self-reported BMI, however, often failed to produce correlations that were sufficient to suggest that they are of use for individual-level assessment, although they may be adequate to study trends on a population basis. This type of evaluation was common in the CoOR review, with 39 papers describing comparisons between self-report (or parental report) and measured height and weight.

Evaluation of SFT was also common in manuscripts identified by CoOR, with 24 studies reporting validating various types of skinfold measurements. Of these, just four studies^{26–29} present strong validity to advocate its use. However, none of these four studies validated against gold standards of the 4C model or TBW. Of the 20 studies evaluating WC reviewed here, 10 reported an adequate level of validity for WC. However, none of these made comparisons with gold standards of the 4C model or TBW.

Diet

A total of 44 studies (within 40 manuscripts) describing 22 different types of dietary assessment methodologies were extracted (see *Appendix 7*). These included 16 different food frequency questionnaires (FFQs), plus other methodologies described in *Figure 6*.

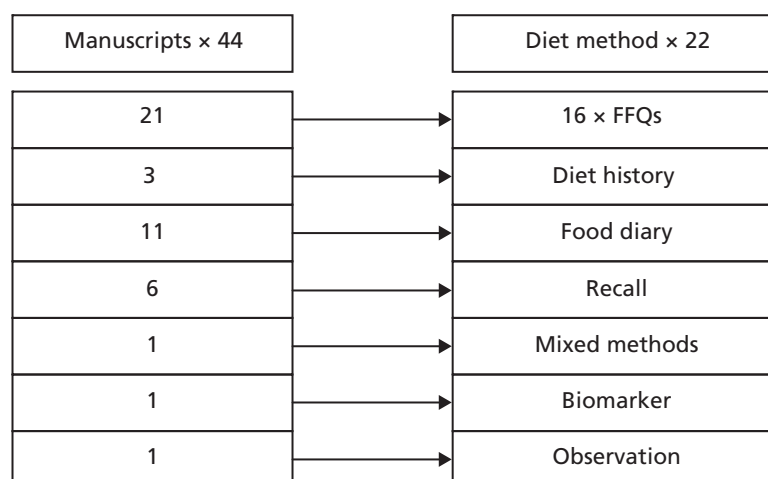


FIGURE 6 Diet methodologies included in appraisal.

Summary findings are shown in *Appendix 7*. Sixteen FFQs/checklists were described in 21 manuscripts,^{30–51} of which 10 assessed TRT reliability, with results varying across studies ($r = 0.16–0.74$). In general, however, most were classed as adequate. Convergent validity was tested in 13 studies, comparing FFQ data with 24-hour recalls and food records (weighed and estimated). Correlations ranged from 0.23 to 0.66, and kappa statistics ranged from 0.08 to 0.67. Criterion validity, comparing against the ‘gold standard’ of direct observation, measure of habitual energy expenditure by doubly labelled water (DLW) or other biomarkers was conducted in four FFQs, with correlations ranging from 0.01 to 0.91. Worryingly, large LOA were often evident in these studies. IC was tested for two FFQs, with both showing strong alpha coefficients ranging from 0.84 to 0.88.^{30–33} Construct validity was evaluated in three papers – with comparisons between the FFQs and (1) screen time,³⁴ (2) BMI,⁵² and (3) diet quality^{35,52} – showing variable, but generally significant, correlations (see *Appendix 7*).

Diet history methods were described in three identified manuscripts.^{53–55} Reliability testing was not reported in any of these papers. Each assessed criterion validity against a gold standard method, but results indicated an impact of BMI on validity. No other evaluation was reported for diet history methods.

Diet diaries were evaluated in 11 of the identified papers.^{56–64} Of these, one⁵⁶ tested inter-rater reliability using a tape-recorded method, with correlations ranging between 0.68 and 0.96. None of the papers assessed TRT reliability. Criterion validity was evaluated in 10 diet diary papers, with many reporting significant effects of weight, BMI or other measures of adiposity on validity.^{55,57–63} The only paper that reported no misreporting by body weight was O’Conner *et al.*⁶⁴ in their study of 45 children. This paper⁶⁴ reports low relative bias [mean difference, energy intake (EI) – total energy expenditure (TEE) = at 118 kJ/day] but with wide LOA (bias plus or minus two standard deviations of the difference) at 118 ± 3345 kJ/day. Bias was associated most strongly with reported fat intake.

Recall methodologies were evaluated in six papers, of which findings for reliability and validity testing were variable. Two papers reported evaluating TRT reliability.^{65,66} Edmunds *et al.*⁶⁶ compared ‘A Day in the Life’ questionnaire (a 24-hour recall method) collected twice, 2 days apart, and found non-significant differences overall, indicating good TRT reliability. Baxter *et al.*⁶⁵ conducted general-linear-model repeated-measures analysis in which diet was recorded over three time periods and included comparisons between different weight status groups. The effect of time period (i.e. repeated measures) was significant, indicating poor repeatability, with a significant interaction by weight status (with greater inaccuracy in overweight children). Comparisons of each of these methods was against direct observation of eating episodes. Two diet recall evaluation papers described different forms of inter-rater reliability,^{56,66} both providing strong evidence. Van Horn *et al.*⁵⁶ compared child report with parental report and show correlations of $r = 0.75$ (range 0.65–0.93). Edmunds *et al.*⁶⁶ made comparisons between coder and reported a kappa range of between 0.82 and 0.92.⁶⁶ Five of the six recall papers evaluated criterion validity, four of which made comparisons with direct observations^{33,65–67} and one with DLW.⁶⁸ Findings from comparisons with direct observation are difficult to compare, as each was conducted using different analytical approaches. In general, however, criterion validity using this type of comparator indicates moderate agreement (see *Appendix 7*). Johnson *et al.*⁶⁸ compared 3-day dietary recalls to data from DLW in 24 children and reported a poor correlation between reported EI and that estimated by DLW ($r = 0.25$, $p = 0.24$). LOA were -4612 ± 3356 kJ/day, with a mean difference of -225.1 kJ/day. Precision, however, was not correlated to body weight.

Three other dietary assessment methodologies meeting eligibility criteria were included. The first describes the measurement of biomarkers insulin-like growth factor 1 (IGF-1), insulin-like growth factor binding protein 1 (IGFBP-1) and insulin-like growth factor binding protein 3 (IGFBP-3),⁶⁹ and reports that, as an indicator of construct validity, overweight children had higher serum levels of IGF-1 and IGFBP-3 but lower levels of IGFBP-1. Consequently, biomarker measurement (especially IGF) is advocated by the authors. The second paper describes the development and preliminary evaluation of a dietary observation method for use within child-care settings⁷⁰ in which trained researchers attend centres to view dietary consumption by children. This paper reports excellent inter-rater reliability between observers of 100% agreement for

most food items observed. One further paper⁷¹ describes the use of a mixed-method approach, including 24-hour recall, FFQ and nutrition, and PA behaviours. The primary purpose of this paper was to compare self-reported EI across weight status groups. However, further evaluation is reported within the methods section (see *Chapter 3*) specifically for evaluation of the recall component (linked to previous abstracts). Findings indicate good reliability [TRT agreement of 77% overall (range = 62–87%) and inter-rater reliability between self-report and dietitian report of $r = 0.55$ – 0.70]. However, comparisons with direct observations (for 24-hour recall data) indicate a systematic under-reporting of dietary intake by gender and weight status.

Eating behaviour

A total of 40 studies (within 39 manuscripts), describing 22 measures of eating behaviours, met the eligibility criteria. A description of data that was extracted from all studies is presented in *Appendix 8*. Of these, one manuscript was written in Portuguese.¹⁴ It was not possible to extract data from this manuscript but the measure that it describes was evaluated by two other manuscripts.^{72,73} It is included in *Appendix 8* for reference purposes only.

Broadly speaking, eating behaviour questionnaires included those that targeted feeding styles/behaviours or those that measured affect/emotions related to eating, although some measures included both. Feeding questionnaires [e.g. Infant Feeding Questionnaire (IFQ) and Preschool Feeding Questionnaire (PFQ),⁷⁴ Child Feeding Questionnaire (CFQ),⁷⁵ Infant Feeding Style Questionnaire (IFSQ)⁷⁶] included constructs such as concern, control, difficulties in feeding, pressure, restriction, etc. Measures of emotional eating [e.g. the Emotional Eating Scale for Children (EES-C),⁷⁷ Dutch Eating Behaviour Questionnaire (DEBQ),^{78,79} Eating in the Absence of Hunger-Children (EAH-C)⁸⁰] included constructs of eating in response to emotions, enjoyment of food, satiety, external eating, etc. However, there was little consistency across measures in the terms/names provided to describe similar constructs. Eligible studies also included those that described the development/evaluation of measures that screened for disordered eating, many of which had been included because they had been previously used in childhood obesity treatment trials. The suitability of these measures was questioned at the point of review, but decisions were left to the expert collaborators group (see *Chapter 3, Expert appraisal*).

Internal consistency assessment was common in eating behaviour questionnaires and was tested in 30 studies (alpha range = 0.54–0.90). Of these, 21 were considered acceptable ($\alpha > 0.70$). TRT reliability was performed in 12 studies,^{73,77,80–89} also demonstrating high correlations ($r = 0.58$ – 0.81). Results from inter-rater reliability testing, however, were less strong; this was assessed in five studies,^{81,90–93} three of which compared child self-report with parent report.^{90–92} Johnson *et al.*⁶⁸ reported a poor agreement of 41% ($\kappa = 0.19$) for their evaluation comparing findings from the Questionnaire of Eating and Weight Patterns (QEWP) reported by adolescents (QEWP-A) and the QEWP reported by parents (QEWP-P). This was later repeated by Steinberg *et al.*,⁹¹ again demonstrating discordance between reports, with children reporting more disordered eating (sensitivity = 24%, specificity = 82% for diagnosis of overeating; sensitivity = 20%, specificity = 80% for diagnosis of eating disorders). Relatively low correlations were also observed by Braet *et al.*⁹² in comparisons between child-reported DEBQ (DEBQ-C) and parent-reported DEBQ (DEBQ-P), with a range in correlations of between $r = 0.35$ and $r = 0.45$. Agreement between parents may be more similar and this was found by Haycraft *et al.*⁹³ in their evaluation of the CFQ ($r = 0.66$, range = 0.53 to 0.78). Similarly, better inter-rater reliability was observed in correlations between interviewers, with Decaluwé and Braet⁸¹ reporting highly correlated responses (mean $r = 0.96$, range = 0.91–0.99). Implications of the poor agreement between child and parent responses may be irrelevant if using these measures as trial outcomes, however, provided that the same reporter is used at baseline and follow-up in all trial arms. Authors would also need to clarify details of reporting to permit cross-trial comparisons.

Internal validity was evaluated in 22 eating behaviour papers, with the total variance ranging from 33% to 67%, and factor loadings ranging from 0.17 to 1.51. Of these, eight papers^{73,74,77,79,80,85,94} had all factors classed as acceptable (> 0.40) (see *Appendix 8*). Where appropriate, findings were used to make alterations to items and/or scales. Criterion validity was assessed in only one study evaluating the CFQ

using the 'gold standard' of direct observation as a comparator.⁹³ Results indicated that fathers ($r = 0.33$) had a greater interpretation of child's eating behaviour than mothers ($r = 0.15$); however, these results are based on a sample size of 46. Convergent validity was more frequently evaluated but was generally restricted to diagnostic measures of eating disorders. Other measures in which convergent validity was assessed included the EES-C⁷⁷ and the Toddler Snack Food Feeding Questionnaire (TSFFQ).⁸² The EES-C was compared with data from the QEWP (Spitzer 1992⁹⁵). Authors reported good convergent validity and show that those with loss of control (LOC) (from QEWP) had higher eating in response to anger, anxiety and frustration and higher depressive symptoms than people without LOC [although results based test for difference (analysis of covariance – ANCOVA), $p < 0.05$]. Convergent validity for the TSFFQ was weak, with correlations with the CRQ of $r = 0.20$ (in toddlers) and $r = 0.21$ (in preschool children) (range = 0.02–0.43). Implications of these findings when the measures are used to assess change is potentially less important and will be based on the choice of comparator measure.

Construct validity was evaluated in 18 manuscripts,^{72,77–80,82–85,90,91,96–101} comparing eating behaviour measures to weight,^{72,77,83,84,96–98} weight concerns^{99,100} and health-related behaviours,^{77–80,82,85,90,91,101} with correlations ranging from 0.03 to 0.59. Of those making comparisons to weight or weight status, correlations were weak: $r = 0.13$ (DEBQ-C⁸³); $r = 0.14$ (CFQ⁹⁶); $r = 0.28$ [Children's Eating Attitudes Test (ChEAT)¹⁰¹]; and $r = 0.07$ (un-named measure of control in parental feeding practices⁸⁴). Higher correlations were seen with weight concerns: $r = 0.59$ for the Youth Eating Disorder Examination-Questionnaire (YEDE-Q)⁹⁹ and $r = 0.4$ for a further study evaluating ChEAT.¹⁰⁰

Physical activity

A total of 45 studies (within 35 manuscripts), describing 24 PA measures, were extracted. A summary of all of these studies is included in *Appendix 9*. Of these, two did not fully meet eligibility criteria (pathway 4) but were deemed relevant by experts during the subsequent external appraisal process (see *Chapter 4, Results of expert appraisal*)^{102,103} and have been added retrospectively.

Objective PA measures included pedometers, accelerometers, monitors and direct observations. Objective methods are considered optimal for quantification of PA and are advantageous over subjective methods through avoidance of reporting bias.¹⁰⁴ Criterion validity was assessed in 12 of the objective studies^{105–114} (with correlations ranging from $r = 0.47$ to $r = 0.82$). Four out of the five studies evaluating accelerometers measured criterion validity (compared with direct observation^{105,106}); VO_2 (oxygen uptake);¹⁰⁷ or heart rate (HR),¹⁰⁸ with strong correlations observed for all (range: $r = 0.71$ – 0.86). Criterion validity of pedometers was also common with a range in correlations between $r = 0.47$ and $r = 0.85$ in studies comparing steps to accelerometers^{112–114} and direct observation.¹⁰⁹ One study, assessing per cent error, however, found that a high degree of error indicated under-reporting.¹¹⁰ Authors commented on the range in validity of measures relating to the type of equipment used; for example, in the assessment of criterion validity of a SenseWear band (BodyMedia® SenseWear, Pittsburgh, PA, USA),¹¹¹ they reported the greatest validity with one specific model (SWA5.1) when comparing against DLW. Convergent validity was evaluated in three studies evaluating objective measures with moderate findings: accelerometers compared with Actiwatch (Actigraph®, Pensacola, FL, USA) data $r = 0.36$;¹⁰⁵ accelerometers compared with activity diaries $r = 0.38$,¹⁰⁸ and SenseWear model SWA5.1 compared with the SWA6.1 model [showing statistically greater estimates of metabolic equivalents (METs) in boys than girls with the SWA5.1 model than in those with the SWA6.1 model].¹¹¹

With regards to external reliability of objective measures, four studies conducted TRT reliability.^{102,110,112,113} With pedometers, one study¹¹² reported high validity ($r = 0.77$) but another¹¹³ found very poor reliability ($r = 0.08$). Another evaluation of pedometer TRT reliability reported a mean difference of 10% between measures.¹¹⁰ The last of the four objective measures evaluating TRT reliability was the System for Observing Children's Activity and Relationships during Play (SOCARP).¹⁰² Findings from this study report per cent agreement ranging from 85% to 93%. Inter-rater reliability analysis was conducted in evaluation of the two direct observation tools. Both reported high levels of reliability with comparison between observers: ($r = 0.96$, $\kappa = 0.87$ ⁸⁴) (89% agreement¹⁰²).

The remaining measures were subjective (questionnaires, recalls, diaries, etc.). Twenty-one of the manuscripts describing subjective PA measures reported evaluating criterion validity, with a resulting range in correlation from $r = 0.04$ for correlations between the Children's Leisure Activities Study Survey (CLASS) and accelerometry¹¹⁵ and $r = 0.53$ for correlations between the Previous Day Physical Activity Recall and accelerometers.¹¹⁶ Overall, criterion validity was lower than that observed by objective measures, with 11 of these studies obtaining correlations of less than the adequate standard of 0.4, and some findings dependent on weight status of the children.¹¹⁷ Convergent validity was assessed in 11 of the subjective studies^{118–125} and correlations were slightly higher ($r = 0.22–0.88$) but most were compared against other subjective methods and thus the high correlations may not necessarily suggest a robust instrument (i.e. it could be interpreted as the tools being equally as poor) (see *Appendix 7*). Construct validity of self-reported measures was poor^{118,119,126,127} (correlations ranging from $r = 0.07$ to $r = 0.33$). Of these, two studies failed to report findings for non-significant correlations and thus the lower end of the range may be less.¹²⁶ Two construct validity studies made comparisons with body weight/weight status.^{126,127} Goran *et al.*¹²⁷ report correlations of 0.24 and 0.33 for findings from two substudies comparing the Physical Activity Questionnaire (PAQ) for Pima Indians with fat mass data from bioimpedance measurement. Moore *et al.*¹²⁶ also report low correlations of $r = 0.10$ for comparisons between the Physical Activity Questionnaire for Older Children (PAQ-C) and percentage body fat, which were also assessed by bioimpedance.

Reliability results of subjective PA measures were generally better than validity findings. Six studies^{126,128,129} reported IC, with a range of alpha values of between 0.66 and 0.84 (with just one reporting alpha values of < 0.70 ¹²⁶). Results of TRT reliability (conducted in 14 studies) were more variable, with correlations ranging from $r = 0.24$ (for child-reported activity in CLASS¹¹⁵ and 0.98 (for the Previous Day Physical Activity Recall¹²⁰). One study¹²⁸ also reported a generalisability coefficient of 0.88. Inter-rater reliability was evaluated by five studies.^{102,103,115,120,130} Again, results are highly variable with correlations as high as $r = 0.99$ for inter-rater reliability of the Previous Day Physical Activity Recall¹²⁰ and as low as $r = 0.19$ for reliability of CLASS.¹¹⁵ Results for inter-rater reliability evaluation may be dependent on the type of activity been assessed. For example, Telford *et al.*¹¹⁵ reported a strong agreement of 87.5% for assessment of soccer, but just 8% agreement for tennis. This type of evaluation may also be dependent on obesity status.¹³⁰

Sedentary behaviour/time

A total of five manuscripts,^{130–134} describing six measures of sedentary time/behaviour met the eligibility criteria for CoOR (see *Appendix 10*).

Of the six measures, three were measures of sedentary time (i.e. time spent being inactive)^{131,132} using activity monitors. The remaining three^{133–135} assessed sedentary behaviours (i.e. frequency or duration spent doing specific low-energy behaviours such as screen time). Measurement of sedentary time in the included studies was by objective measurements compared with those assessing sedentary behaviour, which were all self-reported.

Studies by Reilly *et al.*¹³¹ and Puyau *et al.*¹³² (Study 1) both assessed criterion validity of accelerometers for the measurement of sedentary time using direct observations and room calorimetry, respectively. Both report high validity. Sample sizes for these were low (52 for Reilly *et al.*¹³¹ and 26 for Puyau *et al.*¹³²) but not unusual given the type of measurements used for criterion assessment. Puyau *et al.*¹³² also assessed convergent validity of the accelerometer against another monitor; the Mini-Mitter Actiwatch monitor, with an average correlation of $r = 0.86$ (range = 0.82–0.89). A further study (reported in the same paper) by Puyau *et al.*¹³² (Study 2) also evaluated the Mini-Mitter Actiwatch monitor for criterion validity using room calorimetry and reported a mean correlation between activity and energy expenditure of $r = 0.79$ (range = 0.82–0.89). Other criterion methods of HR monitoring and microwave activity were also used for both the accelerometers and Actiwatch, with good overall findings ($r = 0.57–0.72$ for accelerometers and $r = 0.66–0.83$ for the Actiwatch).

Measures of sedentary behaviour also assessed criterion validity,^{133–135} although comparison was not made against direct observation or measured energy expenditure. Ridley *et al.*¹³³ made comparisons between the Multimedia Activity Recall for Children and Adolescents and accelerometry, and reported an overall correlation of $r = 0.39$ (range = 0.35–0.45). Dunton *et al.*¹³⁴ also used a criterion of accelerometry in their evaluation of the Electronic Momentary Assessment (EMA): a self-report survey on mobile phones – a method by which behaviours are captured in real time by use of mobile phones. Results indicate that the number of steps taken was significantly higher for the EMA surveys reporting active play, sports or exercise than any other type of activity [adjusted Wald test: $F = 22.16$, degrees of freedom (df) = 8, $p < 0.001$]. Epstein *et al.*¹³⁵ also evaluated criterion validity of a measure of Habit books with index cards against a criterion of accelerometers and report correlations of $r = 0.63$ (for average METs) and $r = 0.60$ [for per cent time in moderate to vigorous physical activity (MVPA)]. This study was not the primary aim of the manuscript (which reported trial evaluation results) and was conducted with only 41 participants. TRT reliability was evaluated in only one study¹³³ finding high correlations ($r = 0.92$), although it was also conducted in a small sample of 32 children and adolescents.

Fitness

A total of 14 manuscripts^{136–149} were identified that described 13 fitness outcome measures. A summary of the data extracted for these studies is provided in *Appendix 11*.

The majority (12) of measures described in the included manuscripts assessed aerobic capacity (defined as the maximal amount of physiological work that an individual can do measured by oxygen use). Two^{136,137} assessed general fitness, of which one,¹³⁷ 'Fitnessgram®', also includes measurement of aerobic capacity, in addition to measures of muscular strength; muscular endurance and flexibility; and body composition. This measure was designed as an educational assessment tool for school populations (i.e. it was not designed for obesity research). However, it has been used as an outcome, which is why it met inclusion criteria here.

Seven included measures^{136–142} determined TRT reliability, with correlation results ranging from $r = 0.65$ – 0.91 , kappa statistics ranging from $\kappa = 0.59$ – 0.81 and per cent agreement ranging from 88% to 91%. Thus, all demonstrated at least moderate TRT results, indicating that they can be reliability assessed over multiple time periods.

Inter-rater reliability was evaluated in the Fitnessgram study,¹³⁷ which compared teacher with expert agreement in recording children's fitness scores. Results in agreement (84–87%) and kappa statistics (0.67–0.73) identified adequate robustness of results.

Criterion validity was assessed in 10 studies ($r = 0.03$ – 0.81) in which comparisons were made with measures against a gold standard of measured oxygen consumption [via VO_{2max} (maximum oxygen uptake to the point in which oxygen demands plateau) or VO_{2peak} (highest value of oxygen uptake from a particular test which is limited by tolerance level)].^{138–140,143–149} Of these, four had a sample size of < 50 .^{144–146,149} In the remaining six studies^{138–140,143,147,148} that measured criterion validity, two were evaluations of the 20-m shuttle run;^{139,140} one assessed basal metabolic mass estimates with fat-free mass;¹⁴³ one assessed the 6-minute walk test;¹³⁹ one, the adjustable height step test;¹⁴⁹ and one, bioelectrical impedance-derived VO_{2max} .¹⁴⁸ Correlations for these were mostly moderate but ranged between $r = 0.03$ ¹⁴⁷ and 0.81 .¹⁴⁸ One study¹⁴⁰ reported higher validity in obese children (based on stratified analysis of 126 children). Conversely, Roberts *et al.*¹⁴⁷ assessed bioelectrical impedance-derived VO_{2max} and reported a weight-dependent correlation with measured VO_{2max} (VO_{2max} ml/kg/minute) of $r = 0.03$. Non-weight-dependent correlations (VO_{2max} l/minute) in this sample of 134 obese and overweight adolescents were considerably higher at $r = 0.48$. Thus, although the majority of studies report moderate to high levels of criterion validity, results are varied, with some dependent on weight status and also some conducting analysis on small samples.

Convergent validity was assessed by two studies.^{136,141} Of these Loften *et al.*¹⁴¹ compared different modes of calculating measured VO_{2peak} from either cycle or treadmill, thus it is a rare study within those identified

by CoOR that evaluated the actual gold standard measure. Comparisons between the two approaches reported correlations ranging from $r = 0.48$ to $r = 0.77$. Tests for differences were all non-significant ($p > 0.05$) but the validation study was conducted in only 21 overweight/obese children/adolescents. This study also demonstrated strong TRT correlations – again, in a small sample size. Overall findings indicated that the cycle performed marginally better than the treadmill. However, importantly, children reported higher acceptability of the cycle than the treadmill. The other study assessing convergent validity was a comparison between the International Fitness Scale (IFIS)¹³⁶ and what was described as ‘measured fitness’. However, the measure of fitness was based on a 20-m shuttle run (i.e. not measured VO_{2max} or VO_{2peak}), and has therefore been considered as a convergent validity (not criterion) by CoOR. Authors reported significant positive linear relationships with increased self-report and ‘measured’ fitness. This study also collected a number of cardiovascular outcomes as a means to test construct validity of the IFIS. Findings suggest that obesity was negatively associated with levels of fitness in the IFIS, except for measurement of muscular strength.

Construct validity was assessed in one further paper evaluating aerobic cycling power with insulin and reported a correlation of $r = 0.37$.¹⁴⁶ Similar to many other evaluations of fitness measurement, assessment was conducted in only a small sample of 35 obese adolescents.

No fitness measures conducted a formal assessment of the ability to measure change (responsiveness).

Physiology

A total of 28 papers, describing 12 outcome measures, met inclusion criteria for the physiology domain. A summary of all papers extracted are available in *Appendix 12*. Two included manuscripts were written in languages other than English.^{150,151} Data were partially extracted from each of these, which is included in *Appendix 12* for reference. However, appraisal of these was not conducted (one¹⁵⁰ describes evaluation of ‘indices of insulin sensitivity’, which is evaluated in multiple other included manuscripts).

Of the 28 included manuscripts, the majority described the evaluation of insulin and/or glucose^{150,152–165} or energy expenditure or metabolic rate.^{166–173} Of those assessing criterion validity of measures of insulin or glucose, six made comparisons with the gold standard of the euglycaemic–hyperinsulinaemic clamp (EHC) test^{152,154,156–158,162} reporting correlations ranging from $r = 0.4–0.78$ for varying indices of insulin sensitivity in sample sizes ranging from 31^{156,157} to 323.¹⁶² Criterion validity was evaluated in all of the studies evaluating energy expenditure/metabolic rate, by making comparisons with measures such as direct and indirect calorimetry, but none used the gold standard of DLW. Moderate to high correlations were generally reported, but the primary focus of these studies was usually the development or comparisons of equations used to predict energy expenditure in obese children and adolescents. Except for one study,¹⁷³ sample sizes were high (with 12 studies^{154,158,159,162,166–171} including samples of > 100).

Convergent validity was assessed in four studies,^{152,155,161,174} of which three compared insulin with blood lipids,¹⁵² glucose tolerance¹⁵⁵ and fasting insulin,¹⁶¹ and one examined relationships between glycated haemoglobin (HbA_{1c}) and fasting glucose.¹⁷⁴ Findings for convergent validity of indices of insulin sensitivity were generally high, with correlations ranging between $r = 0.60$ and $r = 0.81$ for insulin. Convergent validity for HbA_{1c} used accuracy testing [receiver operating area under the curve (AUC)], which reported a range of 0.60–0.81 in AUC in 1156 obese adolescents. However, results were influenced by weight status. This study also evaluated the relationship between HbA_{1c} and diabetic status, and demonstrated poor validity with this construct [$\kappa = 0.2$ (95% confidence interval 0.14 to 0.26)]. One further study¹⁷⁵ evaluated construct validity in its assessment of ghrelin in 100 obese children. Results suggest that ghrelin is statistically associated with obesity and cardiovascular outcomes, although correlations are generally weak (ranging from $r = 0.1$ to $r = 0.5$). This study also reported ghrelin pre and post intervention. Tests indicate that it is able to detect change but that changing values levelled off after a period (advocating testing immediately post intervention if used). One other study¹⁷⁰ that met criteria for inclusion to CoOR reported measuring the ability of the measure to detect change. This study¹⁷⁰ evaluated predicted resting energy expenditure and reported a mean difference of 7.45% in resting energy expenditure after weight loss.

Prediction equations for resting energy expenditure in this study¹⁷⁰ involved inclusion of fat-free mass. As weight loss is associated with change in fat-free mass, the authors advocate assessment to be made only during periods of weight stability.

Only 2^{176,177} out of the 26 included studies conducted reliability testing. TRT reliability was evaluated by Libman *et al.*¹⁷⁶ in a study that compared measurement of glucose via fasting and 2-hour samples in 60 overweight/obese adolescents. Results indicated that fasting glucose ($r = 0.73$) had higher reliability than 2-hour glucose ($r = 0.37$) testing. Inter-rater reliability was assessed in one other study¹⁷⁷ comparing radiologists working in three ultrasound units. This study¹⁷⁷ reported high correlations between radiologists ($\kappa \geq 0.8$) in ultrasound analysis of liver echogenicity, although the sample size was small ($n = 11$).

Economic evaluation

The original aim of the CoOR study was to include measures of economic evaluation as one of its outcome domains. However, review of identified manuscripts failed to find any manuscripts that described the development or evaluation of measures used that can assess utility and therefore estimate quality-adjusted life-years (QALYs). The National Institute for Health and Care Excellence (NICE) advocates the conduct of cost-utility analysis using utility measures (with the QALY as the health-related outcome measure for economic evaluation). No such measures were found for use in an obese childhood or adolescent population in this review, although the team are aware of some that are currently under development. Given that the existing CoOR review strategy did not include terms related to measurement of QALYs, a separate 'scoping' search was conducted. A copy of this search can be found in *Appendix 16*. This did not reveal any further appropriate measures of utility. Alternative measures of cost-effectiveness could be considered (although would not fit within NICE guidance), but assessment of cost (of intervention) per unit of weight loss is usually preferred (i.e. those described in the anthropometry domain). HRQoL measures are often used, but CoOR has viewed these as a separate domain, given that they cannot be used to estimate QALYs. Unless the research is focused on quality of life/psychological well-being, such measures are not essential. As measures of HRQoL were identified in the CoOR review from those already used as outcome measures and those that have specifically developed for childhood obesity research, a further domain of HRQoL has been included. This was possible, as the CoOR search did not include specific search terms relative to each outcome domain (i.e. it was designed to be sensitive enough to detect any kind of outcome measure).

Health-related quality of life

A total of 25 papers describing 16 measures were extracted for the HRQoL domain. Of these, four were written in languages other than English,^{15,178–180} which describe measures that have not been evaluated by any other included manuscript. Data have been extracted for three of these.^{178–180} All have been included within the summary table in *Appendix 13* but were not eligible for appraisal.

Seven HRQoL measures were developed specifically for use in a paediatric obese population: (1) Impact of Weight on Quality of Life (IWQoL);^{15,181,182} (2) Sizing Me Up;¹⁸³ (3) Sizing Them Up (a parent-reported version of Sizing Me Up);¹⁸⁴ and the Youth Quality-of-Life Instrument-Weight Module;¹⁸⁵ plus three German HRQoL measures that were developed specifically for obese children.^{178–180} Akin to most of the HRQoL measures, multiple forms of evaluation were conducted on many of these tools. Except for measures described in non-English papers, all assessed IC, reporting alphas ranging from 0.74¹⁸⁴ to 0.92.^{181,185} All report using FA to develop or refine the questionnaires, and all assess convergent validity by comparing against other questionnaires aimed at assessing similar constructs. Comparisons with the Paediatric Quality of Life questionnaire were made with three of the weight specific measures,^{181,183,184} of which the highest correlations were reported with the IWQoL questionnaire ($r = 0.75$).¹⁸¹ Comparisons between Sizing Them Up and the IWQoL questionnaire reported weaker correlations of $r = 0.27$.¹⁸⁴ Additionally, TRT reliability was conducted on each measure, with each demonstrating at least moderate to high reliability (ranging from $r = 0.67$ to $r = 0.82$), although sample sizes were low for two studies.^{182,185} Given that these measures were developed specifically for obese children, correlations with BMI (i.e. construct validity) were surprisingly lower than in other forms of validity, ranging from $r = 0.16$ for

Sizing Me Up¹⁸⁴ to $r = 0.44$ for the Youth Quality-of-Life Instrument-Weight Module.¹⁸¹ Finally, two measures – weight specific – evaluated responsiveness. These were the only studies assessing responsiveness of all included HRQoL measures in CoOR. In evaluation of 80 children and adolescents, Kolotkin *et al.*¹⁸² report a standardised response mean (SRM) of 13.43 [effect size (ES) of 0.75]. A smaller, but significant, SRM of -5.4 was reported in responsiveness testing of Sizing Me Up in 220 obese children and adolescents,¹⁸⁴ both well within acceptable (moderate) levels (described by CoOR as having a SRM of > 0.5) to support their ability to assess change.

Of the remaining studies assessing generic HRQoL measures, a similar level of evaluation was conducted, with many reporting findings from multiple types of evaluation. Average IC findings were high in each of the nine studies^{186–194} conducting this evaluation, with a range of $r = 0.72$ ¹⁸⁶ to 0.86 ¹⁸⁷ in those that presented ‘means’ (and not only ranges). In fact, all of the included measures demonstrated a reasonably high level of reliability and validity, with some variability in findings of convergent validity (see *Appendix 13*). Two measures may be considered redundant, given that newer (or more appropriate) versions are now available. For example, the Paediatric Cancer Quality of Life measure^{188,195} would be less appropriate than a non-cancer version in the evaluation of childhood obesity treatments. Additionally, an older version (V1.0) of the Paediatric Quality of Life questionnaire¹⁹¹ can be substituted for newer versions.^{190,191,196}

Psychological well-being

A total of 20 papers, describing 17 measures were eligible for data extraction. A summary of all papers extracted are available in *Appendix 14*.

Given the nature of these self-reported survey questionnaires, assessment of criterion validity was not anticipated, where ‘gold standard’ measures are unlikely. Some authors reported conducting criterion validity, which was defined as ‘construct’ validity by CoOR (e.g. comparisons with body weight). One study,¹⁹⁷ however, did make comparisons between self-report and direct observations in their evaluation of the Self-Control Rating Scale (SCRS).

Eight of the included psychological well-being studies included evaluation of convergent validity,^{197–204} each making comparisons against different psychological measures of differing constructs (often comparing with more than one other measure). Comparisons of the correlations between these is therefore limited, however, with a range of between $r = 0.06$ in the evaluation of convergent validity of the SCRS against the Delay of Gratification scale¹⁹⁷ and $r = 0.66$ in the evaluation of the Body Esteem Scale against the Piers–Harris Children’s Self-Concept Scale.²⁰⁴ Three studies evaluating convergent validity reported results with a correlation of < 0.40 for all of the included comparator measures.^{197,198,202}

Construct validity was assessed in nine studies. Six of these made comparisons to weight or weight status in children,^{198,200,204–207} of which findings varied between $r = 0.07$ (comparing the Body Shape Questionnaire to WHR²⁰⁶ to $r = 0.55$ (comparing the Body Esteem Scale to weight²⁰⁴). Stein *et al.*²⁰⁷ report significant differences in scores for the Children’s Physical Self-Concept Scale (CPSS) between normal weight and overweight children ($F = 33.91$, $p < 0.001$). Percentage agreement of 90.5% (in obese children) was also reported by Probst *et al.*²⁰⁸ for comparisons between the video distortion measure and BMI.

Test–retest reliability was conducted in 12 studies, with correlations ranging from $r = 0.52$ to $r = 0.91$.^{195,197,199,201,205–211} Thus, all met the criteria ($r > 0.4$), suggesting that psychological well-being measures have strong TRT reliability.

Responsiveness testing was not reported in any of the studies evaluating psychometric well-being that were identified by the CoOR review.

Environment

A total of nine manuscripts,^{212–219} described 10 measures of the environment, met eligibility criteria for the environment domain. A summary of all papers extracted are available in *Appendix 15*.

Two environmental measures assessed child-care environments^{212,213} and seven measured home physical and/or social constructs within the home environment.^{214–219} A further was a measure capturing 'perception' of the built environment.²²⁰

Reliability testing in the form of IC was implemented in six studies,^{215–218,220} all of which demonstrated high levels of internal reliability ($\alpha = 0.75–0.83$). Similarly robust results for TRT reliability were evident in one measure of child-care settings²¹² and seven measures of the home environment,^{214,215,217–220} with mean correlations ranging from $r = 0.59$ of the home PA equipment scale²¹⁹ to $r = 0.85$ of the Family Eating and Activity Habits Questionnaire²¹⁵ (FEAHQ) (with mean $\kappa = 0.57–0.66$). Results for inter-rater reliability testing in six studies^{212,213,215,218,219} were also strong ($r = 0.47–0.88$). Thus, the outcome domain of environmental measures demonstrates high levels of multiple indicators of reliability, with no studies performing no form of reliability.

Internal validity was assessed in two studies,^{216,220} with total variance ranging from 7% to 47%, and factor loadings ranging from 0.31 to 0.88, of which one study²²⁰ reported all loadings to be above the acceptable limit of 0.40. However, providing that necessary amendments are made to questionnaires, this should not preclude the use of measures in which some factor loadings are low. Criterion validity was evaluated in two studies^{212,213} in which the gold standard method was direct observations by researchers. Benjamin *et al.*²¹² evaluated criterion validity of their child-care setting measure, the Nutrition and Physical Activity Self-Assessment for Child Care (NAPSACC), by comparing items to researcher-measured items reported in the Environment and Policy Assessment and Observation (EPAO)²¹³ also included in the CoOR review. Results were variable by item, with kappa ranging from 0.11 to 0.79 (mean $\kappa = 0.37$). The comparator gold standard method by Ward *et al.*²¹³ (EPAO) conducted a study of inter-rater reliability and reported moderate to high correlations (although also variable by item) ($r = 0.63$, range = 0.05–1.0). Bryant *et al.*²¹⁴ compared a parent report home environment measure, the 'Healthy Home Survey' (HHS) to researcher-conducted survey completion in the home and also reported variable findings, with a range in correlations of $r = 0.3$ to $r = 0.88$ (mean $r = 0.62$) and a range in kappa of 0–0.96 (mean $\kappa = 0.55$). This measure appears to be robust, along with strong findings for TRT reliability ($r = 0.72$, $\kappa = 0.66$); however, authors report concern related to the collection of some open-response items (e.g. food availability in the home) and are currently working on a new version – 'HomeSTEAD'.

Convergent validity was assessed in only one study,²¹⁶ in which the Parenting Strategies for Eating and Activity Scale (PEAS) was compared with data from the CFQ.⁶² Findings were low with a mean correlation of $r = 0.22$ (range $r = 0.02–0.65$) in 91 children. Construct validity was evaluated in six studies.^{216–220} Of these, four studies^{216,217,219} assessed correlations with BMI or obesity. Findings are difficult to compare because of inconsistencies in the analytical approaches used. Larios *et al.*²¹⁶ reported very weak correlations between PEAS and BMI z-score in a sample of 714 children ($r = 0.03$, range = 0.03–0.21). McCurdy *et al.*²¹⁷ conducted independent samples *t*-tests to determine whether scores on the Family Food Behaviour Survey (FFBS) varied by child weight status, and found that overweight was related to increased maternal control ($p = 0.052$) and that children were more likely to be of normal weight if there was increased maternal presence at meal and snack times ($p = 0.01$). Sample size for this study, was small, however, with only 28 children included. Both studies by Rosenberg²¹⁹ to assess two brief scales that measure PA and sedentary equipment in the home assessed correlations with BMI z-score using linear regression models. Findings suggest that the electronic equipment scale (specifically, having a television in the bedroom) was significantly and positively associated with BMI z-score.

Responsiveness was assessed in one study, in which Golan *et al.*²¹⁵ reported the ability of the FEAHQ to detect change following a weight loss intervention. Change in child body weight was found to be associated to change in scores from the 'exposure' and 'eating style' scales of the questionnaire in both intervention and the control, with the change in score explaining 27% variance in weight reduction.

Results of internal appraisal

Internal appraisal of all outcome measures resulted in 29 outcome measures being classified into Category 1 (certain, good evidence, fit for purpose). Thirty-five were placed into Category 2 (certain, poor evidence, not fit for purpose) and 121 were placed into Category 3 (uncertain, requiring further consideration). Decisions on certainty, alongside any relevant comments were written in two appraisal forms for (1) anthropometry (primary) outcome measures and (2) all other (secondary) outcome measures (see *Appendices 16* and *27*). These forms were also used by experts in external appraisal. Thus, all final decisions (following internal and external appraisal) are also shown in these appendices. Further details of the internal appraisal are provided below according to outcome domain.

Scores for development and evaluation of secondary outcome measures were assigned and are shown in *Appendices 17–25*.

Anthropometry (1 certainty = '1'; 2 certainty = '2'; 35 certainty = '3')

Appendix 17 provides the internal appraisal results for all included anthropometry measures. Based on the evidence, the only anthropometry measure that was assigned a certainty score of '1' (i.e. deemed fit for inclusion) was ADP. Five^{17,221–224} out of six^{17,219–223,225} studies that evaluated this measure generally advocated its use. The only measures to be assigned a certainty score of '2' (i.e. deemed not fit for inclusion) following internal appraisal were measures of self-reported height and weight, and parent-reported height and weight. These methods were commonly evaluated against a criterion of measured height and weight, with 28 studies evaluating self-report and 14 studies evaluating parent report. However, only two studies^{226,227} of self-reported height and weight concluded that the measure was valid and only one²²⁸ did so for parent report. Findings from the remaining studies were consistent in reporting a poor relationship between measured and self-reported height (for implementation in trials).

All other anthropometry measures were assigned a certainty score of '3' because of inconsistencies between study findings. This score of uncertainty was also assigned for measures in which little evaluation had been conducted.

Diet (3 certainty = '1'; 9 certainty = '2'; 19 certainty = '3')

Scores Two studies evaluating dietary assessment methodologies were assigned a maximum score of four for demonstrating a high degree of quality in the evaluation of TRT reliability; Lanfer *et al.*'s evaluation³⁶ of the Children's Eating Habits Questionnaire food frequency questionnaire (CEHQ-FFQ), and Vance *et al.*'s evaluation⁷¹ of the Food Behaviour Questionnaire (FBQ). Vance *et al.*⁷¹ also conducted inter-rater reliability and received a maximum score of '4' (see *Appendix 18*).

Maximum scores were also assigned for evaluation of the Short-list Youth/Adolescent Questionnaire (Short YAQ)³⁴ for both convergent and construct validity. Robust evaluation and findings were additionally assigned for the convergent validity testing of the YAQ,³⁷ Harvard Service Food Frequency Questionnaire (HSFFQ)³⁸ and familial influence on food intake – FFQ.³⁹ Maximum scores of '4' were provided to 6^{31,53,54,67,70,40} out of 24 studies that evaluated criterion validity of diet measures.

No measures were assigned the minimum score of '1' for the quality of any form of evaluation. However, low scores of '2' were assigned to two assessments of TRT reliability,^{41,42} eight assessments of criterion validity,^{30,33,57–59,61–63} five assessments of convergent validity,^{42–45,52} one assessment of construct validity,⁵² and two assessments of TRT reliability.^{41–42}

Degree of certainty Of the included diet measures, internal appraisal resulted in assigning a degree of certainty score of '1' (i.e. fit for inclusion) to three measures: the YAQ,³⁴ the Australian Child and Adolescent Eating Survey (ACAES)^{32,46} and the New Zealand FFQ.⁴⁷ A certainty score of '2' (i.e. not fit for inclusion) was assigned for nine measures: the Korean FFQ,⁴⁸ the qualitative dietary fat index,⁴² fried food away from home,⁵² the food intake questionnaire,⁴⁹ the Crawford 5-day food frequency

questionnaire (5D FFQ),³³ diet history,^{53–55} the 9-day food diary,⁵⁷ the 2-week food diary^{58,69} and the 7-day food diary.⁶¹ The remaining measures were all assigned a certainty score of '3' (uncertain) (see *Appendix 28*).

[Note: Although there are 22 different types of dietary assessment methods identified, appraisal was made on individual subtypes of methods. For example, a food diary has been considered to be one type of dietary assessment methodology, yet appraisal separated these according to the individual protocols of each (e.g. 3-day food diary appraised separately from 7-day food diary). As such, the total number of measures appraised (30) is not the same as the total number of included measures.²⁰]

Eating behaviours (5 certainty = '1'; 6 certainty = '2'; 11 certainty = '3')

Scores Internal scores for evaluation of eating behaviour studies were generally high, with the majority of studies being assigned a score of '3' or '4' for most types of evaluation. No studies were assigned the lowest score of '1' for any form of evaluation. Only four studies^{93,229–231} received a low score of '2', including one study's evaluation of IC [Child Eating Disorder Examination Questionnaire (ChEDE-Q)],²²⁹ one study's evaluation criterion validity (CFQ),⁹³ and two studies' evaluations of convergent validity [ChEDE-Q,²³⁰ Children's Binge Eating Disorder Scale (C-BEDS)²³¹] (see *Appendix 19*).

Degree of certainty Of the 22 included outcome measures, five were deemed of high quality (fit for purpose, certainty = 1), including the EES-C,⁷⁷ the CFQ,^{75,93,96,97,113,232} the Child Eating Behaviour Questionnaire (CEBQ),^{72,73} the TSFFQ⁸² and EAH-C⁸⁰ (see *Appendix 28*). The internal appraisal judged six measures to be unfit for purpose, including the QEWP-A,^{90,91} ChEAT,^{86,100,101} C-BEDS,²³³ the McKnight Risk Factor Survey-III (MRFS-III)⁸⁷ and an unnamed tool of parental feeding strategies.⁸⁸ The remaining 11 measures were assigned a certainty score of '3' (uncertain, requiring further consideration).

Physical activity (4 certainty = '1'; 9 certainty = '2'; 11 certainty = '3')

Scores Nine evaluations of TRT reliability of PA measures were assigned maximum scores of '4', indicating high-quality reliability evaluation (see *Appendix 20*). However, a score of '4' was generally not common in other forms of evaluation, in which internal appraisal assigned '4' in only one evaluation of criterion validity of the 7-day recall interview,¹²¹ and two forms of evaluation of the PAQ-C (internal validity¹²⁶ and IC).¹²⁸ A minimum score of '1' was assigned to only one study¹¹⁹ evaluating the convergent validity of the Physical Activity Diary. The remaining evaluations were generally assigned quality scores of '3' or '4'.

Degree of certainty Of the 24 included PA measures, the internal appraisal team assigned a degree of certainty score of '1' (i.e. fit for inclusion) to four measures: the accelerometer;^{105–108,234} the 7-day recall interview;¹²¹ the moderate to vigorous PA screener;²³⁵ and the PAQ for Pima Indians^{127,236} (see *Appendix 28*). A further nine measures were deemed unfit for purpose (degree of certainty = 2): HR monitoring;²³⁷ the Activity Questionnaire for Adults and Adolescents;⁹⁷ the Activity Rating Scale;¹²¹ the Activitygram;^{113,115} the National Longitudinal Survey of Children and Youth;¹³⁰ the Outdoor Playtime Checklist;¹²² the Outdoor Playtime Recall;¹²² the Physical Activity Diary;¹¹⁹ and the Youth Risk Behaviour Survey (YRBS).²³⁸ The 11 remaining measures were assigned an uncertainty score of '3'.

Sedentary behaviour/time (0 certainty = '1'; 0 certainty = '2'; 6 certainty = '3')

Scores The only study evaluating measures of sedentary time/behaviour that received a maximum score of '4' (indicating high quality) was for criterion validity evaluation of accelerometry.¹³¹ No studies were assigned the minimum score of '1' but one¹³⁵ was given a score of '2' for the evaluation of criterion validity of Habit books with index card. Remaining evaluations were all assigned a quality score of '3' (see *Appendix 21*).

Degree of certainty All studies evaluating sedentary time/behaviour were assigned a certainty score of '3' (uncertain, requiring further consideration). This was largely due to a lack of identified studies conducting any form of evaluation of sedentary measures for use as outcome measures in childhood obesity treatment intervention evaluations (see *Appendix 28*).

Fitness (1 certainty = '1'; 5 certainty = '2'; 7 certainty = '3')

Scores Eight different types of evaluation from 5^{139,140,148,239,240} out of the 14 included studies were assigned a maximum score of '4' (see *Appendix 22*). Of these, the IFIS²³⁹ was assigned a maximum score for all three evaluations of TRT reliability, convergent validity and construct validity. No studies were assigned the minimum score of '1', but two received a low score of '2', including criterion validity of the submaximal treadmill test,¹⁴⁹ and criterion and construct validity of the aerobic cycling power test.¹⁴⁶

Degree of certainty Only one measure of fitness was assigned an internal certainty score of '1' (i.e. fit for purpose): the IFIS.²³⁹ Five were deemed as unfit for purpose including BIA,¹⁴⁷ the Fitnessgram,²⁴⁰ basal metabolic rate (BMR) with fat-free mass,¹⁴³ estimated maximal oxygen consumption and maximal aerobic power,¹¹⁸ and aerobic cycling power.¹⁴⁶ The remaining fitness measures were assigned an uncertainty score of '3' (see *Appendix 28*).

Physiology (2 certainty = '1'; 0 certainty = '2'; 10 certainty = '3')

Scores Internal appraisal score allocation to studies that evaluated physiological measures were generally high (see *Appendix 23*). The majority of studies (22/26) conducted criterion validity and only two of these scored '2' for quality.^{164,171} Other studies conducting different forms of evaluation that were assigned a low-quality score of '2' included two evaluations of construct validity.^{174,175} A minimum score of '1' was only assigned to one study that conducted responsiveness testing.¹⁷⁰

Degree of certainty Of the 12 different types of measurement, 10 were assigned a certainty score of '3' (uncertain, requiring further consideration) (see *Appendix 28*). Only two were deemed to be fit for purpose based on the evidence, including indices of insulin sensitivity^{152-156,158-162} and DXA lean body mass (LBM) for resting energy expenditure.¹⁷² No measures were considered to be unfit for purpose (degree of certainty = 2).

Health-related quality of life (4 certainty = '1'; 2 certainty = '2'; 6 certainty = '3')

Scores HRQoL measures studies often conducted multiple types of evaluation and the overall scores for these were high, with the majority assigned scores of '3' and '4'. No studies were given the maximum of '4' for all of the forms of evaluation but some demonstrated very good quality overall, including an evaluation of the IWQoL,¹⁸³ Sizing Me Up,¹⁸⁵ Sizing Them Up²¹⁵ and the Youth Quality of Life Instrument-Weight module (YQOL-W).¹⁸⁵ Only one study²⁴¹ was assigned a minimum score of '1' for their assessment of construct validity of the European Quality of Life-5 Dimensions (EQ-5D).

Degree of certainty Of the 12 included measures, four were considered to be of high quality and were assigned a certainty value of '1' (fit for purpose), including the IWQoL,^{181,182} the Paediatric Quality of Life Inventory V4.0,^{190,191,196} Sizing Them Up¹⁸⁴ and the YQOL-W.¹⁸⁵ Only two were assigned a certainty score of '2' (unfit for purpose): the EQ-5D-Y (EQ-5D youth version)²⁴¹⁻²⁴⁴ and the Paediatric Quality of Life Inventory V1.0.¹⁸⁹

Psychological well-being (4 certainty = '1'; 1 certainty = '2'; 12 certainty = '3')

Scores Similar to HRQoL, studies evaluating psychological well-being measures received high scores overall (see *Appendix 25*). In particular, one study evaluating the Social Anxiety Scale for Children²⁰³ was assigned a maximum of four for all tests conducted, which included IC, TRT reliability, internal validity and convergent validity. Of the 20 included studies, none was allocated the minimum quality score of '1' and only three were assigned a low score of '2'.^{197,204,211} Each of these, however, also conducted other forms of evaluation, in which higher scores of '3' and '4' were allocated.

Degree of certainty Four measures were deemed to be of high quality and were assigned a certainty score of '1' (i.e. fit for purpose) by the internal appraisal. These were the Self-Perception Profile for Children (SPPC),^{199,209} the Children's Physical Self-Perception Profile (C-SPSP),^{210,245} the Children's Self-Perceptions of Adequacy in and Predilection for Physical Activity (CSAPPA);²¹¹ and the CPSS.²⁰⁷ Only one measure was

allocated a certainty score of '2' (unfit for purpose): the Self-Report Depression Symptom Scale (CES-D).²⁴⁶ The remaining 12 measures required further consideration and were therefore assigned an uncertainty score of '3'.

Environment (5 certainty = '1'; 1 certainty = '2'; 4 certainty = '3')

Scores Internal appraisal scores were generally high for the evaluations of environmental measures, with 19 out of 33 evaluations receiving the maximum of '4' (see *Appendix 26*). Studies that were assigned maximum scores for all included evaluations were the evaluation of the NAPSACC,²⁴⁷ the EPAO (although reported only inter-rater reliability),²¹³ the electronic equipment scale²¹⁹ and the home PA equipment scale.²¹⁹ No studies were assigned a minimum score of '1' and only two were assigned low scores of '2'.^{216,217}

Degree of certainty Of the 10 included measures, five were deemed to be fit for purpose, and assigned a certainty score of '1'. These were NAPSACC,²⁴⁷ the environment and safety barriers to youth PA measure,²²² the Home Environment Survey (HES),²²⁰ the electronic equipment scale²¹⁹ and the home PA equipment scale.²¹⁹ Only one was deemed unfit for purpose – the HHS²¹⁴ – as this was an earlier version of a tool for which a newer version is currently under development.

Results of expert appraisal

Of the 180 measures that were appraised, a total of 52 outcome measures were recommended for inclusion to the CoOR outcome measures framework shown in *Table 4* (see *Final included studies: results from appraisal*). Information pertaining to the discussion, and key findings, of each measure is presented below according to outcome domain. Additional information, including reasons why some measures were excluded (i.e. internal team and expert's comments), can be found in *Appendix 17* (anthropometry measures) and *Appendix 28* (secondary outcome measures).

Anthropometry

Recommended anthropometric measures from the expert appraisal were (1) BMI and (2) DXA. Although BMI is limited by its inability to assess body composition or fat distribution, it provides an adequate overall proxy for health risks. Importantly, it is widely used and relatively easy to measure, compute and analyse. The ability of BMI to provide consistency between studies that would enable comparisons to be made between interventions is also highly valued. Experts agreed that research to consider thresholds for clinically significant changes would be useful, to encourage greater consideration of ESs and not just statistical significance. It was also clear, both from the evidence, and from agreement with experts, that, although self-reported height and weight may be adequate for some population based research designs, BMI for use in evaluation of interventions ought to be objectively measured.

Despite varied findings for the absolute accuracy of DXA, experts agreed that DXA was sufficiently precise to recommend its use for measuring changes in body composition [although experts admitted that they were basing decisions, in part, on wider evidence (e.g. in adults and/or other study designs that were not included in the CoOR review)]. Furthermore, DXA was considered to be a well-used methodology, with relatively good availability of the required equipment, at least in research and secondary care settings. Costs of DXA measurement, however, may well preclude its use, especially in public health evaluations.

Use of WC was not advocated by the experts, primarily because they felt it offered no benefit over BMI to measure treatment effects and was more subject to measurement error. There is considerable interobserver variability and bias may be related to body size. In addition, evidence gathered by CoOR did not include any validation using gold standard criterion methodologies. Skinfold measures have been extensively used and have been validated against more direct measures of body fatness. However the observer error is high and given the availability of superior methodologies, the CoOR expert group did not advocate using these measurements.

Remaining anthropometric measurements were not recommended primarily owing to a lack of existing validity evidence, with many measurements evaluated in only one study of obese children. Experts agreed that some of these (e.g. predicted thoracic gas volume²⁵¹) may hold potential but that there were insufficient data at present to recommend their use.

Experts emphasised the need to ensure that any anthropometric measurement is performed by trained staff using predefined techniques and standard operating procedures, and that equipment is calibrated on a regular basis. Additionally, it was recognised that there may be significant differences between different manufacturers and models of equipment. Such differences need to be examined and considered in future research. Experts also noted that the search did not identify any evaluations of the gold standard measures of the 4C model or TBW measurement in children.

Diet

Recommended dietary assessment tools are shown in *Table 4* (see *Final included studies: results from appraisal*, below). Of the 22 methodologies appraised, seven were recommended. All of these were FFQs. A total of 16 FFQs were appraised. Those that were deemed to be of a high standard (and were subsequently recommended) included measures with strong evidence in development and evaluation. However, at the time of writing this report (after appraisal), authors of one of these FFQs (the HSFFQ; Blum *et al.*³⁸) sent notification that it had been discontinued owing to maintenance costs. Thus, only six diet measures have now been included in the CoOR outcome measures framework.

Caveats for almost all recommended measures are noted, primarily related to the need to conduct further evaluation for validity and reliability evidence. Akin with all other secondary outcome domains, the specific characteristics of each measure need to be considered prior to deciding which one to use. For example, many have been developed and tested within predefined samples (ages, ethnicities) and are therefore only appropriate for use in similar populations. In the case of diet, the validity and reliability findings usually differ between different nutrients or foods. When choosing an appropriate measure, therefore, it is worth looking more closely at the original manuscript to ensure that it is robust for nutrients or foods that will be targets for change in an intervention.

Experts did not advocate any form of food diary or recall methodology. The decision to exclude these methodologies was initially based on evidence presented by the CoOR review, suggesting that validity of these measures was poor, especially in obese children. Additionally, evidence of reliability was lacking, with no TRT reliability evaluation conducted in the identified food diary studies and in only two studies evaluating recall methodologies. Conversely, 10 out of the 21 studies that evaluated a FFQ assessed TRT reliability in an obese sample. In addition to concerns raised by the evidence, experts also considered diary and recall methodologies to be less feasible, both in terms of participant burden (impacting the quality of data) and in the processing of data from these methodologies. Whereas data FFQ measures can be relatively easily entered, managed and analysed by people with no expertise in nutrition, this is not possible for diaries or recall methodologies, which require trained personnel (preferably a nutritionist/dietitian) for administration, data entry and analysis. Importantly, they are also reliant on having specific software for entry and up-to-date databases of foods and drinks. That said, depending on the specific FFQ, these issues may also be relevant and there is also likely to be a cost incurred for the questionnaire itself.

Overall, it was difficult to identify a measure of diet that all experts agreed they would highly recommend for inclusion into the outcome measures framework. Decisions considered the fact that this was a secondary outcome, specifically in trials evaluating childhood obesity treatment interventions. It was acknowledged that many of the decisions made by experts would not apply in considering other study designs or different populations. For example, experts are not suggesting that methods, such as food diaries, should not be advocated in other studies (especially those with a primary outcome of diet).

Eating behaviours

Twelve out of the 22 measures of eating behaviours that were appraised were recommended for inclusion to the CoOR outcome measures framework. These were chosen, in part, because of strong development and demonstration of reliability and validity, but also because experts were confident in their suitability and feasibility, through their own knowledge of the measures (primarily via previous use in this setting). Constructs that are assessed within these measures are varied (and described in *Table 4*) – see *Final included studies: results from appraisal*. Thus, like diet, the choice of measure should involve consideration of the constructs in which an intervention is expected to target (by the mechanism through which it will influence change). For example, some measures assess parental feeding styles, yet others assess constructs such as emotional eating, restrained eating and eating in the presence of hunger. Additionally, many of these measures are age specific, with questionnaires such as the IFQ specifically designed to assess parental behaviours related to infant feeding.

Although a similar (if not greater) level of evaluation was conducted for eating disorder diagnosis measures. Seven of these measures met eligibility criteria and were subsequently appraised. However, they were not recommended for inclusion to the CoOR outcome measures framework as they were deemed inappropriate for use as an outcome measure in an obesity treatment evaluation (even although many have been used in such designs) primarily because they result in a dichotomous outcome (i.e. presence or absence of a clinically defined eating disorder). In instances when researchers are concerned about the potential of an intervention to induce an eating disorder, these measures may have some potential.

Physical activity

Of the 24 PA outcome measures identified across 35 manuscripts, four were recommended for inclusion to the CoOR framework. These were (1) accelerometers, (2) pedometers, (3) SOCARP¹⁰² and (4) the Observational System for Recording Physical Activity in Children-Preschool version (OSRAC-P).¹⁰³ Although experts agreed that some of the self-reported measures were well developed, they did not advocate any owing to issues with reporting error in samples of obese children. It was recognised that the use of accelerometers may not always be feasible owing to costs and expertise in analysis but this method was viewed as the best measure for assessment of PA. It was acknowledged that data from accelerometers are often dependent on the model of accelerometer, which will improve and change with time. However, given that evidence in this area was outside of the scope of the CoOR review, readers were encouraged to refer to a review by de Vries *et al.*²⁵²

The CoOR evidence for pedometers was less strong but experts agreed it should be included as a less-expensive option, given that it offers objective measurement. Use of pedometers that show the user the number of steps and rely on participant reporting can be overcome by using sealed equipment in which the number of steps is not shown and data are automatically stored for download. However, pedometers should not be used as an outcome measure if they are an integral part of an intervention.

Experts recommended that two observation methodologies for measurement of PA be included in the outcomes framework.^{102,103} These measures did not fully meet CoOR eligibility criteria but were considered to have potential for inclusion. Expert felt that these measures offer an alternative to activity monitors, which are also not reliant on self-report.

One objective measure that was not recommended by experts was HR monitoring.²³⁷ CoOR evidence for this measurement was reliant on a small study of children ($n = 13$), which demonstrated low validity (with large variation in agreement with a gold standard of DLW). However, based on wider evidence from other populations, experts agreed that it may provide useful data when used in conjunction with an accelerometer.

Experts agreed that objective measurement of PA will continue to improve and, dependent on what the new data suggest, newer measures such as Actiheart® (CamNtech Ltd, Cambridge, UK) and SenseWear bands could be recommended.

Sedentary time

Measures identified by the CoOR review included those that assess sedentary behaviour, which would capture specific sedentary activities (e.g. time/frequency of watching television), and sedentary time, which measures the total time spent being inactive. Accelerometry was the only outcome measure – of six reviewed – that was recommended by experts. Accelerometers are not able to measure sedentary behaviours – only sedentary time. Thus, experts have only recommended a measure of sedentary time. In line with other recommendations, data from self-reported measures were deemed to be too affected by reporting bias in samples of obese children.

Similar to measures of PA, experts felt that there are many new and innovative methodologies currently being investigated that permit the objective measurement of sedentary behaviour but that a lack of evidence to date preclude their consideration at the time of writing (e.g. use of webcams and other recording devices/cameras), including those identified by the CoOR review.¹³⁴

Fitness

Only 1 out of the 13 outcome measures appraised in the fitness outcome domain was recommended by experts: measured VO_{2peak} .¹⁴¹ This measure is considered as the gold standard measure for fitness in children as measurement of VO_{2max} is often unacceptable and/or not achievable (based on compliance), especially in obese children. Evidence presented by CoOR was based on one study,¹⁴¹ which conducted evaluations in a small sample of overweight and obese children. However, given the wider evidence of its use in children, experts agreed that it should be included. There was debate, however, about whether the test should be conducted with a treadmill or bike. Lofkin *et al.*¹⁴¹ compared both methods and found the bike to be more acceptable to obese children.

Experts agreed that findings for many of the other outcome measures identified by CoOR were dependent on body weight (e.g. shuttle run, step test, etc.). These tools may be useful for within person comparisons but were not advocated as trial outcomes for the CoOR outcome measures framework. Similar to other domains in which objective measures are available, experts did not recommend self-reported fitness measures.

Physiology

Of the 12 physiological outcomes (described in 26 manuscripts) only one – ‘indices of insulin sensitivity’ – was recommended for inclusion into the framework. Experts stated that physiological outcomes have potential to act as a primary outcome, given that they are indicators of cardiovascular health which is associated with obesity. Furthermore, evidence presented by CoOR and wider evidence outside obesity research indicates that many physiological outcomes can be measured with a high degree of precision (and are often feasible to obtain based on routine clinical measurement). However, based on evidence specific to research in children with obesity, only ‘indices of insulin sensitivity’ offered a sufficient degree of validity evidence (with many studies demonstrating criterion validity comparing against a gold standard of the EHC test). It is important to note that there was considerable debate around use of this outcome measure, as at present there is no evidence related to what constitutes clinical meaningfulness within childhood obesity treatment evaluations. A further scoping search was conducted by the CoOR team, with inclusion of terms specific to all physiological measures and criteria/cut-offs to determine whether wider evidence of what is clinically meaningful existed outside the knowledge of the experts (see *Appendix 16*). However, this did not identify any further data within an obesity paediatric population. Given that other outcome domains also lack information on what is clinically meaningful (e.g. anthropometric outcomes), the team decided to continue to advocate ‘indices of insulin sensitivity’ to the framework. Experts agreed that these offer good surrogates for insulin sensitivity, but pubertal status may affect results, which should therefore be taken into account. There was some concern about the sensitivity of these indices in small samples, and other methods to assess insulin sensitivity may be more appropriate for individuals or small groups (e.g. hyperglycaemic clamp). However, there are clear practical limitations to their use in children.

Eight manuscripts^{151,166–172} within the physiological domain described an evaluation of estimated energy expenditure. These may have been more appropriately added to the fitness outcome domain (as they do not necessarily imply 'metabolic risk'). However, given that none of the energy expenditure measures were advocated, it was agreed to continue to consider energy expenditure within the physiological domain. Results for validation were variable, and one paper that was specifically focused on obese children¹⁷¹ showed a range of correct predictions (comparing predictions to a ventilated hood method) of between 12% and 74%. Overall, study validation results were poor to moderate and this outcome measure was therefore not recommended at present.

Health-related quality of life

Of the 12 HRQoL measures that were appraised by CoOR, 10 were recommended for the CoOR outcome measures framework by experts shown in *Table 4* (see *Final included studies: results from appraisal*, below). These measures were generally well developed and provided evidence of high reliability and validity, with some specific to childhood obesity. The only two measures that were not recommended were earlier versions of the Paediatric Quality of Life Inventory.^{188,189,195} Many of the HRQoL tools had been well used by previous studies within and outside obesity research, and experts noted that any of the included tools could be used subject to context. Similar to other secondary outcome domains, deciding which of the HRQoL measures to use should be based on choosing one that is mostly clearly aligned to the constructs that are expected to change as part of a specific intervention. With this in mind, it would be acceptable to choose a generic HRQoL measure over an obesity-specific measure if appropriate.

Psychological well-being

Of the 17 psychological well-being outcome measures that were appraised by CoOR, experts agreed to include 10 (see *Final included studies: results from appraisal* and *Table 4*). These measures were generally well developed (often involving participants) and demonstrated high-quality evaluation (although results were variable). As they capture a range of different concepts (e.g. self-efficacy, perception of body image, social acceptance, enjoyment, etc.), the decision of which to choose has to be based on the specific requirements of each study. Like other domains, it is important to choose outcomes and corresponding measures that capture what it is that is being targeted by the intervention. There was some debate about the age of some of the measures and whether their language and concepts are remain relevant. This was especially important for the SPPC²⁰⁹ (previously Perceived Competence Scale¹⁹⁹), which had been originally developed in 1982. However, in looking specifically at the scales, experts agreed that they were still current and captured the fundamental domains in a child's life, such as school and appearance, encompassed in global self-worth. This particular measure is well used, and, although some argue that it is a challenge for adults to administer, experts agreed that the majority of children found the style to be highly acceptable. Other 'older' measures were judged on a case-by-case basis to determine whether the scales and/or items remained relevant today.

Excluded measures were not recommended because they were based on poor validity results,^{197,200,210} focused on eating disorders¹⁹⁸ or developed for a completely different population.²⁰⁸

Environment

Of the 10 included environmental measures (described in nine manuscripts^{213–220,236,247}), five were recommended for inclusion into the CoOR framework. The most likely environment targeted for change in childhood treatment interventions is the home environment. The CoOR framework recommended three different measures of the home environment^{218,219} (two studies). The first measure – the 'HES'²¹⁸ – assesses the physical (e.g. food availability) and social environment (e.g. parental role modelling). Two other measures described in a study by Rosenberg *et al.*²¹⁹ are more like checklists of equipment that are available in the home (electronic equipment scale and the home PA equipment scale). Two additional measures that were recommended included one that measures a child-care environment ('NAPSACC'²¹²) and another that is an assessment of parental and child perception of environments related to barriers to PA.²²⁰

The decision to include NAPSACC was debated, as this is a measure, of a child-care environment, which may be more suited as an outcome measure in prevention evaluations. However, experts were aware of existing obesity 'treatment' interventions that target infants at high risk of obesity within child-care environments, which led to its inclusion.

Exclusion of other measures was primarily based on inadequate validity and reliability findings. Experts felt that some demonstrated 'potential' but that more evaluation with larger sample sizes would be required before advocating their use. Although this outcome domain as relatively few recommended measures, interest in this area of research is extremely popular and the experts agreed that there are potentially many more measures that may be appropriate for use that did not meet the inclusion criteria for CoOR. It is likely, for example, that many newly developed measures will be used as trial outcome measures in the future. Although experts are aware that this area of methodology has gained popularity over recent years, this was not demonstrated by the literature probably due, in part, to the CoOR eligibility criteria. Although many environmental measures have been developed for use in obesity research, a majority of these are appropriate for use in the evaluation of obesity prevention interventions (measures of the built environment, community food environments, etc.).

Summary of key findings

- Body mass index and DXA were advocated as primary outcomes. Recommendation of BMI was primarily based on ensuring comparability across studies (plus, ease of use and relatively low measurement error). DXA was advocated as an additional measure to BMI if feasible as a means to estimate adiposity.
- In the diet domain, only FFQs were recommended, which had greater evidence of reliability and validity, and were less dependent on weight status than other methods.
- Although often used (and generally well developed), eating disorder screening questionnaires were not advocated as outcome measures in childhood obesity treatment evaluations.
- Objective measures were recommended by experts where available. Although generally well developed, self-reported measures were deemed to be too much subject to reporting bias in this population.
- Measurement of sedentary behaviour (e.g. television watching) and sedentary time (e.g. time spent inactive) need to be viewed as separate domains.
- Validity findings for many fitness outcomes were poor and/or highly variable. Importantly, many were highly dependent on body weight. Such measures may be of use in within-person comparisons but were not recommended as trial outcomes. $\dot{V}O_{2peak}$ was the only fitness outcome to be recommended by experts.
- Physiological outcomes are indicators of cardiovascular health and therefore have the potential to act as a primary outcome. However, experts felt that further evidence is regarding establishing minimally important difference (MID) in obese children. In this domain, only 'indices of insulin' was recommended by experts, which were considered to offer a more practical approach to assess insulin compared with gold standard methods (i.e. EHC). This recommendation was based on strong evidence of validity.
- The CoOR team are aware of the development of preference-based utility measures that permit assessment of QALYs in obese children. However, manuscripts were not available for review at the time of writing.
- New technologies and innovative ideas are currently being developed that will enable further development and refinement of measures. Data on these measures are insufficient to use in current recommendations.
- Recommendations are specific to evaluation of obesity treatment evaluations in children. These considerations may not be applicable to other types of studies or setting (e.g. surveys, cohorts, intensive experimental interventions and some public health evaluations).

Final included studies: results from appraisal

The CoOR outcome measures framework is shown in *Table 4*. Efforts were made to obtain further information regarding accessing and feasibility for each of these measures and are provided if available (from authors, websites and information from manuscripts). Incomplete information within the table indicates that no further information was obtained from these sources.

TABLE 4 The CoOR outcome measures framework

Measurement name	First author; administration; suitable child age range ^a	Description	Access/feasibility
Anthropometry			
BMI/BMI-SDS	Multiple papers (see <i>Appendix 6</i>)	BMI [weight (kg)/height (m) ²]	Requires scales (regularly calibrated) and a stadiometer to measure height
	Trained researcher/clinical staff	BMI-SDS (age-adjusted BMI)	Existing staff/administrators can be trained to measure with good accuracy
DXA	All age groups		
	Multiple papers (see <i>Appendix 6</i>)	DXA bone density measurement technology, which can estimate body composition (including adiposity)	Requires specialised machinery and staff
	Trained researcher/clinical staff		Cost of each measurement estimate £50–200
	All age groups		
Diet^b			
Short Youth Adolescent Questionnaire (Short YAQ), 26 item	<i>Rockett 2007</i> ³⁴	Fruit, vegetables (carrots only), cereals, white meat, red meat, milk and milk products, snacks, sugar sweetened beverages, non-sugar sweetened beverages.	Access: https://regepi.bwh.harvard.edu/health/KIDS/files
	Self-complete	Note: Most items are presented as 'meals' rather than individual components (e.g. chicken or turkey sandwich')	Copyright: EliteView(TM)
	Suitable for children and adolescents		Cost: Costs incurred for questionnaires and analysis (although can opt to do analysis independently). See website for details Feasibility: No information for Short YAQ. Duration for completion of full YAQ = 20–30 minutes
Youth Adolescent Questionnaire (YAQ), 131 item	<i>Rockett 1995</i> ⁴³	Fruit, vegetables, cereals, white meat, red meat, fish, milk and milk products, snacks, sugar sweetened beverages	Access: Through website https://regepi.bwh.harvard.edu/health/KIDS/files
	<i>Rockett 1997</i> ³⁷		Copyright: EliteView(TM)
	<i>Perks 2000</i> ³⁰		Cost: Costs incurred for questionnaires and analysis (although can opt to do analysis independently). See website for details Feasibility: Duration for completion of full YAQ = 20–30 minutes

continued

TABLE 4 The CoOR outcome measures framework (continued)

Measurement name	First author; administration; suitable child age range ^a	Description	Access/feasibility
Children's Eating Habits Questionnaire (CEHQ-FFQ), 43 item	<i>Lanfer 2011</i> , ³⁶ <i>Huybrechts 2011</i> ³¹ Parent completed Suitable for children	Fruit, vegetables, cereals, white meat, red meat, fish, milk and milk products, snacks, oils/condiments, nuts, sugars, sugar sweetened beverages, non-sugar sweetened beverages, ready-made meals, baked foods	Access: Via author at ahrens@bips.uni-bremen.de Copyright: Intellectual property of study consortium. The paper by Lanfer ³⁶ may serve as a reference Cost: Freely available, but asked to cite paper/book. Costs will be sought if requesting SAS code for managing the data and/or for defining (derived) variables Feasibility: Not evaluated
Australian Child and Adolescent Eating Survey (ACAES), 137 item	<i>Watson 2009</i> , ^{2,46} <i>Burrows 2008</i> ³² Self-complete Suitable for children and adolescents	Fruit, vegetables, red meat, milk and milk products, snacks, oils/condiments, sugar sweetened beverages, non-sugar sweetened beverages, ready-made meals, baked foods	Access: Via Newcastle Innovation at innovation@newcastle.edu.au or www.newcastleinnovationhealth.com.au/research-partners/food-frequency-questionnaires# Copyright: Prior to use, researchers are required to complete a signed agreement. The agreement outlines the terms and conditions of using the ACAES FFQ to ensure it is utilised appropriately and the nutrient data are processed accurately. The agreement can be obtained online at addresses above Cost: Yes – includes scanning, data processing and preparation of a dataset (not analysis). Cost per survey is A\$17, with discounts for > 100 surveys Feasibility: Duration of completion = 20–30 minutes
Diet fat-screening measure, 21 item	<i>Prochaska 2001</i> ⁵⁰ Self-complete Suitable for adolescents	High-fat foods/meals including burgers, pizza, ice cream, whole milk, oils/dressings, etc.	Access: Listed on website (within PACES): http://sallis.ucsd.edu/measure_paceadol.html Copyright: No information Cost: Website indicates that measures are free for research purposes. Links to gnorman@paceproject.org for further information Feasibility: Duration of completion = 5–10 minutes; duration of scoring = 2–3 minutes

TABLE 4 The CoOR outcome measures framework (continued)

Measurement name	First author; administration; suitable child age range ^a	Description	Access/feasibility
New Zealand FFQ, 117 item	<i>Metcalf 2003</i> ²⁴⁷ Parent completed Suitable for children (up to 14 years)	Fruit, vegetables, cereals, white meat, red meat, fish, milk and milk products, snacks, oils/condiments, sugar sweetened beverages, baked foods	Feasibility: (from manuscript): Duration of completion = 20 minutes
Eating behaviours			
Infant Feeding Questionnaire (IFQ), 20 item	<i>Baughcum 2001</i> ⁷⁴ Parent completed Suitable for infants	Concern about infants weight Concern about infant hunger Concern about how much infant eats Control over how much infant eats Using food to calm infant Attention/nurturance by mother during feeding Established feeding schedule Awareness of infants hunger and satiety cues	Access: The instrument is not available online. Scale items are shown (verbatim) in table 1 of the paper Copyright: None Cost: Freely available Feasibility: Not measured
Preschool Feeding Questionnaire (PFQ), 32 item	<i>Baughcum 2001</i> ⁷⁴ Parent completed Suitable for preschoolers (infants and children)	Maternal concern about child weight Structure during feeding interaction Difficulty in child feeding Pushing child to eat more Using food to calm child Child control of feeding interaction Age-inappropriate feeding	Access: The instrument is not available online. Scale items are shown (verbatim) in table 5 of the paper Copyright: None Cost: Freely available Feasibility: Not measured
Dutch Eating Behaviour Questionnaire for Children (DEBQ-C), 20 item	<i>Van Strien 2008</i> , ⁷⁹ <i>Banos 2011</i> , ⁸³ <i>Braet 2007</i> ⁹² Self-complete Suitable for children and adolescents	Emotional eating Restrained eating External eating	Access: Via author: Lien Goossens, Lien.Goossens@UGent.be. If using for commercial purposes, contact Tatjana Van Strien: t.vanstrien@psych.ru.nl Copyright: None for child version Cost: Freely available for non-commercial purposes

continued

TABLE 4 The CoOR outcome measures framework (continued)

Measurement name	First author; administration; suitable child age range ^a	Description	Access/feasibility
Dutch Eating Behaviour Questionnaire for Children (DEBQ-P), 33 item	<i>Caccialanza 2004</i> , ⁹⁸ <i>Braet 1997</i> ⁷⁸ Parent completed Suitable for children and adolescents	Emotional eating Restrained eating External eating	Feasibility: Statistical analysis code available on request to author. Duration of administration ~10–20 minutes Access: Via author: Lien Goossens, Lien.Goossens@UGent.be. If using for commercial purposes, contact Tatjana Van Strien: t.vanstrien@psych.ru.nl Copyright: None for child version Cost: Freely available for non-commercial purposes Feasibility: Statistical analysis code available on request to author. Duration of administration ~10–20 minutes
Emotional Eating Scale for Children and Adolescents (EES-C), 26 item	<i>Tanofsky-Kraff 2007</i> ⁷⁷ Self-complete Suitable for children and adolescents	Eating in response to anger, anxiety and frustration Eating in response to depressive symptoms Eating in response to feeling unsettled	Access: Via author Marian Tanofsky-Kraff, marian.tanofsky-kraff@usuhs.edu Copyright: None, but requested to cite published papers Cost: None Feasibility: None reported/evaluated
Child Feeding Questionnaire (CFQ), 31 item [16-item version also available (Anderson 2005 ⁷⁵)]	<i>Birch 2001</i> , ⁷⁵ <i>Haycraft 2008</i> , ⁹³ <i>Anderson 2005</i> , ⁹⁶ <i>Corsini 2008</i> , ⁹⁷ <i>Polat 2010</i> , ⁹⁴ <i>Boles 2010</i> ²³² Parent completed Suitable for infants and children	Perceived responsibility Parent-perceived weight, perceived child weight Parents concern about child weight Monitoring Pressure to eat Restriction	Access: Via author Sheryl Hughes, shughes@bcm.edu Copyright: None Cost: None Feasibility: See Anderson 2005 ⁹⁶
Infant Feeding Style Questionnaire (IFSQ), 83 item (64-item version available for infants of < 6 months)	<i>Thompson 2009</i> ⁷⁶ Parent completed Suitable for infants	Styles: <ul style="list-style-type: none">● Laissez-faire● Pressuring/controlling● Restrictive/controlling● Responsive● Indulgent	Access: E-mail to althomps@email.unc.edu Copyright: None Cost: None Feasibility: Deemed acceptable based on low levels of missing data. No information on duration

TABLE 4 The CoOR outcome measures framework (continued)

Measurement name	First author; administration; suitable child age range ^a	Description	Access/feasibility
Children's Eating Behaviour Questionnaire (CEBQ), 35 item	<i>Sleddens 2008</i> , ⁷² <i>Wardle 2001</i> ⁷³	Food fussiness	Access: From website: www.ucl.ac.uk/hbr/c/ Copyright: None Cost: None Feasibility: Was perceived as quick and easy by parents. A further version for infants (the Baby Eating Behaviour Questionnaire) is also available on the website. Authors are also currently developing a self-completion version for adolescents
		Enjoyment of food	
	Parent completed	Food responsiveness	
	Suitable for children	Emotional overeating	
		Satiety responsiveness	
		Emotional undereating	
		Desire to drink	
Slowness in eating			
Toddler Snack Food Feeding Questionnaire (TSFFQ), 42 item	<i>Corsini 2010</i> ⁸²	Rules	Access: Email (from manuscript) nadia.corsini@csiro.au
		Parent completed	
	Suitable for children and infants	Self-efficacy	
		Flexibility	
		Allow access	
Kids' Child Feeding Questionnaire (KCFQ), 28 item (16-item version also available)	<i>Monnery-Patris 2011</i> , ⁸⁵ <i>Carper 2000</i> ²⁵⁰	Restriction and pressure to eat	Access: Via author Sandrine Monnery-Patris, Sandrine.Monnery-Patris@dijon.inra.fr or within manuscript: www.ncbi.nlm.nih.gov/pubmed/21565236 www.sciencedirect.com/science/article/pii/S0195666311001358 Copyright: None, but the author would like to be notified of its use Cost: None Feasibility: Completion in 5–10 minutes
		Self-completed	
	Suitable for children		
Un-named (control in parental feeding practices), 29 item	<i>Murashima 2011</i> ⁸⁴	Non-directive, food environmental control, high control, high contingency, child-centred feeding, encouraging nutrient-dense foods, discouraging energy-dense foods, meal-time behaviours, timing of meals	Access: (from manuscript) murashi1@msu.edu Feasibility: Items and details of scoring provided as an appendix within the paper

continued

TABLE 4 The CoOR outcome measures framework (continued)

Measurement name	First author; administration; suitable child age range ^a	Description	Access/feasibility
Eating in the Absence of Hunger questionnaire (EAH-C), 14 item	<i>Tanofsky-Kraff 2008</i> ⁸⁰ Self-completed Suitable for children and adolescents	Negative effect, external eating, fatigue/boredom	Access: Via author Marian Tanofsky-Kraff, marian.tanofsky-kraff@usuhs.edu Copyright: None, but requested to cite published papers Cost: None Feasibility: None reported/evaluated
Physical activity			
Accelerometer	<i>Guinhouya 2009</i> , ²³⁴ <i>Coleman 1997</i> , ¹⁰⁸ <i>Noland 1990</i> , ¹⁰⁶ <i>Pate 2006</i> , ¹⁰⁷ <i>Kelly 2004</i> ¹⁰⁵ Suitable for infants, children and adolescents	Measurement devices for assessment of acceleration forces/movement intensity. Calculates frequency and duration of PA	Monitor (excluding software) costs ~£150–300 each
Pedometer	<i>Duncan 2007</i> , ²⁴⁸ <i>Kilanowski 1999</i> , ¹¹⁴ <i>Treuth 2003</i> , ¹¹³ <i>Jago 2006</i> , ¹¹² <i>Mitre 2009</i> ¹¹⁰ Suitable for infants, children and adolescents	Measures number of steps taken (usually daily)	Sealed equipment costs ~ £10–100 each
Observational System for Recording Physical Activity-Preschool Version (OSRAC-P), eight categories	<i>Brown 2006</i> ¹⁰³ Researcher conducted Suitable for infants and children	Activity level, type of activity, social (e.g. initiator of activity, group composition) and non-social (e.g. child location) environment circumstances	Access: Author provides email in the manuscript: bbrown@gwm.sc.edu Manual available at: www.sph.sc.edu/USC_CPARG/pdf/OSRAC_Manual.pdf Tool download available at: www.sph.sc.edu/usc_cparg/osrac.html Copyright: Authors report no official copyrighted though state it has sufficient documented history of its development and use by the Children's Physical Activity Group to give some intellectual property rights Feasibility: requires systematic training of several weeks to move trained observers to interobserver agreement
System for Observing Children's Activity and Relationships	<i>Ridgers 2010</i> ¹⁰² Researcher conducted Suitable for children	Activity level, group size, activity type, interactions	Access: Protocols manual and observation tool available via email to Nicky Ridgers: nicky.ridgers@deakin.edu.au

TABLE 4 The CoOR outcome measures framework (continued)

Measurement name	First author; administration; suitable child age range ^a	Description	Access/feasibility
During Play (SOCARP)			<p>Copyright: Use of the tool should reference Ridgers <i>et al.</i> (2010)⁸³</p> <p>Cost: None associated with use or analysis, except for staff time taken to establish the interobserver reliability for those who will be collecting the data</p> <p>Feasibility: Training for interobserver reliability = 10–20 hours; Each observation period is 10 minutes in length (one observer during a 60-minute lunchtime would be expected to record data from five to six children)</p>
Accelerometer	Puyau 2002, ¹³² Reilly 2003 ¹³¹	Measurement devices for assessment of acceleration forces/movement intensity. Calculates frequency and duration of PA	Monitor (excluding software) costs ~£150–300 each
Measured VO _{2peak}	Loftin 2004 ¹⁴¹	Measures the amount of oxygen consumed/minute while conducting a graded fitness test on a treadmill, bike or other piece of cardio exercise equipment. VO _{2peak} = peak amount of oxygen used for energy during the test	Specialist equipment to conduct and analyse the data needed, in addition to trained staff. VO _{2peak} is more acceptable to participants than VO _{2max} but still requires full cooperation and a degree of burden
Physiology			
Indices of insulin sensitivity	Rossner 2008, ¹⁶¹ Keskin 2005, ¹⁶⁰ Atabek 2007, ¹⁵⁹ Gungor 2004, ¹⁵⁸ George 2011, ¹⁵⁴ Conwell 2004, ¹⁵³ Yeckel 2004, ¹⁵² Uwaifo 2002, ¹⁵⁶ Schwartz 2008, ¹⁶² Gunczler 2006 ¹⁵⁵	<p><i>Includes indices:</i></p> <p>HOMA-IR</p> <p>QUICKI</p> <p>FGIR</p> <p>FIRI</p> <p>ISI COMP</p> <p>HOMA-B%</p> <p>WBISI</p>	Derived from blood insulin and glucose concentrations under fasting conditions (steady state) or after an oral glucose load (dynamic). Relatively inexpensive surrogates usually taken in (but not restricted to) a clinical setting
Child Health Questionnaire (CHQ), 50 item	Waters 2000, ^{192,193} Landgraf 1998 ¹⁸⁶	<p>Physical functioning</p> <p>Role social emotional</p> <p>Role social physical</p> <p>Bodily pain, mental pain, behaviour</p>	<p>Access: Via website: www.healthactchq.com/</p> <p>Copyright: Registration and licensing required</p> <p>Costs: Yes, fees depend on needs of study</p>

continued

TABLE 4 The CoOR outcome measures framework (continued)

Measurement name	First author; administration; suitable child age range ^a	Description	Access/feasibility
		Self-esteem	Feasibility: Completion in 10–15 minutes
		General health, parent impact – emotional, parent impact – time, family activities, family cohesion	
		Change in health	
DISABKIDS, 37 item	<i>Ravens-Sieberer 2007</i> ¹⁹⁶	Physical well-being	Access: Website: www.child-public-health.org/english/research/
	Self-report and parent-report versions	Psychological well-being	Copyright: See website
	Suitable for children and adolescents	Moods and emotion, self-perception	Feasibility: 12-item version also available, in addition to ‘smiley face’ version for aged 4–6 years. Computer-assisted versions also available
		Autonomy	
		Parent relation and home life	
		Peers and social support	
		School environment	
		Bullying	
		Financial resources	
KIDSCREEN (short), 27 item	<i>Ravens-Sieberer 2007</i> ¹⁹⁴	Physical well-being	Access: Website: www.child-public-health.org/english/research/ Email: Ravens-Sieberer@uke.uni-hamburg.de
	Self-report	Psychological well-being	Copyright: To the KIDSCREEN Group. User agreement required
	Suitable for children and adolescents	Moods/emotions	Cost: Free for non-industry research
		Self-perception	Feasibility: 52-item (long) and 10-item (short) versions also available. Duration for completion ~10–15 minutes
		Autonomy	
		Parent relation and home life	
		Peers and social support	
		School environment	
		Bullying	
		Financial resources	
EQ-5D-Y, 5 item, plus VAS for overall health	<i>Burstrom 2011</i> , ^{241,242} <i>Wille 2010</i> , ²⁴³ <i>Ravens-Sieberer 2010</i> ²⁴⁴	Mobility	Access: English versions via oemar@euroqol.org (EuroQol Business Office)
	Self-report	Self-care	Copyright: Copyrighted and cannot be altered/modified
	Suitable for children or adolescents	Usual activities	Cost: Free to use for non-industry research
		Pain and discomfort	Feasibility: 91–100% complete data obtained in a multinational study (indicating comprehension/acceptability)
		Anxiety and depression	

TABLE 4 The CoOR outcome measures framework (continued)

Measurement name	First author; administration; suitable child age range ^a	Description	Access/feasibility
Impact of Weight on Quality of Life (IWQoL), 27 item	<i>Kolotkin 2006</i> , ¹⁸¹ <i>Modi 2011</i> ¹⁸²	Physical comfort, body esteem	Access: E-mail to Ronette L. Kolotkin: rkolotkin@qualityoflifeconsulting.com or www.qualityoflifeconsulting.com Copyright: Copyright © Ronette L. Kolotkin and Cincinnati Children's Hospital Medical Centre. All commercial rights are owned by Quality of Life Consulting, PLLC, Durham, NC, USA. The questionnaires may not be used without permission and a licence agreement Cost: The licence fee is US\$10/participant for commercially funded studies, US\$5/participant for government-funded, foundation-funded or internally supported studies, or US\$3 per administration for clinical practices Feasibility: Duration for completion = ~8 minutes
	Self-complete or parent-complete versions	Social life	
	Suitable for adolescents	Family relations	
KINDL-R questionnaire, 24 item	<i>Erhart 2009</i> ¹⁸⁷	Physical well-being	Access: Via website http://kindl.org/cms/fragebogen/langswitch_lang/en Copyright: Any duplication or distribution is permitted only with the prior consent of the author, and requests that citations and date are quoted. User agreement required Cost: Free to non-industry research Feasibility: Duration for completion ~10 minutes. Translated in many languages and different versions available for differing age groups
	Self-report	Emotional well-being	
	Suitable for adolescents	Self-worth	
		Well-being in the family	
		Well-being related to friends/peers	
School-related well-being			
Paediatric Quality of Life Inventory V4.0, 23 item	<i>Varni 2001</i> , ¹⁹⁰ <i>Varni 2003</i> , ¹⁹¹ <i>Hughes 2007</i> ¹⁹⁶	Physical	Access: Need to complete a user agreement form: details on-line at www.pedsqol.org . Can also send informal queries to PROinformation@mapi-trust.org Copyright: Reserved to Dr James W Varni
	Self-report/parent report	Emotional	
	Suitable for infants, children and adolescents	Social	
		School	
		Functioning	

continued

TABLE 4 The CoOR outcome measures framework (continued)

Measurement name	First author; administration; suitable child age range ^a	Description	Access/feasibility
Sizing Me Up (self-report)/Sizing Them Up (parent report), 22 item	<i>Modi 2008</i> , ¹⁸⁴ <i>Zeller 2009</i> ¹⁸³ Self-report/parent report Suitable for children and adolescents	Two measures of: Emotional functioning Physical functioning Teasing Positive social attributes Social avoidance (self-report) Mealtime challenges (parent report) School functioning (parent report)	Cost: See website for details. Funded academic research = US\$990 per study (including delivery of one module + US\$330 per additional module + US\$25 for bank expenses). Non-funded = free Feasibility: Duration for completion ≤ 4 minutes Access: Website: www.cincinnatichildrens.org/research/divisions/c/adherence/labs/modi/hrqol/sizing/default/ Email: meg.zeller@cchmc.org Copyright: Copyright agreement (obtained from website) Cost: None (provided agreement is signed) Feasibility: Sizing Them Up and Sizing Me Up can be used together in clinical and research settings. Duration for completion = 15 minutes each
Youth Quality-of-Life Instrument-Weight Module (YQOL-W), 21 item	<i>Morales 2011</i> ¹⁸⁵ Self-completed Suitable for children and adolescents	Self, social and environment scales	Access: Via website: http://depts.washington.edu/seaqol/ Copyright: Yes, a user's agreement is required Costs: US\$500 industry, US\$200 public/university, students free (not including analysis) Feasibility: Duration for completion = 5–10 minutes
Psychological well-being			
Children's Body Image Scale (CBIS) Pictorial (photograph) scale	<i>Truby 2002</i> ¹⁹⁸ Self-completed Suitable for children	Gender-specific self-perception of body image (child identifies image most like their own out of seven images)	[No response: experts that felt this was now out of copyright and is now freely available]
Body figure perception (pictorial), 5 item	<i>Collins 1991</i> ²⁰⁵ Self-completed Suitable for children	Self Ideal self Ideal other child Ideal adult Ideal other adult	Access: Manuscript provides pictures Feasibility: Pictorial scale not dependent on literacy level (and can be used in different languages)

TABLE 4 The CoOR outcome measures framework (continued)

Measurement name	First author; administration; suitable child age range ^a	Description	Access/feasibility
Self-Perception Profile for Children (SPPC), 36 item	<i>Van Dongen-Melman 1993</i> ²⁰⁹	Scholastic competence	Access: Website: www.nlsinfo.org/childya/nlsdocs/guide/assessments/SPPC.htm Copyright: Experts stated that this is freely available (information not provided by authors) Feasibility: Perceived Importance Profile (PIP) is recommended to use in conjunction with the SPPC (Whitehead 1995, ²¹⁰ below)
	Self-completed	Social acceptance	
	Suitable for children	Athletic performance	
		Behavioural conduct	
		Global self-worth	
Perceived Competence Scale (aka SPPC/Harter), 28 item	<i>Harter 1982</i> ¹⁹⁹	Cognitive competence	Access: Website: www.nlsinfo.org/childya/nlsdocs/guide/assessments/SPPC.htm Copyright: Experts stated that this is freely available (information not provided by authors)
	Self-completed	Social competence	
	Suitable for children	Physical competence	
		General self-worth	
Physical Activity Enjoyment Scale (PACES), 12 item	<i>Motl 2001</i> ²⁴⁹	Enjoyment	Access: Lead author, robmotl@illinois.edu ; corresponding author for manuscript, rdishman@coe.uga.edu
	Self-completed	Factors influencing enjoyment in PA	
	Suitable for adolescents		
Children's Physical Self-Perception Profile (C-PSPP), 24 item	<i>Whitehead 1995,</i> ²¹⁰ <i>Eklund 1997</i> ²⁴⁵	Attractive body adequacy	Access: Via author James Whitehead, james.whitehead@email.und.edu Copyright: None Cost: None Feasibility: Note from authors: with the structured alternate response format, it is important to use the sample item to explain it to participants
		Strength competence	
	Self-completed	Condition/stamina	
		Sport competence	
		Physical condition	
	Suitable for children and adolescents	Competence	
		Physical self-worth	
		General self-worth	
	Children's Self Perception of Adequacy in and Predilection for Physical Activity (CSAPPA), 20 item	<i>Hay 1992</i> ²¹¹	
Self-completed		Predilection	
Suitable for children and adolescents		Enjoyment of physical education	
Children's Physical Self-Concept Scale (CPSS), 27 item	<i>Stein 1998</i> ²⁰⁷	Physical performance	Access: Items available within manuscript (Stein 1998 ²⁰⁷)
	Self-completed	Physical appearance	
	Suitable for children	Weight control	

continued

TABLE 4 The CoOR outcome measures framework (continued)

Measurement name	First author; administration; suitable child age range ^a	Description	Access/feasibility
Social Anxiety Scale for children, 22 item	<i>La Greca 1993</i> , ²⁰² <i>1988</i> ²⁰¹ Self-report Suitable for children	Fear of negative evaluation from peers Social avoidance and distress around new peers or in new situations Generalised social avoidance and distress	Access: Enquires to Liz Reyes at ereyes@miami.edu . Website: www.psy.miami.edu/faculty/alagreca/#social_anxiety Copyright: By Annette M. La Greca and may be used only with her written permission Cost: For manual, US\$15.00 Feasibility: The manual for the Social Anxiety Scales contains detailed psychometric and normative information, information on translations, and copies of the scales and their scoring. Adolescent version also available
Body Esteem Scale (BES), 24 item	<i>Mendelson 1982</i> ²⁰⁴ Self-report Suitable for children and adolescents	Appearance, weight and attribution	Access: Via e-mail to: stephen.franzoi@marquette.edu or sashields@psu.edu Copyright: Researchers must forward details of any research conducted with the measure to the author: bev@ego.psych.mcgill.ca
Environment			
Nutrition and Physical Activity Self-Assessment to Child Care (NAPSACC), 56 item	<i>Benjamin 2007</i> ²⁴⁷ Child-care centre staff completed Suitable for infants	<i>Child-care setting:</i> F&V, fried food and high-fat meat, beverages, menu and variety, meals and snacks Foods outside of regular meals and snacks Supporting HE Nutrition education for children, parents and staff Nutrition policy Active play and inactive time Television use and television viewing Play environment PA education for children, parents and staff Supporting PA PA policy	Access: E-mail to dsward@email.unc.edu . Link to associated intervention webpage (with details of researcher conducted version) at www.napsacc.org/ Copyright: None Cost: None Feasibility: Requires completion by child-centre staff

TABLE 4 The CoOR outcome measures framework (continued)

Measurement name	First author; administration; suitable child age range ^a	Description	Access/feasibility
Environment and Safety barriers to Youth Physical Activity Questionnaire, 21 item	<i>Durant 2009</i> ²⁰ Parent and child completed Suitable for adolescents	<i>Child and parent perception of built environment:</i> Street environment Street safety Park environment Park safety	Access: Measure used as part of the Active Wear Study. Details of all measures, including this, at http://sallis.ucsd.edu/measure_activewhere.html
Home Environment Survey (HES), 105 item	<i>Gattshall 2008</i> ¹⁸ Parent completed Suitable for children	<i>Home environment:</i> PA availability PA accessibility PA parental role modelling PA parental policies F&V availability F&V accessibility Fat/sweets availability HE parental role modelling HE parental policies	Access: Via email to Michelle.Gattshall@kp.org Copyright: None Cost: None Feasibility: Not reported
Home electronic equipment scale, 21 item	<i>Rosenberg 2010</i> ¹⁹ Parent and self-completed Suitable for children and adolescents	<i>Home environment:</i> Electronics available in the home Electronics available in the child's or adolescent's bedroom Portable electronics	Access: Measure used as part of the Active Wear? Study. Details of all measures, including this at http://sallis.ucsd.edu/measure_activewhere.html
Home PA equipment scale, 14 item	<i>Rosenberg 2010</i> ¹⁹ Parent and self-completed Suitable for children and adolescents	<i>Home environment:</i> Checklist of availability of 14 types of PA equipment	Access: Measure used as part of the Active Wear? Study. Details of all measures, including this, at http://sallis.ucsd.edu/measure_activewhere.html

CBIS, Children's Body Image Scale; F&V, fruit and vegetable; FGIR, fasting glucose insulin ratio; FIRI, fasting insulin resistance index; HE, healthy eating; HOMA-B%, homeostatic model assessment – pancreatic beta-cell function; HOMA-IR, homeostatic model assessment – insulin resistance; ISI COMP, insulin sensitivity index composite; PACES, Physical Activity Enjoyment Scale; QUICKI, quantitative insulin sensitivity check index; VAS, Visual Analogue Scale; WBISI, whole-body insulin sensitivity index.

- a Age presented are based on CoOR categories of infancy (< 36 months), children (36 months – 12 years) and adolescents (> 12 years). As it is unlikely that the participant characteristics will be comparable, this was set to guide researchers whether the tool is appropriate in terms of the comparability between the age in which the tool was developed/tested and the age of the intended participants.
- b Food frequency questionnaire food categories are CoOR defined, based on availability of items (for consistency). They are not listed as defined by authors.

Chapter 5 Discussion

Summary of evidence

After screening 25,486 manuscripts, the CoOR study identified 379 eligible manuscripts that described the development and/or evaluation of 180 outcome measures for use in the evaluation of childhood obesity treatment interventions. Appraisal of each of these measures resulted in a framework that recommended 52 measures across 10 outcome domains, for use in the future evaluation of childhood weight management programmes. This framework provides clear guidance to researchers about appropriate outcome domains and recommended measures in each of these domains to encourage greater adoption of well-validated tools. This will make it easier to judge clinical effectiveness and enhance the comparability between different studies or treatment interventions.

Outcome measures were identified via a specific methodology search for manuscripts that described the development and/or evaluation of measures, and via citations within manuscripts of childhood obesity treatment trials that describe the outcome measures used. For the latter, a total of 147 citations were identified within 200 trial manuscripts. However, only 56 (13%) of citations were linked to methodology papers that reported the development or evaluation of measures. A majority of citations were incorrect, often referring to a previous study that had used the same measure and not a method development report. This level of inaccuracy in citations is unacceptable and impedes the ability of readers to understand a trial's conduct, analysis and interpretation, and to assess the validity of its results.²⁵³ Authors are advised to adhere to guidance set by the CONSORT statement, specifically related to the statement of measurement of outcomes in trials: 'All outcome measures, whether primary or secondary, should be identified and completely defined. The principle here is that the information provided should be sufficient to allow others to use the same outcomes'.²⁵⁴

Primary outcome measures that were recommended are BMI and DXA. The decision to include BMI was, in part, based on the feasibility of its use and the ability to ensure comparability between evaluations. Fifty-seven per cent of the eligible trials identified by the CoOR review reported using BMI (or a derivative of BMI) as a primary outcome. Although the evidence of validity offered by the methodology studies within the CoOR review was inconsistent for BMI, experts agreed that it can be reliably measured, provided that administrators are well trained and equipment is regularly calibrated. However, the limitations of BMI were also acknowledged. Primarily, BMI does not provide any information about body composition (including adiposity) or fat distribution. This caveat needs to be considered particularly in studies that evaluate interventions focused on PA (especially those with a lot of strength training). However, the majority of childhood obesity programmes are multifaceted, comprising a variety of lifestyle interventions. A further caveat of BMI (which is common to a number of outcome measures) is the lack of evidence regarding the magnitude of change that is clinically meaningful, also referred to as a MID. Evaluations are focused on detecting a statistical difference in change in BMI between treatment arms, but determination of sample sizes to ensure that it is possible to detect differences requires an estimation of an ES that is ideally based on detecting a clinically meaningful change. Limitations in the available evidence lead to arbitrary decisions being made regarding what amount of change is meaningful. Pooled results of a meta-analysis by Luttikhuis *et al.*¹ report a range in change of between -0.06 and -0.014 for BMI-SDS, and of between -3.04 and -3.27 kg/m² for absolute change in BMI for behavioural interventions. Medium- to high-intensity behavioural interventions in a further review by Whitlock *et al.* for the Agency of Health Care Research and Quality²⁵⁵ report mean reductions in BMI of between 1.9 and 3.3 kg/m². Such data are considered by researchers in deciding what would be considered a desirable level of change. However, there remains insufficient evidence to determine the impact of these changes on cardiovascular risk in children (or later in life).

In addition to measurement of BMI, the CoOR framework advocates the use of DXA measurement if feasible. DXA is also a proxy measure of adiposity but is able to provide an estimation that differentiates between fat and lean tissue. The equipment needed to conduct DXA measurements is expensive and, although widely available in hospital settings, may not always be available for research purposes, especially in community settings; thus, the CoOR framework suggests that DXA is supported with measurement of BMI to allow comparisons between intervention evaluations. Similar to BMI, there is limited evidence regarding the magnitude of change in adiposity that is clinically meaningful. Research in adults has suggested a change of at least 5–10% body fat,^{256–257} but this is also somewhat arbitrary and there are no standards in children. Use of DXA may also be limited to measurement of children who are not severely obese, as some of the feasibility evidence found by CoOR suggests that some children were excluded from the analysis owing to issues with obtaining accurate measurements in those children who were too large to measure on the equipment.¹⁶

Secondary outcomes have been recommended for each of the outcome domains. However, researchers are advised to include only measures that will assess what they expect to change following an intervention, or what they believe will mediate such changes. Thus, it is not necessary to include a measure from all outcome domains in every programme evaluation. Similarly, where multiple measures are advocated within an outcome domain, researchers are advised to consider which measures are most closely aligned to the intervention targets and, where available, choose a measure that has been developed in a population most similar to the intended sample.

Experts agreed that objective measurements must be used where available (i.e. use of activity monitors instead of self-reported PA) and where objective measures are available, no self-reported measures have been recommended for inclusion to the framework. Although findings from the CoOR systematic review indicated that some self-reported measures have been well developed,^{121,126,129} the validity evidence was generally less strong than evaluations of objective measurements. The dependence of weight or weight status on reporting was also apparent in CoOR findings from self-reported measures^{53,55,130} and was an issue discussed by experts incorporating wider evidence.^{258,259} For some outcome domains, it is not possible (e.g. psychological well-being) or feasible (e.g. dietary assessment) to use an objective measure.

In the case of assessment of PA, use of pedometers and direct observation methods were recommended in addition to accelerometers. Although the accuracy of pedometers and direct observation is likely to be lower than for accelerometers, they were recommended as alternative measures, which may be more feasible for some researchers. When using pedometers, researchers should opt for sealed equipment that does not display the number of steps and which are not dependent on self-reporting (i.e. should have the capacity to automatically download). Further, the use of pedometers is not recommended in evaluations of programmes that use pedometers as part of the intervention. For all measurements relying on equipment, it is noted that there will be some variability in data produced between different types and models of equipment. This will have an impact on comparability between studies. Thus, researchers should report the name, version and manufacturer of equipment used. Importantly, the same equipment should be used throughout a single study.

Physiological outcomes, such as insulin, blood lipids and blood pressure have the potential to be primary outcomes, as they are measured with high precision and are indicators of cardiovascular health.²⁶⁰ Thus, improvements to such indicators are likely to be more clinically meaningful than reductions in weight alone. However, at present, there is insufficient evidence on what constitutes a clinically meaningful change or which measures are most sensitive to changes in weight. Without further clarification, experts believed it would be premature to advocate their use as primary outcome measures. Based on the validity evidence collated by CoOR, multiple indices of insulin sensitivity were recommended for inclusion into the outcome measures framework. However, limited evaluations (or poor validity) in other physiological measures meant that no other measures were advocated.

In order to determine what constitutes a MID, it is necessary to ascertain whether or not a measure is able to measure change.²⁶¹ The ability of a measure to detect a clinically meaningful change is defined as responsiveness,^{262,263} whereas the interpretation of whether the change is clinically important relates to a MID (the smallest change that would be deemed clinically beneficial). Both factors vary by population and application. For childhood obesity treatment evaluations, responsiveness considers the relationship between changes in demonstrated effectiveness (e.g. weight loss) and changes in scores or values from other outcome measures. Evidence of responsiveness in eligible measures was poor or lacking in CoOR. In order to maximise data, information regarding sensitivity to measure change was also collected, with the key difference being that sensitivity measures change independently of clinical meaningfulness.²⁶² However, this did not lead to substantial improvements to the data.

As previously stated, a concern for the proposed primary outcomes relates to a lack of clarity on MID, although evidence suggests that BMI and DXA can be measured with a good degree of precision. For BMI, wider evidence has indicated that absolute change, rather than standardised change in BMI (i.e. BMI-SDS) may be better, as it is less dependent on baseline BMI (which may have reduced sensitivity in very obese children).²⁶⁴ However, this may be overcome by adjustment for baseline BMI if using standardised BMI-SDS, which will also provide independence from age and gender.

Only six included manuscripts in CoOR reported formally assessing responsiveness.^{66,170,174,181,184,186,215} Importantly, responsiveness was not ascertained for any measures of psychological well-being or eating behaviours, and was assessed in only two HRQoL measures. These measures are most closely related to PRO measures (although, generally, participants in obesity treatment trials are not considered as 'patients'). Guidance in the use of PROs suggest that if there is clear evidence that a patient's (participant's) experience (relative to the intervention) has changed but the PRO scores do not change then either the ability to detect change is inadequate or the measurements' validity should be questioned.⁹ Additionally, if there is evidence that PRO scores are affected by changes that are not specific to the intervention, the validity of the measure may be questioned. Thus, in order to advocate their use, it would be preferable to know if they demonstrate meaningful improvements when used in the evaluation of treatments that have some evidence of effectiveness in childhood obesity. However, following this guidance would mean that the CoOR outcome measures framework would not be able to advocate the majority of the measures that have been included. Instead, it is recommended that responsiveness assessment is considered in future research with an understanding of the caveats of using a measure with no (or little) evidence of responsiveness. It is important to note, however, that a lack of evidence of responsiveness does not necessarily imply that each measure is not able to detect change. Additionally, the eligibility criteria set by CoOR may have excluded wider evidence of the included tools for responsiveness (e.g. assessed in adults).

The CoOR systematic review did not identify any preference-based (utility) measures of quality of life that would permit an estimation of QALYs. These instruments obtain the participants' own values of varying dimensions of their health, which combines the impact of both the quantity and quality of life, and permits a cost-utility analysis.²⁶⁵ In order to generate QALYs, health utilities (or HRQoL weights) are needed. In this model, utilities for health states are based on participants' preferences for varying health states, with more desirable health states receiving greater weights.²⁶⁶ Utilities are measured on an interval scale of 0–1, where '0' equates to death and '1' indicates full health (although negative scores are also possible). Current guidance by NICE states that the QALY should be used to estimate outcomes in economic evaluation of competing health interventions in order to allow consistent decision making (NICE 2008²⁶⁷).

All identified quality-of-life measures in the CoOR review lacked preference weights and are therefore not able to calculate QALYs. Instead, these measures derive scores for varying dimensions of health statuses. They have been defined as HRQoL measures that are recommended to be considered for inclusion in future evaluations in line with other secondary outcomes with the CoOR framework. But they should not be considered as outcome measures specifically for economic evaluation unless used in cost-effectiveness evaluations of interventions with a primary target on quality of life. However, for evaluations of childhood obesity interventions, a more likely measure to establish cost-effectiveness is that of the primary outcome

(i.e. cost per unit of reduction in BMI). The CoOR team are aware of research in which utility measures are being developed for use in obese paediatric populations. Unfortunately, these were not available at the time of the review.

How to use the Childhood obesity Outcomes Review outcome measures framework

Figure 7 provides guidance as to how the CoOR outcome measures framework should be used by researchers. Importantly, researchers need to first consider which (if any) secondary outcome domains are most closely aligned to the targets of the intervention under investigation, including those that are expected to change, those that are expected to mediate this change (if appropriate), and any that may indicate an adverse event (if appropriate). Researchers are then advised to view the recommended outcome measures within each of the chosen outcome domains in the framework. Any selected measure needs to be aligned to the intervention targets, developed for use in a similar population and feasible to implement. In deciding the similarity between populations, validation of a measure is relevant to only really the population in which it was evaluated. However, given that it is unlikely that there will be a tool that has been developed for use, and evaluated within all populations, researchers should make informed decisions regarding whether the characteristics of their populations are sufficiently close to the population in which the tool was developed. For example, it would not be advisable to use a tool that was developed within a white middle-class population of America in a South Asian lower-class population of the UK. Similarly, tools that were developed to be self-completed by adolescents are unlikely to be relevant for completion by parents.

Further details on each measure within the framework can be accessed from the CoOR summary tables (see *Appendices 6–15*). If a measure that fulfils these criteria does not exist then the researcher may also choose to locate an alternative measure from the summary tables, with the caveat that these were not recommended for inclusion to the framework.

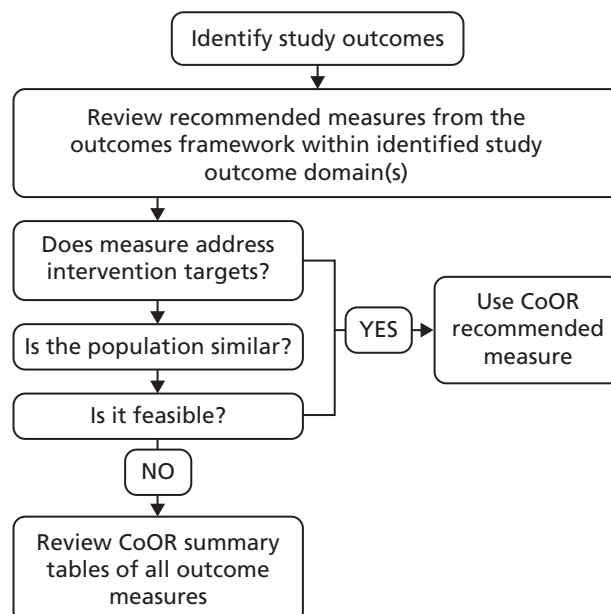


FIGURE 7 Using the CoOR outcome measures framework.

Limitations of the research

The recommendations within the CoOR outcome measures framework are specifically intended for use as outcome measures within studies that evaluate childhood obesity treatment evaluations. These may or may not be suitable for other study designs. It could be argued that some measures that were not recommended are equally, if not more, valid than those that were advocated for other populations or treatment evaluations. However, all decisions were focused on the intended population and study design, which means that some popular, commonly used measures were not deemed appropriate. For example, within the diet domain only FFQs were recommended. This is perhaps of some surprise, given that collection of diet data using diet diaries allows the detailed capture of information about food intake, as well as contextual factors, such as when and where the food was consumed and with whom. It is possible that they are appropriate for use in trials in which diet is a primary outcome (i.e. not obesity trials). It is possible that they are appropriate for use in trials in which change in diet or eating behaviours is the primary outcome. However, the validity evidence presented by CoOR demonstrated a significant impact of body weight on reporting in food diaries, making it unsuitable for studies in this population. As a secondary outcome, it was therefore decided that food diaries should not be advocated.

It is important to note that data that have been presented in tables *in Appendices 6–15* are based on mean values for validity and reliability. A questionnaire with multiple scales should report validity results for each scale in addition to an overall mean. The CoOR review extracted all data for each scale, but it was not feasible to report this volume of data. Where available, means (and ranges) were extracted as presented by authors. If not available, the CoOR team generated mean values from the available data. A limitation of this approach is that it does not permit readers to understand whether some scales performed better than others. This may have a particular impact on measures of dietary assessment, for which there will be variability in the validity and reliability data across different foods or nutrients. Researchers wishing to assess a particular food or nutrient are advised to read the original article of the proposed outcome measure to ensure that there is sufficient evidence of validity/reliability in relation to specific foods or nutrient. Additionally, a copy of the database (which presents findings for individual scales/categories) is available on request.

A further limitation of the data presented by the CoOR review is that validity and reliability findings are often presented as correlation coefficients (with variability in inter- and intraclass correlations used). This type of analysis produces an average correlation across all possible orderings of pairs into X and Y. The reliance on correlations may be sufficient in the case of repeatability (i.e. evaluation of reliability), as it infers a ratio of the variability between participants (or times) over the total variability.²⁶⁸ This method assumes that the measurement error is the same for each repeated assessment, which is likely if assessments are chosen at random with a sufficient sample size. However, this is not the case when comparing two methods in which there is likely to be variability between participant responses. The greater this variability, the greater the correlation coefficient. Conversely, lower between-participant variability would lead to a lower correlation coefficient, which does not necessarily imply that the methods do not agree. Ideally, analysis of validity should consider the differences and the standard deviations of the difference between measurements. Provided differences within the observed LOA are not clinically important we could use the two measurement methods interchangeably²⁶⁸ (although this is difficult to judge with insufficient evidence of a MID). Outcome domains of anthropometry and diet were most likely to use this form of analysis but this was not common to other domains. Further, there was little consideration of whether it would be more appropriate to conduct alternative non-parametric assessments of agreement for differences that are not evenly distributed.²⁶⁹ Lastly, although the CoOR study set standards for what should be considered as a 'gold standard' method within each outcome domain, it is acknowledged that this does not imply that these measures are without error.

The CoOR team recognises that there are likely to be other manuscripts describing the evaluation of eligible measures (i.e. wider evidence of existing measures). These may provide additional evidence regarding the robustness of the measure. However, only those that were evaluated in an obese paediatric

sample (or with results stratified by weight status) were included unless the underlying theoretical framework of the measure was for childhood obesity research. Given the size of the CoOR study, it was not deemed feasible to search for all studies that had conducted evaluation on the included measures outside the predefined eligibility criteria.

Ultimate decisions for the inclusion of each measure were based on agreement by the CoOR experts. Discussions surrounding each measure were made, partly on the data presenting by the CoOR review (including the internal scores for the conduct, reporting and findings of studies). However, final decisions for inclusion were based on the expertise and experience of the CoOR experts, which incorporated wider evidence and feasibility issues. Thus, although some measures were deemed to be of high quality by the internal appraisal (e.g. the IFIS¹³⁶) they were not necessarily advocated by the experts (i.e. no self-reported measures of fitness were recommended).

Future recommendations

It is acknowledged that the output of the CoOR study is somewhat transient, given that new measures are being continually developed. Recommendations from the study, however, suggest that (for the majority of domains) existing measures are appropriate for use, negating the need to develop new measures. Instead, future work should focus on further evaluation and refinement of these measures in different populations. For some outcome domains, new measures are imminent, including utility-based quality-of-life measures and measures of PA and sedentary behaviour that use new technologies. These were not available at the time of writing. It is therefore recommended that the CoOR study is updated every 5–10 years, although the ability of this is dependent on availability of funding.

Chapter 6 Conclusions

The CoOR outcome measures framework provides clear guidance to researchers regarding recommended measures for use in their evaluations of childhood obesity treatment interventions. This should encourage a greater adoption of well-validated tools and ensure comparability between different studies or treatment interventions. Details of the validity of each of the recommended outcome tools provide an evidence base on which to base more accurate reporting of these measures in future studies. In addition, further details of other measures that may be appropriate for other settings are provided to inform decision-making.

It is recommended that further research should be conducted in the development and evaluation of preference-based measures for cost–utility analysis in line with NICE guidance. The CoOR team are aware of some measures currently being developed. Further research is also recommended to ascertain responsiveness of the recommended measures. This would be possible to conduct as part of future trials of childhood obesity treatments. Ascertainment of a MID is also recommended and should be based on consensus by clinical and academic experts and by children and their parents. Finally, there is also a lack of consistency within measures used in the evaluation of treatment of obesity in adults, and it is suggested that similar work to CoOR is conducted to fill this gap in evidence.

Acknowledgements

The CoOR team are extremely grateful to the following expert collaborators who gave their time and expertise in deciding which outcome measures to recommend for inclusion to the CoOR outcome measures framework:

- Professor John Reilly, Professor of Paediatric Energy Metabolism, Royal Hospital for Sick Children, Glasgow.
- Professor Ashley Cooper, Professor of Exercise and Health Science, University of Bristol.
- Professor Paul Kind, Honorary Professor of Economics, University of Leeds.
- Professor Carolyn Summerbell, Professor of Human Nutrition, School of Medicine & Health, and Fellow of the Wolfson Research Institute, Durham University Queen's Campus.
- Professor Julian Hamilton-Shield, Professor in Diabetes and Metabolic Endocrinology, University of Bristol.
- Professor Ulf Ekelund, Professor of Physical Activity and Public Health, MRC Epidemiology Unit, Cambridge.
- Professor Andrew Hill, Professor of Medical Psychology, University of Leeds.
- Dr Lucy Griffiths, Senior Research Fellow at the Institute of Child Health, University College London.
- Professor Steven Cummings, Professor of Population Health & National Institute for Health Research Senior Fellow, London School of Hygiene and Tropical Medicine.
- Dr Claudia Gorecki, Research Fellow in Psychometrics, University of Leeds.

We are also very grateful to colleagues, at the University of Leeds, who volunteered to extract data from manuscripts written in languages other than English, including Elizabeth Mawer (Clinical Trials Research Unit), Ge Yu (Institute of Health Sciences), Roberta Longo (Institute of Health Sciences) and Sandy Tubeuf (Institute of Health Sciences).

Contributions of authors

Dr Maria Bryant (Senior Research Fellow) designed and led the study throughout, including overall management, contribution to literature reviewing, co-leading the expert meeting and leading the publication.

Mr Lee Ashton (Research Assistant) reviewed the literature, co-led the expert meeting and contributed to the interpretation and publication writing.

Professor Julia Brown (Professor of Clinical Trials Research and Director of the Leeds Institute of Clinical Trials Research) contributed intellectually (providing input into design and study procedures throughout), and contributed to interpretation of statistical results within review papers and publication writing.

Professor Susan Jebb (Professor in Diet and Population Health) contributed intellectually (providing input into design and study procedures throughout) and contributed to publication writing.

Ms Judy Wright (Senior Information Specialist) led the search strategy and literature-reviewing process, and contributed to publication writing.

Ms Katharine Roberts (Senior Public Health Analyst) contributed intellectually (providing input into design and study procedures throughout), advised on public health relevance and contributed to publication writing.

Professor Jane Nixon (Professor of Tissue Viability and Clinical Trials, and Deputy Director of the Leeds Institute of Clinical Trials Research) contributed intellectually (providing input into design and study procedures throughout), provided expertise with the expert meeting methodology and contributed to publication writing.

References

1. Oude Luttikhuis H, Baur L, Jansen H, Shrewsbury VA, O'Malley C, Stolk RP, *et al.* Interventions for treating obesity in children. *Cochrane Database Syst Rev* 2009;**1**:CD001872. <http://dx.doi.org/10.1002/14651858.CD001872.pub2>
2. National Institute for Health and Care Excellence (NICE). *Guide to the Methods of Technology Appraisal*. URL: www.nice.org.uk/media/B52/A7/TAMethodsGuideUpdatedJune2008.pdf
3. Holloway RG, Dick AW. Clinical trial end points: on the road to nowhere? *Neurology* 2002;**58**:679–86. <http://dx.doi.org/10.1212/WNL.58.5.679>
4. Sinha I, Jones L, Smyth RL, Williamson PR. A systematic review of studies that aim to determine which outcomes to measure in clinical trials in children. *PLOS Med* 2008;**5**:e96. <http://dx.doi.org/10.1371/journal.pmed.0050096>
5. Roberts K, Cavill N, Rutter H. *Standard Evaluation Framework for Weight Management Interventions*. Oxford: National Obesity Observatory (NOO); 2009.
6. Bryant M, Lucove J, Evenson K, Marshall S. Measurement of television viewing in children and adolescents: a systematic review. *Obes Rev* 2007;**8**:197–209. <http://dx.doi.org/10.1111/j.1467-789X.2006.00295.x>
7. Terwee CB, Jansma EP, Riphagen II, de Vet HCW. Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. *Qual Life Res* 2009;**18**:1115–23. <http://dx.doi.org/10.1007/s11136-009-9528-5>
8. Must A, Anderson SE. Body mass index in children and adolescents: considerations for population-based applications. *Int J Obes* 2006;**30**:590–4.
9. US Department of Health and Human Services FDA. Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims. *Health Qual Life Outcomes* 2006;**4**:79. <http://dx.doi.org/10.1186/1477-7525-4-79>
10. Scientific Advisory Committee of the Medical Outcomes Trust. Assessing health status and quality of life instruments: attributes and review criteria. *Qual Life Res* 2002;**11**:193–205. <http://dx.doi.org/10.1023/A:1015291021312>
11. Gropper SS, Acosta PB. The therapeutic effect of fiber in treating obesity. *J Am Coll Nutr* 1987;**6**:533–5. <http://dx.doi.org/10.1080/07315724.1987.10720213>
12. McCallum Z, Wake M, Gerner B, Baur LA, Gibbons K, Gold L, *et al.* Outcome data from the LEAP (Live, Eat and Play) trial: a randomized controlled trial of a primary care intervention for childhood overweight/mild obesity. *Int J Obes* 2007;**31**:630–6.
13. Ozcetin M, Yilmaz R, Erkorkmaz U, Esmeray H. Reliability and validity study of parental feeding style questionnaire. *Turk Pediatri Arsivi* 2010;**45**:124–31.
14. Viana V, Sinde S. Validation of the Child Eating Behavior Questionnaire (CEBQ) in a Portuguese sample. *Analise Psicologica* 2008;**26**:111–20.
15. Wouters EJM, Geenen R, Kolotkin RL, Vingerhoets AJJM. Body-weight-related quality of life in adolescents: psychometric quality of the Dutch translation of the verse IWQOL-Kids. *Tijdschr Kindergeneesk* 2010;**78**:119–25. <http://dx.doi.org/10.1007/BF03089888>
16. Wells JCK, Haroun D, Williams JE, Wilson C, Darch T, Viner RM, *et al.* Evaluation of DXA against the four-component model of body composition in obese children and adolescents aged 5–21 years. *Int J Obes (Lond)* 2010;**34**:649–55. <http://dx.doi.org/10.1038/ijo.2009.249>

17. Gately PJ, Radley D, Cooke CB, Carroll S, Oldroyd B, Truscott JG, *et al.* Comparison of body composition methods in overweight and obese children. *J Appl Physiol* 2003;**95**:2039–46.
18. Williams JE, Wells JCK, Wilson CM, Haroun D, Lucas A, Fewtrell MS. Evaluation of Lunar Prodigy dual-energy X-ray absorptiometry for assessing body composition in healthy persons and patients by comparison with the criterion 4-component model. *Am J Clin Nutr* 2006;**83**:1047–54.
19. Ramirez E, Valencia ME, Moya Camarena SY, Aleman-Mateo H, Mendez RO. Estimation of body fat by DXA and the four compartment model in Mexican youth. *Arch Latinoam Nutr* 2010;**60**:240–6.
20. Alvero-Cruz JR, Carnero EA, Fernandez-Garcia JC, Exposito JB, De Albornoz Gil MC, Sardinha LB. Validity of body mass index and fat mass index as indicators of overweight status in Spanish adolescents: Esccola Study. *Med Clin (Barc)* 2010;**135**:8–14. <http://dx.doi.org/10.1016/j.medcli.2010.01.017>
21. Rush EC, Puniani K, Valencia ME, Davies PS, Plank LD. Estimation of body fatness from body mass index and bioelectrical impedance: comparison of New Zealand European, Maori and Pacific Island children. *Eur J Clin Nutr* 2003;**57**:1394–401. <http://dx.doi.org/10.1038/sj.ejcn.1601701>
22. Wickramasinghe VP, Cleghorn GJ, Edmiston KA, Murphy AJ, Abbott RA, Davies PSW. Validity of BMI as a measure of obesity in Australian white Caucasian and Australian Sri Lankan children. *Ann Hum Biol* 2005;**32**:60–71. <http://dx.doi.org/10.1080/03014460400027805>
23. Wickramasinghe VP, Lamabadusuriya SP, Cleghorn GJ, Davies PSW. Validity of currently used cutoff values of body mass index as a measure of obesity in Sri Lankan children. *Ceylon Med J* 2009;**54**:114–19. <http://dx.doi.org/10.4038/cmj.v54i4.1451>
24. Wabitsch M, Braun U, Heinze E, Muehe R, Mayer H, Teller W, *et al.* Body composition in 5–18-year-old obese children and adolescents before and after weight reduction as assessed by deuterium dilution and bioelectrical impedance analysis. *Am J Clin Nutr* 1996;**64**:1–6.
25. Haroun D, Croker H, Viner RM, Williams JE, Darch TS, Fewtrell MS, *et al.* Validation of BIA in obese children and adolescents and re-evaluation in a longitudinal study. *Obesity (Silver Spring)* 2009;**17**:2245–50. <http://dx.doi.org/10.1038/oby.2009.98>
26. Marshall JD, Hazlett CB, Spady DW, Conger PR, Quinney HA. Validity of convenient indicators of obesity. *Hum Biol* 1991;**63**:137–53.
27. Marshall JD, Hazlett CB, Spady DW, Quinney HA. Comparison of convenient indicators of obesity. *Am J Clin Nutr* 1990;**51**:22–8.
28. Ayvaz DNC, Klinc FN, Pac FA, Cakal E. Anthropometric measurements and body composition analysis of obese adolescents with and without metabolic syndrome. *Turk J Med Sci* 2011;**41**:267–74.
29. Sardinha LB, Going SB, Teixeira PJ, Lohman TG. Receiver operating characteristic analysis of body mass index, triceps skinfold thickness, and arm girth for obesity screening in children and adolescents. *Am J Clin Nutr* 1999;**70**:1090–5.
30. Perks SM, Roemmich JN, Sandow-Pajewski M, Clark PA, Thomas E, Weltman A, *et al.* Alterations in growth and body composition during puberty. IV. Energy intake estimated by the youth-adolescent food-frequency questionnaire: validation by the doubly labeled water method. *Am J Clin Nutr* 2000;**72**:1455–60.
31. Huybrechts IHI, Bornhorst C, Pala V, Moreno LA, Barba G, Lissner L, *et al.* Evaluation of the Children's Eating Habits Questionnaire used in the IDEFICS study by relating urinary calcium and potassium to milk consumption frequencies among European children. *Int J Obes* 2011;**35**:S69–78. <http://dx.doi.org/10.1038/ijo.2011.37>

32. Burrows T, Warren JM, Baur LA, Collins CE. Impact of a child obesity intervention on dietary intake and behaviors. *Int J Obes* 2008;**32**:1481–8. <http://dx.doi.org/10.1038/ijo.2008.96>
33. Crawford PB, Obarzanek E, Morrison J, Sabry ZI. Comparative advantage of 3-day food records over 24-hour recall and 5-day food frequency validated by observation of 9-year-old and 10-year-old girls. *J Am Diet Assoc* 1994;**94**:626–30. [http://dx.doi.org/10.1016/0002-8223\(94\)90158-9](http://dx.doi.org/10.1016/0002-8223(94)90158-9)
34. Rockett HRH, Berkey CS, Colditz GA. Comparison of a short food frequency questionnaire with the Youth/Adolescent Questionnaire in the Growing Up Today Study. *Int J Pediatr Obes* 2007;**2**:31–9. <http://dx.doi.org/10.1080/17477160601095417>
35. Golley RK, Hendrie GA, McNaughton SA. Scores on the dietary guideline index for children and adolescents are associated with nutrient intake and socio-economic position but not adiposity. *J Nutr* 2011;**41**:1340–7. <http://dx.doi.org/10.3945/jn.110.136879>
36. Lanfer ALA, Hebestreit A, Ahrens W, Krogh V, Sieri S, Lissner L, et al. Reproducibility of food consumption frequencies derived from the Children's Eating Habits Questionnaire used in the IDEFICS study. *Int J Obes* 2011;**35**:S61–8. <http://dx.doi.org/10.1038/ijo.2011.36>
37. Rockett HRH, Colditz GA. Assessing diets of children and adolescents. *Am J Clin Nutr* 1997;**65**(Suppl. 4):1116–22.
38. Blum RE, Wei EK, Rockett HR, Langeliers JD, Leppert J, Gardner JD, et al. Validation of a food frequency questionnaire in Native American and Caucasian children 1 to 5 years of age. *Matern Child Health J* 1999;**3**:167–72. <http://dx.doi.org/10.1023/A:1022350023163>
39. Vereecken C, Covents M, Maes L. Comparison of a food frequency questionnaire with an online dietary assessment tool for assessing preschool children's dietary intake. *J Hum Nutr Diet* 2010;**23**:502–10. <http://dx.doi.org/10.1111/j.1365-277X.2009.01038.x>
40. Burrows TL, Warren JM, Colyvas K, Garg ML, Collins CE. Validation of overweight children's fruit and vegetable intake using plasma carotenoids. *Obesity (Silver Spring)* 2009;**17**:162–8. <http://dx.doi.org/10.1038/oby.2008.495>
41. Yaroch AL, Resnicow K, Davis M, Davis A, Smith M, Khan LK. Development of a modified picture-sort food frequency questionnaire administered to low-income, overweight, African-American adolescent girls. *J Am Diet Assoc* 2000;**100**:1050–6. [http://dx.doi.org/10.1016/S0002-8223\(00\)00306-0](http://dx.doi.org/10.1016/S0002-8223(00)00306-0)
42. Yaroch AL, Resnicow K, Petty AD, Khan LK. Validity and reliability of a modified qualitative dietary fat index in low-income, overweight, African American adolescent girls. *J Am Diet Assoc* 2000;**100**:1525–9. [http://dx.doi.org/10.1016/S0002-8223\(00\)00422-3](http://dx.doi.org/10.1016/S0002-8223(00)00422-3)
43. Rockett HR, Wolf AM, Colditz GA. Development and reproducibility of a food frequency questionnaire to assess diets of older children and adolescents. *J Am Diet Assoc* 1995;**95**:336–40. [http://dx.doi.org/10.1016/S0002-8223\(95\)00086-0](http://dx.doi.org/10.1016/S0002-8223(95)00086-0)
44. Nelson MC, Lytle LA. Development and evaluation of a brief screener to estimate fast-food and beverage consumption among adolescents. *J Am Diet Assoc* 2009;**109**:730–4. <http://dx.doi.org/10.1016/j.jada.2008.12.027>
45. Davis JN, Nelson MC, Ventura EE, Lytle LA, Goran MI. A brief dietary screener: appropriate for overweight Latino adolescents? *J Am Diet Assoc* 2009;**109**:725–9. <http://dx.doi.org/10.1016/j.jada.2008.12.025>
46. Watson JF, Collins CE, Sibbritt DW, Dibley MJ, Garg ML. Reproducibility and comparative validity of a food frequency questionnaire for Australian children and adolescents. *Int J Behav Nutr Phys Act* 2009;**6**:62. <http://dx.doi.org/10.1186/1479-5868-6-62>

47. Metcalf PA, Scragg RK, Sharpe S, Fitzgerald ED, Schaaf D, Watts C. Short-term repeatability of a food frequency questionnaire in New Zealand children aged 1–14 years. *Eur J Clin Nutr* 2003;**57**:1498–503. <http://dx.doi.org/10.1038/sj.ejcn.1601717>
48. Lee S, Ahn H-S. Comparison of major dish item and food group consumption between normal and obese Korean children: application to development of a brief food frequency questionnaire for obesity-related eating behaviors. *Nutr Res Pract* 2007;**1**:313–20. <http://dx.doi.org/10.4162/nrp.2007.1.4.313>
49. Epstein LH, Gordy CC, Raynor HA, Beddome M, Kilanowski CK, Paluch R. Increasing fruit and vegetable intake and decreasing fat and sugar intake in families at risk for childhood obesity. *Obesity* 2000;**9**:171–8. <http://dx.doi.org/10.1038/oby.2001.18>
50. Prochaska JJ, Sallis JF, Rupp J. Screening measure for assessing dietary fat intake among adolescents. *Prev Med* 2001;**33**:699–706. <http://dx.doi.org/10.1006/pmed.2001.0951>
51. Taveras EM, Rifas-Shiman S, Berkey CS, Rockett HRH, Field AE, Frazier AL, *et al.* Family dinner and adolescent overweight. *Obes Res* 2005;**13**:900–6. <http://dx.doi.org/10.1038/oby.2005.104>
52. Taveras EM, Berkey CS, Rifas-Shiman SL, Ludwig DS, Rockett HR, Field AE, *et al.* Association of consumption of fried food away from home with body mass index and diet quality in older children and adolescents. *Pediatrics* 2005;**116**:e518–24. <http://dx.doi.org/10.1542/peds.2004-2732>
53. Sjöberg A, Slinde F, Arvidsson D, Ellegård L, Gramatkovski E, Hallberg L, *et al.* Energy intake in Swedish adolescents: validation of diet history with doubly labelled water. *Eur J Clin Nutr* 2003;**57**:1643–52. <http://dx.doi.org/10.1038/sj.ejcn.1601892>
54. Waling MU, Larsson CL. Energy intake of Swedish overweight and obese children is underestimated using a diet history interview. *J Nutr* 2009;**139**:522–7. <http://dx.doi.org/10.3945/jn.108.101311>
55. Maffei C, Schutz Y, Zaffanello M, Piccoli R, Pinelli L. Elevated energy expenditure and reduced energy intake in obese prepubertal children: paradox of poor dietary reliability in obesity? *J Pediatr* 1994;**124**:348–54. [http://dx.doi.org/10.1016/S0022-3476\(94\)70355-8](http://dx.doi.org/10.1016/S0022-3476(94)70355-8)
56. Van Horn LV, Gernhofer N, Moag-Stahlberg A, Farris R, Hartmuller G, Lasser VI, *et al.* Dietary assessment in children using electronic methods: telephones and tape recorders. *J Am Diet Assoc* 1990;**90**:412–16.
57. Singh R, Martin BR, Hickey Y, Teegarden D, Campbell WW, Craig BA, *et al.* Comparison of self-reported, measured, metabolizable energy intake with total energy expenditure in overweight teens. *Am J Clin Nutr* 2009;**89**:1744–50. <http://dx.doi.org/10.3945/ajcn.2008.26752>
58. Bandini LG, Schoeller DA, Cyr HN, Dietz WH. Validity of reported energy intake in obese and nonobese adolescents. *Am J Clin Nutr* 1990;**52**:421–5.
59. Bandini LG, Vu D, Must A, Cyr H, Goldberg A, Dietz WH. Comparison of high-calorie, low-nutrient-dense food consumption among obese and non-obese adolescents. *Obes Res* 1999;**7**:438–43. <http://dx.doi.org/10.1002/j.1550-8528.1999.tb00431.x>
60. Lindquist CH, Cummings T, Goran MI. Use of tape-recorded food records in assessing children's dietary intake. *Obes Res* 2000;**8**:2–11. <http://dx.doi.org/10.1038/oby.2000.2>
61. Bratteby LE, Sandhagen B, Enghardt H, Fan H, Samuelson G. Validity of dietary intake measurements in adolescents: three validation studies. *Scand J Nutr Näringsforskning* 1998;**42**:29–30.

62. Champagne CM, Baker NB, DeLany JP, Harsha DW, Bray GA. Assessment of energy intake underreporting by doubly labeled water and observations on reported nutrient intakes in children. *J Am Diet Assoc* 1998;**98**:426–33. [http://dx.doi.org/10.1016/S0002-8223\(98\)00097-2](http://dx.doi.org/10.1016/S0002-8223(98)00097-2)
63. Champagne CM, Delany JP, Harsha DW, Bray GA. Underreporting of energy intake in biracial children is verified by doubly labeled water. *J Am Diet Assoc* 1996;**96**:707. [http://dx.doi.org/10.1016/S0002-8223\(96\)00193-9](http://dx.doi.org/10.1016/S0002-8223(96)00193-9)
64. O'Connor J, Ball EJ, Steinbeck KS, Davies PSW, Wishart C, Gaskin KJ, et al. Comparison of total energy expenditure and energy intake in children aged 6–9 years. *Am J Clin Nutr* 2001;**74**:643–9.
65. Baxter SD, Smith AF, Nichols MD, Guinn CH, Hardin JW. Children's dietary reporting accuracy over multiple 24-hour recalls varies by body mass index category. *Nutr Res* 2006;**26**:241–8. <http://dx.doi.org/10.1016/j.nutres.2006.05.005>
66. Edmunds L, Ziebland S. Development and validation of the Day in the Life Questionnaire (DILQ) as a measure of fruit and vegetable questionnaire for 7–9 year olds. *Health Educ Res* 2002;**17**:211–20. <http://dx.doi.org/10.1093/her/17.2.211>
67. Lytle LA, Murray DM, Perry CL, Eldridge AL. Validating fourth-grade students' self-report of dietary intake: results from the 5 A Day Power Plus program. *J Am Diet Assoc* 1998;**98**:570–2. [http://dx.doi.org/10.1016/S0002-8223\(98\)00127-8](http://dx.doi.org/10.1016/S0002-8223(98)00127-8)
68. Johnson RK, Driscoll P, Goran MI. Comparison of multiple-pass 24-hour recall estimates of energy intake with total energy expenditure determined by the doubly labeled water method in young children. *J Am Diet Assoc* 1996;**96**:1140–4. [http://dx.doi.org/10.1016/S0002-8223\(96\)00293-3](http://dx.doi.org/10.1016/S0002-8223(96)00293-3)
69. Martinez de Icaya P, Fernandez C, Vazquez C, del Olmo D, Alcazar V, Hernandez M. IGF-1 and its binding proteins IGFBP-1 and 3 as nutritional markers in prepubertal children. *Ann Nutr Metab* 2000;**44**:139–43. <http://dx.doi.org/10.1159/000012836>
70. Ball SC, Benjamin SE, Ward DS. Development and reliability of an observation method to assess food intake of young children in child care. *J Am Diet Assoc* 2007;**107**:656–61. <http://dx.doi.org/10.1016/j.jada.2007.01.003>
71. Vance VA, Woodruff SJ, McCargar LJ, Husted J, Hanning RM. Self-reported dietary energy intake of normal weight, overweight and obese adolescents. *Public Health Nutr* 2008;**12**:222–7. <http://dx.doi.org/10.1017/S1368980008003108>
72. Sleddens EFC, Kremers SPJ, Thijs C. The Children's Eating Behaviour Questionnaire: factorial validity and association with Body Mass Index in Dutch children aged 6–7. *Int J Behav Nutr Phys Act* 2008;**5**:49–57. <http://dx.doi.org/10.1186/1479-5868-5-49>
73. Wardle J, Guthrie CA, Sanderson S, Rapoport L. Development of the children's eating behaviour questionnaire. *J Child Psychol Psychiatry* 2001;**42**:963–70. <http://dx.doi.org/10.1111/1469-7610.00792>
74. Baughcum AE, Powers SW, Johnson SB, Chamberlin LA, Deeks CM, Jain A, et al. Maternal feeding practices and beliefs and their relationships to overweight in early childhood. *J Dev Behav Pediatr* 2001;**22**:391–408. <http://dx.doi.org/10.1097/00004703-200112000-00007>
75. Birch LL, Fisher JO, Grimm-Thomas K, Markey CN, Sawyer R, Johnson SL. Confirmatory factor analysis of the Child Feeding Questionnaire: a measure of parental attitudes, beliefs and practices about child feeding and obesity proneness. *Appetite* 2001;**36**:201–10. <http://dx.doi.org/10.1006/appe.2001.0398>
76. Thompson AL, Mendez MA, Borja JB, Adair LS, Zimmer CR, Bentley ME. Development and validation of the Infant Feeding Style Questionnaire. *Appetite* 2009;**53**:210–21. <http://dx.doi.org/10.1016/j.appet.2009.06.010>

77. Tanofsky-Kraff M, Theim KR, Yanovski SZ, Bassett AM, Burns NP, Ranzenhofer LM, *et al.* Validation of the emotional eating scale adapted for use in children and adolescents (EES-C). *Int J Eat Disord* 2007;**40**:232–40. <http://dx.doi.org/10.1002/eat.20362>
78. Braet C, Van Strien T. Assessment of emotional, externally induced and restrained eating behaviour in nine to twelve-year-old obese and non-obese children. *Behav Res Ther* 1997;**35**:863–73. [http://dx.doi.org/10.1016/S0005-7967\(97\)00045-4](http://dx.doi.org/10.1016/S0005-7967(97)00045-4)
79. Van Strien T, Oosterveld P. The children's DEBQ for assessment of restrained, emotional, and external eating in 7- to 12-year-old children. *Int J Eat Disord* 2008;**41**:72–81. <http://dx.doi.org/10.1002/eat.20424>
80. Tanofsky-Kraff M, Ranzenhofer LM, Yanovski SZ, Schvey NA, Faith M, Gustafson J, *et al.* Psychometric properties of a new questionnaire to assess eating in the absence of hunger in children and adolescents. *Appetite* 2008;**51**:148–55. <http://dx.doi.org/10.1016/j.appet.2008.01.001>
81. Decaluwé V, Braet C. Assessment of eating disorder psychopathology in obese children and adolescents: interview versus self-report questionnaire. *Behav Res Ther* 2004;**42**:799–811. <http://dx.doi.org/10.1016/j.brat.2003.07.008>
82. Corsini N, Wilson C, Kettler L, Danthiir V. Development and preliminary validation of the Toddler Snack Food Feeding Questionnaire. *Appetite* 2010;**54**:570–8. <http://dx.doi.org/10.1016/j.appet.2010.03.001>
83. Banos RM, Cebolla A, Etchemendy E, Felipe S, Rasal P, Botella C. Validation of the Dutch eating behavior questionnaire for children (DEBQ-C) for use with Spanish children. *Nutr Hosp* 2011;**26**:890–8. <http://dx.doi.org/10.1590/S0212-16112011000400032>
84. Murashima M, Hoerr SL, Hughes SO, Koplowitz S. Confirmatory factor analysis of a questionnaire measuring control in parental feeding practices in mothers of Head Start children. *Appetite* 2011;**56**:594–601. <http://dx.doi.org/10.1016/j.appet.2011.01.031>
85. Monnery-Patris S, Rigal N, Chabanet C, Boggio V, Lange C, Cassuto DA, *et al.* Parental practices perceived by children using a French version of the Kids' Child Feeding Questionnaire. *Appetite* 2011;**57**:161–6. <http://dx.doi.org/10.1016/j.appet.2011.04.014>
86. Maloney MJ, McGuire JB, Daniels SR. Reliability testing of a children's version of the Eating Attitude Test. *J Am Acad Child Adolesc Psychiatry* 1988;**27**:541–3. <http://dx.doi.org/10.1097/00004583-198809000-00004>
87. Shisslak CM, Renger R, Sharpe T, Crago M, McKnight KM, Gray N, *et al.* Development and evaluation of the McKnight Risk Factor Survey for assessing potential risk and protective factors for disordered eating in preadolescent and adolescent girls. *Int J Eat Disord* 1999;**25**:195–214. [http://dx.doi.org/10.1002/\(SICI\)1098-108X\(199903\)25:2<195::AID-EAT9>3.0.CO;2-B](http://dx.doi.org/10.1002/(SICI)1098-108X(199903)25:2<195::AID-EAT9>3.0.CO;2-B)
88. Kröller K, Warschburger P. Associations between maternal feeding style and food intake of children with a higher risk for overweight. *Appetite* 2008;**51**:166–72. <http://dx.doi.org/10.1016/j.appet.2008.01.012>
89. Childress AC, Brewerton TD, Hodges EL, Jarrell MP. The kids eating disorders survey (KEDS): a study of middle school students. *J Am Acad Child Adolesc Psychiatry* 1993;**32**:843–50. <http://dx.doi.org/10.1097/00004583-199307000-00021>
90. Johnson WG, Grieve FG, Adams CD, Sandy J. Measuring binge eating in adolescents: adolescent and parent versions of the questionnaire of eating and weight patterns. *Int J Eat Disord* 1999;**26**:301–14. [http://dx.doi.org/10.1002/\(SICI\)1098-108X\(199911\)26:3<301::AID-EAT8>3.0.CO;2-M](http://dx.doi.org/10.1002/(SICI)1098-108X(199911)26:3<301::AID-EAT8>3.0.CO;2-M)

91. Steinberg E, Tanofsky-Kraff M, Cohen ML, Elberg J, Freedman RJ, Semega-Janneh M, *et al.* Comparison of the child and parent forms of the Questionnaire on Eating and Weight Patterns in the assessment of children's eating-disordered behaviors. *Int J Eat Disord* 2004;**36**:183–94. <http://dx.doi.org/10.1002/eat.20022>
92. Braet C, Soetens B, Moens E, Mels S, Goossens L, Van Vlierberghe L. Are two informants better than one? Parent-child agreement on the eating styles of children who are overweight. *Eur Eat Disord Rev* 2007;**15**:410–17. <http://dx.doi.org/10.1002/erv.798>
93. Haycraft EL, Blissett JM. Maternal and paternal controlling feeding practices: reliability and relationships with BMI. *Obesity (Silver Spring)* 2008;**16**:1552–8. <http://dx.doi.org/10.1038/oby.2008.238>
94. Polat S, Erci B. Psychometric Properties of the Child Feeding Scale in Turkish Mothers. *Asian Nurs Res* 2010;**4**:111–21. [http://dx.doi.org/10.1016/S1976-1317\(10\)60011-4](http://dx.doi.org/10.1016/S1976-1317(10)60011-4)
95. Spitzer RL, Devlin M, Walsh BT, Hasin D, Wing R, Marcus M, *et al.* Binge eating disorder: a multi-site field trial of the diagnostic criteria. *Int J Eat Disord* 1992;**11**:191–203. [http://dx.doi.org/10.1002/1098-108X\(199204\)11:3<191::AID-EAT2260110302>3.0.CO;2-S](http://dx.doi.org/10.1002/1098-108X(199204)11:3<191::AID-EAT2260110302>3.0.CO;2-S)
96. Anderson CB, Hughes SO, Fisher JO, Nicklas TA. Cross-cultural equivalence of feeding beliefs and practices: the psychometric properties of the child feeding questionnaire among Blacks and Hispanics. *Prev Med* 2005;**41**:521–31. <http://dx.doi.org/10.1016/j.ypmed.2005.01.003>
97. Corsini N, Danthiir V, Kettler L, Wilson C. Factor structure and psychometric properties of the Child Feeding Questionnaire in Australian preschool children. *Appetite* 2008;**51**:474–81. <http://dx.doi.org/10.1016/j.appet.2008.02.013>
98. Caccialanza R, Nicholls D, Cena H, Maccarini L, Rezzani C, Antonioli L, *et al.* Validation of the Dutch Eating Behaviour Questionnaire parent version (DEBQ-P) in the Italian population: a screening tool to detect differences in eating behaviour among obese, overweight and normal-weight preadolescents. *Eur J Clin Nutr* 2004;**58**:1217–22. <http://dx.doi.org/10.1038/sj.ejcn.1601949>
99. Goldschmidt AB, Doyle AC, Wilfley DE. Assessment of binge eating in overweight youth using a questionnaire version of the Child Eating Disorder Examination with Instructions. *Int J Eat Disord* 2007;**40**:460–7. <http://dx.doi.org/10.1002/eat.20387>
100. Smolak L, Levine MP. Psychometric properties of the Children's Eating Attitudes Test. *Int J Eat Disord* 1994;**16**:275–82. [http://dx.doi.org/10.1002/1098-108X\(199411\)16:3<275::AID-EAT2260160308>3.0.CO;2-U](http://dx.doi.org/10.1002/1098-108X(199411)16:3<275::AID-EAT2260160308>3.0.CO;2-U)
101. Ranzenhofer LM, Tanofsky-Kraff M, Menzie CM, Gustafson JK, Rutledge MS, Keil MF, *et al.* Structure analysis of the Children's Eating Attitudes Test in overweight and at-risk for overweight children and adolescents. *Eat Behav* 2008;**9**:218–27. <http://dx.doi.org/10.1016/j.eatbeh.2007.09.004>
102. Ridgers ND, Stratton G, McKenzie TL. Reliability and validity of the System for Observing Children's Activity and Relationships during Play (SOCARP). *J Phys Act Health* 2010;**7**:17–25.
103. Brown WH, Pfeiffer KA, McIver KL, Dowda M, Almeida M, Pate RR. Assessing preschool children's physical activity: the observational system for recording physical activity in children-preschool version. *Res Q Exerc Sport* 2006;**77**:167–76.
104. Reilly JJ, Penpraze V, Hislop J, Davies G, Grant S, Paton JY. Objective measurement of physical activity and sedentary behaviour: review with new data. *Arch Dis Child* 2008;**93**:614–19. <http://dx.doi.org/10.1136/adc.2007.133272>
105. 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;**16**:324–33.

106. Noland M, Danner F, DeWalt K, McFadden M, Kotchen JM. The measurement of physical activity in young children. *Res Q Exerc Sport* 1990;**61**:146–53. <http://dx.doi.org/10.1080/02701367.1990.10608668>
107. Pate RR, Almeida MJ, McIver KL, Pfeiffer KA, Dowda M. Validation and calibration of an accelerometer in preschool children. *Obesity (Silver Spring)* 2006;**14**:2000–6. <http://dx.doi.org/10.1038/oby.2006.234>
108. Coleman KJ, Saelens BE, Wiedrich-Smith MD, Finn JD, Epstein LH. Relationships between TriTrac-R3D vectors, heart rate, and self-report in obese children. *Med Sci Sports Exerc* 1997;**29**:1535–42. <http://dx.doi.org/10.1097/00005768-199711000-00022>
109. Duncan JS, Schofield G, Duncan EK, Hinckson EA. Effects of age, walking speed, and body composition on pedometer accuracy in children. *Res Q Exerc Sport* 2007;**78**:420–8. <http://dx.doi.org/10.1080/02701367.2007.10599442>
110. Mitre N, Lanningham-Foster L, Foster R, Levine JA. Pedometer accuracy for children: can we recommend them for our obese population? *Pediatrics* 2009;**123**:e127–31. <http://dx.doi.org/10.1542/peds.2008-1908>
111. Backlund C, Sundelin G, Larsson C. Problems in enhancing physical activity among overweight and obese children. 11th International Congress on Obesity, 11–15 July 2010, Stockholm, Sweden. *Obes Rev* 2010;**11**:78–9.
112. Jago R, Watson K, Baranowski T, Zakeri I, Yoo S, Baranowski J, et al. Pedometer reliability, validity and daily activity targets among 10- to 15-year-old boys. *J Sports Sci* 2006;**24**:241–51. <http://dx.doi.org/10.1080/02640410500141661>
113. Treuth MS, Sherwood NE, Butte NF, McClanahan B, Obarzanek E, Zhou A, et al. Validity and reliability of activity measures in African-American girls for GEMS. *Med Sci Sports Exerc* 2003;**35**:532–9. <http://dx.doi.org/10.1249/01.MSS.0000053702.03884.3F>
114. Kilanowski CK, Consalvi AR, Epstein LH. Validation of an electronic pedometer for measurement of physical activity in children. *Pediatr Exerc Sci* 1999;**11**:63–8.
115. 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;**16**:64–78.
116. Welk GJ, Dziewaltowski DA, Hill JL. Comparison of the computerized ACTIVITYGRAM instrument and the previous day physical activity recall for assessing physical activity in children. *Res Q Exerc Sport* 2004;**75**:370–80. <http://dx.doi.org/10.1080/02701367.2004.10609170>
117. Slootmaker SM, Schuit AJ, Chinapaw MJM, Seidell JC, van Mechelen W. Disagreement in physical activity assessed by accelerometer and self-report in subgroups of age, gender, education and weight status. *Int J Behav Nutr Phys Act* 2009;**6**:17. <http://dx.doi.org/10.1186/1479-5868-6-17>
118. Kowalski K, Crocker P, Faulkner R. Validation of the physical activity questionnaire for older children. *Pediatr Exerc Sci* 1997;**9**:174–86.
119. Epstein LH, Paluch RA, Coleman KJ, Vito D, Anderson K. Determinants of physical activity in obese children assessed by accelerometer and self-report. *Med Sci Sports Exerc* 1996;**28**:1157–64. <http://dx.doi.org/10.1097/00005768-199609000-00012>
120. Weston AT, Petosa R, Pate RR. Validation of an instrument for measurement of physical activity in youth. *Med Sci Sports Exerc* 1997;**29**:138–43. <http://dx.doi.org/10.1097/00005768-199701000-00020>
121. Sallis JF, Buono MJ, Roby JJ, Micale FG, Nelson JA. 7-day recall and other physical activity self-reports in children and adolescents. *Med Sci Sports Exerc* 1993;**25**:99–108. <http://dx.doi.org/10.1249/00005768-199301000-00014>

122. Burdette HL, Whitaker RC, Daniels SR. Parental report of outdoor playtime as a measure of physical activity in preschool-aged children. *Arch Pediatr Adolesc Med* 2004;**158**:353. <http://dx.doi.org/10.1001/archpedi.158.4.353>
123. Booth ML, Okely AD, Chey TN, Bauman A. The reliability and validity of the Adolescent Physical Activity Recall Questionnaire. *Med Sci Sports Exerc* 2002;**34**:1986–95. <http://dx.doi.org/10.1097/00005768-200212000-00019>
124. Welk GJ, Schaben JA, Shelley M. Physical activity and physical fitness in children schooled at home and children attending public schools. *Pediatr Exerc Sci* 2004;**16**:310–23.
125. Kowalski K, Crocker P, Kowalski N. Convergent validity of the physical activity questionnaire for adolescents. *Pediatr Exerc Sci* 1997;**9**:342–52.
126. Moore JB, Hanes JC, Jr, Barbeau P, Gutin B, Trevino RP, Yin Z. Validation of the physical activity questionnaire for older children in children of different races. *Pediatr Exerc Sci* 2007;**19**:6–19.
127. Goran MI, Hunter G, Nagy TR, Johnson R. Physical activity related energy expenditure and fat mass in young children. *Int J Obes (Lond)* 1997;**21**:171–8. <http://dx.doi.org/10.1038/sj.ijo.0800383>
128. Crocker PR, Bailey DA, Faulkner RA, Kowalski KC, McGrath R. Measuring general levels of physical activity: preliminary evidence for the Physical Activity Questionnaire for Older Children. *Med Sci Sports Exerc* 1997;**29**:1344–9. <http://dx.doi.org/10.1097/00005768-199710000-00011>
129. Janz KF, Lutuchy EM, Wenthe P, Levy SM. Measuring activity in children and adolescents using self-report: PAQ-C and PAQ-A. *Med Sci Sports Exerc* 2008;**40**:767. <http://dx.doi.org/10.1249/MSS.0b013e3181620ed1>
130. Sithole F, Veugelers PJ. Parent and child reports of children's activity. *Health Rep* 2008;**19**:19–24.
131. 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;**11**:1155–8. <http://dx.doi.org/10.1038/oby.2003.158>
132. Puyau MR, Adolph AL, Vohra FA, Butte NF. Validation and calibration of physical activity monitors in children. *Obes Res* 2002;**10**:150–7. <http://dx.doi.org/10.1038/oby.2002.24>
133. Ridley K, Olds TS, Hill A. The Multimedia Activity Recall for Children and Adolescents (MARCA): development and evaluation. *Int J Behav Nutr Phys Act* 2006;**3**:10. <http://dx.doi.org/10.1186/1479-5868-3-10>
134. Dunton GF, Liao Y, Intille SS, Spruijt-Metz D, Pentz M. Investigating children's physical activity and sedentary behavior using ecological momentary assessment with mobile phones. *Obesity (Silver Spring)* 2011;**19**:1205–12. <http://dx.doi.org/10.1038/oby.2010.302>
135. Epstein LH, Paluch RA, Kilanowski CK, Raynor HA. The effect of reinforcement or stimulus control to reduce sedentary behavior in the treatment of pediatric obesity. *Health Psychol* 2004;**23**:371. <http://dx.doi.org/10.1037/0278-6133.23.4.371>
136. Ortega FB, Ruiz JR, Espaa-Romero V, Vicente-Rodriguez G, Martinez-Gmez D, Manios Y, et al. International Fitness Scale (IFIS): self-reported fitness and obesity in youth: the HELENA study. 1st IDEFICS Symposium and Workshop Child Health in Europe – The IDEFICS Study: Towards a Better Understanding of Obesity, 8–9 November 2010, Zaragoza, Spain. *Int J Obes (Lond)* 2011;**35**:S160.
137. Morrow JR, Martin SB, Jackson AW. Reliability and validity of the FITNESSGRAM (R): quality of teacher-collected health-related fitness surveillance data. *Res Q Exerc Sport* 2010;**81**:S24–30. <http://dx.doi.org/10.1080/02701367.2010.10599691>

138. Morinder G, Mattsson E, Sollander C, Marcus C, Larsson UE. Six-minute walk test in obese children and adolescents: reproducibility and validity. *Physiother Res Int* 2009;**14**:91–104. <http://dx.doi.org/10.1002/pri.428>
139. Leger LA, Mercier D, Gadoury C, Lambert J. The multistage 20 metre shuttle run test for aerobic fitness. *J Sports Sci* 1988;**6**:93–101. <http://dx.doi.org/10.1080/02640418808729800>
140. Suminski RR, Ryan ND, Poston CS, Jackson AS. Measuring aerobic fitness of Hispanic youth 10 to 12 years of age. *Int J Sports Med* 2004;**25**:61–7. <http://dx.doi.org/10.1055/s-2003-45230>
141. Loftin M, Sothorn M, Warren B, Udall J. Comparison of VO₂ peak during treadmill and cycle ergometry in severely overweight youth. *J Sports Sci Med* 2004;**3**:254–60.
142. Meyers CR. A study of the reliability of the Harvard step test. *Res Q* 1969;**40**:423.
143. Drinkard B, Roberts MD, Ranzenhofer LM, Han JC, Yanoff LB, Merke DP, *et al.* Oxygen-uptake efficiency slope as a determinant of fitness in overweight adolescents. *Med Sci Sports Exerc* 2007;**39**:1811–16. <http://dx.doi.org/10.1249/mss.0b013e31812e52b3>
144. Aucouturier J, Rance M, Meyer M, Isacco L, Thivel D, Fellmann N, *et al.* Determination of the maximal fat oxidation point in obese children and adolescents: validity of methods to assess maximal aerobic power. *Eur J Appl Physiol* 2009;**105**:325–31. <http://dx.doi.org/10.1007/s00421-008-0907-3>
145. Rowland TW, Rambusch JM, Staab JS, Unnithan VB, Siconolfi SF. Accuracy of physical working capacity (PWC170) in estimating aerobic fitness in children. *J Sports Med Phys Fitness* 1993;**33**:184–8.
146. Carrel AL, Sledge JS, Ventura SJ, Clark RR, Peterson SE, Eickhoff J, *et al.* Measuring aerobic cycling power as an assessment of childhood fitness. *J Strength Cond Res* 2007;**21**:685–8.
147. Roberts MD, Drinkard B, Ranzenhofer LM, Salaita CG, Sebring NG, Brady SM, *et al.* Prediction of maximal oxygen uptake by bioelectrical impedance analysis in overweight adolescents. *J Sports Med Phys Fitness* 2009;**49**:240–5.
148. Francis K, Feinstein R. A simple height-specific and rate-specific step test for children. *South Med J* 1991;**84**:169–74. <http://dx.doi.org/10.1097/00007611-199102000-00005>
149. Nemeth BA, Carrel AL, Eickhoff J, Clark RR, Peterson SE, Allen DB. Submaximal treadmill test predicts VO₂max in overweight children. *J Pediatr* 2009;**154**:677–81. <http://dx.doi.org/10.1016/j.jpeds.2008.11.032>
150. Wang CL, Liang L, Fu JF, Hong F. Comparison of methods to detect insulin resistance in obese children and adolescents. *Zhejiang Da Xue Xue Bao Yi Xue Ban* 2005;**34**:316–19.
151. Thiel C, Claussnitzer G, Vogt L, Banzer W. Energy expenditure estimation by flex heart rate method in obese children. *Dtsch Z Sportmed* 2007;**58**:78–82.
152. Yeckel CW, Weiss R, Dziura J, Taksali SE, Dufour S, Burgert TS, *et al.* Validation of insulin sensitivity indices from oral glucose tolerance test parameters in obese children and adolescents. *J Clin Endocrinol Metab* 2004;**89**:1096–101. <http://dx.doi.org/10.1210/jc.2003-031503>
153. Conwell LS, Trost SG, Brown WJ, Batch JA. Indexes of insulin resistance and secretion in obese children and adolescents: a validation study. *Diabetes Care* 2004;**27**:314–19. <http://dx.doi.org/10.2337/diacare.27.2.314>
154. George L, Bacha F, Lee S, Tfayli H, Andreatta E, Arslanian S. Surrogate estimates of insulin sensitivity in obese youth along the spectrum of glucose tolerance from normal to prediabetes to diabetes. *J Clin Endocrinol Metab* 2011;**96**:2136–45. <http://dx.doi.org/10.1210/jc.2010-2813>

155. Gunczler P, Lanes R. Relationship between different fasting-based insulin sensitivity indices in obese children and adolescents. *J Pediatr Endocrinol* 2006;**19**:259–65. <http://dx.doi.org/10.1515/JPEM.2006.19.3.259>
156. Uwaifo GI, Fallon EM, Chin J, Elberg J, Parikh SJ, Yanovski JA. Indices of insulin action, disposal, and secretion derived from fasting samples and clamps in normal glucose-tolerant black and white children. *Diabetes Care* 2002;**25**:2081–7. <http://dx.doi.org/10.2337/diacare.25.11.2081>
157. Uwaifo GI, Parikh SJ, Keil M, Elberg J, Chin J, Yanovski JA. Comparison of insulin sensitivity, clearance, and secretion estimates using euglycemic and hyperglycemic clamps in children. *J Clin Endocrinol Metab* 2002;**87**:2899–905. <http://dx.doi.org/10.1210/jcem.87.6.8578>
158. Gungor N, Saad R, Janosky J, Arslanian S. Validation of surrogate estimates of insulin sensitivity and insulin secretion in children and adolescents. *J Pediatr* 2004;**144**:47–55. <http://dx.doi.org/10.1016/j.jpeds.2003.09.045>
159. Atabek ME, Pirgon O. Assessment of insulin sensitivity from measurements in fasting state and during an oral glucose tolerance test in obese children. *J Pediatr Endocrinol* 2007;**20**:187–95. <http://dx.doi.org/10.1515/JPEM.2007.20.2.187>
160. Keskin M, Kurtoglu S, Kendirci M, Atabek ME, Yazici C. Homeostasis model assessment is more reliable than the fasting glucose/insulin ratio and quantitative insulin sensitivity check index for assessing insulin resistance among obese children and adolescents. *Pediatrics* 2005;**115**:e500–3. <http://dx.doi.org/10.1542/peds.2004-1921>
161. Rossner SM, Neovius M, Montgomery SM, Marcus C, Norgren S. Alternative methods of insulin sensitivity assessment in obese children and adolescents. *Diabetes Care* 2008;**31**:802–4. <http://dx.doi.org/10.2337/dc07-1655>
162. Schwartz B, Jacobs DR, Moran A, Steinberger J, Hong CP, Sinaiko AR. Measurement of insulin sensitivity in children comparison between the euglycemic–hyperinsulinemic clamp and surrogate measures. *Diabetes Care* 2008;**31**:783–8. <http://dx.doi.org/10.2337/dc07-1376>
163. Cambuli VM, Incani M, Pilia S, Congiu T, Cavallo MG, Cossu E, *et al.* Oral glucose tolerance test in Italian overweight/obese children and adolescents results in a very high prevalence of impaired fasting glycaemia, but not of diabetes. *Diabetes Metab Res Rev* 2009;**25**:528–34. <http://dx.doi.org/10.1002/dmrr.980>
164. Libman IM, Barinas-Mitchell E, Bartucci A, Arslanian S. Reproducibility of the oral glucose tolerance test (OGTT) in overweight children: does it provide meaningful information? *Diabetes* 2008;**57**:A493.
165. Jetha MM, Nzekwu U, Lewanczuk RZ, Ball GDC. A novel, non-invasive ¹³C-glucose breath test to estimate insulin resistance in obese prepubertal children. *J Pediatr Endocrinol* 2009;**22**:1051–9. <http://dx.doi.org/10.1515/JPEM.2009.22.11.1051>
166. Molnar D, Jeges S, Erhardt E, Schutz Y. Measured and predicted resting metabolic rate in obese and nonobese adolescents. *J Pediatr* 1995;**127**:571–7. [http://dx.doi.org/10.1016/S0022-3476\(95\)70114-1](http://dx.doi.org/10.1016/S0022-3476(95)70114-1)
167. Rodriguez G, Moreno LA, Sarria A, Fleta J, Bueno M. Resting energy expenditure in children and adolescents: agreement between calorimetry and prediction equations. *Clin Nutr* 2002;**21**:255–60. <http://dx.doi.org/10.1054/clnu.2001.0531>
168. Lazzer S, Agosti F, De Col A, Sartorio A. Development and cross-validation of prediction equations for estimating resting energy expenditure in severely obese Caucasian children and adolescents. *Br J Nutr* 2006;**96**:973–9. <http://dx.doi.org/10.1017/BJN20061941>

169. Firouzbakhsh S, Mathis RK, Dorchester WL, Oseas RS, Groncy PK, Grant KE, *et al.* Measured resting energy-expenditure in children. *J Pediatr Gastroenterol Nutr* 1993;**16**:136–42. <http://dx.doi.org/10.1097/00005176-199302000-00007>
170. Derumeaux-Burel H, Meyer M, Morin L, Boirie Y. Prediction of resting energy expenditure in a large population of obese children. *Am J Clin Nutr* 2004;**80**:1544–50.
171. Hofsteenge GH, Chinapaw MJM, Delemarre-van de Waal HA, Weijs PJM. Validation of predictive equations for resting energy expenditure in obese adolescents. *Am J Clin Nutr* 2010;**91**:1244–54. <http://dx.doi.org/10.3945/ajcn.2009.28330>
172. Schmelzle H, Schroder C, Armbrust S, Unverzagt S, Fusch C. Resting energy expenditure in obese children aged 4 to 15 years: measured versus predicted data. *Acta Paediatr* 2004;**93**:739–46.
173. Dietz WH, Bandini LG, Schoeller DA. Estimates of metabolic rate in obese and nonobese adolescents. *J Pediatr* 1991;**118**:146–19. [http://dx.doi.org/10.1016/S0022-3476\(05\)81870-0](http://dx.doi.org/10.1016/S0022-3476(05)81870-0)
174. Nowicka P, Santoro N, Liu H, Lartaud D, Shaw MM, Goldberg R, *et al.* Utility of hemoglobin A(1c) for diagnosing prediabetes and diabetes in obese children and adolescents. *Diabetes Care* 2011;**34**:1306–11. <http://dx.doi.org/10.2337/dc10-1984>
175. Kelishadi R, Hashemipour M, Mohammadifard N, Alikhassy H, Adeli K. Short- and long-term relationships of serum ghrelin with changes in body composition and the metabolic syndrome in prepubescent obese children following two different weight loss programmes. *Clin Endocrinol (Oxf)* 2008;**69**:721–9. <http://dx.doi.org/10.1111/j.1365-2265.2008.03220.x>
176. Libman IM, Barinas-Mitchell E, Bartucci A, Robertson R, Arslanian S. Reproducibility of the oral glucose tolerance test in overweight children. *J Clin Endocrinol Metab* 2008;**93**:4231–7. <http://dx.doi.org/10.1210/jc.2008-0801>
177. Soder RB, Baldisserotto M, Duval da Silva V. Computer-assisted ultrasound analysis of liver echogenicity in obese and normal-weight children. *Am J Roentgenol* 2009;**192**:W201–5. <http://dx.doi.org/10.2214/AJR.08.2061>
178. Warschburger P, Buchholz HT, Petermann F. Development of a disease-specific interview method to assess the quality of life of obese children and teenagers. *Z Klin Psychol Psychiatr Psychother* 2001;**49**:247–61.
179. Warschburger P, Fromme C, Petermann F. Weight-specific quality of life in school-children: validity of the GW-LQ-KJ. *Zeitschrift fur Gesundheitspsychologie* 2004;**2**:159–66. <http://dx.doi.org/10.1026/0943-8149.12.4.159>
180. Warschburger P, Fromme C, Petermann F. Conception and analysis of a weight-specific quality of life questionnaire for overweight and obese children and adolescents (GW-LQ-KJ). *Z Klin Psychol Psychiatr Psychother* 2005;**53**:356–69.
181. Kolotkin RL, Zeller M, Modi AC, Samsa GP, Quinlan NP, Yanovski JA, *et al.* Assessing weight-related quality of life in adolescents. *Obesity (Silver Spring)* 2006;**14**:448–57. <http://dx.doi.org/10.1038/oby.2006.59>
182. Modi AC, Zeller MH. The IWQOL-Kids: establishing minimal clinically important difference scores and test-retest reliability. *Int J Pediatr Obes* 2011;**6**:e94–6. <http://dx.doi.org/10.3109/17477166.2010.500391>
183. Zeller MH, Modi AC. Development and initial validation of an obesity-specific quality-of-life measure for children: Sizing Me Up. *Obesity (Silver Spring)* 2009;**17**:1171–7. <http://dx.doi.org/10.1038/oby.2009.47>
184. Modi AC, Zeller MH. Validation of a parent-proxy, obesity-specific quality-of-life measure: sizing them up. *Obesity (Silver Spring)* 2008;**16**:2624–33. <http://dx.doi.org/10.1038/oby.2008.416>

185. Morales LS, Edwards TC, Flores Y, Barr L, Patrick DL. Measurement properties of a multicultural weight-specific quality-of-life instrument for children and adolescents. *Qual Life Res* 2011;**20**:215–24. <http://dx.doi.org/10.1007/s11136-010-9735-0>
186. Landgraf JM, Maunsell E, Speechley KN, Bullinger M, Campbell S, Abetz L, *et al.* Canadian-French, German and UK versions of the Child Health Questionnaire: methodology and preliminary item scaling results. *Qual Life Res* 1998;**7**:433–45. <http://dx.doi.org/10.1023/A:1008810004694>
187. Erhart M, Ellert U, Kurth B-M, Ravens-Sieberer U. Measuring adolescents' HRQoL via self reports and parent proxy reports: an evaluation of the psychometric properties of both versions of the KINDL-R instrument. *Health Qual Life Outcomes* 2009;**7**:77. <http://dx.doi.org/10.1186/1477-7525-7-77>
188. Varni JW, Katz ER, Seid M, Quiggins DJL, Friedman-Bender A. The pediatric cancer quality of life inventory-32 (PCQL-32). *Cancer* 1998;**82**:1184–96. [http://dx.doi.org/10.1002/\(SICI\)1097-0142\(19980315\)82:6<1184::AID-CNCR25>3.0.CO;2-1](http://dx.doi.org/10.1002/(SICI)1097-0142(19980315)82:6<1184::AID-CNCR25>3.0.CO;2-1)
189. Varni JW, Seid M, Rode CA. The PedsQL: measurement model for the pediatric quality of life inventory. *Med Care* 1999;**37**:126–39. <http://dx.doi.org/10.1097/00005650-199902000-00003>
190. Varni JW, Burwinkle TM, Seid M, Skarr D. The PedsQL™* 4.0 as a pediatric population health measure: feasibility, reliability, and validity. *Ambul Pediatr* 2003;**3**:329–41. [http://dx.doi.org/10.1367/1539-4409\(2003\)003<0329:TPAAPP>2.0.CO;2](http://dx.doi.org/10.1367/1539-4409(2003)003<0329:TPAAPP>2.0.CO;2)
191. Varni JW, Seid M, Kurtin PS. PedsQL (TM) 4.0: Reliability and validity of the Pediatric Quality of Life Inventory (TM) version 4.0 Generic Core Scales in healthy and patient populations. *Med Care* 2001;**39**:800. <http://dx.doi.org/10.1097/00005650-200108000-00006>
192. Waters E, Salmon L, Wake M, Hesketh K, Wright M. The Child Health Questionnaire in Australia: reliability, validity and population means. *Aust N Z J Public Health* 2000;**24**:207–10. <http://dx.doi.org/10.1111/j.1467-842X.2000.tb00145.x>
193. Waters E, Salmon L, Wake M. The parent-form Child Health Questionnaire in Australia: comparison of reliability, validity, structure, and norms. *J Pediatr Psychol* 2000;**25**:381–91. <http://dx.doi.org/10.1093/jpepsy/25.6.381>
194. Ravens-Sieberer USS, Gosch A, Erhart M, Petersen C, Bullinger M. Measuring subjective health in children and adolescents: results of the European KIDSCREEN/DISABKIDS Project. Emotional and external eating behavior. *Psychosoc Med* 2007;**4**:8.
195. Varni JW, Katz ER, Seid M, Quiggins DJL, Friedman-Bender A, Castro CM. The Pediatric Cancer Quality of Life Inventory (PCQL). I. Instrument development, descriptive statistics, and cross-informant variance. *J Behav Med* 1998;**21**:179–204. <http://dx.doi.org/10.1023/A:1018779908502>
196. Hughes AR, Farewell K, Harris D, Reilly J. Quality of life in a clinical sample of obese children. *Int J Obes* 2007;**31**:39–44. <http://dx.doi.org/10.1038/sj.ijo.0803410>
197. Kendall PC, Wilcox LE. Self-control in children: development of a rating scale. *J Consult Clin Psychol* 1979;**47**:1020–9. <http://dx.doi.org/10.1037/0022-006X.47.6.1020>
198. Truby H, Paxton SJ. Development of the Children's Body Image Scale. *Br J Clin Psychol* 2002;**41**:185–203. <http://dx.doi.org/10.1348/014466502163967>
199. Harter S. The perceived competence scale for children. *Child Dev* 1982;**53**:87–97. <http://dx.doi.org/10.2307/1129640>
200. Janicke DM, Storch EA, Novoa W, Silverstein JH, Samyn MM. The pediatric barriers to a healthy diet scale. *Child Health Care* 2007;**36**:155–68. <http://dx.doi.org/10.1080/02739610701334996>

201. La Greca AM, Dandes SK, Wick P, Shaw K, Stone WL. Development of the Social Anxiety Scale for Children: reliability and concurrent validity. *J Clin Child Psychol* 1988;**17**:84–91. http://dx.doi.org/10.1207/s15374424jccp1701_11
202. La Greca AM, Stone WL. Social Anxiety Scale for Children-revised: factor structure and concurrent validity. *J Clin Child Psychol* 1993;**22**:17–27. http://dx.doi.org/10.1207/s15374424jccp2201_2
203. Nowicki S, Strickland BR. A locus of control scale for children. *J Consult Clin Psychol* 1973;**40**:148. <http://dx.doi.org/10.1037/h0033978>
204. Mendelson BK, White DR. Relation between body-esteem and self-esteem of obese and normal children. *Percept Mot Skills* 1982;**54**:899–905. <http://dx.doi.org/10.2466/pms.1982.54.3.899>
205. Collins ME. Body figure perceptions and preferences among preadolescent children. *Int J Eat Disord* 1991;**10**:199–208. [http://dx.doi.org/10.1002/1098-108X\(199103\)10:2<199::AID-EAT2260100209>3.0.CO;2-D](http://dx.doi.org/10.1002/1098-108X(199103)10:2<199::AID-EAT2260100209>3.0.CO;2-D)
206. Conti MA, Cordas TA, Latorre MdRDdO. A study of the validity and reliability of the Brazilian version of the Body Shape Questionnaire (BSQ) among adolescents. *Rev Bras Saúde Matern Infant* 2009;**9**:331–8. <http://dx.doi.org/10.1590/S1519-38292009000300012>
207. Stein RJ, Bracken BA, Haddock CK, Shadish WR. Preliminary development of the Children's Physical Self-Concept Scale. *J Dev Behav Pediatr* 1998;**19**:1–8. <http://dx.doi.org/10.1097/0004703-199802000-00001>
208. Probst M, Braet C, Vandereycken W, De Vos P, Van Coppenolle H, Verhofstadt-Deneve L. Body size estimation in obese children: a controlled study with the video distortion method. *Int J Obes Relat Metab Disord* 1995;**19**:820–4.
209. Van Dongen-Melman J, Koot H, Verhulst F. Cross-cultural validation of Harter's self-perception profile for children in a Dutch sample. *Educ Psychol Meas* 1993;**53**:739–53. <http://dx.doi.org/10.1177/0013164493053003018>
210. Whitehead JR. A study of children's physical self-perceptions using an adapted physical self-perception profile questionnaire. *Pediatr Exerc Sci* 1995;**7**:132.
211. Hay JA. Adequacy in and predilection for physical activity in children. *Clin J Sport Med* 1992;**2**:192. <http://dx.doi.org/10.1097/00042752-199207000-00007>
212. Benjamin SE, Ammerman A, Sommers J, Dodds J, Neelon B, Ward DS. Nutrition and Physical Activity Self-Assessment for Child Care (NAP SACC): results from a child care pilot intervention. *J Nutr Educ Behav* 2007;**39**:142–9. <http://dx.doi.org/10.1016/j.jneb.2006.08.027>
213. Ward D, Hales D, Haverly K, Marks J, Benjamin S, Ball S, et al. An instrument to assess the obesogenic environment of child care centers. *Am J Health Behav* 2008;**32**:380–6. <http://dx.doi.org/10.5993/AJHB.32.4.5>
214. Bryant MJ, Ward DS, Hales D, Vaughn A, Tabak RG, Stevens J. Reliability and validity of the Healthy Home Survey: a tool to measure factors within homes hypothesized to relate to overweight in children. *Int J Behav Nutr Phys Act* 2008;**5**:23. <http://dx.doi.org/10.1186/1479-5868-5-23>
215. Golan M, Weizman A. Reliability and validity of the Family Eating and Activity Habits Questionnaire. *Eur J Clin Nutr* 1998;**52**:771–7. <http://dx.doi.org/10.1038/sj.ejcn.1600647>
216. Larios SE, Ayala GX, Arredondo EM, Baquero B, Elder JP. Development and validation of a scale to measure Latino parenting strategies related to children's obesigenic behaviors. The Parenting strategies for Eating and Activity Scale (PEAS). *Appetite* 2009;**52**:166–72. <http://dx.doi.org/10.1016/j.appet.2008.09.011>

217. McCurdy K, Gorman KS. Measuring family food environments in diverse families with young children. *Appetite* 2010;**54**:615–18. <http://dx.doi.org/10.1016/j.appet.2010.03.004>
218. Gattshall ML, Shoup JA, Marshall JA, Crane LA, Estabrooks PA. Validation of a survey instrument to assess home environments for physical activity and healthy eating in overweight children. *Int J Behav Nutr Phys Act* 2008;**5**:3. <http://dx.doi.org/10.1186/1479-5868-5-3>
219. Rosenberg DE, Sallis JF, Kerr J, Maher J, Norman GJ, Durant N, *et al.* Brief scales to assess physical activity and sedentary equipment in the home. *Int J Behav Nutr Phys Act* 2010;**7**:10. <http://dx.doi.org/10.1186/1479-5868-7-10>
220. Durant N, Kerr J, Harris SK, Saelens BE, Norman GJ, Sallis JF. Environmental and safety barriers to youth physical activity in neighborhood parks and streets: reliability and validity. *Pediatr Exerc Sci* 2009;**21**:86–99.
221. Nicholson JC, McDuffie JR, Bonat SH, Russell DL, Boyce KA, McCann S, *et al.* Estimation of body fatness by air displacement plethysmography in African American and white children. *Pediatr Res* 2001;**50**:467–73. <http://dx.doi.org/10.1203/00006450-200110000-00008>
222. Sampei J, McDuffie JR, Sebring NG, Salaita C, Keil M, Robotham D, *et al.* Comparison of methods to assess change in children's body composition. *Am J Clin Nutr* 2004;**80**:64–9.
223. Lazzer S, Bedogni G, Agosti F, De Col A, Mornati D, Sartorio A. Comparison of dual-energy X-ray absorptiometry, air displacement plethysmography and bioelectrical impedance analysis for the assessment of body composition in severely obese Caucasian children and adolescents. *Br J Nutr* 2008;**100**:918–24. <http://dx.doi.org/10.1017/S0007114508922558>
224. Mello MTd, Damaso AR, Antunes HKM, Siqueira KO, Castro ML, Bertolino SV, *et al.* Body composition evaluation in obese adolescents: the use of two different methods. *Rev Bras Med Esporte* 2005;**11**:262–6.
225. Radley D, Gately PJ, Cooke CB, Carroll S, Oldroyd B, Truscott JG. Estimates of percentage body fat in young adolescents: a comparison of dual-energy X-ray absorptiometry and air displacement plethysmography. *Eur J Clin Nutr* 2003;**57**:1402–10. <http://dx.doi.org/10.1038/sj.ejcn.1601702>
226. Goodman E, Hinden BR, Khandelwal S. Accuracy of teen and parental reports of obesity and body mass index. *Pediatrics* 2000;**106**:52–8. <http://dx.doi.org/10.1542/peds.106.1.52>
227. Strauss RS. Comparison of measured and self-reported weight and height in a cross-sectional sample of young adolescents. *Int J Obes Relat Metab Disord* 1999;**23**:904–8. <http://dx.doi.org/10.1038/sj.ijo.0800971>
228. Scholtens S, Brunekreef B, Visscher TLS, Smit HA, Kerkhof M, de Jongste JC, *et al.* Reported versus measured body weight and height of 4-year-old children and the prevalence of overweight. *Eur J Public Health* 2007;**17**:369–74. <http://dx.doi.org/10.1093/eurpub/ckl253>
229. Jansen E, Mulkens S, Hamers H, Jansen A. Assessing eating disordered behaviour in overweight children and adolescents: Bridging the gap between a self-report questionnaire and a gold standard interview. *Neth J Psychol* 2007;**63**:102–6. <http://dx.doi.org/10.1007/BF03061070>
230. Tanofsky-Kraff M, Morgan CM, Yanovski SZ, Marmarosh C, Wilfley DE, Yanovski JA. Comparison of assessments of children's eating-disordered behaviors by interview and questionnaire. *Int J Eat Disord* 2003;**33**:213–24. <http://dx.doi.org/10.1002/eat.10128>
231. Shapiro JR, Woolson SL, Hamer RM, Kalarchian MA, Marcus MD, Bulik CM. Evaluating binge eating disorder in children: Development of the Children's Binge Eating Disorder Scale (C-BEDS). *Int J Eat Disord* 2007;**40**:82–9. <http://dx.doi.org/10.1002/eat.20318>

232. Boles RE, Nelson TD, Chamberlin LA, Valenzuela JM, Sherman SN, Johnson SL, *et al.* Confirmatory factor analysis of the Child Feeding Questionnaire among low-income African American families of preschool children. *Appetite* 2010;**54**:402–5. <http://dx.doi.org/10.1016/j.appet.2009.12.013>
233. Kramer MS, Matush L, Vanilovich I, Platt RW, Bogdanovich N, Sevkovskaya Z, *et al.* Effects of prolonged and exclusive breastfeeding on child height, weight, adiposity, and blood pressure at age 6.5 years: evidence from a large randomized trial. *Am J Clin Nutr* 2007;**86**:1717–21.
234. Guinhoya CB, Apete GK, Hubert H. Diagnostic quality of Actigraph-based physical activity cut-offs for children: what overweight/obesity references can tell? *Pediatr Int* 2009;**51**:568–73. <http://dx.doi.org/10.1111/j.1442-200X.2008.02801.x>
235. Prochaska JJ, Sallis JF, Long B. A physical activity screening measure for use with adolescents in primary care. *Arch Pediatr Adolesc Med* 2001;**155**:554. <http://dx.doi.org/10.1001/archpedi.155.5.554>
236. Kriska AM, Knowler WC, LaPorte RE, Drash AL, Wing RR, Blair SN, *et al.* Development of questionnaire to examine relationship of physical activity and diabetes in Pima Indians. *Diabetes Care* 1990;**13**:401–11. <http://dx.doi.org/10.2337/diacare.13.4.401>
237. Maffei C, Pinelli L, Zaffanello M, Schena F, Iacumin P, Schutz Y. Daily energy expenditure in free-living conditions in obese and non-obese children: comparison of doubly labelled water (2H₂(18)O) method and heart-rate monitoring. *Int J Obes Relat Metab Disord* 1995;**19**:671–7.
238. Troped PJ, Wiecha JL, Fragala MS, Matthews CE, Finkelstein DM, Kim J, *et al.* Reliability and validity of YRBS physical activity items among middle school students. *Med Sci Sports Exerc* 2007;**39**:416–25. <http://dx.doi.org/10.1249/mss.0b013e31802d97af>
239. Ortega FB, Ruiz JR, Espana-Romero V, Vicente-Rodriguez G, Martinez-Gomez D, Manios Y, *et al.* The International Fitness Scale (IFIS): usefulness of self-reported fitness in youth. *Int J Epidemiol* 2011;**40**:701–11. <http://dx.doi.org/10.1093/ije/dyr039>
240. Morrow JR, Jr, Martin SB, Welk GJ, Zhu W, Meredith MD. Overview of the Texas Youth Fitness Study. *Res Q Exerc Sport* 2010;**81**(Suppl. 3):1–5. <http://dx.doi.org/10.1080/02701367.2010.10599688>
241. Burstrom K, Svartengren M, Egmar AC. Testing a Swedish child-friendly pilot version of the EQ-5D instrument: initial results. *Eur J Public Health* 2011;**21**:178–83. <http://dx.doi.org/10.1093/eurpub/ckq042>
242. Burstrom K, Egmar AC, Lugner A, Eriksson M, Svartengren M. A Swedish child-friendly pilot version of the EQ-5D instrument: the development process. *Eur J Public Health* 2011;**21**:171–7. <http://dx.doi.org/10.1093/eurpub/ckq037>
243. Wille N, Bullinger M, Holl R, Hoffmeister U, Mann R, Goldapp C, *et al.* Health-related quality of life in overweight and obese youths: results of a multicenter study. *Health Qual Life Outcomes* 2010;**8**:36. <http://dx.doi.org/10.1186/1477-7525-8-36>
244. Ravens-Sieberer U, Wille N, Badia X, Bonsel G, Burstrom K, Cavrini G, *et al.* Feasibility, reliability, and validity of the EQ-5D-Y: results from a multinational study. *Qual Life Res* 2010;**19**:887–97. <http://dx.doi.org/10.1007/s11136-010-9649-x>
245. Eklund RC, Whitehead JR, Welk GJ. Validity of the children and youth physical self-perception profile: a confirmatory factor analysis. *Res Q Exerc Sport* 1997;**68**:249–56. <http://dx.doi.org/10.1080/02701367.1997.10608004>
246. Radloff LS. The use of the Center for Epidemiologic Studies Depression Scale in adolescents and young adults. *J Youth Adolesc* 1991;**20**:149–66. <http://dx.doi.org/10.1007/BF01537606>

247. Benjamin SE, Neelon B, Ball SC, Bangdiwala SI, Ammerman AS, Ward DS. Reliability and validity of a nutrition and physical activity environmental self-assessment for child care. *Int J Behav Nutr Phys Act* 2007;**4**:29. <http://dx.doi.org/10.1186/1479-5868-4-29>
248. Duncan MJ, Al-Nakeeb Y, Woodfield L, Lyons M. Pedometer determined physical activity levels in primary school children from central England. *Prev Med* 2007;**44**:416–20. <http://dx.doi.org/10.1016/j.ypmed.2006.11.019>
249. Motl RW, Dishman RK, Saunders R, Dowda M, Felton G, Pate RR. Measuring enjoyment of physical activity in adolescent girls. *Am J Prev Med* 2001;**21**:110–17. [http://dx.doi.org/10.1016/S0749-3797\(01\)00326-9](http://dx.doi.org/10.1016/S0749-3797(01)00326-9)
250. Carper JL, Orlet Fisher J, Birch LL. Young girls' emerging dietary restraint and disinhibition are related to parental control in child feeding. *Appetite* 2000;**35**:121–9. <http://dx.doi.org/10.1006/appe.2000.0343>
251. Radley D, Fields DA, Gately PJ. Validity of thoracic gas volume equations in children of varying body mass index classifications. *Int J Pediatr Obes* 2007;**2**:180–7. <http://dx.doi.org/10.1080/17477160701191710>
252. de Hof SI, Bakker I, Hopman-Rock M, Hirasing RA, van Mechelen W. Clinimetric review of motion sensors in children and adolescents. *J Clin Epidemiol* 2006;**59**:670–80. <http://dx.doi.org/10.1016/j.jclinepi.2005.11.020>
253. Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, et al. CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ* 2010;**340**:c869. <http://dx.doi.org/10.1136/bmj.c869>
254. Schulz K, Altman D, Moher D, Group tC. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *BMC Med* 2010;**8**:18. <http://dx.doi.org/10.1186/1741-7015-8-18>
255. Agency for Healthcare Research and Quality. *Effectiveness of Weight Management Programs in Children and Adolescents*. Report 170. Rockville, MD: US Department of Health and Human Services; 2008.
256. Blackburn G. Effect of degree of weight loss on health benefits. *Obes Res* 1995;**3**(Suppl. 2):211–16. <http://dx.doi.org/10.1002/j.1550-8528.1995.tb00466.x>
257. Institute of Medicine National Academy of Sciences. *Weighing the Options: Criteria for Evaluating Weight Management Programs*. Washington DC: National Academy Press; 1995.
258. Klesges LM, Baranowski T, Beech B, Cullen K, Murray DM, Rochon J, et al. Social desirability bias in self-reported dietary, physical activity and weight concerns measures in 8- to 10-year-old African-American girls: results from the Girls health Enrichment Multisite Studies (GEMS). *Prev Med* 2004;**38**:78–87. <http://dx.doi.org/10.1016/j.ypmed.2003.07.003>
259. Goran MI. Measurement issues related to studies of childhood obesity: Assessment of body composition, body fat distribution, physical activity and food intake. *Pediatrics* 1998;**101**:505–18.
260. van Emmerik NMA, Renders CM, van de Veer M, van Buuren S, van der Baan-Slootweg OH, Kist-van Holthe JE, et al. High cardiovascular risk in severely obese young children and adolescents. *Arch Dis Child* 2012;**97**:818–21. <http://dx.doi.org/10.1136/archdischild-2012-301877>
261. Wells G, Li T, Maxwell L, Maclean R, Tugwell P. Responsiveness of patient reported outcomes including fatigue, sleep quality, activity limitation, and quality of life following treatment with abatacept for rheumatoid arthritis. *Ann Rheum Dis* 2008;**67**:260–5. <http://dx.doi.org/10.1136/ard.2007.069690>

262. Liang MH. Longitudinal construct validity: establishment of clinical meaning in patient evaluative instruments. *Med Care* 2000;**39**(Suppl. 9):84–90.
263. Liang MH, Lew R, Stucki G, Fortin PR, Daltroy L. Measuring clinically important changes with patient-oriented questionnaires. *Med Care* 2002;**40**(Suppl. 4):1145–51. <http://dx.doi.org/10.1097/00005650-200204001-00008>
264. Cole TJ, Faith MS, Pietrobelli A, Heo M. What is the best measure of adiposity change in growing children: BMI, BMI%, BMI z-score or BMI centile? *Euro J Clin Nutr* 2005;**59**:419–25. <http://dx.doi.org/10.1038/sj.ejcn.1602090>
265. Whitehead SJ, Ali S. Health outcomes in economic evaluation: the QALY and utilities. *Br Med Bull* 2010;**96**:5–21. <http://dx.doi.org/10.1093/bmb/ldq033>
266. Bakker C, van der Linden S. Health related utility measurement: an introduction. *J Rheumatol* 1995;**22**:1197–9.
267. National Institute for Health and Care Excellence (NICE). *Guide to the Methods of Technology Appraisal*. URL: www.nice.org.uk/media/B52/A7/TAMethodsGuideUpdatedJune2008.pdf
268. Bland JM, Altman DG. A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. *Comput Biol Med* 1990;**20**:337–40. [http://dx.doi.org/10.1016/0010-4825\(90\)90013-F](http://dx.doi.org/10.1016/0010-4825(90)90013-F)
269. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res* 1999;**8**:135–60. <http://dx.doi.org/10.1191/096228099673819272>
270. Savgan-Gurol E, Bredella M, Russell M, Mendes N, Klibanski A, Misra M. Waist to hip ratio and trunk to extremity fat (DXA) are better surrogates for IMCL and for visceral fat respectively than for subcutaneous fat in adolescent girls. *Nutr Metab* 2010;**7**:86. <http://dx.doi.org/10.1186/1743-7075-7-86>
271. Semiz S, Ozgoren E, Sabir N. Comparison of ultrasonographic and anthropometric methods to assess body fat in childhood obesity. *Int J Obes* 2007;**31**:53–8. <http://dx.doi.org/10.1038/sj.ijo.0803414>
272. Rolland-Cachera MF, Brambilla P, Manzoni P, Akrouf M, Sironi S, Del Maschio A, et al. Body composition assessed on the basis of arm circumference and triceps skinfold thickness: a new index validated in children by magnetic resonance imaging. *Am J Clin Nutr* 1997;**65**:1709–13.
273. Shaikh MG, Crabtree NJ, Shaw NJ, Kirk JMW. Body fat estimation using bioelectrical impedance. *Horm Res* 2007;**68**:8–10. <http://dx.doi.org/10.1159/000098481>
274. Azcona C, Koek N, Fruhbeck G. Fat mass by air-displacement plethysmography and impedance in obese/non-obese children and adolescents. *Int J Pediatr Obes* 2006;**1**:176–82. <http://dx.doi.org/10.1080/17477160600858740>
275. Okasora K, Takaya R, Tokuda M, Fukunaga Y, Oguni T, Tanaka H, et al. Comparison of bioelectrical impedance analysis and dual energy X-ray absorptiometry for assessment of body composition in children. *Pediatr Int* 1999;**41**:121–5. <http://dx.doi.org/10.1046/j.1442-200X.1999.4121048.x>
276. Loftin M, Nichols J, Going S, Sothorn M, Schmitz KH, Ring K, et al. Comparison of the validity of anthropometric and bioelectric impedance equations to assess body composition in adolescent girls. *Int J Body Compos Res* 2007;**5**:1–8.
277. Iwata K, Satou Y, Iwata F, Hara M, Fuchigami S, Kin H, et al. Assessment of body composition measured by bioelectrical impedance in children. *Acta Paediatr Jpn* 1993;**35**:369–72. <http://dx.doi.org/10.1111/j.1442-200X.1993.tb03074.x>

278. Guida B, Pietrobelli A, Trio R, Laccetti R, Falconi C, Perrino NR, *et al.* Body mass index and bioelectrical vector distribution in 8-year-old children. *Nutr Metab Cardiovasc Dis* 2008;**18**:133–41. <http://dx.doi.org/10.1016/j.numecd.2006.08.008>
279. Asayama K, Oguni T, Hayashi K, Dobashi K, Fukunaga Y, Kodera K, *et al.* Critical value for the index of body fat distribution based on waist and hip circumferences and stature in obese girls. *Int J Obes Relat Metab Disord* 2000;**24**:1026–31. <http://dx.doi.org/10.1038/sj.ijo.0801355>
280. Lazzer S, Boirie Y, Meyer M, Vermorel M. Evaluation of two foot-to-foot bioelectrical impedance analysers to assess body composition in overweight and obese adolescents. *Br J Nutr* 2003;**90**:987–92. <http://dx.doi.org/10.1079/BJN2003983>
281. Eisenkolbl J, Kartasurya M, Widhalm K. Underestimation of percentage fat mass measured by bioelectrical impedance analysis compared to dual energy X-ray absorptiometry method in obese children. *Eur J Clin Nutr* 2001;**55**:423–9. <http://dx.doi.org/10.1038/sj.ejcn.1601184>
282. Hannon JC, Ratliffe T, Williams DP. Agreement in body fat estimates between a hand-held bioelectrical impedance analyzer and skinfold thicknesses in African American and Caucasian adolescents. *Res Q Exerc Sport* 2006;**77**:519–26. <http://dx.doi.org/10.1080/02701367.2006.10599387>
283. Goran MI, Driscoll P, Johnson R, Nagy TR, Hunter G. Cross-calibration of body-composition techniques against dual-energy X-ray absorptiometry in young children. *Am J Clin Nutr* 1996;**63**:299–305.
284. Ellis KJ. Measuring body fatness in children and young adults: comparison of bioelectric impedance analysis, total body electrical conductivity, and dual-energy X-ray absorptiometry. *Int J Obes Relat Metab Disord* 1996;**20**:866–73.
285. Fernandes RA, Rosa CSC, Buonani C, De Oliveira AR, Freitas IF Jr. The use of bioelectrical impedance to detect excess visceral and subcutaneous fat. *J Pediatr (Rio J)* 2007;**83**:529–34. <http://dx.doi.org/10.2223/JPED.1722>
286. Widhalm K, Schonegger K, Huemer C, Auterith A. Does the BMI reflect body fat in obese children and adolescents? A study using the TOBEC method. *Int J Obes Relat Metab Disord* 2001;**25**:279–85. <http://dx.doi.org/10.1038/sj.ijo.0801511>
287. Gaskin PS, Walker SP. Obesity in a cohort of black Jamaican children as estimated by BMI and other indices of adiposity. *Eur J Clin Nutr* 2003;**57**:420–6. <http://dx.doi.org/10.1038/sj.ejcn.1601564>
288. Warner JT, Cowan FJ, Dunstan FD, Gregory JW. The validity of body mass index for the assessment of adiposity in children with disease states. *Ann Hum Biol* 1997;**24**:209–15. <http://dx.doi.org/10.1080/03014469700004942>
289. Pietrobelli A, Faith MS, Allison DB, Gallagher D, Chiumello G, Heymsfield SB. Body mass index as a measure of adiposity among children and adolescents: a validation study. *J Pediatr* 1998;**132**:204–10. [http://dx.doi.org/10.1016/S0022-3476\(98\)70433-0](http://dx.doi.org/10.1016/S0022-3476(98)70433-0)
290. Glaner MF. Body mass index as indicative of body fat compared to the skinfolds. *Rev Brasil Med Esporte* 2005;**11**:243–6.
291. Reilly JJ, Dorosty AR, Emmett PM, Avon Longitudinal Study of P, Childhood Study T. Identification of the obese child: adequacy of the body mass index for clinical practice and epidemiology. *Int J Obes Relat Metab Disord* 2000;**24**:1623–7. <http://dx.doi.org/10.1038/sj.ijo.0801436>
292. Potter JA, Laws CJ, Candy DC. Classification of body composition in 11–14 year olds by both body mass index and bioelectrical impedance. *Int J Pediatr Obes* 2007;**2**:126–8. <http://dx.doi.org/10.1080/17477160701207276>

293. Ochiai H, Shirasawa T, Nishimura R, Morimoto A, Shimada N, Ohtsu T, *et al.* Relationship of body mass index to percent body fat and waist circumference among schoolchildren in Japan: the influence of gender and obesity: a population-based cross-sectional study. *BMC Public Health* 2010;**10**:493. <http://dx.doi.org/10.1186/1471-2458-10-493>
294. Morrissey SL, Whetstone LM, Cummings DM, Owen LJ. Comparison of self-reported and measured height and weight in eighth-grade students. *J Sch Health* 2006;**76**:512–15. <http://dx.doi.org/10.1111/j.1746-1561.2006.00150.x>
295. Molina M del C, de Faria CP, Montero P, Cade NV. Correspondence between children's nutritional status and mothers' perceptions: a population-based study. *Cad Saude Pública* 2009;**25**:2285–90. <http://dx.doi.org/10.1590/S0102-311X2009001000018>
296. Maynard LM, Galuska DA, Blanck HM, Serdula MK. Maternal perceptions of weight status of children. *Pediatrics* 2003;**111**:1226–31.
297. Mast M, Langnase K, Labitzke K, Bruse U, Preuss U, Muller MJ. Use of BMI as a measure of overweight and obesity in a field study on 5–7 year old children. *Eur J Nutr* 2002;**41**:61–7. <http://dx.doi.org/10.1007/s003940200009>
298. Malina RM, Katzmarzyk PT. Validity of the body mass index as an indicator of the risk and presence of overweight in adolescents. *Am J Clin Nutr* 1999;**70**:S131–6.
299. Ellis KJ, Abrams SA, Wong WW. Monitoring childhood obesity: assessment of the weight/height index. *Am J Epidemiol* 1999;**150**:939–46. <http://dx.doi.org/10.1093/oxfordjournals.aje.a010102>
300. Duncan JS, Duncan EK, Schofield G. Accuracy of body mass index (BMI) thresholds for predicting excess body fat in girls from five ethnicities. *Asia Pac J Clin Nutr* 2009;**18**:404–11.
301. Bartok CJ, Marini ME, Birch LL. High body mass index percentile accurately reflects excess adiposity in white girls. *J Am Diet Assoc* 2011;**111**:437–41. <http://dx.doi.org/10.1016/j.jada.2010.11.015>
302. El Taguri A, Dabbas-Tyan M, Goulet O, Ricour C. The use of body mass index for measurement of fat mass in children is highly dependant on abdominal fat. *East Mediterr Health J* 2009;**15**:563–73.
303. Yoo S, Lee SY, Kim KN, Sung E. Obesity in Korean pre-adolescent school children: comparison of various anthropometric measurements based on bioelectrical impedance analysis. *Int J Obes (Lond)* 2006;**30**:1086–90. <http://dx.doi.org/10.1038/sj.ijo.0803327>
304. Eto C, Komiya S, Nakao T, Kikkawa K. Validity of the body mass index and fat mass index as an indicator of obesity in children aged 3-5 years. *J Physiol Anthropol Appl Human Sci* 2004;**23**:25–30. <http://dx.doi.org/10.2114/jpa.23.25>
305. Rolland-Cachera MF, Sempe M, Guilloud-Bataille M, Patois E, Pequignot-Guggenbuhl F, Fautrad V. Adiposity indices in children. *Am J Clin Nutr* 1982;**36**:178–84.
306. Sampei MA, Novo NF, Juliano Y, Sigulem DM. Comparison of the body mass index to other methods of body fat evaluation in ethnic Japanese and Caucasian adolescent girls. *Int J Obes Relat Metab Disord* 2001;**25**:400–8. <http://dx.doi.org/10.1038/sj.ijo.0801558>
307. Mei ZG, Grummer-Strawn LM, Pietrobelli A, Goulding A, Goran MI, Dietz WH. Validity of body mass index compared with other body-composition screening indexes for the assessment of body fatness in children and adolescents. *Am J Clin Nutr* 2002;**75**:978–85.
308. Himes JH. Agreement among anthropometric indicators identifying the fattest adolescents. *Int J Obes Relat Metab Disord* 1999;**23**(Suppl. 2):18–21. <http://dx.doi.org/10.1038/sj.ijo.0800854>

309. Nuutinen EM, Turtinen J, Pokka T, Kuusela V, Dahlstrom S, Viikari J, *et al.* Obesity in children, adolescents and young adults. *Ann Med* 1991;**23**:41–6. <http://dx.doi.org/10.3109/07853899109147929>
310. Mei Z, Grummer-Strawn LM, Wang J, Thornton JC, Freedman DS, Pierson RN Jr, *et al.* Do skinfold measurements provide additional information to body mass index in the assessment of body fatness among children and adolescents? *Pediatrics* 2007;**119**:e1306–13. <http://dx.doi.org/10.1542/peds.2006-2546>
311. Glasser N, Zellner K, Kromeyer-Hauschild K. Validity of body mass index and waist circumference to detect excess fat mass in children aged 7–14 years. *Eur J Clin Nutr* 2011;**65**:151–9. <http://dx.doi.org/10.1038/ejcn.2010.245>
312. Neovius M, Linne Y, Rossner S. BMI, waist-circumference and waist-hip ratio as diagnostic tests for fatness in adolescents. *Int J Obes (Lond)* 2005;**29**:163–9. <http://dx.doi.org/10.1038/sj.jjo.0802867>
313. Adegboye ARA, Andersen LB, Froberg K, Sardinha LB, Heitmann BL. Linking definition of childhood and adolescent obesity to current health outcomes. *Int J Pediatr Obes* 2010;**5**:130–42. <http://dx.doi.org/10.3109/17477160903111730>
314. Jung C, Fischer N, Fritzenwanger M, Pernow J, Brehm BR, Figulla HR. Association of waist circumference, traditional cardiovascular risk factors, and stromal-derived factor-1 in adolescents. *Pediatr Diabetes* 2009;**10**:329–35. <http://dx.doi.org/10.1111/j.1399-5448.2008.00486.x>
315. Fujita Y, Kouda K, Nakamura H, Iki M. Cut-off values of body mass index, waist circumference, and waist-to-height ratio to identify excess abdominal fat: population-based screening of Japanese school children. *J Epidemiol* 2011;**21**:191–6. <http://dx.doi.org/10.2188/jea.JE20100116>
316. Rosenberg M, Greenberger S, Rawal A, Latimer-Pierson J, Thundiyil J. Comparison of Broselow tape measurements versus physician estimations of pediatric weights. *Am J Emerg Med* 2011;**29**:482–8. <http://dx.doi.org/10.1016/j.ajem.2009.12.002>
317. Killion L, Hughes SO, Wendt JC, Pease D, Nicklas TA. Minority mothers' perceptions of children's body size. *Int J Pediatr Obes* 2006;**1**:96–102. <http://dx.doi.org/10.1080/17477160600684286>
318. Fors H, Gelerander L, Bjarnason R, Albertsson-Wikland K, Bosaeus I. Body composition, as assessed by bioelectrical impedance spectroscopy and dual-energy X-ray absorptiometry, in a healthy paediatric population. *Acta Paediatr* 2002;**91**:755–60. <http://dx.doi.org/10.1111/j.1651-2227.2002.tb03323.x>
319. Springer F, Eehalt S, Sommer J, Ballweg V, Machann J, Binder G, *et al.* Assessment of relevant hepatic steatosis in obese adolescents by rapid fat-selective GRE imaging with spatial-spectral excitation: a quantitative comparison with spectroscopic findings. *Eur Radiol* 2011;**21**:816–22. <http://dx.doi.org/10.1007/s00330-010-1975-4>
320. Ball GDC, Huang TTK, Cruz ML, Shaibi GQ, Weigensberg MJ, Goran MI. Predicting abdominal adipose tissue in overweight Latino youth. *Int J Pediatr Obes* 2006;**1**:210–16. <http://dx.doi.org/10.1080/17477160600913578>
321. O'Connor DP, Gugenheim JJ. Comparison of measured and parents' reported height and weight in children and adolescents. *Obesity (Silver Spring)* 2011;**19**:1040–6. <http://dx.doi.org/10.1038/oby.2010.278>
322. Rasmussen F, Eriksson M, Nordquist T. Bias in height and weight reported by Swedish adolescents and relations to body dissatisfaction: the COMPASS study. *Eur J Clin Nutr* 2007;**61**:870–6. <http://dx.doi.org/10.1038/sj.ejcn.1602595>
323. Lu K, Quach B, Tong TK, Lau PWC. Validation of leg-to-leg bio-impedance analysis for assessing body composition in obese Chinese children. *J Exerc Sci Fit* 2003;**1**:97–103.

324. Dubois L, Girad M. Accuracy of maternal reports of pre-schoolers' weights and heights as estimates of BMI values. *Int J Epidemiol* 2007;**36**:132–8. <http://dx.doi.org/10.1093/ije/dyl281>
325. Gillis L, Bar-Or O, Calvert R. Validating a practical approach to determine weight control in obese children and adolescents. *Int J Obes Relat Metab Disord* 2000;**24**:1648–52. <http://dx.doi.org/10.1038/sj.ijo.0801458>
326. Nafiu OO, Burke C, Lee J, Voepel-Lewis T, Malviya S, Tremper KK. Neck circumference as a screening measure for identifying children with high body mass index. *Pediatrics* 2010;**126**:e306–10. <http://dx.doi.org/10.1542/peds.2010-0242>
327. Akinbami LJ, Ogden CL. Childhood Overweight Prevalence in the United States: The Impact of Parent-reported Height and Weight. *Obesity (Silver Spring)* 2009;**17**:1574–80. <http://dx.doi.org/10.1038/oby.2009.1>
328. Huybrechts I, De Bacquer D, Van Trimpont I, De Backer G, De Henauw S. Validity of parentally reported weight and height for preschool-aged children in Belgium and its impact on classification into body mass index categories. *Pediatrics* 2006;**118**:2109–18. <http://dx.doi.org/10.1542/peds.2006-0961>
329. Huybrechts I, Himes JH, Ottevaere C, De Vriendt T, De Keyzer W, Cox B, et al. Validity of parent-reported weight and height of preschool children measured at home or estimated without home measurement: a validation study. *BMC Pediatr* 2011;**11**:63. <http://dx.doi.org/10.1186/1471-2431-11-63>
330. Garcia-Marcos L, Valverde-Molina J, Sanchez-Solis M, Soriano-Perez MJ, Baeza-Alcaraz A, Martinez-Torres A, et al. Validity of parent-reported height and weight for defining obesity among asthmatic and nonasthmatic schoolchildren. *Int Arch Allergy Immunol* 2006;**139**:139–45. <http://dx.doi.org/10.1159/000090389>
331. Jones AR, Parkinson KN, Drewett RF, Hyland RM, Pearce MS, Adamson AJ. Parental perceptions of weight status in children: The Gateshead Millennium Study. *Int J Obes (Lond)* 2011;**35**:953–62. <http://dx.doi.org/10.1038/ijo.2011.106>
332. Vuorela N, Saha MT, Salo MK. Parents underestimate their child's overweight. *Acta Paediatrica, Int J Pediatr* 2010;**99**:1374–9. <http://dx.doi.org/10.1111/j.1651-2227.2010.01829.x>
333. Tschamler JM, Conn KM, Cook SR, Halterman JS. Underestimation of children's weight status: views of parents in an urban community. *Clin Pediatr (Phila)* 2010;**49**:470–6. <http://dx.doi.org/10.1177/0009922809336071>
334. Wen X, Hui S. Chinese parents' perceptions of their children's weights and their relationship to parenting behaviours. *Child Care Health Dev* 2011;**37**:343–51. <http://dx.doi.org/10.1111/j.1365-2214.2010.01166.x>
335. Akerman A, Williams ME, Meunier J. Perception versus reality: an exploration of children's measured body mass in relation to caregivers' estimates. *J Health Psychol* 2007;**12**:871–82. <http://dx.doi.org/10.1177/1359105307082449>
336. van Vliet JS, Kjolhede EA, Duchon K, Rasanen L, Nelson N. Waist circumference in relation to body perception reported by Finnish adolescent girls and their mothers. *Acta Paediatr* 2009;**98**:501–6. <http://dx.doi.org/10.1111/j.1651-2227.2008.01112.x>
337. Seghers J, Claessens AL. Bias in self-reported height and weight in preadolescents. *J Pediatr* 2010;**157**:911–16. <http://dx.doi.org/10.1016/j.jpeds.2010.06.038>
338. Jansen W, van de Looij-Jansen PM, Ferreira I, de Wilde EJ, Brug J. Differences in measured and self-reported height and weight in Dutch adolescents. *Ann Nutr Metab* 2006;**50**:339–46. <http://dx.doi.org/10.1159/000094297>

339. Zhou X, Dibley MJ, Cheng Y, Ouyang X, Yan H. Validity of self-reported weight, height and resultant body mass index in Chinese adolescents and factors associated with errors in self-reports. *BMC Public Health* 2010;**10**:190. <http://dx.doi.org/10.1186/1471-2458-10-190>
340. Yan AF, Zhang G, Wang MQ, Stoesen CA, Harris BM. Weight perception and weight control practice in a multiethnic sample of US adolescents. *South Med J* 2009;**102**:354–60. <http://dx.doi.org/10.1097/SMJ.0b013e318198720b>
341. Fonseca H, Silva AM, Matos MG, Esteves I, Costa P, Guerra A, et al. Validity of BMI based on self-reported weight and height in adolescents. *Acta Paediatr* 2010;**99**:83–8. <http://dx.doi.org/10.1111/j.1651-2227.2009.01518.x>
342. Enes CC, Fernandez PMF, Voci SM, Toral N, Romero A, Slater B. Validity and reliability of self-reported weight and height measures for the diagnoses of adolescent's nutritional status. *Rev Brasil Epidemiol* 2009;**12**:627–35. <http://dx.doi.org/10.1590/S1415-790X2009000400012>
343. Crawley HF, Portides G. Self-reported versus measured height, weight and body-mass index amongst 16–17-year-old british teenagers. *Int J Obes* 1995;**19**:579–84.
344. Linhart Y, Romano-Zelekha O, Shohat T. Validity of self-reported weight and height among 13–14 year old schoolchildren in Israel. *Isr Med Assoc J* 2010;**12**:603–5.
345. Lee K, Valeria B, Kochman C, Lenders CM. Self-assessment of height, weight, and sexual maturation: validity in overweight children and adolescents. *J Adolesc Health* 2006;**39**:346–52. <http://dx.doi.org/10.1016/j.jadohealth.2005.12.016>
346. Wang Z, Patterson CM, Hills AP. A comparison of self-reported and measured height, weight and BMI in Australian adolescents. *Aust N Z J Public Health* 2002;**26**:473–8. <http://dx.doi.org/10.1111/j.1467-842X.2002.tb00350.x>
347. Tsigilis N. Can secondary school students' self-reported measures of height and weight be trusted? An effect size approach. *Eur J Public Health* 2006;**16**:532–5. <http://dx.doi.org/10.1093/eurpub/ckl050>
348. Tokmakidis SP, Christodoulos AD, Mantzouranis NI. Validity of self-reported anthropometric values used to assess body mass index and estimate obesity in Greek school children. *J Adolesc Health* 2007;**40**:305–10. <http://dx.doi.org/10.1016/j.jadohealth.2006.10.001>
349. Shields M, Gorber SC, Tremblay MS. Estimates of obesity based on self-report versus direct measures. *Health Rep* 2008;**19**:61–76.
350. Abalkhail BA, Shawky S, Soliman NK. Validity of self-reported weight and height among Saudi school children and adolescents. *Saudi Med J* 2002;**23**:831–7.
351. Hauck FR, White L, Cao G, Woolf N, Strauss K. Inaccuracy of self-reported weights and heights among American Indian adolescents. *Ann Epidemiol* 1995;**5**:386–92. [http://dx.doi.org/10.1016/1047-2797\(95\)00036-7](http://dx.doi.org/10.1016/1047-2797(95)00036-7)
352. Bae J, Joung H, Kim JY, Kwon KN, Kim Y, Park SW. Validity of self-reported height, weight, and body mass index of the Korea Youth Risk Behavior Web-based Survey questionnaire. *J Prev Med Public Health* 2010;**43**:396–402. <http://dx.doi.org/10.3961/jpmph.2010.43.5.396>
353. De Vriendt T, Huybrechts I, Ottevaere C, Van Trimpont I, De Henauw S. Validity of self-reported weight and height of adolescents, its impact on classification into BMI categories and the association with weighing behaviour. *Int J Environ Res Public Health* 2009;**6**:2696–711. <http://dx.doi.org/10.3390/ijerph6102696>
354. Ambrosi-Randic N, Bulian AP. Self-reported versus measured weight and height by adolescent girls: a Croatian sample. *Percept Mot Skills* 2007;**104**:79–82. <http://dx.doi.org/10.2466/pms.104.1.79-82>

355. Field AE, Aneja P, Rosner B. The validity of self-reported weight change among adolescents and young adults. *Obesity (Silver Spring)* 2007;**15**:2357–64. <http://dx.doi.org/10.1038/oby.2007.279>
356. Elgar FJ, Roberts C, Tudor-Smith C, Moore L. Validity of self-reported height and weight and predictors of bias in adolescents. *J Adolesc Health* 2005;**37**:371–5. <http://dx.doi.org/10.1016/j.jadohealth.2004.07.014>
357. Brener ND, McManus T, Galuska DA, Lowry R, Wechsler H. Reliability and validity of self-reported height and weight among high school students. *J Adolesc Health* 2003;**32**:281–7. [http://dx.doi.org/10.1016/S1054-139X\(02\)00708-5](http://dx.doi.org/10.1016/S1054-139X(02)00708-5)
358. Bekkers MBM, Brunekreef B, Scholtens S, Kerkhof M, Smit HA, Wijga AH. Parental reported compared with measured waist circumference in 8-year-old children. *Int J Pediatr Obes* 2011;**6**:e78–86. <http://dx.doi.org/10.3109/17477166.2010.490266>
359. Watts K, Naylor LH, Davis EA, Jones TW, Beeson B, Bettenay F, et al. Do skinfolds accurately assess changes in body fat in obese children and adolescents? *Med Sci Sports Exerc* 2006;**38**:439–44. <http://dx.doi.org/10.1249/01.mss.0000191160.07893.2d>
360. Rowe DA, Dubose KD, Donnelly JE, Mahar MT. Agreement between skinfold-predicted percent fat and percent fat from whole-body bioelectrical impedance analysis in children and adolescents. *Int J Pediatr Obes* 2006;**1**:168–75. <http://dx.doi.org/10.1080/17477160600881296>
361. Rodriguez G, Moreno LA, Blay MG, Blay VA, Fleta J, Sarria A, et al. Body fat measurement in adolescents: comparison of skinfold thickness equations with dual-energy X-ray absorptiometry. *Eur J Clin Nutr* 2005;**59**:1158–66. <http://dx.doi.org/10.1038/sj.ejcn.1602226>
362. Morrison JA, Barton BA, Obarzanek E, Crawford PB, Guo SS, Schreiber GB, et al. Racial differences in the sums of skinfolds and percentage of body fat estimated from impedance in black and white girls, 9 to 19 years of age: the National Heart, Lung, and Blood Institute Growth and Health Study. *Obes Res* 2001;**9**:297–305. [Erratum published in *Obes Res* 2001;**9**:510.] <http://dx.doi.org/10.1038/oby.2001.37>
363. Jorga J, Marinkovic J, Kentric B, Hetherington M. Alternative methods of nutritional status assessment in adolescents. *Coll Antropol* 2007;**31**:413–18.
364. Hager ER, McGill AE, Black MM. Development and validation of a toddler silhouette scale. *Obesity (Silver Spring)* 2010;**18**:397–401. <http://dx.doi.org/10.1038/oby.2009.293>
365. Battistini N, Brambilla P, Virgili F, Simone P, Bedogni G, Morini P, et al. The prediction of total body water from body impedance in young obese subjects. *Int J Obes Relat Metab Disord* 1992;**16**:207–12.
366. Pineau J-C, Lally L, Bocquet M, Guihard-Costa A-M, Polak M, Frelut M-L, et al. Ultrasound measurement of total body fat in obese adolescents. *Ann Nutr Metab* 2010;**56**:36–44. <http://dx.doi.org/10.1159/000265849>
367. Garnett SP, Cowell CT, Baur LA, Shrewsbury VA, Chan A, Crawford D, et al. Increasing central adiposity: the Nepean longitudinal study of young people aged 7–8 to 12–13 years. *Int J Obes (Lond)* 2005;**29**:1353–60. <http://dx.doi.org/10.1038/sj.ijo.0803038>
368. Taylor RW, Jones IE, Williams SM, Goulding A. Evaluation of waist circumference, waist-to-hip ratio, and the conicity index as screening tools for high trunk fat mass, as measured by dual-energy X-ray absorptiometry, in children aged 3–19 years. *Am J Clin Nutr* 2000;**72**:490–5.
369. Weili Y, He B, Yao H, Dai J, Cui J, Ge D, et al. Waist-to-height ratio is an accurate and easier index for evaluating obesity in children and adolescents. *Obesity (Silver Spring)* 2007;**15**:748–52. <http://dx.doi.org/10.1038/oby.2007.601>

370. Hitze B, Bosy-Westphal A, Bielfeldt F, Settler U, Monig H, Muller MJ. Measurement of waist circumference at four different sites in children, adolescents, and young adults: concordance and correlation with nutritional status as well as cardiometabolic risk factors. *Obes Facts* 2008;**1**:243–9. <http://dx.doi.org/10.1159/000157248>
371. Reilly JJ, Dorosty AR, Ghomizadeh NM, Sheriff A, Wells JC, Ness AR. Comparison of waist circumference percentiles versus body mass index percentiles for diagnosis of obesity in a large cohort of children. *Int J Pediatr Obes* 2010;**5**:151–6. <http://dx.doi.org/10.3109/17477160903159440>
372. Mazicioglu MM, Hatipoglu N, Ozturk A, Cicek B, Ustunbas HB, Kurtoglu S. Waist circumference and mid-upper arm circumference in evaluation of obesity in children aged between 6 and 17 years. *J Clin Res Pediatr Endocrinol* 2010;**2**:144–50. <http://dx.doi.org/10.4274/jcrpe.v2i4.144>
373. Candido AP, Freitas SN, Machado-Coelho GL. Anthropometric measurements and obesity diagnosis in schoolchildren. *Acta Paediatr* 2011;**100**:e120–4. <http://dx.doi.org/10.1111/j.1651-2227.2011.02296.x>
374. Stettler N, Zomorodi A, Posner JC. Predictive value of weight-for-age to identify overweight children. *Obesity (Silver Spring)* 2007;**15**:3106–12. <http://dx.doi.org/10.1038/oby.2007.370>
375. Himes JH, Bouchard C. Validity of anthropometry in classifying youths as obese. *Int J Obes (Lond)* 1989;**13**:183–93.
376. Zheng XF, Tang QY, Tao YX, Lu W, Cai W. Clinical value of methods for analyzing the abdominal fat levels of obese children and adolescents. *Obes Metab* 2010;**6**:105–10.
377. Yamborisut U, Kijboonchoo K, Wimonpeerapattana W, Srichan W, Thasanasuwan W. Study on different sites of waist circumference and its relationship to weight-for-height index in Thai adolescents. *J Med Assoc Thai* 2008;**91**:1276–84.
378. Campanozzi A, Dabbas M, Ruiz JC, Ricour C, Goulet O. Evaluation of lean body mass in obese children. *Eur J Pediatr* 2008;**167**:533–40. <http://dx.doi.org/10.1007/s00431-007-0546-4>
379. Goldfield GS, Cloutier P, Mallory R, Prud'homme D, Parker T, Doucet E. Validity of foot-to-foot bioelectrical impedance analysis in overweight and obese children and parents. *J Sports Med Phys Fitness* 2006;**46**:447–53.
380. Guntsche Z, Guntsche EM, Saravi FD, Gonzalez LM, Lopez Avellaneda C, Ayub E, et al. Umbilical waist-to-height ratio and trunk fat mass index (DXA) as markers of central adiposity and insulin resistance in Argentinean children with a family history of metabolic syndrome. *J Pediatr Endocrinol* 2010;**23**:245–56. <http://dx.doi.org/10.1515/JPEM.2010.23.3.245>
381. Hatipoglu N, Mazicioglu MM, Kurtoglu S, Kendirci M. Neck circumference: an additional tool of screening overweight and obesity in childhood. *Eur J Pediatr* 2010;**169**:733–9. <http://dx.doi.org/10.1007/s00431-009-1104-z>
382. Johnston FE. Validity of triceps skinfold and relative weight as measures of adolescent obesity. *J Adolesc Health Care* 1985;**6**:185–90. [http://dx.doi.org/10.1016/S0197-0070\(85\)80015-2](http://dx.doi.org/10.1016/S0197-0070(85)80015-2)
383. Kurth B-M, Ellert U. Estimated and measured BMI and self-perceived body image of adolescents in Germany: part 1 – general implications for correcting prevalence estimations of overweight and obesity. *Obes Facts* 2010;**3**:181–90. <http://dx.doi.org/10.1159/000314638>
384. Lewy VD, Danadian K, Arslanian S. Determination of body composition in African-American children: validation of bioelectrical impedance with dual energy X-ray absorptiometry. *J Pediatr Endocrinol* 1999;**12**:443–8. <http://dx.doi.org/10.1515/JPEM.1999.12.3.443>

385. Moore WE, Yeh J, Knehans AW, Eichner JE, Lee ET. Intermethod agreement and body fat estimates using skinfolds and a footpad-style bioelectrical impedance device. *Meas Phys Educ Exerc Sci* 1999;**3**:51–62. http://dx.doi.org/10.1207/s15327841mpee0301_4
386. Owens S, Litaker M, Allison J, Riggs S, Ferguson M, Gutin B. Prediction of visceral adipose tissue from simple anthropometric measurements in youths with obesity. *Obes Res* 1999;**7**:16–22. <http://dx.doi.org/10.1002/j.1550-8528.1999.tb00386.x>
387. Tsang TW, Briody J, Kohn M, Chow CM, Singh MF. Abdominal fat assessment in adolescents using dual-energy X-ray absorptiometry. *J Pediatr Endocrinol* 2009;**22**:781–94. <http://dx.doi.org/10.1515/JPEM.2009.22.9.781>
388. Williams J, Wake M, Campbell M. Comparing estimates of body fat in children using published bioelectrical impedance analysis equations. *Int J Pediatr Obes* 2007;**2**:174–9. <http://dx.doi.org/10.1080/17477160701408783>
389. Malina RM, Zavaleta AN, Little BB. Estimated overweight and obesity in Mexican American school children. *Int J Obes (Lond)* 1986;**10**:483–91.
390. Brambilla P, Manzoni P, Sironi S, Simone P, Del Maschio A, di Natale B, et al. Peripheral and abdominal adiposity in childhood obesity. *Int J Obes Relat Metab Disord* 1994;**18**:795–800.
391. Pecoraro P, Guida B, Caroli M, Trio R, Falconi C, Principato S, et al. Body mass index and skinfold thickness versus bioimpedance analysis: fat mass prediction in children. *Acta Diabetol* 2003;**40**(Suppl. 1):S278–81. <http://dx.doi.org/10.1007/s00592-003-0086-y>
392. Taylor RW, Williams SM, Grant AM, Ferguson E, Taylor BJ, Goulding A. Waist circumference as a measure of trunk fat mass in children aged 3 to 5 years. *Int J Pediatr Obes* 2008;**3**:226–33. <http://dx.doi.org/10.1080/17477160802030429>
393. Freedman DS, Ogden CL, Berenson GS, Horlick M. Body mass index and body fatness in childhood. *Curr Opin Clin Nutr Metab Care* 2005;**8**:618–23. <http://dx.doi.org/10.1097/01.mco.0000171128.21655.93>
394. Freedman DS, Sherry B. The validity of BMI as an indicator of body fatness and risk among children. *Pediatrics* 2009;**124**(Suppl. 1):23–34. <http://dx.doi.org/10.1542/peds.2008-3586E>
395. Kayhan G, Ersoz G. [Comparison of the different methods of measurement used in the detection of body fat rate and diagnosis of obesity in adolescents aged from 15 up to 18.] *Turk Klin Spor Bilim* 2009;**1**:107–16.
396. Majcher A, Pyrzak B, Czerwonogrodzka A, Kucharska A. Body fat percentage and anthropometric parameters in children with obesity. *Med Wieku Rozwoj* 2008;**12**:493–8.
397. Zambon MP, Zanolli MdL, Marmo DB, Magna LA, Guimarey LM, Morcillo AM. Body mass index and triceps skinfold correlation in children from Paulinia city, Sao Paulo, SP. *Rev Assoc Med Bras* 2003;**49**:137–40. <http://dx.doi.org/10.1590/S0104-42302003000200029>
398. Zaragozano JF, Frenne LMd, Aznar LM, Sanchez MB. Anthropometric criteria used in the assessment of obesity in childhood. *Rev Esp Pediatr* 1998;**54**:407–13.
399. Behbahani BH, Dorosty AR, Eshraghian MR. Assessment of obesity in children: fat mass index versus body mass index. *Tehran Univ Med J* 2009;**67**:408–14.
400. Chiara V, Sichieri R, Martins PcD. Sensitivity and specificity of overweight classification of adolescents, Brazil. *Rev Saude Publica* 2003;**37**:226–31.
401. da Silva KS, Lopes AD, da Silva FM. Sensitivity and specificity of different classification criteria for excess weight in schoolchildren from Joao Pessoa, Paraiba, Brazil. *Rev Nutr* 2010;**23**:27–35.

402. Giugliano R, Melo ALP. Diagnosis of overweight and obesity in schoolchildren: utilization of the body mass index international standard. *J Pediatr (Rio J)* 2004;**80**:129–34. <http://dx.doi.org/10.2223/JPED.1152>
403. Jakubowska-Pietkiewicz E, Prochowska A, Fendler W, Szadkowska A. Comparison of body fat measurement methods in children. *Pediatr Endocrinol Diabetes Metab* 2009;**15**:246–50.
404. Perez BM, Landaeta-Jimenez M, Amador J, Vasquez M, Marrodan MD. Sensitivity and specificity of anthropometric indicators of adiposity and fat distribution in Venezuelan children and adolescents. *Interciencia* 2009;**34**:84–90.
405. Rodriguez DP, Bermudez EF, Rodriguez GS, Spina MA, Zeni SN, Friedman SM, *et al.* Body composition by simple anthropometry, bioimpedance and DXA in preschool children: inter-relationships among methods. *Arch Argent Pediatr* 2008;**106**:102–9. <http://dx.doi.org/10.1590/S0325-007520080002000003>
406. Schonhaut BL, Rodriguez OL, Pizarro QT, Kohn BJ, Merino LD, Lopez OA, *et al.* [Concordance in nutritional diagnosis between the healthcare and school teachers teams, using the body mass index (BMI) in the borough of Colina.] *Revista Chil Pediatr* 2004;**75**:32–5.
407. Stein D, Koch S, Ingrisch S, Bauer CP, Ulm K, Schuster T. [Child and adolescent obesity. Long term results at weight loss programs, child obesity, BMI, BMI-SDS, SDS-difference, weight %.] *Pediatr Prax* 2006;**68**:293–302.
408. Zhang Q, Du WJ, Hu XQ, Liu AL, Pan H, Ma GS. The relation between body mass index and percentage body fat among Chinese adolescent living in urban Beijing. *Zhonghua liuxingbingxue zazhi* 2004;**25**:113–16.
409. Rockett HR, Colditz JA. Assessing diets of children and adolescents. *AJCN* 1997;**65**:S1116–22.
410. Bratteby LE, Sandhagen B, Fan H, Enghardt H, Samuelson G. Total energy expenditure and physical activity as assessed by the doubly labeled water method in Swedish adolescents in whom energy intake was underestimated by 7-d diet records. *Am J Clin Nutr* 1998;**67**:905–11.
411. Bryant-Waugh RJ, Cooper PJ, Taylor CL, Lask BD. The use of the eating disorder examination with children: a pilot study. *Int J Eat Disord* 1996;**19**:391–7. [http://dx.doi.org/10.1002/\(SICI\)1098-108X\(199605\)19:4<391::AID-EAT6>3.0.CO;2-G](http://dx.doi.org/10.1002/(SICI)1098-108X(199605)19:4<391::AID-EAT6>3.0.CO;2-G)
412. Goossens L, Braet C. Screening for eating pathology in the pediatric field. *Int J Pediatr Obes* 2010;**5**:483–90. <http://dx.doi.org/10.3109/17477160903571995>
413. Tanofsky-Kraff M, Yanovski SZ, Yanovski JA. Comparison of child interview and parent reports of children's eating disordered behaviors. *Eat Behav* 2005;**6**:95–9. <http://dx.doi.org/10.1016/j.eatbeh.2004.03.001>
414. Wells JE, Coope PA, Gabb DC, Pears RK. The factor structure of the Eating Attitudes Test with adolescent schoolgirls. *Psychol Med* 1985;**15**:141–6. <http://dx.doi.org/10.1017/S0033291700021000>
415. Birch LL, Davison KK. Family environmental factors influencing the developing behavioral controls of food intake and childhood overweight. *Pediatr Clin North Am* 2001;**48**:893–907. [http://dx.doi.org/10.1016/S0031-3955\(05\)70347-3](http://dx.doi.org/10.1016/S0031-3955(05)70347-3)
416. Backlund CS, Larsson C. Validity of armband measuring energy expenditure in overweight and obese children. *Med Sci Sports Exerc* 2010;**42**:1154–61. <http://dx.doi.org/10.1249/MSS.0b013e3181c84091>
417. Pate RR, Dowda M, Trost S, Sirard JR. Validation of a three-day physical activity recall instrument in female youth. *Pediatr Exerc Sci* 2003;**15**:257–65.

418. Trost S, Ward D, McGraw B, Pate R. Validity of the Previous Day Physical Activity Recall (PDPAR) in fifth-grade children. *Pediatr Exerc Sci* 1999;**11**:341–8.
419. McMurray RG, Ward DS, Elder JP, Lytle LA, Strikmiller PK, Baggett CD, *et al*. Do overweight girls overreport physical activity? *Am J Health Behav* 2008;**32**:538–46. <http://dx.doi.org/10.5993/AJHB.32.5.9>
420. Russoniello CV, Pougatchev V, Zhirnov E, Mahar MT. A measurement of electrocardiography and photoplethysmography in obese children. *Appl Psychophysiol Biofeedback* 2010;**35**:257–9. <http://dx.doi.org/10.1007/s10484-010-9136-8>
421. Riva G, Molinari E. Replicated factor analysis of the Italian Version of the Body Image Avoidance Questionnaire. *Percept Mot Skills* 1998;**86**:1071–4. <http://dx.doi.org/10.2466/pms.1998.86.3.1071>
422. Asayama K, Dobashi K, Hayashibe H, Kodera K, Uchida N, Nakane T, *et al*. Threshold values of visceral fat measures and their anthropometric alternatives for metabolic derangement in Japanese obese boys. *Int J Obes Relat Metab Disord* 2002;**26**:208–13. <http://dx.doi.org/10.1038/sj.ijo.0801865>

Appendix 1 Search 1 search strategy

**Database: Ovid MEDLINE(R) 1948 to August Week 2 2011
(modified and repeated in 10 other databases; available
on request)**

#	Searches	Results
	clinical trial/ or clinical trial, phase i/ or clinical trial, phase ii/ or clinical trial, phase iii/ or clinical trial, phase iv/ or controlled clinical trial/ or multicenter study/ or randomized controlled trial/	653,759
	exp Clinical Trials as Topic/	247,291
	Evaluation studies/	155,147
	Meta-analysis/	30,113
	Validation studies/	51,814
	research design/ or cross-over studies/ or double-blind method/ or matched-pair analysis/ or random allocation/ or "reproducibility of results"/ or sample size/ or exp "sensitivity and specificity"/ or single-blind method/ or Early Termination of Clinical Trials/ or control groups/	743,384
	(pre post or pre test or post test or non-randomi?ed or quasi experiment).tw.	11,816
	Feasibility studies/	33,415
	Intervention studies/	4941
	Pilot projects/	67,278
	placebo*.tw.	131,751
	(random* adj3 (study or studies or trial or trials)).tw.	190,936
	(random* adj3 (allocation or assign* or allocate*)).tw.	72,994
	(study adj (pilot or feasibility or evaluation or validation)).tw.	571
	(studies adj (pilot or feasibility or evaluation or validation)).tw.	177
	((blind* or mask*) adj2 (singl* or doubl* or trebl* or tripl*)).tw.	109,924
	(matched adj (communities or schools or populations)).tw.	141
	(control adj group*).tw.	219,781
	((trial or trials) adj2 (clinical or controlled)).tw.	236,788
	("outcome study" or "outcome studies" or quasiexperimental or "quasi experimental" or quasi-experimental or "pseudo experimental").tw.	8374
	(meta-analysis or crossover* or "cross over*" or cross-over*).tw.	77,609
	((cluster or factorial) adj2 trial*).tw.	1240
	or/1-22	1,851,224
	((child* or adolescen* or teen or teens or teenager* or youth or youths or girl or girls or boy or boys or p?ediatric* or juvenil*) adj4 (obesity or obese or adiposity)).tw.	11,629
	((child* or adolescen* or teen or teens or teenager* or youth or youths or girl or girls or boy or boys or p?ediatric* or juvenil*) adj4 (overweight or overeat* or "over weight" or "over eat*")).tw.	4580
	((child* or adolescen* or teen or teens or teenager* or youth or youths or girl or girls or boy or boys or p?ediatric* or juvenil*) adj4 ((weight or bmi or "body mass index") adj2 (gain* or change* or increas* or loss))).tw.	1865
	((infant or infants or "young people" or "young person" or "young adult" or " young men" or "young women" or "schoolchild*") adj4 (obesity or obese or adiposity)).tw.	716

#	Searches	Results
	((infant or infants or "young people" or "young person" or "young adult" or " young men" or "young women" or "schoolchild*") adj4 (overweight or overeate* or "over weight" or "over eat*")).tw.	275
	((infant or infants or "young people" or "young person" or "young adult" or " young men" or "young women" or "schoolchild*") adj4 ((weight or bmi or "body mass index") adj2 (gain* or change* or increas* or loss))).tw.	935
	or/24-29	16,320
	Weight Gain/	18,875
	weight loss/	19,028
	Body Weight Changes/	4
	Ideal Body Weight/	41
	Adiposity/	2784
	Overweight/	6841
	obesity/ or obesity hypoventilation syndrome/ or obesity, abdominal/ or obesity, morbid/ or prader-willi syndrome/	112,599
	Adolescent behavior/	16,695
	exp Child behavior/	12,609
	adolescent/	1,436,947
	child/	1,235,275
	child, preschool/	682,135
	infant/	578,637
	38 or 39 or 40 or 41 or 42 or 43	2,333,340
	31 or 32 or 33 or 34 or 35 or 36 or 37	141,780
	44 and 45	32,263
	30 or 46	36,196
	23 and 47	6705
	addresses/ or lectures/ or anecdotes/ or biography/ or interview/ or comment/ or directory/ or editorial/ or legal cases/ or case reports/ or legislation/ or letter/ or news/ or newspaper article/ or patient education handout/	2,804,108
	48 not 49	6519

Appendix 2 Search 2 search strategy

Database(s): Ovid MEDLINE(R) 1948 to August Week 2 2011 (modified and repeated in 10 other databases; available on request)

#	Searches	Results
	((child* or adolescen* or teen or teens or teenager* or youth or youths or girl or girls or boy or boys or p?ediatic* or juvenil*) adj4 (obesity or obese or adiposity)).tw.	11,629
	((child* or adolescen* or teen or teens or teenager* or youth or youths or girl or girls or boy or boys or p?ediatic* or juvenil*) adj4 (overweight or overeat* or "over weight" or "over eat*")).tw.	4580
	((child* or adolescen* or teen or teens or teenager* or youth or youths or girl or girls or boy or boys or p?ediatic* or juvenil*) adj4 ((weight or bmi or "body mass index") adj2 (gain* or change* or increas* or loss))).tw.	1865
	((infant or infants or "young people" or "young person" or "young adult" or " young men" or "young women" or "schoolchild*") adj4 (obesity or obese or adiposity)).tw.	716
	((infant or infants or "young people" or "young person" or "young adult" or " young men" or "young women" or "schoolchild*") adj4 (overweight or overeat* or "over weight" or "over eat*")).tw.	275
	((infant or infants or "young people" or "young person" or "young adult" or " young men" or "young women" or "schoolchild*") adj4 ((weight or bmi or "body mass index") adj2 (gain* or change* or increas* or loss))).tw.	935
	or/1-6	16,320
	obesity/	102,151
	obesity hypoventilation syndrome/	565
	obesity, abdominal/	545
	obesity, morbid/	8517
	prader-willi syndrome/	2048
	Weight Gain/	18,875
	weight loss/	19,028
	body weight changes/	4
	Ideal Body Weight/	41
	adiposity/	2784
	Overweight/	6841
	or/8-18	141,780
	Adolescent behavior/	16,695
	exp Child behavior/	12,609
	adolescent/	1,436,947
	child/	1,235,275
	child, preschool/	682,135
	infant/	578,637
	or/20-25	2,333,340
	19 and 26	32,263
	7 or 27	36,196
	exp validation studies/	51,814

#	Searches	Results
	exp reproducibility of results/ reproducib*.tw.	219,955 88,002
	exp psychometrics/ psychometr*.tw.	45,677 19,152
	clin#metr*.tw.	372
	observer variation/ "observer variation".tw.	26,493 740
	discriminant analysis/ reliab*.tw.	6053 235,792
	valid*.tw.	274,556
	coefficient.tw.	92,800
	"internal consistency".tw.	11,083
	((cronbach* or cronback*) adj5 (alpha or alphas)).tw.	6847
	"item correlation?".tw.	253
	"item selection?".tw.	239
	"item reduction?".tw.	253
	agreement.tw.	123,618
	precision.tw.	50,812
	imprecision.tw.	3056
	"precise values".tw.	112
	(test adj2 retest).tw.	11,223
	(reliab* adj2 (test or retest)).tw.	11,612
	stability.tw.	172,904
	(intrarater or "intra rater").tw.	1438
	(interrater or "inter rater" or interator).tw.	7185
	(intertester or "inter tester").tw.	275
	(intratester or "intra tester").tw.	217
	(interobserver or "inter observer").tw.	11,243
	(intraobserver or "intraobserver").tw.	3641
	(intertechnician or "inter technician").tw.	16
	(inratechnician or "intra technician").tw.	5
	(interexaminer or "inter examiner").tw.	889
	(intraexaminer or "intra examiner").tw.	549
	(interassay or "inter assay").tw.	5086
	(intraassay or "intra assay").tw.	3259
	(interindividual or "inter individual").tw.	14,867
	(intraindividual or "intra individual").tw.	6112
	(interparticipant or "inter participant").tw.	27
	(intraparticipant or "intra participant").tw.	21

#	Searches	Results
	kappa?.tw.	75,369
	"coefficient of variation".tw.	13,427
	repeatable*.tw.	13,573
	(replicable* adj2 (measure? or findings or result? or test?)).tw.	128
	(repeated adj2 (measure? or findings or result? or test?)).tw.	20,667
	generalisability*.tw.	18,375
	concordance.tw.	19,838
	(intraclass adj5 correlation*).tw.	8176
	discriminative.tw.	8148
	"known group".tw.	314
	"factor analysis".tw.	19,413
	"factor structure?".tw.	4819
	dimensionality.tw.	3020
	subscale*.tw.	17,508
	"multitrait scaling analysis".tw.	63
	"item discriminant".tw.	63
	"interscale correlation?".tw.	64
	(error? adj3 (measure* or correlat* or evaluat* or accuracy or accurate or precision or mean)).tw.	22,281
	(variability adj (individual or interval or rate analysis)).tw.	23
	(uncertainty adj3 (measurement or measuring)).tw.	657
	"standard error of measurement".tw.	492
	sensitivity*.tw.	780,838
	responsiveness*.tw.	141,921
	(limit adj3 detection).tw.	30,895
	"minimal detectable concentration".tw.	68
	interpretable*.tw.	3824
	(small* adj5 ((real or detectable) adj3 (change* or difference))).tw.	247
	"meaningful change".tw.	320
	"minimal* important change".tw.	38
	"minimal* important difference".tw.	202
	"minimal* detectable change".tw.	152
	"minimal* detectable difference".tw.	19
	"minimal* real change".tw.	0
	"minimal* real difference".tw.	0
	"ceiling effect".tw.	700
	"floor effect".tw.	187
	"item response model".tw.	48
	"item response theory".tw.	803
	(irt adj3 model*).tw.	146

#	Searches	Results
	rasch.tw.	1387
	"differen* item function* ".tw.	460
	"computer* adaptive test* ".tw.	236
	"item bank ".tw.	132
	"cross cultural equivalence ".tw.	65
	or/29-112	1,969,251
	"conceptual framework ".tw.	5202
	Concept Formation/	8423
	conceptuali#ation.tw.	4355
	operationali#ation.tw.	658
	"construct development ".tw.	38
	"pre testing ".tw.	237
	"cognitive interview* ".tw.	231
	"patient interview* ".tw.	1529
	Consensus/	3480
	"item pooling ".tw.	2
	"content development ".tw.	62
	"cognitive theory ".tw.	883
	"cognitive debrief* ".tw.	87
	tourangeau.tw.	6
	"survey development? ".tw.	74
	interviews as topic/	32,428
	or/114-129	56,595
	113 or 130	2,015,462
	(measure* or test or tests or scale or scales or rate or rates or rating*).tw.	3,753,606
	(inventory or inventories or score* or index or indexes or instrument or instruments or tool or tools or questionnaire* or survey*).tw.	1,366,656
	"Outcome Assessment (Health Care)"/	39,977
	exp Health Status Indicators/	158,577
	Questionnaires/	241,283
	or/132-136	4,559,127
	28 and 131 and 137	3741
	addresses/ or lectures/ or anecdotes/ or biography/ or comment/ or directory/ or editorial/ or legal cases/ or case reports/ or legislation/ or letter/ or news/ or newspaper article/ or patient education handout/	2,784,229
	138 not 139	3707

Appendix 3 Search 1 references (included childhood obesity treatment trials)

The following list of references includes eligible search 1 trials, from which citations of outcome measures used were obtained.

1. Adamo KB, Rutherford JA, Goldfield GS. Effects of interactive video game cycling on overweight and obese adolescent health. *Appl Physiol Nutr Metab* 2010;**35**:805–15.
2. Albala C, Ebbeling CB, Cifuentes M, Lera L, Bustos N, Ludwig DS. Effects of replacing the habitual consumption of sugar-sweetened beverages with milk in Chilean children. *Am J Clin Nutr* 2008;**88**:605–11.
3. Andelman MB, Jones C, Nathan S. Treatment of obesity in underprivileged adolescents. Comparison of diethylpropion hydrochloride with placebo in a double-blind study. *Clin Pediatr (Phila)* 1967;**6**:327–30.
4. Aragona J, Cassady J, Drabman RS. Treating overweight children through parental training and contingency contracting. *J Appl Behav Anal* 1975;**8**:269–78.
5. Atabek ME, Pirgon O. Use of metformin in obese adolescents with hyperinsulinemia: a 6-month, randomized, double-blind, placebo-controlled clinical trial. *J Pediatr Endocrinol* 2008;**21**:339–48.
6. Bacon GE, Lowrey GH. A clinical trial of fenfluramine in obese children. *Curr Ther Res Clin Exp* 1967;**9**:626–30.
7. Barkin SL, Gesell SB, Poe EK, Ip EH. Changing overweight Latino preadolescent body mass index: the effect of the parent–child dyad. *Clin Pediatr (Phila)* 2011;**50**:29–36.
8. Bathrellou E, Yannakoulia M, Papanikolaou K, Pehlivanidis A, Pervanidou P, Kanaka-Gantenbein C, *et al*. Parental involvement does not augment the effectiveness of an intense behavioral program for the treatment of childhood obesity. *Hormones* 2010;**9**:171–5.
9. Bauer S, de Niet J, Timman R, Kordy H. Enhancement of care through self-monitoring and tailored feedback via text messaging and their use in the treatment of childhood overweight. *Patient Educ Couns* 2010;**79**:315–19.
10. Bean MK, Mazzeo SE, Stern M, Bowen D, Ingersoll K. A values-based Motivational Interviewing (MI) intervention for pediatric obesity: study design and methods for MI values. *Contemp Clin Trials* 2011;**32**:667–74.
11. Berkowitz RI, Fujioka K, Daniels SR, Hoppin AG, Owen S, Perry AC, *et al*. Effects of sibutramine treatment in obese adolescents: a randomized trial. *Ann Intern Med* 2006;**145**:81–90.
12. Berkowitz RI, Wadden TA, Gehrman CA, Bishop-Gilyard CT, Moore RH, Womble LG, *et al*. Meal replacements in the treatment of adolescent obesity: a randomized controlled trial. *Obesity (Silver Spring)* 2011;**19**:1193–9.
13. Berkowitz RI, Wadden TA, Tershakovec AM, Cronquist JL. Behavior therapy and sibutramine for the treatment of adolescent obesity: a randomized controlled trial. *JAMA* 2003;**289**:1805–12.

14. Berry D, Savoye M, Melkus G, Grey M. An intervention for multiethnic obese parents and overweight children. *Appl Nurs Res* 2007;**20**:63–71.
15. Boutelle KN, Cafri G, Crow SJ. Parent-only treatment for childhood obesity: A randomized controlled trial. *Obesity (Silver Spring)* 2011;**19**:574–80.
16. Bravender T, Russell A, Chung RJ, Armstrong SC. A 'novel' intervention: a pilot study of children's literature and healthy lifestyles. *Pediatrics* 2010;**125**:e513–17.
17. Brownell KD, Kelman JH, Stunkard AJ. Treatment of obese children with and without their mothers: changes in weight and blood pressure. *Pediatrics* 1983;**71**:515–23.
18. Burgert TS, Duran EJ, Goldberg-Gell R, Dziura J, Yeckel CW, Katz S, *et al.* Short-term metabolic and cardiovascular effects of metformin in markedly obese adolescents with normal glucose tolerance. *Pediatr Diabetes* 2008;**9**:567–76.
19. Carrel AL, Clark RR, Peterson SE, Nemeth BA, Sullivan J, Allen DB. Improvement of fitness, body composition, and insulin sensitivity in overweight children in a school-based exercise program: a randomized, controlled study. *Arch Pediatr Adolesc Med* 2005;**159**:963–8.
20. Chandra RK. Obesity in childhood: a clinical trial of low-calorie 'limical'. *Indian J Pediatr* 1968;**35**:23–6.
21. Chang C, Liu W, Zhao X, Li S, Yu C. Effect of supervised exercise intervention on metabolic risk factors and physical fitness in Chinese obese children in early puberty. *Obes Rev* 2008;**9**(Suppl. 1):135–41.
22. Chanoine JP, Hampl S, Jensen C, Boldrin M, Hauptman J. Effect of orlistat on weight and body composition in obese adolescents: a randomized controlled trial. *JAMA* 2005;**293**:2873–83.
23. Clarson CL, Mahmud FH, Baker JE, Clark HE, McKay WM, Schauteet VD, *et al.* Metformin in combination with structured lifestyle intervention improved body mass index in obese adolescents, but did not improve insulin resistance. *Endocrine* 2009;**36**:141–6.
24. Coates TJ, Jeffery RW, Slinkard LA, Killen J, Danaher BG. Frequency of contact and monetary reward in weight loss, lipid change, and blood pressure reduction with adolescents. *Behav Ther* 1982;**13**:175–85. URL: www.mrw.interscience.wiley.com/cochrane/clcentral/articles/912/CN-00183912/frame.html.
25. Coppins DF, Margetts BM, Fa JL, Brown M, Garrett F, Huelin S. Effectiveness of a multi-disciplinary family-based programme for treating childhood obesity (The Family Project). *Eur J Clin Nutr* 2011;**65**:903–9.
26. Daniels SR, Long B, Crow S, Styne D, Sothorn M, Vargas-Rodriguez I, *et al.* Cardiovascular effects of sibutramine in the treatment of obese adolescents: results of a randomized, double-blind, placebo-controlled study. *Pediatrics* 2007;**120**:e147–57.
27. Danielsson P, Janson A, Norgren S, Marcus C. Impact sibutramine therapy in children with hypothalamic obesity or obesity with aggravating syndromes. *J Clin Endocrinol Metab* 2007;**92**:4101–6.
28. Davis AM, James RL, Boles RE, Goetz JR, Belmont J, Malone B. The use of TeleMedicine in the treatment of paediatric obesity: feasibility and acceptability. *Matern Child Nutr* 2011;**7**:71–9.

29. Davis JN, Tung A, Chak SS, Ventura EE, Byrd-Williams CE, Alexander KE, *et al.* Aerobic and strength training reduces adiposity in overweight Latina adolescents. *Med Sci Sports Exerc* 2009;**41**:1494–503.
30. Demol S, Yackobovitch-Gavan M, Shalitin S, Nagelberg N, Gillon-Keren M, Phillip M. Low-carbohydrate (low and high-fat) versus high-carbohydrate low-fat diets in the treatment of obesity in adolescents. *Acta Paediatr* 2009;**98**:346–51.
31. Diaz RG, Esparza-Romero J, Moya-Camarena SY, Robles-Sardin AE, Valencia ME. Lifestyle intervention in primary care settings improves obesity parameters among Mexican youth. *J Am Diet Assoc* 2010;**110**:285–90.
32. Doyle AC, Goldschmidt A, Huang C, Winzelberg AJ, Taylor CB, Wilfley DE. Reduction of overweight and eating disorder symptoms via the Internet in adolescents: a randomized controlled trial. *J Adolesc Health* 2008;**43**:172–9.
33. Duckworth LC, Gately PJ, Radley D, Cooke CB, King RF, Hill AJ. RCT of a high-protein diet on hunger motivation and weight-loss in obese children: an extension and replication. *Obesity (Silver Spring)* 2009;**17**:1808–10. URL: www.mrw.interscience.wiley.com/cochrane/clcentral/articles/987/CN-00718987/frame.html
34. Duffy G, Spence SH. The effectiveness of cognitive self-management as an adjunct to a behavioural intervention for childhood obesity: a research note. *J Child Psychol Psychiatry* 1993;**34**:1043–50.
35. Duggins M, Cherven P, Carrithers J, Messamore J, Harvey A. Impact of family YMCA membership on childhood obesity: a randomized controlled effectiveness trial. *J Am Board Fam Med* 2010;**23**:323–33.
36. Dunshea-Mooij C, Wall C, King C. 'Games Galore'; a feasibility study to investigate the effect of a physical activity and a nutrition education programme for 10–14 year old New Zealand overweight and obese children. *Proc Nutr Soc New Zeal* 2003;**28**:71–4.
37. Ebbeling CB, Leidig MM, Sinclair KB, Hangen JP, Ludwig DS. A reduced-glycemic load diet in the treatment of adolescent obesity. *Arch Pediatr Adolesc Med* 2003;**157**:773–9.
38. Edwards C, Nicholls D, Croker H, Van Zyl S, Viner R, Wardle J. Family-based behavioural treatment of obesity: acceptability and effectiveness in the UK. *Eur J Clin Nutr* 2006;**60**:587–92.
39. Elloumi M, Makni E, Ounis OB, Zbidi A, Lac G, Tabka Z. Six-minute walking test to assess exercise tolerance in Tunisian obese adolescents over two-months individualized program training. *Sci Sports* 2007;**22**:289–92.
40. Elmahgoub SM, Lambers S, Stegen S, Van Laethem C, Cambier D, Calders P. The influence of combined exercise training on indices of obesity, physical fitness and lipid profile in overweight and obese adolescents with mental retardation. *Eur J Pediatr* 2009;**168**:1327–33.
41. Epstein LH, Paluch R, Kilanowski CK, Raynor HA. Effects of family-based behavioral treatment on obese 5-to-8-year-old children. *Behav Ther* 1985;**16**:205–12.
42. Epstein LH, McKenzie SJ, Valoski A, Klein KR, Wing RR. Effects of mastery criteria and contingent reinforcement for family-based child weight control. *Addict Behav* 1994;**19**:135–45.

43. Epstein LH, Nudelman S, Wing RR. Long-term effects of family-based treatment for obesity on non-treated family members. *Behav Ther* 1987;**18**:147–52.
44. Epstein LH, Paluch RA, Beecher MD, Roemmich JN. Increasing healthy eating vs. reducing high energy-dense foods to treat pediatric obesity. *Obesity (Silver Spring)* 2008;**16**:318–26.
45. Epstein LH, Paluch RA, Gordy CC, Dorn J. Decreasing sedentary behaviors in treating pediatric obesity. *Arch Pediatr Adolesc Med* 2000;**154**:220–6.
46. Epstein LH, Paluch RA, Gordy CC, Saelens BE, Ernst MM. Problem solving in the treatment of childhood obesity. *J Consult Clin Psychol* 2000;**68**:717–21.
47. Epstein LH, Paluch RA, Kilanowski CK, Raynor HA. The effect of reinforcement or stimulus control to reduce sedentary behavior in the treatment of pediatric obesity. *Health Psychol* 2004;**23**:371–80. URL: www.mrw.interscience.wiley.com/cochrane/clcentral/articles/322/CN-00490322/frame.html.
48. Epstein LH, Paluch RA, Raynor HA. Sex differences in obese children and siblings in family-based obesity treatment. *Obes Res* 2001;**9**:746–53.
49. Epstein LH, Roemmich JN, Stein RI, Paluch RA, Kilanowski CK. The challenge of identifying behavioral alternatives to food: clinic and field studies. *Ann Behav Med* 2005;**30**:201–9.
50. Epstein LH, Valoski AM, Vara LS, McCurley J, Wisniewski L, Kalarchian MA, et al. Effects of decreasing sedentary behavior and increasing activity on weight change in obese children. *Health Psychol* 1995;**14**:109–15.
51. Epstein LH, Wing RR, Koeske R, Andrasik F, Ossip DJ. Child and parent weight loss in family-based behavior modification programs. *J Consult Clin Psychol* 1981;**49**:674–85.
52. Epstein LH, Wing RR, Koeske R, Valoski A. Effects of diet plus exercise on weight change in parents and children. *J Consult Clin Psychol* 1984;**52**:429–37. URL: www.mrw.interscience.wiley.com/cochrane/clcentral/articles/924/CN-00193924/frame.html
53. Epstein LH, Wing RR, Penner BC, Kress MJ. Effect of diet and controlled exercise on weight loss in obese children. *J Pediatr* 1985;**107**:358–61.
54. Estabrooks PA, Shoup JA, Gattshall M, Dandamudi P, Shetterly S, Xu S. Automated telephone counseling for parents of overweight children: a randomized controlled trial. *Am J Prev Med* 2009;**36**:35–42.
55. Figueroa-Colon R, Franklin FA, Lee JY, von Almen TK, Suskind RM. Feasibility of a clinic-based hypocaloric dietary intervention implemented in a school setting for obese children. *Obes Res* 1996;**4**:419–29.
56. Figueroa-Colon R, von Almen TK, Franklin FA, Schuftan C, Suskind RM. Comparison of two hypocaloric diets in obese children. *Am J Dis Child* 1993;**147**:160–6.
57. Flodmark CE, Ohlsson T, Ryden O, Sveger T. Prevention of progression to severe obesity in a group of obese schoolchildren treated with family therapy. *Pediatrics* 1993;**91**:880–4.

58. Ford AL, Bergh C, Södersten P, Sabin MA, Hollinghurst S, Hunt LP, *et al.* Treatment of childhood obesity by retraining eating behaviour: randomised controlled trial [published online ahead of print]. *BMJ* 2010;**340**:b5388. URL: www.mrw.interscience.wiley.com/cochrane/clcentral/articles/155/CN-00735155/frame.html (accessed 26 June 2014).
59. Freemark M, Bursley D. The effects of metformin on body mass index and glucose tolerance in obese adolescents with fasting hyperinsulinemia and a family history of type 2 diabetes. *Pediatrics* 2001;**107**:E55.
60. Garcia-Morales LM, Berber A, Macias-Lara CC, Lucio-Ortiz C, Del-Rio-Navarro BE, Dorantes-Alvarez LM. Use of sibutramine in obese mexican adolescents: a 6-month, randomized, double-blind, placebo-controlled, parallel-group trial. *Clin Ther* 2006;**28**:770–82.
61. Garipagaoglu M, Sahip Y, Darendeliler F, Akdikmen O, Kopuz S, Sut N. Family-based group treatment versus individual treatment in the management of childhood obesity: randomized, prospective clinical trial. *Eur J Pediatr* 2009;**168**:1091–9.
62. Gately PJ, King NA, Greatwood HC, Humphrey LC, Radley D, Cooke CB, *et al.* Does a high-protein diet improve weight loss in overweight and obese children? *Obesity (Silver Spring)* 2007;**15**:1527–34.
63. Ghayour-Mobarhan M, Sahebkar A, Vakili R, Safarian M, Nematy M, Lotfian E, *et al.* Investigation of the effect of high dairy diet on body mass index and body fat in overweight and obese children. *Indian J Pediatr* 2009;**76**:1145–50.
64. Gillis D, Brauner M, Granot E. A community-based behavior modification intervention for childhood obesity. *J Pediatr Endocrinol* 2007;**20**:197–203.
65. Godoy-Matos A, Carraro L, Vieira A, Oliveira J, Guedes EP, Mattos L, *et al.* Treatment of obese adolescents with sibutramine: a randomized, double-blind, controlled study. *J Clin Endocrinol Metab* 2005;**90**:1460–5.
66. Golan M, Fainaru M, Weizman A. Role of behaviour modification in the treatment of childhood obesity with the parents as the exclusive agents of change. *Int J Obes Relat Metab Disord* 1998;**22**:1217–24.
67. Golan M, Kaufman V, Shahar DR. Childhood obesity treatment: targeting parents exclusively v. parents and children. *Br J Nutr* 2006;**95**:1008–15.
68. Golan M, Weizman A, Apter A, Fainaru M. Parents as the exclusive agents of change in the treatment of childhood obesity. *Am J Clin Nutr* 1998;**67**:1130–5.
69. Goldfield GS, Epstein LH, Kilanowski CK, Paluch RA, Kogut-Bossler B. Cost-effectiveness of group and mixed family-based treatment for childhood obesity. *Int J Obes Relat Metab Disord* 2001;**25**:1843–9.
70. Golley RK, Magarey AM, Baur LA, Steinbeck KS, Daniels LA. Twelve-month effectiveness of a parent-led, family-focused weight-management program for prepubertal children: a randomized, controlled trial. *Pediatrics* 2007;**119**:517–25.
71. Gropper SS, Acosta PB. The therapeutic effect of fiber in treating obesity. *J Am Coll Nutr* 1987;**6**:533–5.

72. Grugni G, Guzzaloni G, Ardizzi A, Moro D, Morabito F. Dexfentluramine in the treatment of juvenile obesity. *Minerva Pediatr* 1997;**49**:109–17.
73. Gunnarsdottir T, Sigurdardottir ZG, Njardvik U, Olafsdottir AS, Bjarnason R. A randomized-controlled pilot study of Epstein's family-based behavioural treatment for childhood obesity in a clinical setting in Iceland. *Nord Psychol* 2011;**63**:6–19.
74. Gutin B, Barbeau P, Owens S, Lemmon CR, Bauman M, Allison J, *et al.* Effects of exercise intensity on cardiovascular fitness, total body composition, and visceral adiposity of obese adolescents. *Am J Clin Nutr* 2002;**75**:818–26.
75. Gutin B, Owens S. Role of exercise intervention in improving body fat distribution and risk profile in children. *Am J Hum Biol* 1999;**11**:237–47.
76. Gutin B, Owens S, Okuyama T, Riggs S, Ferguson M, Litaker M. Effect of physical training and its cessation on percent fat and bone density of children with obesity. *Obes Res* 1999;**7**:208–14.
77. Herrera EA, Johnston CA, Steele RG. A comparison of cognitive and behavioral treatments for pediatric obesity. *Child Health Care* 2004;**33**:151–67.
78. Hills AP, Parker AW. Obesity management via diet and exercise intervention. *Child Care Health Dev* 1988;**14**:409–16.
79. Hughes AR, Stewart L, Chapple J, McColl JH, Donaldson MD, Kelnar CJ, *et al.* Randomized, controlled trial of a best-practice individualized behavioral program for treatment of childhood overweight: Scottish Childhood Overweight Treatment Trial (SCOTT). *Pediatrics* 2008;**21**:e539–46.
80. Israel AC, Guile CA, Baker JE, Silverman WK. An evaluation of enhanced self-regulation training in the treatment of childhood obesity. *J Pediatr Psychol* 1994;**19**:737–49.
81. Israsena T, Israngkura M, Srivuthana S. Treatment of childhood obesity. *J Med Assoc Thai* 1980;**63**:433–7.
82. Janicke DM, Lim CS, Perri MG, Bobroff LB, Mathews AE, Brumback BA, *et al.* The Extension Family Lifestyle Intervention Project (E-FLIP for Kids): design and methods. *Contemp Clin Trials* 2011;**32**:50–8.
83. Janicke DM, Sallinen BJ, Perri MG, Lutes LD, Huerta M, Silverstein JH, *et al.* Comparison of parent-only vs. family-based interventions for overweight children in underserved rural settings: outcomes from project STORY. *Arch Pediatr Adolesc Med* 2008;**162**:1119–25.
84. Jansen E, Mulkens S, Jansen A. Tackling childhood overweight: treating parents exclusively is effective. *Int J Obes (Lond)* 2011;**35**:501–9.
85. Jelalian E, Lloyd-Richardson EE, Mehlenbeck RS, Hart CN, Flynn-O'Brien K, Kaplan J, *et al.* Behavioral weight control treatment with supervised exercise or peer-enhanced adventure for overweight adolescents. *J Pediatr* 2010;**157**:923–8.
86. Jelalian E, Mehlenbeck R, Lloyd-Richardson EE, Birmaher V, Wing RR. 'Adventure therapy' combined with cognitive-behavioral treatment for overweight adolescents. *Int J Obes (Lond)* 2006;**30**:31–9.
87. Jiang JX, Xia XL, Greiner T, Lian GL, Rosenqvist U. A two-year family based behaviour treatment for obese children. *Arch Dis Child* 2005;**90**:1235–8.

88. Johnston CA, Steele RG. Treatment of pediatric overweight: an examination of feasibility and effectiveness in an applied clinical setting. *J Pediatr Psychol* 2007;**32**:106–10.
89. Johnston CA, Tyler C, Fullerton G, Poston WS, Haddock CK, McFarlin B, *et al.* Results of an intensive school-based weight loss program with overweight Mexican American children. *Int J Pediatr Obes* 2007;**2**:144–52.
90. Johnston CA, Tyler C, McFarlin BK, Poston WS, Haddock CK, Reeves R, *et al.* Weight loss in overweight Mexican American children: a randomized, controlled trial. *Pediatrics* 2007;**120**:e1450–7.
91. Jones M, Luce KH, Osborne MI, Taylor K, Cuning D, Doyle AC, *et al.* Randomized, controlled trial of an internet-facilitated intervention for reducing binge eating and overweight in adolescents. *Pediatrics* 2008;**121**:453–62.
92. Kalarchian MA, Levine MD, Arslanian SA, Ewing LJ, Houck PR, Cheng Y, *et al.* Family-based treatment of severe pediatric obesity: randomized, controlled trial. *Pediatrics* 2009;**124**:1060–8.
93. Kalavainen MP, Korppi MO, Nuutinen OM. Clinical efficacy of group-based treatment for childhood obesity compared with routinely given individual counseling. *Int J Obes (Lond)* 2007;**31**:1500–8.
94. Kay JP, Alemzadeh R, Langley G, D'Angelo L, Smith P, Holshouser S. Beneficial effects of metformin in normoglycemic morbidly obese adolescents. *Metabolism* 2001;**50**:1457–61.
95. Kelishadi R, Zemel MB, Hashemipour M, Hosseini M, Mohammadifard N, Poursafa P. Can a dairy-rich diet be effective in long-term weight control of young children? *J Am Coll Nutr* 2009;**28**:601–10.
96. Kirschenbaum DS, Harris ES, Tomarken AJ. Effects of parental involvement in behavioral weight loss therapy for preadolescents. *Behav Ther* 1984;**15**:485–500.
97. Kirscht JP, Becker MH, Haefner DP, Maiman LA. Effects of threatening communications and mothers health beliefs on weight change in obese children. *J Behav Med* 1978;**1**:147–57.
98. Krebs NF, Gao D, Gralla J, Collins JS, Johnson SL. Efficacy and safety of a high protein, low carbohydrate diet for weight loss in severely obese adolescents. *J Pediatr* 2010;**157**:252–8.
99. Kwapiszewski RM, Lee Wallace A. A pilot program to identify and reverse childhood obesity in a primary care clinic. *Clin Pediatr (Phila)* 2011;**50**:630–5.
100. Lau PWC, Yu CW, Lee A, Sung RYT. The physiological and psychological effects of resistance training on Chinese obese adolescents. *J Exerc Sci Fit* 2004;**2**:115–20.
101. Lazzer S, Lafortuna C, Busti C, Galli R, Agosti F, Sartorio A. Effects of low- and high-intensity exercise training on body composition and substrate metabolism in obese adolescents. *J Endocrinol Invest* 2011;**34**:45–52.
102. Lorber J. Obesity in childhood. A controlled trial of anorectic drugs. *Arch Dis Child* 1966;**41**:309–12.
103. Lorber J, Rendle-Short J. Obesity in childhood: a controlled trial of phenmetrazine, amphetamine resinate, and diet. *Q Rev Pediatr* 1961;**16**:93–6.
104. Love-Osborne K, Sheeder J, Zeitler P. Addition of metformin to a lifestyle modification program in adolescents with insulin resistance. *J Pediatr* 2008;**152**:817–22.

105. Lustig RH, Hinds PS, Ringwald-Smith K, Christensen RK, Kaste SC, Schreiber RE, *et al.* Octreotide therapy of pediatric hypothalamic obesity: a double-blind, placebo-controlled trial. *J Clin Endocrinol Metab* 2003;**88**:2586–92.
106. Maahs D, de Serna DG, Kolotkin RL, Ralston S, Sandate J, Qualls C, *et al.* Randomized, double-blind, placebo-controlled trial of orlistat for weight loss in adolescents. *Endocr Pract* 2006;**12**:18–28.
107. Maddison R, Foley L, Mhurchu CN, Jull A, Jiang Y, Prapavessis H, *et al.* Feasibility, design and conduct of a pragmatic randomized controlled trial to reduce overweight and obesity in children: the electronic games to aid motivation to exercise (eGAME) study. *BMC Public Health* 2009;**9**:146.
108. Maddison R, Foley L, Ni Mhurchu C, Jiang Y, Jull A, Prapavessis H, *et al.* Effects of active video games on body composition: a randomized controlled trial. *Am J Clin Nutr* 2011;**94**:156–63.
109. Maddison R, Mhurchu CN, Foley L, Epstein L, Jiang Y, Tsai M, *et al.* Screen-time Weight-loss Intervention Targeting Children at Home (SWITCH): a randomized controlled trial study protocol. *BMC Public Health* 2011;**11**:524.
110. Magarey AM, Perry RA, Baur LA, Steinbeck KS, Sawyer M, Hills AP, *et al.* A parent-led family-focused treatment program for overweight children aged 5 to 9 years: the PEACH RCT. *Pediatrics* 2011;**127**:214–22.
111. Makkes S, Halberstadt J, Renders CM, Bosmans JE, van der Baan-Slootweg OH, Seidell JC. Cost-effectiveness of intensive inpatient treatments for severely obese children and adolescents in the Netherlands; a randomized controlled trial (HELIOS). *BMC Public Health* 2011;**11**:518.
112. Matsuyama T, Tanaka Y, Kamimaki I, Nagao T, Tokimitsu I. Catechin safely improved higher levels of fatness, blood pressure, and cholesterol in children. *Obesity (Silver Spring)* 2008;**16**:1338–48.
113. McCallum Z, Wake M, Gerner B, Baur LA, Gibbons K, Gold L, *et al.* Outcome data from the LEAP (Live, Eat and Play) trial: a randomized controlled trial of a primary care intervention for childhood overweight/mild obesity. *Int J Obes (Lond)* 2007;**31**:630–6.
114. McDuffie JR, Calis KA, Uwaifo GI, Sebring NG, Fallon EM, Frazer TE, *et al.* Efficacy of orlistat as an adjunct to behavioral treatment in overweight African American and Caucasian adolescents with obesity-related co-morbid conditions. *J Pediatr Endocrinol* 2004;**17**:307–19.
115. Mellin LM, Slinkard LA, Irwin CE, Jr. Adolescent obesity intervention: validation of the SHAPEDOWN program. *J Am Diet Assoc* 1987;**87**:333–8.
116. Melnyk BM, Small L, Morrison-Beedy D, Strasser A, Spath L, Kreipe R, *et al.* The COPE Healthy Lifestyles TEEN program: feasibility, preliminary efficacy, and lessons learned from an after school group intervention with overweight adolescents. *J Pediatr Health Care* 2007;**21**:315–22.
117. Molnar D, Torok K, Erhardt E, Jeges S. Safety and efficacy of treatment with an ephedrine/caffeine mixture. The first double-blind placebo-controlled pilot study in adolescents. *Int J Obes Relat Metab Disord* 2000;**24**:1573–8.
118. Munsch S, Roth B, Michael T, Meyer AH, Biedert E, Roth S, *et al.* Randomized controlled comparison of two cognitive behavioral therapies for obese children: mother versus mother-child cognitive behavioral therapy. *Psychother Psychosom* 2008;**77**:235–46.

119. Naar-King S, Ellis D, Kolmodin K, Cunningham P, Jen KL, Saelens B, *et al.* A randomized pilot study of multisystemic therapy targeting obesity in African-American adolescents. *J Adolesc Health* 2009;**45**:417–19.
120. Nemet D, Barkan S, Epstein Y, Friedland O, Kowen G, Eliakim A. Short- and long-term beneficial effects of a combined dietary-behavioral-physical activity intervention for the treatment of childhood obesity. *Pediatrics* 2005;**115**:e443–9.
121. Nemet D, Barzilay-Teeni N, Eliakim A. Treatment of childhood obesity in obese families. *J Pediatr Endocrinol* 2008;**21**:461–7.
122. Nova A, Russo A, Sala E. Long-term management of obesity in paediatric office practice: experimental evaluation of two different types of intervention. *Ambulatory Child Health* 2001;**7**:239–47.
123. Nowicka P, Lanke J, Pietrobelli A, Apitzsch E, Flodmark C-E. Sports camp with six months of support from a local sports club as a treatment for childhood obesity. *Scand J Public Health* 2009;**37**:793–800.
124. O'Brien PE, Sawyer SM, Laurie C, Brown WA, Skinner S, Veit F, *et al.* Laparoscopic adjustable gastric banding in severely obese adolescents: a randomized trial. *JAMA* 2010;**303**:519–26. [Erratum published in *JAMA* 2010;**303**:2357.]
125. O'Connor J, Steinbeck K, Hill A, Booth M, Kohn M, Shah S, *et al.* Evaluation of a community-based weight management program for overweight and obese adolescents: the Loozit study. *Nutr Diet* 2008;**65**:121–7.
126. Okely AD, Collins CE, Morgan PJ, Jones RA, Warren JM, Cliff DP, *et al.* Multi-site randomized controlled trial of a child-centered physical activity program, a parent-centered dietary-modification program, or both in overweight children: the HIKCUPS study. *J Pediatr* 2010;**157**:388–94.
127. Ornstein RM, Copperman NM, Jacobson MS. Effect of weight loss on menstrual function in adolescents with polycystic ovary syndrome. *J Pediatr Adolesc Gynecol* 2011;**24**:161–5.
128. Ounis OB, Elloumi M, Amri M, Zbidi A, Tabka Z, Lac G. Impact of diet, exercise and diet combined with exercise programs on plasma lipoprotein and adiponectin levels in obese girls. *J Sports Sci Med* 2008;**7**:437–45.
129. Owens S, Gutin B, Allison J, Riggs S, Ferguson M, Litaker M, *et al.* Effect of physical training on total and visceral fat in obese children. *Med Sci Sports Exerc* 1999;**31**:143–8.
130. Ozkan B, Bereket A, Turan S, Keskin S. Addition of orlistat to conventional treatment in adolescents with severe obesity. *Eur J Pediatr* 2004;**163**:738–41.
131. Park TG, Hong HR, Lee J, Kang HS. Lifestyle plus exercise intervention improves metabolic syndrome markers without change in adiponectin in obese girls. *Ann Nutr Metab* 2007;**51**:197–203.
132. Pedersen MH, Molgaard C, Hellgren LI, Matthiessen J, Holst JJ, Lauritzen L. The effect of dietary fish oil in addition to lifestyle counselling on lipid oxidation and body composition in slightly overweight teenage boys [published online ahead of print July 9 2011]. *J Nutr Metab* 2011. doi:10.1155/2011/348368
133. Pena L, Pena M, Gonzalez J, Claro A. A comparative study of two diets in the treatment of primary exogenous obesity in children. *Acta Paediatr Acad Sci Hung* 1979;**20**:99–103.

134. Racine NM, Watras AC, Carrel AL, Allen DB, McVean JJ, Clark RR, *et al.* Effect of conjugated linoleic acid on body fat accretion in overweight or obese children. *Am J Clin Nutr* 2010;**91**:1157–64.
135. Rao G, Krall J, Loewenstein G. An internet-based pediatric weight management program with and without financial incentives: a randomized trial. *Child Obes* 2011;**7**:122–8.
136. Rauh JL, Lipp R. Chlorphentermine as an anorexigenic agent in adolescent obesity. Report of its efficacy in a double-blind study of 30 teenagers. *Clin Pediatr (Phila)* 1968;**7**:138–40.
137. Reinehr T, Schaefer A, Winkel K, Finne E, Toschke AM, Kolip P. An effective lifestyle intervention in overweight children: findings from a randomized controlled trial on 'Obeldicks light'. *Clin Nutr* 2010;**29**: 331–6.
138. Rendleshort J. Obesity in childhood: a clinical trial of phenmetrazine. *Br Med J* 1960;**1**:703–4.
139. Resnick EA, Bishop M, O'Connell A, Hugo B, Isern G, Timm A, *et al.* The CHEER study to reduce BMI in Elementary School students: a school-based, parent-directed study in Framingham, Massachusetts. *J Sch Nurs* 2009;**25**:361–72.
140. Resnicow K, Taylor R, Baskin M, McCarty F. Results of Go Girls: a weight control program for overweight African-American adolescent females. *Obes Res* 2005;**13**:1739–48.
141. Rezvanian H, Hashemipour M, Kelishadi R, Tavakoli N, Poursafa P. A randomized, triple masked, placebo-controlled clinical trial for controlling childhood obesity. *World J Pediatr* 2010;**6**:317–22.
142. Robertson W, Friede T, Blissett J, Rudolf MCJ, Wallis M, Stewart-Brown S. Pilot of 'Families for Health': community-based family intervention for obesity. *Arch Dis Child* 2008;**93**:921–6.
143. Rodearmel SJ, Wyatt HR, Barry MJ, Dong F, Pan D, Israel RG, *et al.* A family-based approach to preventing excessive weight gain. *Obesity (Silver Spring)* 2006;**14**:1392–401.
144. Rodearmel SJ, Wyatt HR, Stroebele N, Smith SM, Ogden LG, Hill JO. Small changes in dietary sugar and physical activity as an approach to preventing excessive weight gain: the America on the Move family study. *Pediatrics* 2007;**120**:e869–79.
145. Rolland-Cachera MF, Thibault H, Souberbielle JC, Soulie D, Carbonel P, Deheeger M, *et al.* Massive obesity in adolescents: dietary interventions and behaviours associated with weight regain at 2 y follow-up. *Int J Obes Relat Metab Disord* 2004;**28**:514–19.
146. Rooney BL, Gritt LR, Havens SJ, Mathiason MA, Clough EA. Growing healthy families: family use of pedometers to increase physical activity and slow the rate of obesity. *WMJ* 2005;**104**:54–60.
147. Rosado JL, del R Arellano M, Montemayor K, Garcia OP, Caamano MdC. An increase of cereal intake as an approach to weight reduction in children is effective only when accompanied by nutrition education: a randomized controlled trial. *Nutr J* 2008;**7**:28.
148. Rotatori AF, Fox R. The effectiveness of a behavioral weight reduction program for moderately retarded adolescents. *Behav Ther* 1980;**11**:410–16.
149. Rotatori AF, Fox RA, Matson J, Mehta S, Baker A. Changes in biomedical and physical correlates in behavioral weight loss with retarded youths. *J Obes Weight Regul* 1986;**5**:17–27.

150. Rotatori AF, Switzky H. A successful behavioral weight-loss program for moderately-retarded teenagers. *Int J Obes (Lond)* 1979;**3**:223–8.
151. Rudolf M, Christie D, McElhone S, Sahota P, Dixey R, Walker J, *et al.* WATCH IT: a community based programme for obese children and adolescents. *Arch Dis Child* 2006;**91**:736–9.
152. Sabet Sarvestani R, Jamalfard MH, Kargar M, Kaveh MH, Tabatabaee HR. Effect of dietary behaviour modification on anthropometric indices and eating behaviour in obese adolescent girls. *J Adv Nurs* 2009;**65**:1670–5.
153. Sacher PM, Chadwick P, Wells JCK, Williams JE, Cole TJ, Lawson MS. Assessing the acceptability and feasibility of the MEND Programme in a small group of obese 7–11-year-old children. *J Hum Nutr Diet* 2005;**18**:3–5.
154. Sacher PM, Kolotourou M, Chadwick PM, Cole TJ, Lawson MS, Lucas A, *et al.* Randomized controlled trial of the MEND program: a family-based community intervention for childhood obesity. *Obesity (Silver Spring)* 2010;**18**(Suppl. 1):62–8.
155. Saelens BE, Grow HM, Stark LJ, Seeley RJ, Roehrig H. Efficacy of increasing physical activity to reduce children's visceral fat: a pilot randomized controlled trial. *Int J Pediatr Obes* 2011;**6**:102–12.
156. Saelens BE, Sallis JF, Wilfley DE, Patrick K, Cella JA, Buchta R. Behavioral weight control for overweight adolescents initiated in primary care. *Obes Res* 2002;**10**:22–32.
157. Satoh A, Menzawa K, Lee S, Hatakeyama A, Sasaki H. Dietary guidance for obese children and their families using a model nutritional balance chart. *Jpn J Nurs Sci* 2007;**4**:95–102.
158. Savoye M, Shaw M, Dziura J, Tamborlane WV, Rose P, Guandalini C, *et al.* Effects of a weight management program on body composition and metabolic parameters in overweight children: a randomized controlled trial. *JAMA* 2007;**297**:2697–704.
159. Schwingshandl J, Sudi K, Eibl B, Wallner S, Borkenstein M. Effect of an individualised training programme during weight reduction on body composition: a randomised trial. *Arch Dis Child* 1999;**81**:426–8.
160. Senediak C, Spence SH. Rapid versus gradual scheduling of therapeutic contact in a family based behavioural weight control programme for children. *Behav Psychother* 1985;**13**:265–87.
161. Shalitin S, Ashkenazi-Hoffnung L, Yackobovitch-Gavan M, Nagelberg N, Karni Y, Hershkovitz E, *et al.* Effects of a twelve-week randomized intervention of exercise and/or diet on weight loss and weight maintenance, and other metabolic parameters in obese preadolescent children. *Horm Res* 2009;**72**:287–301.
162. Shelton D, Le Gros K, Norton L, Stanton-Cook S, Morgan J, Masterman P. Randomised controlled trial: a parent-based group education programme for overweight children. *J Paediatr Child Health* 2007;**43**:799–805.
163. Shrewsbury VA, O'Connor J, Steinbeck KS, Stevenson K, Lee A, Hill AJ, *et al.* A randomised controlled trial of a community-based healthy lifestyle program for overweight and obese adolescents: the Loozit (R) study protocol. *BMC Public Health* 2009;**9**:119.

164. Sondike SB, Copperman N, Jacobson MS. Effects of a low-carbohydrate diet on weight loss and cardiovascular risk factor in overweight adolescents. *J Pediatr* 2003;**142**:253–8.
165. Srinivasan S, Ambler GR, Baur LA, Garnett SP, Tepsa M, Yap F, *et al.* Randomized, controlled trial of metformin for obesity and insulin resistance in children and adolescents: improvement in body composition and fasting insulin. *J Clin Endocrinol Metab* 2006;**91**:2074–80.
166. Stark LJ, Spear S, Boles R, Kuhl E, Ratcliff M, Scharf C, *et al.* A pilot randomized controlled trial of a clinic and home-based behavioral intervention to decrease obesity in preschoolers. *Obesity (Silver Spring)* 2011;**19**:134–41.
167. St-Onge MP, Goree LL, Gower B. High-milk supplementation with healthy diet counseling does not affect weight loss but ameliorates insulin action compared with low-milk supplementation in overweight children. *J Nutr* 2009;**39**:933–8.
168. Sun M-X, Huang X-Q, Yan Y, Li B-W, Zhong W-J, Chen J-F, *et al.* One-hour after-school exercise ameliorates central adiposity and lipids in overweight Chinese adolescents: a randomized controlled trial. *Chin Med J* 2011;**124**:323–9.
169. Suttapreyasri D, Suthontan N, Kanpoem J, Krainam J, Boonsuya C. Weight-control training-models for obese pupils in Bangkok. *J Med Assoc Thai* 1990;**73**:394–400.
170. Tan S, Yang C, Wang J. Physical training of 9- to 10-year-old children with obesity to lactate threshold intensity. *Pediatr Exerc Sci* 2010;**22**:477–85.
171. Taveras EM, Gortmaker SL, Hohman KH, Horan CM, Kleinman KP, Mitchell K, *et al.* Randomized controlled trial to improve primary care to prevent and manage childhood obesity: the High Five for Kids study. *Arch Pediatr Adolesc Med* 2011;**165**:714–22.
172. Toruner EK, Savaser S. A controlled evaluation of a school-based obesity prevention in Turkish school children. *J Sch Nurs* 2010;**26**:473–82.
173. Truby H, Baxter K, Elliott S, Warren J, Davies P, Batch J. Adolescents seeking weight management: who is putting their hand up and what are they looking for? *J Paediatr Child Health* 2011;**47**:2–4.
174. Truby H, Baxter KA, Barrett P, Ware RS, Cardinal JC, Davies PS, *et al.* The Eat Smart Study: a randomised controlled trial of a reduced carbohydrate versus a low fat diet for weight loss in obese adolescents. *BMC Public Health* 2010;**10**:464.
175. Tsang TW, Kohn M, Chow C, Singh MF. A randomised placebo-exercise controlled trial of Kung Fu training for improvements in body composition in overweight/obese adolescents: the 'martial fitness' study. *J Sports Sci Med* 2009;**8**:97–106.
176. Tsiros MD, Sinn N, Brennan L, Coates AM, Walkley JW, Petkov J, *et al.* Cognitive behavioral therapy improves diet and body composition in overweight and obese adolescents. *Am J Clin Nutr* 2008;**87**:1134–40.
177. Van Mil EG, Westerterp KR, Kester AD, Delemarre-van de Waal HA, Gerver WJ, Saris WH. The effect of sibutramine on energy expenditure and body composition in obese adolescents. *J Clin Endocrinol Metab* 2007;**92**:1409–14.

178. Viccari Sabia R, dos Santos JE, Pessa Ribeiro RP. Effect of physical activity associated with nutritional orientation for obese adolescents: comparison between aerobic and anaerobic exercise. *Rev Bras Med Esporte* 2004;**10**:356–61.
179. Vido L, Facchin P, Antonello I, Gobber D, Rigon F. Childhood obesity treatment: double blinded trial on dietary fibres (glucomannan) versus placebo. *Pediatr Padol* 1993;**28**:133–6.
180. Vissers D, De Meulenaere A, Vanroy C, Vanherle K, Van de Sompel A, Truijien S, *et al*. Effect of a multidisciplinary school-based lifestyle intervention on body weight and metabolic variables in overweight and obese youth. *E Spen Eur E J Clin Nutr Metab* 2008;**3**:e196–202.
181. Vos RC, Wit JM, Pijl H, Kruyff CC, Houdijk ECAM. The effect of family-based multidisciplinary cognitive behavioral treatment in children with obesity: study protocol for a randomized controlled trial. *Trials* 2011;**49**:3104–11.
182. Wadden TA, Stunkard AJ, Rich J, Rubin CJ, Sweidel G, McKinney S. Obesity in black adolescent girls: a clinical trial of treatment by diet, behaviour modification, and parental support. *Pediatrics* 1990;**85**:345–52.
183. Wafa SW, Talib RA, Hamzaid NH, McColl JH, Rajikan R, Ng LO, *et al*. Randomized controlled trial of a good practice approach to treatment of childhood obesity in Malaysia: Malaysian Childhood Obesity Treatment Trial (MASCOT). *Int J Pediatr Obes* 2011;**6**:e62–9.
184. Wake M, Baur LA, Gerner B, Gibbons K, Gold L, Gunn J, *et al*. Outcomes and costs of primary care surveillance and intervention for overweight or obese children: the LEAP 2 randomised controlled trial. *BMJ* 2009;**339**:b3308.
185. Warschburger P, Fromme C, Petermann F, Wojtalla N, Oepen J. Conceptualisation and evaluation of a cognitive-behavioural training programme for children and adolescents with obesity. *Int J Obes Relat Metab Disord* 2001;**25**(Suppl. 1):93–5.
186. Weigel C, Kokocinski K, Lederer P, Dötsch J, Rascher W, Knerr I. Childhood obesity: concept, feasibility, and interim results of a local group-based, long-term treatment program. *J Nutr Educ Behav* 2008;**40**:369–73.
187. Weintraub DL, Tirumalai EC, Haydel KF, Fujimoto M, Fulton JE, Robinson TN. Team sports for overweight children: the Stanford Sports to Prevent Obesity Randomized Trial (SPORT). *Arch Pediatr Adolesc Med* 2008;**162**:232–7.
188. West F, Sanders MR, Cleghorn GJ, Davies PS. Randomised clinical trial of a family-based lifestyle intervention for childhood obesity involving parents as the exclusive agents of change. *Behav Res Ther* 2010;**48**:1170–9.
189. White MA, Martin PD, Newton RL, Walden HM, York-Crowe EE, Gordon ST, *et al*. Mediators of weight loss in a family-based intervention presented over the internet. *Obes Res* 2004;**12**:1050–9.
190. Williams CL, Strobino BA, Brotanek J. Weight control among obese adolescents: a pilot study. *Int J Food Sci Nutr* 2007;**58**:217–30.
191. Williamson DA, Martin PD, White MA, Newton R, Walden H, York-Crowe E, *et al*. Efficacy of an internet-based behavioral weight loss program for overweight adolescent African-American girls. *Eat Weight Disord* 2005;**10**:193–203.

192. Williamson DA, Walden HM, White MA, York-Crowe E, Newton RL, Jr, Alfonso A, *et al.* Two-year internet-based randomized controlled trial for weight loss in African-American girls. *Obesity (Silver Spring)* 2006;**14**:1231–43.
193. Wilson AJ, Prapavessis H, Jung ME, Cramp AG, Vascotto J, Lenhardt L, *et al.* Lifestyle modification and metformin as long-term treatment options for obese adolescents: study protocol. *BMC Public Health* 2009;**9**:434.
194. Wilson DM, Abrams SH, Aye T, Lee PD, Lenders C, Lustig RH, *et al.* Metformin XR for treating adolescent obesity. *Brown Uni Child Adolescent Psychopharmac Update* 2010;**12**:3–4.
195. Wong PC, Chia MY, Tsou IY, Wansaicheong GK, Tan B, Wang JC, *et al.* Effects of a 12-week exercise training programme on aerobic fitness, body composition, blood lipids and C-reactive protein in adolescents with obesity. *Ann Acad Med Singapore* 2008;**4**:286–93. URL: www.mrw.interscience.wiley.com/cochrane/clcentral/articles/232/CN-00667232/frame.html.
196. Yackobovitch-Gavan M, Nagelberg N, Demol S, Phillip M, Shalitin S. Influence of weight-loss diets with different macronutrient compositions on health-related quality of life in obese youth. *Appetite* 2008;**51**:697–703.
197. Yanovski JA, Krakoff J, Salaita CG, McDuffie JR, Kozlosky M, Sebring NG, *et al.* Effects of metformin on body weight and body composition in obese insulin-resistant children: a randomized clinical trial. *Diabetes* 2011;**60**:477–85.
198. Yin TJ, Wu FL, Liu YL, Yu S. Effects of a weight-loss program for obese children: a ‘mix of attributes’ approach. *J Nurs Res* 2005;**13**:21–30.
199. Ylitalo VM. Treatment of obese schoolchildren. *Klin Padiatr* 1982;**194**:310–14.
200. Zakus G, Chin ML, Cooper H, Jr, Makovsky E, Merrill C. Treating adolescent obesity: a pilot project in a school. *J Sch Health* 1981;**51**:663–6.

Appendix 4 Data extraction form for search 1

1. Ref Man ID:
2. Reviewer initials:
3. Authors: [put * next to contact author]
4. Year:

[Unless otherwise stated, tick relevant box(s)]

5. Study design:

5.1	Pilot study
5.2	Feasibility study
5.3	Phase III RCT
5.4	Pre-post
5.5	Other (<i>please write in</i>)

6. Type of intervention:

6.1	Lifestyle
6.2	Diet
6.3	Physical activity
6.4	Sedentary behaviour
6.5	Drug/surgical
6.6	Other (<i>please write in</i>)

7. Intervention delivered to:

7.1	Child only
7.2	Parent/caregiver only
7.3	Child and parent(s)/caregiver
7.4	Other (<i>please write in</i>)

8. Sample size (final):

8.1	Individual
8.2	Family

9. Ethnicity (continents and subcategories):

9.1	Europe
9.2	UK
9.3	Ireland
9.4	Eastern Europe
9.5	Scandinavian
9.6	Spain
9.7	France
9.8	Germany
9.9	Italy
9.10	Antarctica
9.11	Asia
9.12	South Asia (Indian, Pakistani, Bangladeshi)
9.13	Middle East
9.14	China
9.15	Japan
9.16	Other Asian background
9.17	North America
9.18	USA
9.19	Canada
9.20	Mexico
9.21	Central America and Caribbean islands
9.22	South America
9.23	Brazil
9.24	Argentina
9.25	Australia
9.26	Australia
9.27	New Zealand
9.28	Africa
9.29	North Africa
9.30	South Africa
9.31	Other (<i>please state</i>)
9.32	Not stated

Note: Turkey is part of Asia (Middle East) and Europe; Russia is part of Europe and Asia.

10. Ethnicity:

10.1	White
10.2	Black
10.3	Caribbean
10.4	African
10.5	African American
10.6	Any other black background (<i>please write in</i>)
10.7	South Asian
10.8	Indian
10.9	Pakistani
10.10	Bangladeshi
10.11	Any other Asian background (<i>please write in</i>)
10.12	Northeast Asian
10.13	China
10.14	Korean
10.15	Japan
10.16	Southeast Asian or South Mongoloid
10.17	Thailand
10.18	Malaysia
10.19	Indonesia
10.20	Philippines
10.21	Turanid (Kazakhstan, Hungary, Turkey)
10.22	Bambutid race (African Pygmies)
10.23	Hispanic or Latino
10.24	Native Hawaiian or Other Pacific Islander
10.25	Alaska Native or American Indian
10.26	Australian Aborigines
10.27	Melanesian (New Guinea, Papua, Solomon islands)
10.28	Mixed ethnic groups
10.29	Ethnicity not defined
10.30	Other (<i>please write in</i>)
10.31	Other (<i>but not stated what other is</i>)

11. Sample age:

11.1	Infant (< 36 months)
11.2	Child (36 months to 12 years)
11.3	Adolescent (> 12 years)
11.4	Infant and children
11.5	Children and adolescents
11.6	All ages

12. Primary outcome measure:

	Name of tool	Author	Year
12.1	BMI/BMI-SDS/%BMI (self-report)		
12.2	BMI/BMI-SDS/%BMI (measured)		
12.3	Weight (self-report)		
12.4	Weight (measured)		
12.5	SFT		
12.6	Waist circumference		
12.7	Waist-hip ratio		
12.8	Mid-arm circumference		
12.9	DXA		
12.10	BIA		
12.11	Hydrodensitometry weighing		
12.12	Near infrared interactance (NIR)		
12.13	BOD POD (air displacement)		
12.14	Total body electrical conductivity (TOBEC)		
12.15	Magnetic resonance imaging (MRI)		
12.16	Computed tomography (CT)		
12.17	Other measure of obesity (<i>please specify</i>)		
12.18	Not reported		

13. Secondary outcome measures (*if more than one type of measure within each outcome, report name of tool and first author for each measure*)

	Outcome	Type of measure	Name of tool	First author	Year
13.1	Anthropometry	BMI (self-report)			
		BMI (measured)			
		Weight (self-report)			
		Weight (measured)			
		Waist circumference			
		Waist-to-hip ratio (WHR)			
		Skinfold thickness (multiple sites or one site – measured with calipers)			
		Mid-arm circumference			
		Dual energy X-ray absorptiometry (DXA)			
		Bioelectrical impedance (BIA)			
		Hydrodensitometry weighing			
		Near infrared interactance (NIR)			
		BOD POD (air displacement)			
		Total body electrical conductivity (TOBEC)			
		Magnetic resonance imaging (MRI)			
		Computed tomography (CT)			
		Other (<i>please write in</i>)			
13.2	Other measure/ proxy of adiposity				
13.3	Diet	Weighed food diary/record			
		Estimated food diary/record			
		FFQ			
		Semiquantitative FFQ			
		Multiple-pass dietary recall			
		24-hour dietary recall			
		Food intake checklist [i.e. specific food/ groups (e.g. fruit and vegetable intake checklist)]			
		Diet history			
		Diet observation (DVD or direct observation)			
		Doubly labelled water			
		Dietary nitrogen			
		Other (<i>please write in</i>)			

	Outcome	Type of measure	Name of tool	First author	Year
13.4	Eating behaviour	Eating behaviour checklists Eating disorders questionnaires/observations Other (<i>please write in</i>)			
13.5	PA	Activity monitor/movement sensors Activity diaries Retrospective questionnaires Activity recalls Direct observation (recorded or researcher conducted) Other (<i>please write in</i>)			
13.6	Sedentary behaviour	TV questionnaire Screen time questionnaires Activity monitor/movement sensors Direct observation (recorded or researcher conducted)			
13.7	Psychological well-being	Self-esteem Self-perception Depression Anxiety Behaviour Psychiatric dysfunction Perceived competence Body image General well-being Other (<i>please write in</i>)			
13.8	Economics	Direct costs Quality-of-life scales Other (<i>please write in</i>)			
13.9	Environment	Geospatial (food/retail outlets) Built environment (e.g. neighbourhood layout) Home environment [physical (e.g. food availability) and social (e.g. rules and policies)] School/nursery environment [physical (e.g. food availability) and social (e.g. rules and policies)] Other (<i>please write in</i>)			

	Outcome	Type of measure	Name of tool	First author	Year
13.10	Fitness	Heart rate (resting and/or recovery) Aerobic capacity/agility (step test, shuttle runs, sprints, timed/endurance runs/walk/bike) Room calorimetry (CO ₂ /VO ₂ , energy expenditure) Indirect calorimetry (CO ₂ /VO ₂ , energy expenditure) Doubly labelled water Respiratory exchange ratio Packed cell volume Muscular strength Muscular endurance Flexibility, other (<i>please write in</i>)			
13.11	Physiological	Blood pressure Metabolic markers (e.g. lipids, glucose, insulin, leptin, adipocytokines) Other (<i>please write in</i>)			
13.11	Other (<i>please write in</i>)				
13.12	Not reported				

14. Comments:

Appendix 5 Data extraction form for search 2: dietary assessment

1a. Ref Man ID:

1b. Manuscript type:

Primary development paper Original used and evaluated Modified and evaluated

1c. Category of measurement tool.

- Questionnaires/surveys with scales or categories with pre-defined terms
- Diaries, recalls, direct observations or monitors with open responses/recall/observation
- Biochemical or anthropometric measures or assays

2. Reviewer initials:

3. First author: [put * next to contact author]

4. Year:

Outcome measure details

5a. Full name of measure:

5b. Acronym of measure: [mark N/A where appropriate]

6. Type of measurement:

6.1	Weighed food diary/record
6.2	Estimated food diary/record
6.3	FFQ
6.4	Semi-quantitative FFQ
6.5	24-hour dietary recall
6.6	Food intake checklist [i.e. specific food/groups (e.g. fruit and vegetable intake checklist)]
6.7	Diet history
6.8	Diet observation (DVD or researcher)
6.9	Dietary patterns
6.10	Other

Provide details:

7. Mode of administration:

- | | |
|------|--|
| 7.1 | Self-completed |
| 7.2 | Parent completed |
| 7.3 | Interview administered in person – parent |
| 7.4. | Interview administered over telephone – parent |
| 7.5 | Interview administered in person – child |
| 7.6 | Interview administered over telephone – child |
| 7.7 | Interview administered in person – parent and child |
| 7.8 | Interview administered over telephone – parent and child |
| 7.9 | Researcher conducted/observed (direct measures) |
| 8.0 | Other |

Provide details:

7b. Method of data collection:

- | | |
|-------|---------------------------------------|
| 7b.1 | Pen and paper |
| 7b.2 | Personal digital assistants |
| 7b.3 | Smart phones |
| 7b.4. | Web-based tools |
| 7b.5 | Download data |
| 7b.6 | Biochemical (e.g. blood, urine, etc.) |
| 7b.6 | Other |

Provide details:

8a. Sample age:

- | | |
|------|-------------------------------|
| 8a.1 | Infant (< 36 months) |
| 8a.2 | Child (36 months to 12 years) |
| 8a.3 | Adolescent (> 12 years) |
| 8a.4 | Infant and children |
| 8a.5 | Children and adolescents |
| 8a.6 | All ages |
-

8b Sample weight status:

8b.1	All obese
8b.2	Obese and overweight
8b.3	Overweight
8b.4	Mixed (stratified)
8b.5	Mixed (non-stratified)

9. Ethnicity (continents and subcategories):

9.1	Europe
9.2	UK
9.3	Ireland
9.4	Eastern Europe
9.5	Scandinavian
9.6	Spain
9.7	France
9.8	Germany
9.9	Italy
9.10	Antarctica
9.11	Asia
9.12	South Asia (Indian, Pakistani, Bangladeshi)
9.13	Middle East
9.14	China
9.15	Japan
9.16	Other Asian background
9.17	North America
9.18	USA
9.19	Canada
9.20	Mexico
9.21	Central America and Caribbean islands
9.22	South America
9.23	Brazil
9.24	Argentina
9.25	Australia
9.26	New Zealand
9.27	Africa
9.28	North Africa
9.29	South Africa
9.30	Other (<i>please state</i>)
9.31	Not stated

9b. Race

9b.1	White
9b.2	Black
9b.3	Caribbean
9b.4	African
9b.5	African American
9b.6	Any other black background (<i>please write in</i>)
9b.7	South Asian
9b.8	Indian
9b.9	Pakistani
9b.10	Bangladeshi
9b.11	Any other Asian background (<i>please write in</i>)
9b.12	Northeast Asian
9b.13	China
9b.14	Korean
9b.15	Japan
9b.16	Southeast Asian or South Mongoloid
9b.17	Thailand
9b.18	Malaysia
9b.19	Indonesia
9b.20	Philippines
9b.21	Turanid (Kazakhstan, Hungary, Turkey)
9b.22	Bambutid race (African Pygmies)
9b.23	Hispanic or Latino
9b.24	Native Hawaiian or Other Pacific Islander
9b.25	Alaska Native or American Indian
9b.26	Australian Aborigines
9b.27	Melanesian (New Guinea, Papua, Solomon islands)
9b.28	Mixed ethnic groups
9b.29	Race not defined
9b.30	Other (<i>please write in</i>)
9b.31	Other (but not stated what other is)

10. Number of items: [mark N/A where appropriate]

11. Categories/domains:

11.1	No categories/domains
11.2	Fruits
11.3	Vegetables
11.4	Cereals and cereal products
11.5	Meat: white meat
11.6	Meat: red and processed meat
11.7	Meat: fish and other proteins
11.6	Milk and milk products
11.8	Beans and pulses
11.9	Snack foods
11.10	Oils, spreads and condiments
11.11	Nuts and seeds
11.12	Sugars and preserves
11.13	Baby foods
11.14	Sugar-sweetened beverages
11.15	Non-sugar sweetened beverages
11.16	Ready-made foods (including takeaway and frozen)
11.17	Baked goods
11.18	Macronutrients
11.19.	Protein
11.20	Carbohydrate
11.21	Fat
11.22	Micronutrients
11.23	Energy intake
	Other

Provide details:

12A. Tool development/theoretical framework

Question	Response options	Score
12A1. The concept to be measured was clearly stated (rationale and description)	4 = strongly agree (concepts are named and clearly defined) 3 = agree (concepts are named and general described) 2 = disagree (concepts only named but not defined) 1 = strongly disagree (concepts are not clearly named or defined)	

Question	Response options	Score
12A2. Was a theoretical or conceptual framework used or referenced?	<p>4 = strongly agree (theory/framework used as a basis for development)</p> <p>3 = agree (theory/framework named and incorporated)</p> <p>2 = disagree (theory/framework named but not used)</p> <p>1 = strongly disagree (no theory/framework described)</p> <p>0 = N/A = (biochemical/anthropometry, direct measures/observations)</p>	
12A3. Populations that the measure was intended for were adequately described	<p>4 = strongly agree (describes at least four characteristics including: age, gender, race/ethnicity and SES)</p> <p>3 = agree (three characteristics reported)</p> <p>2 = disagree (two characteristics reported)</p> <p>1 = strongly disagree (no characteristics reported)</p>	
12A4. Were the populations that the measure was intended for involved in measurement development?	<p>4 = strongly agree (at least three methods of involvement including: part of study team, steering committee, pilot testing, cognitive interviews/focus groups)</p> <p>3 = agree (involved using at least two methods)</p> <p>2 = disagree (populations minimally involved in one method)</p> <p>1 = strongly disagree (populations not involved)</p> <p>0 = N/A (biochemical/anthropometry)</p>	
<i>If response to 12A4 is 1 or 0, skip to A5</i>		
12A4a.1. Please specify how they were involved	<ul style="list-style-type: none"> ● Steering/advisory committee <input type="checkbox"/> ● Pilot test <input type="checkbox"/> ● Focus group <input type="checkbox"/> <p>Other:</p>	
12A5. Determination of items?	<p>Subject specific (e.g. from literature) <input type="checkbox"/></p> <p>Data driven (e.g. analysis of existing dietary database) <input type="checkbox"/></p> <p>Combination of subject specific and data driven <input type="checkbox"/></p> <p>Item from existing tool <input type="checkbox"/></p> <p>N/A (e.g. diary/recall methods) <input type="checkbox"/></p> <p>Other:</p>	
12A6a. Did they start with a larger pool and then narrow down items included?	<p>Yes <input type="checkbox"/></p> <p>No <input type="checkbox"/></p> <p>Not reported <input type="checkbox"/></p>	
12A6b. Was a systematic process used to generate a pool of items	<p>4 = strongly agree (expert and/or clinical input/review, data driven approach and user input)</p> <p>3 = agree (two of the three approach for strongly agree)</p> <p>2 = disagree (one of the three approaches for strongly agree)</p> <p>1 = strongly disagree (no clear methodology reported)</p> <p>0 = N/A (all non-itemised questionnaires/surveys)</p>	

Tool evaluation

12B. Reliability testing: internal consistency

Question	Response options	Score
12B1. Was internal consistency measured? <i>If answer to B1 is no, skip to section C</i>	Y/N	
12B2. Results for internal consistency Scale domain/name	Cronbach's alpha	KR-20 Split half R
12B3. Results for full tool	Yes <input type="checkbox"/> No <input type="checkbox"/>	
12B4. Scale results provided at a range? <i>(if 'no' please work out range)</i>	Yes <input type="checkbox"/> No <input type="checkbox"/>	
12B5. Other statistics?	Yes <input type="checkbox"/> Statistical name(s) and result(s) No <input type="checkbox"/>	
12B6. Sample size	N=	
12B7. Robustness	4 = Strongly agree (adequate sample size, reported by scale category, appropriate stats, adequate results, for example: <ul style="list-style-type: none"> ● sample size > 50 ● alpha > 0.7 ● KR-20 > 0.7) ● Split half: 3 = Agree (3 of 4) 2 = disagree (2 of 4) 1 = strongly disagree (< 2 of 4)	

12C. Reliability: reproducibility

Question	Response options	Score
12C1. Was reproducibility measured? <i>If answer to C1 is no, skip to section D</i>	Y/N	
12C2. How was reproducibility measured? <i>If answer to 12C3 is test-retest fill out section 12C3a, if answered inter-rater go to C3b</i>	<i>Tick all that apply:</i> TRT <input type="checkbox"/> Inter-rater <input type="checkbox"/>	

12C3a. Results for TRT

12C3. Interval between tests	... years ... weeks ... days ... hours	
12C4. Scale domain/name	t-test (or non-para equivalent)	Correlation Pearson's/ ICC/rho Kappa
12C5. Results for full tool	Yes <input type="checkbox"/> No <input type="checkbox"/>	
12C6. Scale results provided at a range (if 'no' please work out range)	Yes <input type="checkbox"/> No <input type="checkbox"/>	
12C7. Other statistics?	Yes <input type="checkbox"/> No <input type="checkbox"/>	Statistical name(s) and result(s):
12C8. Sample size?	N=	

Question	Response options	Score
12C9. Robustness	<p>4 = Strongly agree (adequate sample size, reported by scale category, appropriate stats, adequate results e.g.</p> <ul style="list-style-type: none"> ● sample size ≥ 50 ● $\kappa \geq 0.4$ ● Spearman/Pearson ≥ 0.4 <p>3 = Agree (3 of 4)</p> <p>2 = Disagree (2 of 4)</p> <p>1 = Strongly disagree (< 2 of 4)</p>	
12C3b. Results for inter-rater		
12C1b. Scale domain/name	% agreement	Correlation Pearson's/ ICC/rho
		Kappa
		Krippendorff's alpha
12C2b. Results for full tool	Yes <input type="checkbox"/>	
	No <input type="checkbox"/>	
12C3b. Scale results provided at a range (if no, please work out range)	Yes <input type="checkbox"/>	
	No <input type="checkbox"/>	
12C4b. Other statistics?	Yes <input type="checkbox"/>	
	No <input type="checkbox"/>	
12C5b. Sample size?	N=	
12C5b. Robustness	<p>4 = Strongly agree (adequate sample size, reported by scale category, appropriate stats, adequate results, e.g.</p> <ul style="list-style-type: none"> ● Sample size – study specific ● Pearson's/ICC/rho ≥ 0.40 ● $\kappa \geq 0.40$ ● Krippendorff's alpha ≥ 0.80 <p>3 = Agree (3 of 4)</p> <p>2 = Disagree (2 of 4)</p> <p>1 = Strongly disagree (< 2 of 4)</p>	

12D. Internal validity testing

D. Internal validity

Question

Response

12D1. Was internal validity testing performed?

Y/N

If answer to 12D1 is 'no' go to section 12E

12D2. Type of analysis

Principle components analysis Principle factor analysis Confirmatory factor analysis
(structural equation modelling) Cluster analysis Indexed-based analysis Varimax rotation

Other:

12D3. Identified factors

Factor
loadingRange of
factor
loadingsNo.
of
items

Eigenvalue

Coefficient

% total
variance

12D4. Results for full tool?

12D5. Scale results provided as a range?

12D6. Other stats? Sensitivity,
specificity, discriminate validity testing?Yes Statistical name and resultsNo

12D7. Sample size?

N=

12D8. Robustness

4 = Strongly agree (adequate sample
size, reported by scale category,
appropriate stats, adequate results,
e.g. sample size of five participants
per item:

- Eigenvalue ≥ 1
- Factor loading = High > 0.6 , Low < 0.4
- Range of factor loading
- No. of items
- Coefficient ≥ 0.5
- % total variance

12E. External validity testing

Question	Response options
12E1. Was validity testing performed?	Y/N
If answer to E1 is 'no' skip to F1	
12E2. What statistical tests were used?	<i>Tick all that apply</i>
	Criterion validity <input type="checkbox"/>
	Convergent validity <input type="checkbox"/>
	Construct validity <input type="checkbox"/>
	Content validity <input type="checkbox"/>
	Face validity <input type="checkbox"/>

Depending on what validity test was done please fill out results in appropriate section

12E3 Criterion validity

12E3i. Gold standard reference method	DLW with PABA (para-aminobenzoic acid)	<input type="checkbox"/>
	DLW without PABA	<input type="checkbox"/>
	Goldberg cut-off = energy intake: BMR (lab measured)	<input type="checkbox"/>
	Goldberg cut-off = energy intake: BMR (estimated)	<input type="checkbox"/>
	Goldberg cut-off (measured) with physical activity (objective)	<input type="checkbox"/>
	Goldberg cut-off (measured) with physical activity (self-report)	<input type="checkbox"/>
	Goldberg cut-off (estimated) with physical activity (objective)	<input type="checkbox"/>
	Goldberg cut-off (estimated) with physical activity (self-report)	<input type="checkbox"/>
	Dietary nitrogen–urinary nitrogen (multiple measures with PABA)	<input type="checkbox"/>
	Dietary nitrogen–urinary nitrogen (single measure with PABA)	<input type="checkbox"/>
	Dietary nitrogen–urinary nitrogen (multiple measures no PABA)	<input type="checkbox"/>
	Dietary nitrogen–urinary nitrogen (single measure no PABA)	<input type="checkbox"/>
	Direct observation	<input type="checkbox"/>

Other:

12E3ii. Results for scales/domain

Pearson's/ Spearman's	Regression coefficient	t-test	Agreement (%)	Agreement (kappa)
--------------------------	---------------------------	--------	------------------	----------------------

Question	Response options										
12E3 iii. Results for full tool											
12E3iv. Scale provided as a range (if 'no' please work out range)	Yes <input type="checkbox"/> No <input type="checkbox"/>										
12E3v. Other stats? Sensitivity, specificity, discriminate validity testing?	Yes <input type="checkbox"/> Statistical name and results No <input type="checkbox"/>										
12E3vi. Sample size?	$N =$										
12E3vii. Robustness	4 = Strongly agree (adequate sample size, reported by scale category, appropriate stats, adequate results, e.g.: <ul style="list-style-type: none"> ● Sample size: adequate: > 100 ● Pearson's/Spearman's ≥ 0.4 ● Regression coefficient = $\rho > 0.5$ or $r \geq 0.50$ ● Agreement ● $\kappa \geq 0.4$ ● t-test $p > 0.05$, t-value > 1 ● AUC > 0.7 3 = Agree (3 of 4) 2 = Disagree (2 of 4) 1 = Strongly disagree (< 2 of 4)										
12E4 Convergent validity											
12E4i. Comparison method	Weighed food diary/record <input type="checkbox"/> Estimated food diary/record <input type="checkbox"/> FFQ <input type="checkbox"/> Semi-quantitative FFQ <input type="checkbox"/> 24-hour diet recall <input type="checkbox"/> Multiple-pass dietary recall <input type="checkbox"/> Food intake checklist (e.g. fruit and vegetable intake checklist) <input type="checkbox"/> Diet history <input type="checkbox"/> Food purchase record <input type="checkbox"/> Electronic observations (e.g. mobile phone photographs) <input type="checkbox"/> Other:										
12E4ii. Results for scales/domain	<table border="0"> <thead> <tr> <th>Pearson's/ Spearman's</th> <th>Regression coefficient</th> <th>t-test</th> <th>Agreement (%)</th> <th>Agreement (kappa)</th> </tr> </thead> <tbody> <tr> <td></td> <td></td> <td></td> <td></td> <td></td> </tr> </tbody> </table>	Pearson's/ Spearman's	Regression coefficient	t-test	Agreement (%)	Agreement (kappa)					
Pearson's/ Spearman's	Regression coefficient	t-test	Agreement (%)	Agreement (kappa)							

Question	Response options
12E4iii. Results for full tool	
12E4iv. Scale provided as a range (if 'no' please work out range)	Yes <input type="checkbox"/> No <input type="checkbox"/>
12E4v. Other stats? Sensitivity, specificity, discriminate validity testing?	Yes <input type="checkbox"/> Statistical name and results No <input type="checkbox"/>
12E4vi. Sample size?	$N =$
12E4vii. Robustness	4 = Strongly agree (adequate sample size, reported by scale category, appropriate stats, adequate results, for example: <ul style="list-style-type: none"> ● Sample size: > 100 ● Pearson's/Spearman's ≥ 0.4 ● Regression coefficient = $p > 0.5$ or $r \geq 0.50$ ● Agreement ● $\kappa \geq 0.4$ ● t-test $p > 0.05$, t-value > 1. ● AUC > 0.7 3 = Agree (3 of 4) 2 = Disagree (2 of 4) 1 = Strongly disagree (< 2 of 4)
12E5 Construct validity	
12E5i. Construct	Obesity <input type="checkbox"/> Eating behaviour <input type="checkbox"/> Screen time <input type="checkbox"/> Physical activity <input type="checkbox"/> Disease outcome <input type="checkbox"/> Other:
12E5ii. Results for scales/domain	Pearson's/ Spearman's Regression coefficient t-test Agreement

Question	Response options
12E5iii. Results for full tool	Yes <input type="checkbox"/> No <input type="checkbox"/>
12E5iv. Scale provided as a range (if 'no' please work out range)	Yes <input type="checkbox"/> No <input type="checkbox"/>
12E5v. Other stats? Sensitivity, specificity, discriminate validity testing?	Yes <input type="checkbox"/> Statistical name and results No <input type="checkbox"/>
12E5vi. Sample size?	$N=$
12E5vii. Robustness	4 = Strongly agree (adequate sample size, reported by scale category, appropriate stats, adequate results, for example: <ul style="list-style-type: none"> ● Sample size: > 100 ● Pearson's/Spearman's ≥ 0.4 ● $\kappa \geq 0.4$ ● Regression coefficient = $\rho > 0.5$ or $r \geq 0.50$ ● Agreement ● t-test $p > 0.05$, t-value > 1 ● AUC > 0.7 3 = Agree (3 of 4) 2 = Disagree (2 of 4) 1 = Strongly disagree (< 2 of 4)
12E6 Content validity	
12E6i. Stakeholders	Experts – general review/consensus <input type="checkbox"/> Experts – content validity ratio <input type="checkbox"/> Other:
12E6ii. Method	Consensus methodology <input type="checkbox"/> Focus groups <input type="checkbox"/> Interviews <input type="checkbox"/> Other:
12E6iii. Results for scales/domain	Content validity ratio Content validity index Other
12E6iiib. Open response for results	
12E6iv. Sample size?	$N=$
12E6v. Robustness	Not applicable

Question	Response options
12E7 Face validity	
12E7i. Stakeholders	Experts – general review/consensus <input type="checkbox"/> Experts – content validity ratio <input type="checkbox"/> Other:
12E7ii. Method	Consensus methodology <input type="checkbox"/> Focus groups <input type="checkbox"/> Interviews <input type="checkbox"/> Other:
12E7iii. Results for scales/domain	
12E7iiib. Open response for results	
12E7iv. Sample size	N =
12E7v. Robustness	N/A

12F. Responsiveness

12F1a. Was responsiveness testing performed?	Y/N
<i>If answer to F1 is 'no' skip to G1</i>	
F2a. Results for responsiveness test(s)	
12F1b. Time interval	... years ... weeks ... days ... hours
12F2. Method	Change over time (non-intervention dependent) <input type="checkbox"/> Change following an intervention <input type="checkbox"/>
12F3. Results for scales/domains	Standardised response means Effect size Other

Question	Response options
12F4. Results for full tool?	Yes <input type="checkbox"/>
	No <input type="checkbox"/>
12F5. Scale results provided as a range (<i>if 'no' please work out range</i>)	Yes <input type="checkbox"/>
	No <input type="checkbox"/>
12F6. Other stats? Sensitivity, specificity, discriminate validity testing?	Yes <input type="checkbox"/> Statistical name and results
	No <input type="checkbox"/>
12F7. Sample size?	N=
12F8. Robustness	4 = Strongly agree (adequate sample size, clear report of with or without intervention, appropriate stats by scale (if applicable), adequate results (e.g.?)
	3 = Agree (3 of 4 for strongly agree)
	2 = Disagree (2 of 4 for strongly agree)
	1 = Strongly disagree (< 2 of 4 for strongly agree)

13A. Cultural Language adaptations or translations

Question	Response options
13A1. Has the measure been adapted and/or translated for use in different cultures/languages?	Yes <input type="checkbox"/>
	No <input type="checkbox"/>
<i>If answer to G1 is 'no' skip to H1</i>	
13A2. What language is it in?	
13A3. If 'yes' specify languages or cultures.	
13A4. Methods used for translation and/or adaptation	

14A. Scoring/cut-offs

Question	Response option	Score
14A1. Does the paper provide sufficient detail on how data should be reported?	4 = Strongly agree (must include information on response options and scoring/cut-offs AND interpretation of scoring)	
	3 = Agree (includes information on response options and scoring/cut-offs)	
	2 = Disagree (includes information on response options or scoring/cut-offs)	
	1 = Strongly disagree (scoring/cut-offs or interpretation not reported)	
<i>If answer to H1 is 1, skip to I1</i>		
14A2. Is there a published manuscript?	Yes <input type="checkbox"/> [<i>citation if differs from current paper</i>]	
	No <input type="checkbox"/>	

Question	Response option	Score
14A3. Is there a website?	Yes <input type="checkbox"/>	
	[URL]	
	No <input type="checkbox"/>	
14A4. Is there an author contact	Yes <input type="checkbox"/>	
	[author contact]	
	No <input type="checkbox"/>	

15. Burden

Question	Response options	Reviewer response
15a1. Is the administrative burden discussed?	Y/N	
<i>If answer to 15a1 is 'no' skip to 15b</i>		
15a2. What sources of burden are addressed?	Time required to administer <input type="checkbox"/>	
	Training requirements for those administering <input type="checkbox"/>	
	Other	
	<i>(report all that apply)</i>	
15a3. For each source provide range (or summary) of results		
15a4. Was burden considered acceptable?	4 = Strongly agree <input type="checkbox"/>	
	3 = Agree <input type="checkbox"/>	
	2 = Disagree <input type="checkbox"/>	
	1 = Strongly disagree <input type="checkbox"/>	
15b1. Do they report on the level of cognitive ability on behalf of participant (e.g. reading level)?	Y/N	
<i>If answer to 15b1 is 'no' skip to 16</i>		
15b2. If yes, what level of ability was required?		
16a. Is information on cost available?	Y/N	<i>(If applicable please provide cost per participant)</i>
16b. Is information on copyright available?	Y/N	<i>(If possible please provide link/details)</i>

Appendix 6 Anthropometry studies: summary table

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
1	Nicholson 2001 ²²¹	ADP	Child	Mixed (stratified)	119	USA	White, African American	<p>Comparator = DXA</p> <p>ADP had strong correlation with DXA ($r = 0.95$)</p> <p>ADP using the Siri equation underestimated %BF by 1.9%</p> <p>Determination of %BF from ADP using the Siri model slightly underestimates %BF as determined by DXA in girls but appears to be superior to existing field methods both in accuracy and LOA</p> <p>Comparator = DXA (%BF)</p> <p>Evaluated 'change' in all measures</p> <p>All methods highly correlated</p> <p>No mean bias for estimates of %BF change in ADP</p> <p>Magnitude bias was present for ADP relative to DXA ($p < 0.01$)</p> <p>Estimates of change in %BF were systematically overestimated by BIA ($1.37 \pm 6.98\%$, $p < 0.001$)</p> <p>TSF accounted for only 13% of the variance in %BF change</p> <p>Conclusion: None of the methods measured change as well as DXA, but ADP performed better than did TSF or BIA</p>
2	Elberg 2004 ²²²	ADP, TSF and BIA	Children and adolescents	Mixed (stratified)	86	USA	White, African American	<p>Comparator = MRI and IH-MRS measures of VAT, SAT, IMCL</p> <p>WHR (anthropometry), and per cent trunk fat and TEFR (DXA) are good surrogates for IMCL ($r = 0.66$, $p = 0.0004$) and for VAT ($r = 0.83$ and 0.82, $p = 0.0001$), respectively, in adolescent girls</p>
3	Savgan-Gurol 2010 ²⁷⁰	Anthropometry: WC-UC, WC-IC, WHR, WHtR, DXA = total fat, %BF, per cent trunk fat, TEFR, as surrogates for IMCLs and VAT	Adolescents	Mixed (stratified)	30 (15 obese, 15 normal weight)	USA	White, African American	

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
4	Semiz 2007 ⁷¹	Anthropometric measures [BMI, WC, WHR, triceps and subscapular (SFT)] of body fat	Children	Mixed (stratified)	84	Turkey	Not defined	<p>Comparators = Ultrasound measurements of visceral, preperitoneal and subcutaneous fat layers at maximum and minimum thickness sites</p> <p>In the obese group, BMI was significantly correlated with ultrasound measurements of fat thicknesses, except minimum preperitoneal and subscapular, in which the control group BMI was significantly correlated with all ultrasound fat measurements</p> <p>Multiple regression analyses using VAT as the dependent variable, and anthropometric parameters, gender and obese/non-obese as the independent variable, revealed that BMI was the best single predictor of V ($R^2 = 0.53$)</p> <p><i>Conclusion:</i> The validity of anthropometric SFT in children is low – BMI provides best estimate of body fat. WHR in children and adolescents is not a good index to show intra-abdominal fat deposition</p> <p>Comparator = MRI</p>
5	Rolland-Cachera 1997 ⁷²	Arm circumference SFT (triceps)	Children and adolescents	Mixed (stratified)	28	Europe	Not defined	<p>MRI used to validate new equation for calculating body composition from upper arm circumference and TSF</p> <p>Correlations between MRI and UFA (existing equation result) and MRI and UFE (new equation result) were similar ($r = 0.96$ for both correlations in control group and $r = 0.84$ and 0.82 in obese group), but the areas assessed by MRI (13.8 cm^2) were closer to UFE (12.4 cm^2) than to UFA (11.2 cm^2) in the control group as well as in the obese group (MRI = 48.7 cm^2, UFE = 46.6 cm^2, UFA = 38.5 cm^2)</p> <p>LOA between MRI and anthropometry were $5.7 \pm 5.8 \text{ cm}^2$ for UFA and $0.6 \pm 5.0 \text{ cm}^2$ for UFE, showing that UFA is not acceptable in most cases. Conclude that UFE is simple and accurate index for measuring body composition</p>

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
6	Shaikh 2007 ²⁷³	BIA	Children	Obese	46	UK		<p>Comparator = DXA</p> <p>Highly significant correlation shown between BIA and DXA (Pearson's $r = 0.971$, $p < 0.001$) in total body fat mass, %BF ($r = 0.832$, $p = 0.001$)</p> <p>95% confidence intervals for LOA were 2.4 ± 6.0 kg (-3.6 to 8.3 kg) and $5.3 \pm 9.6\%$ (-3.8% to 15.4%), respectively</p> <p>Correlation between BMI and fat mass determined by DXA was 0.855 ($p < 0.001$) and between BIA and BMI was 0.847 ($p < 0.001$)</p> <p>Fat mass measured using BIA was 2.4 kg lower than measurement using DXA</p> <p>Gold standard = ADP</p>
7	Azcona 2006 ²⁷⁴	BIA	Children and adolescents (5–22 years)	Mixed (stratified)	187	Europe	White	<p>BIA and ADP estimates of fat mass and fat-free mass are highly correlated for both obese and non-obese children [$R_c = 0.79$ (95% confidence interval 0.73 to 0.83)] and [$R_c = 0.96$ (95% confidence interval 0.95 to 0.97)]</p> <p>However, the LOA were -13.70 to 6.90 for fat mass and 1.40 to 7.60 for fat-free mass, suggesting that these methods should not be used interchangeably</p> <p>Gold standard = 3C model</p> <p>Compared with 3C model, BIA (Tanita) equations overestimated fat-free mass by 2.7 kg ($p < 0.001$)</p> <p>Authors derived a new equation (fat-free mass = $2.211 + 1.115$ (HT²/Z), with r^2 of 0.96, standard error of the estimate 2.3 kg, which showed no significant bias in fat mass or fat-free mass, or change in fat mass or fat-free mass</p>
8	Haroun 2009 ²⁵	BIA	Children and adolescents	Obese	77	UK	White	<p>Gold standard = 3C model</p> <p>Compared with 3C model, BIA (Tanita) equations overestimated fat-free mass by 2.7 kg ($p < 0.001$)</p> <p>Authors derived a new equation (fat-free mass = $2.211 + 1.115$ (HT²/Z), with r^2 of 0.96, standard error of the estimate 2.3 kg, which showed no significant bias in fat mass or fat-free mass, or change in fat mass or fat-free mass</p>

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
9	Okasora 1999 ²⁷⁵	BIA	Children and adolescents	Mixed (stratified)	104	Japan	Race not defined	<p>Comparator = DXA</p> <p>The %fat, fat-free mass and body fat content showed a close correlation when measured by BIA and DXA, with the correlation coefficients being 0.90, 0.95 and 0.95, respectively</p> <p>%Fat value determined by BIA tended to be lower than that determined by DXA in the overweight group; the same trend was also seen in obese children before and after therapy with exercise and diet</p>
10	Lofin 2007 ²⁷⁶	BIA	Children and adolescents	Mixed (stratified)	166	USA	African American, Hispanic, white, mixed race	<p>Comparator = DXA</p> <p>BIA was significantly related to DXA body composition parameters, but data in results section not stratified by obese, but states in discussion, BIA underestimated per cent fat in the overweight children</p>
11	Iwata 1993 ²⁷⁷	BIA	Children and adolescents	Mixed (stratified)	1216	Japan	Not defined	<p>Comparator = SFT</p> <p>%BF correlated strongly with %OB</p> <p>Sensitivity of %BF to predict %OB = 0.4–0.8 (but reduced with increasing %OB cut-off)</p> <p>Specificity of %BF to predict %OB = 0.66–0.97 (but increased with increasing %OB cut-off)</p> <p>Conclusion: BIA is a reliable way of assessing lipid storage in children</p>
12	Wabitsch 1996 ²⁴	BIA (change)	Children and adolescents	All obese	146	Switzerland	Not defined	<p>Gold standard = TBW by deuterium dilution and resistance index (BIA)</p> <p>Measured before and after weight loss programme</p> <p>Cross-sectional comparisons showed good agreement between BIA and TBW, but correlations were poor ($r = 0.21$) with change, where BIA was not accurate at predicting small changes in TBW</p>

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
13	Guida 2008 ²⁷⁸	BIA (vector distribution)	Children	Mixed (stratified)	464	Europe	Not defined	<p>No gold standard</p> <p>Compared with anthropometry and conventional BIA</p> <p>Fat measurement using tricep skinfold thickness and BIA were comparable within the different BMI ranges</p> <p>Concludes that although BMI is a reliable measure to grade overweight, it cannot differentiate whether weight change is due to variation of fat mass, fat-free mass or water</p> <p>It is important to estimate paediatric body composition using very precise and accurate measurements. The bioelectrical impedance vector analysis method may therefore be of clinical utility to enable discrimination between fat mass, fat-free mass and ECW</p> <p>Compared with 'biochemical complications'</p> <p>Only (WHR/Ht)-SDS showed high sensitivity and specificity to predict metabolic derangement. Concludes that only (WHR/Ht)-SDS can serve in the diagnostic criterion than classifies obesity in Japanese adolescent girls into two types</p> <p>Comparators = DXA (fat mass) and HF BIA</p> <p>HF BIA underestimated fat mass more than both FF. However, LOA between DXA and FF-Tanita or FF-Tefal were much greater than those obtained with the HF BIA (-7.7 and +4.3, -12.0 and +10.6 vs. 2.1 and 6.7 kg, respectively)</p> <p>Differences between FF BIA and DXA increased with WHR</p> <p>Major limiting factor was the interindividual variability in fat mass estimates of FF BIA estimates</p>
14	Asayama 2000 ²⁷⁹	%OW, WC, WHR and (WHR/Ht)-SDS	Children and adolescents	Obese	124	Japan	Not defined	
15	Lizzer 2003 ²⁸⁰	BIA (x2 FF), Tanita and Tefal	Adolescent	Overweight and obese	53	Europe	Not defined	

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
16	Eisenkolbl 2001 ²⁸¹	BIA and DXA	Children and adolescents	All obese	27	Austria	Not defined	No gold standard: %fat by BIA = ~10% lower than DXA ($r=0.91$) Biggest difference in boys; <i>t</i> -test showed significant differences Overall, concern with differences, especially in boys (three times higher) Considered DXA more accurate and suggest use of correction formula if using BIA No gold standard criterion
17	Hannon 2006 ²⁸²	BIA, SFT (triceps and calf)	Adolescents	Mixed (stratified)	198	USA	White, African American, Hispanic, Asian, multicultural, Native American	In each of gender- and race-specific groups the %BF from BIA was lower, on average, than from SFT: Caucasian girls 27.5 ± 6.5 vs. 31.9 ± 8.3 , $p < 0.001$; African American girls 30.1 ± 7.8 vs. 32.1 ± 11.2 , $p = 0.002$; Caucasian boys 20.3 ± 9.1 vs. 24.9 ± 10.5 , $p < 0.001$; African American boys 20.5 ± 8.6 vs. 22.3 ± 11.6 , $p = 0.012$ When expressed as mean difference ± 2 SD, LOA of %BF between BIA and SFT methods ranged from -11.6 to $+2.9$ in Caucasian girls, from -12.4 to $+8.2$ in African American girls, from -12.7 to $+3.4$ in Caucasian boys, and from -10.9 to $+7.3$ in African American boys <i>Conclusion</i> : Caution should be used in recommending segmental BIA devices over SFT to predict BF in adolescents <i>Comparator</i> = DXA
18	Goran 1996 ²⁸³	BIA, SFT indices, BMI	Children	Mixed (stratified)	98	USA	White ($n = 94$), Native American ($n = 4$)	Analysis failed to cross-validate existing techniques against DXA measures Authors have developed new anthropometric equations that provide accurate estimates of body fat

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
19	Ellis 1996 ²⁸⁴	BIA, TOBEC and BIS	Children and adolescents	Mixed (stratified)	99	USA	White, African American, Hispanic	<p>Comparator = DXA (%BF)</p> <p>Each method differed with respect to accuracy depending on the specific outcome</p> <p>If comparing ability to detect obese vs. non-obese, BIS identified fewer children as obese ($\chi^2 = 9.1$, $p < 0.005$). But TOBEC and DXA were similar ($\chi^2 = 5.79$, $p > 0.05$)</p> <p>If comparing overweight vs. non-overweight, BIS and DXA were similar ($\chi^2 = 0.38$, $p > 0.30$) but TOBEC differed by identifying more overweight than DXA ($\chi^2 = 7.23$, $p = 0.03$)</p> <p>Comparator = WC</p>
20	Fernandes 2007 ²⁸⁵	Bioimpedance	Adolescent	Mixed (stratified)	811	Brazil	Not defined	<p>Sensitivity of BIA to identify excess VAT = 81% (boys), 63% (girls)</p> <p>Specificity = 93% (boys), 94% (girls)</p> <p>AUC = 0.87 (boys), 0.79 (girls)</p> <p>Similar high sensitivity and specificity and AUC for identification of excess fat associated with overweight/obesity</p> <p>Also correlated well with subcutaneous fat</p>
21	Wickramasinghe 2005 ²²	BMI	Children and adolescents	Mixed (stratified)	138	Australia	White, Sri Lankan	<p>Fat mass and BMI = strongly correlated in white and Sri Lankan participants ($r \geq 0.8$), but obesity cut-offs for BMI were very poor at detecting obesity as defined by fat mass (very poor sensitivity, range = 3.5–20%)</p> <p>Gold standard = isotope dilution (deuterium D_2O – fat-free mass with 20% = obese in boys and 30% in girls)</p>

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
22	Wickramasinghe 2009 ²³	BMI	Children and adolescents	Mixed (stratified)	282	Australia	Sri Lankan	Gold standard = isotope dilution (deuterium D ₂ O – fat mass > 30% in girls, > 25% in boys) Fat mass and BMI closely correlated ($r = 0.82$ in girls, $r = 0.87$ in boys). But, although specificity was high (100%), sensitivity was very low (8–23.6%) = poor predictive ability for obesity Comparator = TOBEC (%BF) BMI and %BF, $r = 0.65$
23	Widhalm 2001 ²⁸⁶	BMI	Children and adolescents	All obese	204	Austria	White	In boys < 10 years, 73% of variance in %BF was explained by BMI (63% in girls) Poorer in older children and increased variation; therefore, not a good indicator on an individual basis but OK on population basis Comparator = SFT
24	Gaskin 2003 ²⁸⁷	BMI	Child	Mixed (stratified)	306	Jamaica	Not defined	High degree of misclassification with low sensitivity (2–38% in 7- to 8-year-old boys) and higher specificity = (100% in 7- to 8-year-old boys) Higher sensitivity in girls (10–66%) and older children (67–86%) but still a high degree of misclassification Comparators = DXA and skin fold
25	Warner 1997 ²⁸⁸	BMI	Children and adolescents	Mixed (non-stratified)	143	UK	Not defined	Assessment in children in disease states that are expected to alter body composition Sensitivity = 66%, specificity = 94% (with DXA) Sensitivity = 50%, specificity = 100% (with skin fold) Similar in children with and without disease but both stating BMI underpredictions

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
26	Pietrobelli 1998 ²⁸⁹	BMI	Children and adolescents	Mixed (stratified)	188	Europe	White	Comparator = DXA BMI was strongly associated with total body fat [$r^2 = 0.85$ (boys), $r^2 = 0.89$ (girls)] and %BF [$r^2 = 0.63$ (boys), 0.69 (girls)]
27	Glaner 2005 ²⁹⁰	BMI	Children and adolescents	Mixed (stratified)	1410	Brazil	Not defined	Confidence limits on BMI-fatness association were wide, with individuals of similar BMI showing large differences in total body fat and %BF Comparator = SFT (TR + CA) Kappa index showed weak agreement between the three classifications of body fat as estimated by BMI and categorised by SFT
28	Reilly 2000 ²⁹¹	BMI	Child	Mixed (stratified)	4175	UK	Race not defined	Only 48.98% of girls and 57.32% of boys were classified correctly or concomitantly by both procedures Conclusion: BMI does not present consistence in order to classify girls and boys in relations to body fat Comparator: BIA Obesity definition based on BMI (95th centile) had moderately high sensitivity (88%) and high specificity (94%) Sensitivity and specificity did not differ significantly between boys and girls Receiver operating curve analysis showed that lower cut-offs applied to the BMI improved sensitivity with no marked loss of specificity; the optimum combination of sensitivity (92%) and specificity (92%) was at a BMI cut-off equivalent to the 92nd centile The IOTF cut-off was much lower leading to potential underestimation of obesity prevalence

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
29	Potter 2007 ²⁹²	BMI	Children and adolescents	Mixed (stratified)	1671	UK	Caucasian	<p>Comparator: BIA</p> <p>Using BMI, 5.6% males and 6.1% females were identified as obese</p> <p>BIA (%fat) gave higher values for obesity: 11.9% males, 15.3% females</p> <p>Conclusion: BMI underestimates</p>
30	Ochiai 2010 ²⁹³	BMI	Children and adolescents	Mixed (stratified)	3750	Japan	Race not defined	<p>Comparator: %Fat by BIA</p> <p>In fourth graders, correlation in boys was 0.74 and 0.97 for girls</p> <p>Similar results were obtained for seventh graders</p> <p>However, when stratified by obese correlations for boys were < 0.5 but for girls were > 0.7</p> <p>The study also compared BMI to WC ($r = 0.94$ in boys and $r = 0.90$ in girls)</p>
31	Morrissey 2006 ²⁹⁴	BMI	Adolescents	Mixed (stratified)	416	USA	White, African American, other (not stated)	<p>Conclusions: BMI is positively correlated with BIA and WC but results are influenced by obesity</p> <p>Comparator: Measured BMI</p> <p>Mean self-reported BMI (22.8 kg/m^2) was significantly lower than mean measured BMI (23.3 kg/m^2)</p> <p>Students who were at risk for overweight and those who were overweight were more likely to underestimate their BMI than students who were normal weight</p> <p>Approximately 17% of students were misclassified in BMI categories when self-reported data were used</p>

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
32	Molina 2009 ²⁹⁵	Parent-reported BMI	Child	Mixed (stratified)	538	Brazil	White, non-white	Comparator: Measured BMI Kappa value between parent report and actual BMI was 0.217 ($p < 0.000$)
33	Maynard 2003 ²⁹⁶	Parent-reported BMI	Child	Mixed (stratified)	5500	USA	White, African American, Hispanic	Only 33% of overweight children were correctly classified and only 10.4% of obese were correctly classified as obese Comparator: Measured BMI 65% of overweight boys and 69% overweight girls were correctly classified by their mothers
34	Mast 2002 ²⁹⁷	BMI	Child	Mixed (stratified)	2286	Germany	Race not defined	Nearly one-third of mothers misclassify overweight children as being lower than their measured weight status Comparators: SFT and BIA
35	Malina 1999 ²⁹⁸	BMI	Children and adolescents	Mixed (stratified)	1570	USA	White, Hispanic, African American, Asian	By contrast, BMI reached higher sensitivity to screen for obese children: 0.83 to 0.85 for boys, and 0.62 to 0.80 for girls, at a concomitant specificity of 0.95 to 0.98 for boys, and 0.96 to 0.97 for girls, as defined by assessment of body fat mass Comparator: TSF and %BF from densitometry BMI had high specificities (86.1–98.8% for risk of overweight and 96.3–100% for presence of overweight) and lower but variable sensitivities (4.3–75.0% for risk of overweight and 14.3–60% for presence of overweight), and so those at risk of overweight or who were overweight were not correctly identified as measured by BMI

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
36	Ellis 1999 ²⁹⁹	BMI	Children and adolescents	Mixed (stratified)	979	USA	White, African American, Hispanic	Comparator = DXA (%fat) $R^2 = 0.34-0.70$ ($p < 0.0005$), SE for % fat = 4.7–7.3% of body weight – indicating a poor prediction at the individual level, but good for population-based level (results differed by gender and ethnicity)
37	Duncan 2009 ³⁰⁰	BMI (IOTF and CDC cut-offs)	Children and adolescents	Mixed (stratified)	1676	New Zealand	European, Pacific Island, Maori, East Asian, South Asian	Areas under ROC curves ranged from 89.9% to 92.4%, suggesting that BMI is an acceptable screening tool for identifying excess adiposity. However, IOTF and CDC thresholds showed low sensitivity for predicting excess %BF in South Asian and East Asian girls, with low specificity in Pacific Island and Maori girls <i>Conclusion:</i> BMI can be an acceptable proxy measure of excess fatness in girls from diverse ethnicities, especially when ethnic-specific BMI reference points are implemented
38	Rush 2003 ²¹	BMI and BIA	Children and adolescents	Mixed (stratified)	172	New Zealand	New Zealand European, Maori, Pacific Island	Gold standard = TBW by deuterium dilution (fat-free mass) Regression analyses provided an equation to determine body fatness from BIA that was more suitable and robust than BMI across this sample Comparator = DXA Sensitivity = 69–96%; specificity = 83–96%
39	Bartok 2011 ³⁰¹	BMI percentile for estimation of fat mass	Children and adolescents	All obese	197	USA	White (all girls)	Relative fat mass is fairly constant between 0 and the 40th BMI percentile but then increases as BMI percentile increases thereafter <i>Conclusion:</i> Age-specific BMI percentile is a useful clinical and research tool for classifying white girls as either over-fat or obese during childhood and adolescence

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
40	El Taguri 2009 ³⁰²	BMI z-score	Children and adolescents	All obese	748	France	White, other not defined	Comparator = DXA (fat mass) Predicted fat mass in high agreement (99.8%) with measured fat mass FMI (DXA) and BMI were correlated ($R^2 = 0.77$)
41	Yoo 2006 ³⁰³	BMI, PWH	Child	Mixed (stratified)	892	Korea	Korean	All correlations (by age and gender) were high Comparator = BIA (%BF with > 35% = obese) BMI and %BF, $r = 0.91$; PWH and %BF $r = 0.92$ PWH sensitivity = 0.91, specificity = 0.88 BMI (IOTF) sensitivity = 0.46, specificity = 0.99
42	Eto 2004 ³⁰⁴	BMI, FMI	Child	Mixed (stratified)	486	Japan	Not defined	Local BMI cut-off sensitivity = 0.7, specificity = 0.79 Comparator = BIA (%fat mass) Obesity defined at $\geq 20\%$ fat mass (boys) $\geq 25\%$ (girls) BMI sensitivity = 30.4–37.5%, specificity = 95.5–96.4% FMI sensitivity = 42.9–68%, specificity = 99.5–100%
43	Rolland-Cachera 1982 ³⁰⁵	BMI, height/weight ² , height/weight ³	Children and adolescents	Mixed (stratified)	117	Europe	Not defined	BMI should be used in caution because of poor sensitivity FMI may be better Comparator: Subscapular SFT Conclusion: The Quetelet index (height/weight ²) is better for estimating adiposity in children of both sexes than height/weight or height/weight ³ The authors identify a number of caveats, however

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
44	Sampei 2001 ³⁰⁶	BMI, NIR and Slaughter's skinfold equation)	Children and adolescents	Mixed (stratified)	436	Brazil	Japanese, Caucasian	No gold standard In 10- to 11-year-old girls BMI was significantly correlated with other methods In Japanese: BMI × NIR = 82.3%, BMI × BIA = 85.7% In Caucasian adolescents: BMI × NIR = 80.7%, BMI × BIA = 87.4% In the 16- to 17-year-old adolescents, the BMI demonstrated low or no correlation with other methods <i>Conclusions:</i> BMI can be used in place of other methods in 10- to 11-year-olds, although it may underestimate obesity. In 16- to 17-year-olds it is not a suitable index focusing on identification of obesity <i>Comparator:</i> DXA or SFT
45	Mei 2002 ³⁰⁷	BMI, Rohrer index and weight-for-height index	Children and adolescents	Mixed (stratified)	920	USA, Italy, New Zealand	White, black	BMI for age was significantly better than were weight-for-height index and Rohrer index for age in detecting overweight when average SFTs were used as the standard BMI for age was significantly better than was Rohrer index for age in detecting overweight when DXA was standard, but there was no difference between BMI and weight-for-height index

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
46	Sardinha 1999 ²⁹	BMI, SFT (triceps), arm girth	Children and adolescents (10–15 years)	Mixed (stratified)	328	Europe	White	<p>Comparator: DXA</p> <p>In assessing the ability of the anthropometric variables to discriminate obesity from non-obesity as assessed by DXA with cut-offs – true-positive rates ranged from 67% to 87% and from 50% to 100% in girls and boys, respectively, and false-positive rates ranged from 0% to 19% and from 5% to 26%, respectively</p> <p>For children aged 10–11 years, the AUCs for ROCs were close to 1.0, suggesting very good accuracy</p> <p>For older boys and girls, AUCs for triceps SFT were similar to, or greater than, AUCs for BMI and upper arm girth</p> <p><i>Conclusions:</i> Triceps SFT gives the best results for obesity screening in adolescents aged 10–15 years. BMI and upper arm girth were reasonable alternatives, except in 14- to 15-year-old boys in whom both indices were only marginally able to discriminate obesity</p> <p>Comparators: BIA and %BF</p>
47	Himes 1999 ^{30E}	BMI, SFT indices, WC	Adolescents	Mixed (stratified)	625	USA	White	<p>The fat-test youth in each age and gender group were considered those in > 80th centile for the indicator</p> <p>Agreement determined by kappa coefficients</p> <p>Kappa among indicators range from 0.57 to 0.85 for males, and from 0.56 to 0.79 for females</p> <p>Categorical agreement with the fat-test youth by %BF changes considerably with age for most indicators, suggesting that relationships among indicators change during adolescence</p> <p><i>Conclusions:</i> Difference indicators may identify different subpopulations as the fat-test – therefore caution should be used in interpretation of results from different indicators</p>

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
48	Nuutinen 1991 ³⁰⁹	BMI, triceps skinfold, subscapular skinfold	Children and adolescents	Mixed (stratified)	3596	Finland	White	No gold standard Results found that fewer children were classified as obese when two criteria were used together than when they were used individually BMI and triceps or subscapular skinfolds vary in sensitivity and specificity as indicators of obesity <i>Comparator:</i> DXA
49	Mei 2007 ³¹⁰	BMI, triceps, and subscapular skinfold	Children and adolescents	Mixed (stratified)	1196	USA	White, African American, Hispanic, Asian	All three measurements did well in ROC curve in identifying excess body fat defined by either the 85th or 95th percentile of %BF by DXA. But if BMI for age was already known, and was > 95th percentile, the additional measurement of skinfolds did not significantly increase the sensitivity or specificity in the identification of excess body fat
50	Glasser 2011 ³¹¹	BMI, WC	Children and adolescents	Mixed (stratified)	2132	Europe	Not defined	Skinfold measurements do not seem to provide additional information about excess body fat beyond BMI for age alone if the BMI for age is 95th percentile <i>Comparator</i> = SFT ROC curves to evaluate performance of BMI and WC in reflecting excess fatness – AUCs > 0.9 for both sexes indicating good performance The specificity for all references systems were high for both sexes (95–98%). However, sensitivities were low (53–67% in boys; 51–67% in girls) <i>Conclusion:</i> Results support use of BMI-based references for monitoring in epidemiological studies but sample based cut-offs should be refined for clinical use on national level

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
51	Neovius 2005 ³¹²	BMI, WC and WHR	Adolescents	Mixed (stratified)	474	Sweden	Race not defined	Gold standard: ADP For overweight and obesity in boys and obesity in girls, the AUC ROC curve was high (0.96–0.99) for BMI and WC WHR was not significantly better than chance as diagnostic test for obesity in girls For BMI and WC, highly sensitive and specific cut-offs for obesity could be derived Conclusion: BMI and WC were found to perform well as diagnostic tests for fatness, whereas WHR was less useful Comparators = Cardiovascular and metabolic risk factors BMI cut-offs for overweight: sensitivity = 58.8–75%, specificity = 60–71.2% Cut-offs for obesity sensitivity = 9.3–52.6%, specificity = 94.4–99.7% High specificity for BMI to predict obesity but not sensitivity Prediction of obesity via cardiovascular risk factors (blood lipids, blood pressure, CRP, metabolic syndrome) Sensitivity and specificity analysis Good AUC for all except WHC, best = BMI
52	Adegboye 2010 ³¹³	BMI, WC, WHR	Children and adolescents	Mixed (stratified)	2835	Denmark, Portugal, Estonia	Not defined	
53	Jung 2009 ³¹⁴	BMI, WC, WHR	Adolescent (all boys)	Mixed (stratified)	79	Germany	White	

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
54	Fujita 2011 ³¹⁵	BMI, WC, WHtR	Children	Mixed (stratified)	422	Japan	Not defined	<p>Comparator: DXA</p> <p>AUCs were ≥ 0.98 for BMI, WC, and WHtR as indicators of excess abdominal fat (≥ 95th percentile) for both sexes</p> <p><i>Conclusion:</i> Sensitivity and specificity of BMI, WC and WHtR as indicators of excess abdominal fat were high for both sexes</p>
55	Marshall 1990 ²⁷	BMI, weight, O-Scale, SFT	Children and adolescents	Mixed (stratified)	533	Canada	Race not defined	<p>Comparator: Visual inspection</p> <p>All measures show good accuracy $> 93\%$</p> <p>BMI was most sensitive¹⁰¹ and the O-Scale was most specific (98.1)</p> <p>Comparator: BMI</p> <p>Broselow estimates were within 10% of actual weight 63% of the time, physician estimates were within 10% of the actual weight 43% of the time and hybrid estimates 55% of the time</p> <p>Based on average mean per cent error, compared with actual weight, Broselow estimates differed by 10.8% (95% confidence interval 9.7% to 12%), hybrid estimate by 11.3% (95% confidence interval 10.3% to 12.2%) and physician estimate by 16.2% (95% confidence interval 14.7% to 17.7%)</p> <p>The Broselow estimates were significantly worse than physician estimate for obese patients: 26.4% (95% confidence interval 19.7% to 33.1%) vs. 16% (95% confidence interval 12.3% to 19.8%)</p> <p><i>Conclusion:</i> Broselow tape generally has greater agreement with actual weight than physician visual estimate, except for obese children</p>
56	Rosenberg 2011 ³¹⁶	Broselow tape measurement	Children and adolescents	Mixed (stratified)	372	USA	White, black, Hispanic	

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
57	Killion 2006 ³¹⁷	Child figure silhouettes	Child	Mixed (stratified)	192	USA	African American, Hispanic	<p>Comparator: Measured BMI</p> <p>Estimated BMI from silhouettes mathematically</p> <p>Mothers perceived BMI (mean = 15.0, standard deviation = 0.66) of their children were less than the actual BMI (mean = 16.7, standard deviation = 1.84) of their children ($t = 15.77$; $p = 0.0001$). This was dependent on the actual weight status of children ($\chi^2 = 7.13$, $p = 0.008$)</p> <p>Gold standard: DXA</p>
58	Lazzer 2008 ²²³	ADP and BIA	Children and adolescents	All obese	58	Italy	Not defined	<p>ADP body fat estimated from body density using equations [Siri (ADPSiri) and Lohman (ADPLOhman)]</p> <p>Bland-Altman test: showed that ADPSiri and ADPLOhman underestimated %fat mass by 2.1% and 3.8% ($p = 0.001$). BIA underestimated %fat mass by 5.8% ($p = 0.001$). A new prediction equation [fat-free mass (kg) = 0.87 (stature squared/body impedance) + 3.1] was developed and cross-validated on an external group of obese children and adolescents ($n = 61$)</p> <p>Difference between predicted and measured fat-free mass in the external group was 21.6 kg ($p = 0.001$) and fat-free mass was predicted accurately (error, 5%) in 75% of subjects</p> <p>Gold standard: 4C model</p> <p>21 children were too big to be scanned</p> <p>DXA overestimated fat mass and underestimated LBM</p> <p>LOA were wide in change ($n = 66$ had second measure 1 year later)</p> <p>%Variance explained by DXA was 76% for change in fat mass and 43% for change in LBM</p>
59	Wells 2010 ¹¹⁶	DXA	Adolescent (<21 years)	All obese	174	UK	White, black, Asian	

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
60	Gately 2003 ¹⁷	DXA, ADP (Siri and Loh), TBW (Siri and Loh)	Adolescent	Overweight and obese	30	UK	Not defined	Gold standard: 4C model All estimates of % fat were highly correlated with that of the 4C model ($r \geq 0.95$, $p < 0.001$; $SE \leq 2.14$). For % fat, the total error and mean difference $\pm 95\%$ LOA compared with 4C model were 2.5, 1.8 \pm 3.5 (ADPSiri); 1.82, 0.04 \pm 3.6 (ADPLLoH); 2.86, -2.0 \pm 4.1 (TBW73); 1.9, -0.3 \pm 3.8 (TBWLoH) and 2.74, 1.9 \pm 4.0 (DXA)
61	Fors 2002 ³¹⁸	DXA, BIA and multifrequency bioelectrical impedance spectroscopy (BIS)	Children and adolescents	Mixed (stratified)	61	Sweden	Not defined	No gold standard Estimated fat-free mass, body fat mass and per cent fat Correlations between measures for all of these were high ($r = 0.73$ – 0.96) but with wide LOA
62	Springer 2011 ³¹⁹	GRE imaging for ILC	Adolescents	All obese	29	Germany	Race not defined	BIA overestimated fat mass in lean and underestimated fat mass in overweight subjects more than BIS, compared with DXA Comparator: MRS Correlations $r = 0.78$ – 0.86 , with no regional differences Ability of GRE to accurately predict ILC content of > 5% was good, with positive likelihood ratio of 11.8 and negative likelihood ratio of 0.05

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
63	Ball 2006 ³²⁰	Height, weight SFT, WC, hip circumference as predictors of VAT/SAT	Children and adolescents	Overweight and obese	196	USA	Latino	<p>Comparator: MRI (VAT and SAT)</p> <p>Strongest univariate correlate for VAT was WC ($r = 0.65$, $p < 0.01$), where strongest correlate for SAT was hip circumference ($r = 0.88$, $p < 0.001$)</p> <p>Regression analyses showed 50% of the variance in VAT was explained by WC (43.8%), Tanner stage (4.3%) and calf skinfold (1.7%)</p> <p>Variance in the SAT model was explained by WC (77.8%), triceps skinfold (4.2%) and gender (2.3%)</p> <p>Although mean differences between measured and predicted VAT and SAT were small, there was a large degree of variability at the individual level, especially for VAT</p> <p><i>Conclusions:</i> Both VAT and SAT prediction equations performed well at group level but the relatively high degree of variability suggest limited clinical utility of the VAT equation. MRI is needed to derive an accurate measure of VAT at the individual level</p> <p><i>Comparators:</i> Measured height and weight</p> <p>Mean weight error increased with age ($p < 0.001$), was higher among girls and black children, and mean weight error also increased with age-specific BMI z-score ($r = 0.32$, $p < 0.001$)</p> <p><i>Conclusion:</i> Twenty-one per cent of obese children would not be identified by using parent-reported data to calculate the BMI</p>
64	O'Connor 2011 ³²¹	Parent-reported height and weight	Children and adolescents	Mixed (stratified)	1430	USA	White, black, Hispanic, Asian	

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
65	Rasmussen 2007 ³²²	Self-report height and weight	Adolescents	Mixed (stratified)	2726	Sweden	Race not defined	<p>Comparators: Measured height and weight</p> <p>Obese boys under-reported their weight (5.2 kg) more than obese girls (3.8 kg)</p> <p>Agreement between self-reported and measured BMI-categories (obese, overweight and normal), as estimated by weighted kappa, was 0.77 for girls and 0.74 for boys</p> <p>Obese girls and boys sensitivity of self-reports were 0.65 and 0.52</p> <p>Conclusion: Thirty-five per cent of obese girls and 48% of obese boys would remain undetected from self-reported data</p>
66	Asayama 2000 ²⁷⁹	Height, weight, BW, WC, hip circumference, triceps and subscapular SFT. CT: TAF, VAT, SAT	Children and adolescents	Obese	75	Japan	Not defined	<p>Comparators: Blood biochemistry indicators of metabolic derangement</p> <p>VAT area was the best diagnostic criterion, although this was an age-dependent variable. VAT/SAT was a little less sensitive and was less closely associated with blood biochemistry than VAT area was but was independent of age</p> <p>Conclusion: Results suggest that the threshold values for VAT and TAF areas, VAT/SAT and sagittal diameter can be used for classifying the obese boys into two types – those with medical problems and those without</p>

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
67	Lu 2003 ³²³	Leg-leg	Children and adolescents	All obese	64	China	Race not defined	<p>Comparator: DXA</p> <p>In all subjects, estimates of fat-free mass, fat mass and %BF were highly correlated ($r = 0.85-0.95$) between the two methods</p> <p>Bland-Altman comparison showed wide LOA between the methods</p> <p>Despite the high correlations comparing with DXA, the leg-leg BIA might overestimate the fat mass and %BF in serious obese children</p> <p>Comparator: Measured</p>
68	Dubois 2007 ³²⁴	Maternal report of height and weight (BMI)	Children	Mixed (stratified)	1464	Canada		<p>This study indicates that mothers overestimate their children's weight more than their height, resulting in an overestimation of overweight children of > 3% in the studied population</p> <p>Conclusion: The results emphasise the importance of collecting measured data in childhood studies of overweight and obesity at the population level</p> <p>Comparator: BIA</p>
69	Gillis 2000 ³²⁵	Mathematical index for assessing changes in body composition	Children and adolescents	Obese	67	Canada	Not defined	<p>The mathematical index was valid for assessing changes in %BF of obese children and adolescents over time</p> <p>Conclusion: The index could be used by clinicians who lack body composition equipment to need a quick method to analyse effectiveness of a weight control programme in obese children and adolescents</p>

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
70	Nafiu 2010 ³²⁶	Neck circumference	Children and adolescents	Mixed (stratified)	1 102	USA	Race not defined	<p>Comparators: BMI and WC</p> <p>Neck circumference was significantly correlated with BMI (0.73) and WC (0.73) in both boys and girls</p> <p>Optimal neck circumference cut-off, indicative of high BMI in boys, ranged from 28.5 to 39.0 cm; corresponding values in girls ranged from 27.0 to 34.6 cm</p>
71	Akinbami 2009 ³²⁷	Parent-reported height and weight	Children and adolescents	Mixed (stratified)	12261	USA	White, black, Hispanic	<p>Comparators: Measured height and weight</p> <p>Parents overestimate in younger children but underestimate in older children</p> <p>Largest discrepancies were with height</p> <p>Conclusion: Parents are poor indicators</p>
72	Huybrechts 2006 ³²⁸	Parent-reported height and weight	Child	Mixed (stratified)	297	Belgium	Belgian	<p>Comparators: Measured height and weight</p> <p>Sensitivity = 47% (national BMI cut-off) and 44% (international BMI cut-off for overweight)</p> <p>Specificity = 94% and 95%</p> <p>> 50% overweight children and > 75% of the obese children would be missed with the use of parentally reported weight and height values; 70% of underweight children could be encouraged wrongly to gain weight</p> <p>The bias of parent-reported BMI values = significantly greater when weight and height were both guessed, rather than being measured at home</p>

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
73	Huybrechts 2011 ³²⁹	Parent-reported height and weight	Child	Mixed (stratified)	297	Belgium	Belgian	<p>Comparators: Measured height and weight</p> <p>Sensitivity = for underweight and overweight/obesity were, respectively, 73% and 47% when parents measured their child's height and weight, and 55% and 47% when parents estimated values without measurement</p> <p>Specificity for underweight and overweight/obesity = respectively 82% and 97% when parents measured the children, and 75% and 93% with parent estimations</p> <p>Conclusion: Parents measurements at home are better than estimations</p>
74	Garcia-Marcos 2006 ³³⁰	Parent-reported height and weight for defining obesity	Children	Mixed (stratified)	818	Europe	Country of origin: Spain	<p>Comparators: Measured height and weight</p> <p>Bias (minus reported real) was, respectively, for non-asthmatics and asthmatics: weight 0.42 kg (95% confidence interval 0.24 to 0.59 kg) vs. 0.97 kg (0.50 to 1.44 kg); height 2.37 cm (2.06 to 2.68 cm) vs. 2.87 cm (1.87 to 3.87 cm); BMI -0.39 kg/m^2 (-0.52 to 0.23 kg/m^2) vs. 0.23 kg/m^2 (-0.58 to 0.13 kg/m^2)</p> <p>Conclusions: Reported weights and heights had large biases, comparable between parents of both asthmatic and those of non-asthmatic children. However, this information could be reasonably valid for classifying children as obese or non-obese in large epidemiological studies</p>
75	Jones 2011 ³³¹	Parent-reported weight status	Child	Mixed (stratified)	536	UK	White	<p>Comparator: Measured BMI/obesity (IOTF)</p> <p>7.3% of children perceived as overweight/very overweight compared with 23.7% measured</p> <p>69.3% of parents of overweight or obese children identified their child as being of normal weight</p>

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
76	Vuorela 2010 ³³²	Parent-reported weight status	Child	Mixed (stratified)	606	Finland	Not defined	<p>Comparator: Measured weight (obesity with IOTF criteria)</p> <p>In 5-year-olds and 11-year-olds</p> <p>Accuracy to detect normal weight was high, but most parents of overweight in 5-year-olds misclassified as normal weight</p> <p>50% misclassified in 11-year-olds</p> <p>Similar with WC (i.e. good specificity but poor sensitivity)</p> <p>Comparators: Measured height and weight</p> <p>31% of parents underestimated weight status (46% of the parents of overweight children)</p> <p>Comparators: Measured height and weight</p> <p>Pearson's correlation coefficients between measured and reported were 0.91, 0.92 and 0.79 for body weight, height and BMI, respectively</p> <p>> 92% of the parents reported body weight of their child within 10% of measured body weight and 72% within 5% of measured body weight</p> <p>Almost 99% of the parents reported height of their child within 5% of measured height</p> <p>15.1% of girls and 11.8% of boys were overweight when measured data used; 11.9% of girls and 7.1% of boys were overweight when reported data used</p> <p>Conclusion: Overweight prevalence rates in children are underestimated when based on reported weight and height</p>
77	Tschamler 2010 ³³³	Parent-reported weight status	Infants and children	Mixed (stratified)	193	USA	White, African American, Hispanic, other (not defined)	
78	Scholtens 2007 ²²⁸	Parental report height and weight	Children	Mixed (stratified)	864	Europe	Not defined	

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
79	Wen 2011 ³³⁴	Parental reported height and weight	Adolescent	Mixed (stratified)	2143	China	Chinese	<p>Comparators: Measured height and weight $\kappa = 0.22$ (poor) and affected by gender (of child and parent) and perception of own weight</p> <p>Comparator: Measured height and weight</p> <p>ANOVA = highly significant variance between difference between parental report BMI and measured BMI and the actual weight status classification [$F(3, 1173) = 40.13, p < 0.001$ with a strong linear component]</p> <p>The absolute percentile BMI raw score differences were largest among underweight children [M (means statistic for a Games-Howell post-hoc analysis) = 27.21] and grew progressively smaller among normal (M = 20.7), at risk (M = 12.5) and overweight (M = 6.95) children</p> <p>Relationship between perceived and actual BMI percentiles scores was strongest for those children who classified as normal ($r(606) = 0.45, p < 0.001$)</p> <p>No relationship to be found for those who are classified as underweight or at risk – significant, although weaker, relationship for overweight children</p> <p>Comparators: Measured height and weight, WC</p> <p>Girls overestimated body size compared with BMI but not WC</p> <p>Parents report of body size (BMI) was more accurate</p> <p>Estimates of WC was more accurate than BMI. WC agreed best with perception of body size</p> <p>Authors advocate the use of WC</p>
80	Akerman 2007 ³³⁵	Parent-reported height and weight	Children and adolescents	Mixed (stratified)	1205	USA	African Americans, Caucasians, Hispanics, other	
81	VanVliet 2009 ³³⁶	Self-report and parent-reported height and weight and WC	Adolescent (all girls)	Mixed (stratified)	304	Finland	Not defined	

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
82	Goodman 2000 ²⁶	Self-report and parental report of BMI	Adolescents	Mixed (stratified)	11 495	USA	White, black, Hispanic, Asian/Pacific Islander	<p>Comparators: Measured height and weight</p> <p>Correlation between measured and self-reported height was 0.94, weight was 0.95 and BMI was 0.92 ($p < 0.0005$)</p> <p>Specificity, sensitivity, positive predictive value and negative predictive value were all high (0.996, 0.722, 0.860, 0.978, respectively)</p> <p>Conclusion: Studies can use self-reported height and weight to understand teen obesity</p>
83	Seghers 2010 ³⁷	Self-report height and weight	Children	Mixed (stratified)	798	Europe	Not defined	<p>Comparators: Measured height and weight</p> <p>The <i>t</i>-tests between measured and self-reported height, weight and BMI – significant differences except for height in girls. BMI derived from self-reported data was underestimated by $0.47 \pm 1.79 \text{ kg/m}^2$. Children who were overweight or obese underestimated their weight and BMI to a greater degree than normal weight/underweight children. Cohen's <i>d</i> values were all < 0.20</p> <p>Conclusion: Children aged 8–11 years were not able to accurately estimate their actual height and weight, leading to erroneous estimating rates of their weight status</p>
84	Jansen 2006 ³⁸	Self-reported height and weight (BMI)	Adolescent	Mixed (stratified)	499	Europe	Country of origin: Dutch, Surinam, Dutch Antillean, Moroccan, Turkish	<p>Comparators: Measured height and weight</p> <p>Self-report weight, height and BMI were considerably underestimated ($r = 0.85$, $r = 0.8$, 0.75, respectively, $p < 0.001$)</p> <p>Underestimation was higher in pupils who regarded themselves as more fat, those who were of non-Dutch origin and in lower education levels</p> <p>An adjustment could be applied, but new formulae need to be drawn up for each new sample</p>

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
85	Zhou 2010 ³³⁹	Self-reported height and weight	Adolescents	Mixed (stratified)	1761	China	Chinese	Comparators: Measured height and weight Sensitivity= 56.1%, specificity = 98.6% Even although correlations were high (r = 0.91 for height, r = 0.94 for BMI), overall, self-report is a poor measure because of sensitivity
86	Yan 2009 ³⁴⁰	Self-reported height and weight	Adolescents	Mixed (stratified)	2195	USA	White, black, Hispanic, other not defined	Comparators: Measured height and weight Weight status misclassified in 25% of girls and 33% of boys $\kappa = 0.31$ (boys) 0.5 (girls)
87	Fonseca 2010 ³⁴¹	Self-reported height and weight	Adolescent	Mixed (stratified)	462	Portugal	Not defined	Misclassification varied by age, gender and marital status of parent Comparators: Measured height and weight Prevalence of normal weight, overweight and obesity based on self-report compared with that of measured values was not significantly different for boys and girls, and among age groups but BMI was underestimated, with large LOA Self-report not suggested on an individual level

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
88	Enes 2009 ³⁴²	Self-reported height and weight	Children and adolescents	Mixed (stratified)	360	Brazil	Not defined	<p>Comparators: Measured height and weight</p> <p>Sensitivity of estimated BMI based on reported measures to classify obese subjects = boys (87.5%) girls (60.9%)</p> <p>Specificity = girls (92.7%) = boys (80.6%)</p> <p>Positive predictive value was high only for classification of normal-weight adolescents</p> <p>10% of obese boys and 40% of obese girls remained unidentified using only self-reported measures</p> <p><i>Conclusion:</i> Self-reported in adolescents do not present valid measures</p> <p><i>Comparators:</i> Measured height and weight</p>
89	Crawley 1995 ³⁴³	Self-reported height and weight	Adolescents	Mixed (stratified)	1211	UK	Not defined	<p>Self-reported data used to calculate BMI would result in a lower estimate of overweight</p> <p>Self-assessment of body fatness (but no other personal or demographic variable) was influential on the height and weight reporting of females in this study</p> <p><i>Comparators:</i> Measured height and weight</p>
90	Linhardt 2010 ³⁴⁴	Self-reported height and weight	Adolescents	Mixed (stratified)	517	Israel	Jews, Non-Arab Christians and Arabs	<p>Comparators: Measured height and weight</p> <p>Only 54.9% of overweight/obesity children classified correctly, whereas 6.3% of normal-weight children were wrongly classified as overweight/obese</p> <p>Largest difference in BMI = obese females (4.40 ± 4.34) followed by overweight females (2.18 ± 1.95)</p> <p>Similar findings were observed for males, where the largest difference was found among obese (2.83 ± 3.44)</p>

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
91	Lee 2006 ³⁴⁵	Self-reported height and weight	Children and adolescents	All obese	77	USA	White, Hispanic	Comparators: Measured height and weight Intraclass correlation coefficient = 0.64 to 0.95 (boys, with bias -1.6 ± 6.7); 0.49 to 0.84 (girls with bias 0.2 ± 9.2); papers also evaluated self-assessment of pubertal development
92	Wang 2002 ³⁴⁶	Self-reported height and weight	Adolescent	Mixed (stratified)	572	Australia	Not defined	This obese sample sign underestimated height, but reproducibility of the self-reported weight or height was good or excellent Comparators: Measured height and weight Height over-reported, weight under-reported (both significantly different) Differences were greater in overweight/obese
93	Tsigilis 2006 ³⁴⁷	Self-reported height and weight	Adolescent	Mixed (stratified)	300	Greece	Not defined	Misclassification = 31% (boys) and 30% (girls) Comparators: Measured height and weight High correlation between estimated and measured, but large bias for weight (0.36) and BMI (0.31), with overweight/obese underestimating both
94	Tokmakidis 2007 ³⁴⁸	Self-reported height and weight	Children and adolescents	Mixed (stratified)	676	Greece	Greek, Albanian	Comparators: Measured height and weight Prevalence estimates for overweight = 23.1% and obese = 4.3% Measured = 28.8% and 9.5%, respectively
95	Strauss 1999 ²²⁷	Self-reported height and weight	Adolescent	Mixed (stratified)	1657	USA	White, African American, Hispanic, other (not defined)	Comparators: Measured height and weight Good correlations in boys and girls (but girls less accurate): all $r > 0.8$ Greater misclassification in obese but overall correct classification = 94%

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
96	Shields 2008 ³⁴⁹	Self-reported height and weight	Adolescents	Mixed (stratified)	4535	Canada	Not defined	<p>Comparators: Measured height and weight</p> <p>Sensitivity to predict obesity = 56.6%, specificity = 99%</p> <p>Paper describes many correlations and includes adults. These are specific to prediction of obesity in age 12–24 years</p> <p>Comparators: Measured height and weight</p>
97	Abalkhail 2002 ³⁵⁰	Self-reported height and weight	Children and adolescents (9–21 years)	Mixed (stratified)	1167	Saudi Arabia	Not defined	<p>In all students, mean weight was significantly under-reported ($p < 0.05$) and mean height significantly over-reported ($p < 0.001$)</p> <p>Underestimation of weight differed with age, sex, nutritional status and maternal educational level. Females were more likely to under-report their weight than males. Underestimation of weight was reported by obese girls, in the 6- to 21-year group, in those with high SES and born from highly educated mothers</p> <p>Comparators: Measured height and weight</p>
98	Hauck 1995 ³⁵¹	Self-reported height and weight	Adolescents	Mixed (stratified)	806	USA	American Indian	<p>Comparators: Measured height and weight</p> <p>Pearson's correlation between measured and self-reported weight, height and BMI were high for males (0.95, 0.83 and 0.88, respectively)</p> <p>For females, the correlation between measured and reported weight was high (0.90) but for height the correlation was low (0.62), resulting in an intermediate correlation for BMI (0.79)</p> <p>Conclusions: Self-reported weights and heights should not be asked in surveys of American Indian adolescents when the purpose of the survey is to obtain accurate estimates of the prevalence of overweight and other weight categories. Self-reported weights and heights may be used cautiously for other analytical purposes</p>

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
99	Bae 2010 ³²	Self-reported height and weight (BMI)	Children and adolescents	Mixed (stratified)	379	Korea	Not defined	<p>Comparators: Measured height and weight</p> <p>Self-reported weight and BMI tended to be underestimated</p> <p>The prevalence estimate of obesity based on self-report data (10.6%) was lower than that based on directly measured data (15.3%)</p> <p>The estimated sensitivity of obesity based on self-reported data was 69% and the specificity was 100%</p> <p>The value of kappa was 0.79 (95% confidence interval 0.70 to 0.88)</p> <p>Comparators: Measured height and weight</p>
100	De Vriendt 2009 ⁵³	Self-reported height and weight (BMI)	Adolescents	Mixed (stratified)	982	Europe	Not defined	<p>Intraclass correlation coefficients between the self-reported and measured weight, height and BMI were, respectively, 0.961, 0.949 and 0.899 ($p < 0.01$), indicating a high level of agreement between self-reported and measured values. The <i>t</i>-tests showed that there were significant differences between self-reported and measured BMI in girls ($p < 0.001$) but not for boys; however, Cohen's <i>d</i> values indicated that the magnitude of these differences was trivial</p> <p>Bland–Altman plots showed that at individual level these differences can be quite large, indicating limited usefulness of self-reported values on individual level</p> <p>Conclusion: Self-report cannot replace measured values for categorising adolescents</p>

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
101	Ambrosi-Randic 2007 ³⁵⁴	Self-reported height and weight (girls)	Children and adolescents	Mixed (stratified)	234	Croatia	Not defined	<p><i>Comparators:</i> Measured height and weight</p> <p>Pearson's correlation between measured and self-reported weight, height and BMI were high (ranging from 0.94 to 0.99)</p> <p>ANOVA = overweight girls had significantly greater differences between self-reported and measured weight when compared with normal and underweight girls</p> <p><i>Conclusion:</i> Self-reported data may be appropriate for group self-comparisons over time but should not be used to assess body size in clinical settings for the purposes of diagnostic and therapeutic decision</p> <p><i>Comparators:</i> Measured height and weight</p>
102	Field 2007 ³⁵⁵	Self-reported weight change	Adolescent	Mixed (stratified)	4760	USA	White, African American, Hispanic, other	<p>Self-report was slightly lower than measured weight but weight change was accurate by 2.1 pounds (girls) and 2.8 pounds (boys)</p> <p>Overweight and obese = under-report but did so consistently so that the change values were similar. Discrepancies not related to ethnicity, weight loss effects, television or PA</p> <p><i>Comparators:</i> Measured height and weight</p>
103	Elgar 2005 ³⁵⁶	Self-reported height and weight	Adolescent	Mixed (stratified)	418	Europe	Not defined	<p><i>Comparators:</i> Measured height and weight</p> <p>Under-reported weight by 0.52 kg</p> <p>13.9% of self-reported overweight compared with 18.7% of measured (obese = 2.8 vs. 4.4)</p> <p>Self-report not recommended for individual measurement</p> <p>Underestimate overweight by 4.8% and obesity by 1.6%</p> <p>Poor sensitivity (52.2% overweight and 55.6% obese)</p>

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
104	Brener 2003 ³⁵⁷	Self-reported height and weight	Adolescent	Mixed (stratified)	4619 (reliability); 2032 (validity)	USA	White, African American, Hispanic	<p>Comparators: Measured height and weight</p> <p>TRT: $\kappa = 0.87$ (categorised as overweight both times); $\kappa = 0.77$ (categorised as at risk both times)</p> <p>Mean self-reported BMI = 23.5 kg/m², lower than measured height and weight (26.2 kg/m²), $r = 0.89$</p> <p>White females most likely to under-report</p> <p>Comparator: Measured WC</p> <p>Comparison $r = 0.83$ (also compared measured and reported BMI $r = 0.9$)</p> <p>22.7% of overweight children were classified as being normal weight based on reported WC compared with measured (BMI misclassified 23.7%)</p> <p>Conclusion: Reported WC is of value</p> <p>Comparator: BIA (%fat mass, FMI)</p> <p>Subscapular skinfold more accurate than triceps skinfold</p> <p>Other results relate to differences between those with and without metabolic syndrome</p> <p>Conclusion: Subscapular skinfold (also correlated well with WC and WHR) is the best marker</p>
105	Bekkers 2011 ³⁵⁸	Self-reported waist circumference	Child	Mixed (stratified)	1292	The Netherlands	Not defined	
106	Ayvaz 2011 ²⁵	SF, WC, Hip, WHR, BMI	Children and adolescents	Mixed (stratified)	64	Turkey	Not defined	

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
107	Watts 2006 ³⁵⁹	SFT	Children and adolescents	All obese	38	Australia	Not defined	<p>Comparator: DXA (total fat)</p> <p>R = 0.83 (weight), r = 0.86 (BMI), r = 0.81 (waist), r = 0.88 (hip), r = 0.76 (six skinfolds)</p> <p>Similar for DXA abdominal fat</p> <p>Sum of SF and %BF from SF were not independent predictors of DXA total fat or %BF</p> <p>Change following an exercise intervention – SFT (both sum and percentage) were not able to predict change in total fat or change in abdominal fat by DXA – therefore not a good measure in exercise interventions</p>
108	Rowe 2006 ³⁶⁰	SFT	Children and adolescents	Mixed (stratified)	1254	USA	Not defined	<p>Comparator: BIA (two %BF equations)</p> <p>All Pearson's correlations between BMI and two methods of estimating %BF were significant ($p < 0.05$)</p> <p>Size of correlation was moderate to high in boys (r = 0.77) and girls (r = 0.79)</p> <p>Bland-Altman analyses revealed fixed and proportional bias, and 95% LOA covered a range of > 20% BF</p> <p>Agreement of obesity classification was moderately high in boys ($\kappa = 0.77$) and girls (0.81) but fewer children were classified as obese via %BF-BIA (14.5%) than via %BF-SF (19.8%)</p> <p>Conclusions: Results indicate that whole-body BIA provides %BF estimates that are systematically different from %BF estimates from skinfolds in children and adolescents</p>

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
109	Rodriguez 2005 ³⁶¹	SFT equations	Adolescents	Mixed (stratified)	238	Spain	Race not defined	<p>Comparator: DXA</p> <p>Most equations did not demonstrate good agreement compared with DXA. Correlations in females ranged from 0.00 (Brook equation) to 0.67 (Wilmore and Behnke) and in males 0.02 (Slaughter) to 0.74 (Deurenberg)</p> <p>In addition, %fat mass is overestimated in lean subjects and underestimated in obese subjects</p> <p>Comparator: BIA (%BF)</p> <p>The correlation coefficient between subscapular skinfold and %BF was 0.79, and there was good agreement between %BF and subscapular skinfold in separating high (> 85th percentile) from not high ($\kappa = 0.60$ for white people and $\kappa = 0.66$ for black people). Per cent agreement between subscapular skinfold and %BF was lower in overweight/obese (64%) than normal weight (94%) in white people and black people (65% vs. 94%)</p> <p>Comparators: Measured height and weight</p> <p>Most normal weight adolescents accurately reported body size</p> <p>Percentage of under-reporters was significantly higher in the overweight/obese group than in the normal weight group ($\chi^2 = 9.741, p = 0.003$)</p> <p>Correlation between BMI, both measured and self-reported, and perceived body size was positive and highly significant ($p < 0.001$)</p> <p>Self-reported weight and height = acceptable for estimating weight status in normal-weight adolescents, but not in those who are overweight or obese</p>
110	Morrison 2001 ³⁶²	SFT	Children and adolescents	Mixed (stratified)	2379	USA	White, African American	
111	Jorga 2007 ³⁶³	Silhouette rating scale	Adolescents (> 11 years)	Mixed (stratified)	245	UK/Serbia (not clear – four central Belgrade communities)	Serbian	

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
112	Radley 2007 ²⁵¹	Thoracic gas volume equations (predicted) and converted to %BF	Children and adolescents	Mixed (stratified)	258	UK	Race not defined	<p><i>Comparator:</i> Thoracic gas volume (measured) When converted to %BF, the mean %BF (Fields) estimates were within 1% of the measured value in all groups, except obese males (1.1%), whereas the mean %BF (Crapo) estimates were > 1% in all groups, except lean males (0.5%). Using either prediction equation, Bland–Altman analysis revealed that the greatest %BF + 95% LOA were in the lean and overweight groups and lowest in the obese groups</p> <p><i>Conclusion:</i> Thoracic gas volume (Fields) greater than thoracic gas volume (Crapo) in providing accurate %BF estimates</p> <p>Scale development: silhouettes (similar to Stunkards) for toddlers</p> <p>Content validity showed good ability to correctly order picture and interobserver agreement for weight status classification was high ($\kappa=0.7$, $r=0.8$)</p> <p>Health professionals agreed scale was ethnically and gender neutral</p> <p>Inter-rater reliability (matched to photos) $r=0.78$</p> <p>Cronbach's α 0.855. Validity with weight for length $r=0.63$ </p> <p>Gold standard: Deuterium oxide dilution</p> <p>TBW underestimated in obese. BMI accounted for >40% of the interindividual variability, suggesting that body size was not taken sufficiently into consideration by the predictive formulae used</p> <p>Authors developed own equation using body surface area [TBW = 1.156 x (surface area/body impedance) - 2.356; R = 0.96] but this was not validated</p>
113	Hager 2010 ³⁶⁴	Toddler Silhouette scale	Infants	Mixed (stratified)	129 parents/ 10 health visitors	USA	Not defined	
114	Battistini 1992 ³⁶⁵	TBW prediction from BIA	Children and adolescents	Mixed (stratified)	29	Italy	Not defined	

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
115	Pineau 2010 ³⁶⁶	Ultrasound measurement	Children and adolescents	All obese	94	France	Race not defined	Comparator: DXA BF by ultrasound correlated closely with BF by DXA, in both females ($r = 0.958$) and males ($r = 0.981$)
116	Garnett 2005 ³⁶⁷	Waist circumference	Children and adolescents	Mixed (stratified)	342	Australia	Not defined	Longitudinal study (7–8 years) and 12–13 years. WC increased by 0.74 compared with BMI z-score (0.18). Kappa value between measures in detecting obesity was 0.68 in younger children and 0.64 in older children
117	Taylor 2000 ³⁶⁸	WC, WHR, conicity index	Children and adolescents	Mixed (stratified)	580	New Zealand	White	WC identified more children as overweight/obese than BMI (i.e. increased prevalence of obesity defined by WC compared with that defined by BMI) ROC curves, and AUCs for the ROCs, were calculated to compare the relative abilities of the anthropometric measure to correctly identify children with high trunk fat mass
118	Weili 2007 ³⁶⁹	WHtR	Children and adolescents	Mixed (stratified)	4187	China	Han and Uygur	The 80th percentile for WC correctly identified 89% of girls and 87% of boys with high trunk fat mass, and this measure performed significantly better as an index of trunk fat mass than WHR or the conicity index. (AUCs for waist circumference in girls and boys = 0.97 and 0.97, respectively; AUCs for conicity index in girls and boys = 0.8 and 0.81, respectively; AUCs for WHR in girls and boys = 0.73 and 0.71, respectively) The authors provide cut-offs for high trunk fat mass and high waist circumference for both sexes for each year of age Comparator: BMI AUC for WHtR to define overweight/obese > 0.90. WHtR cut-off defined at 0.445 (sensitivity and specificity > 0.8) <i>Author's conclusion:</i> This is a simple accurate tool

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
119	Hitze 2008 ³⁷⁰	WC	Children and adolescents	Mixed (stratified)	180	Germany	White	<p>Comparator: BOD POD (overwaist \geq 90th centile of WC in Dutch population reference)</p> <p>All sites were well correlated with BMI, per cent fat mass and metabolic risk factors, but all at significant difference levels in different genders</p> <p>Strongest correlations in boys = beneath lowest rib (waist to chest ratio) and BMI ($r = 0.93$; in girls = above iliac crest and per cent fat mass (0.63). Differences advocate consensus on measurement area</p> <p>Comparator: DXA</p>
120	Reilly 2010 ³⁷¹	WC and BMI percentiles	Child	Mixed (stratified)	7722	UK	Race not defined	<p>The area under the ROC curve = slightly higher for BMI percentile (0.92 in boys and 0.94 in girls) than WC percentile (0.89 in boys and 0.81 in girls)</p> <p>Specificity of BMI percentile was slightly but significantly higher than that of WC percentile for both sexes ($p = 0.05$ in each case). WC percentile has no advantage over BMI percentile for diagnosis of high fat mass</p> <p>Comparator: BMI</p>
121	Mazicioglu 2010 ³⁷²	WC and MUAC	Children and adolescents	Mixed (stratified)	2358	Turkey	Race not defined	<p>Differences between area under curve (AUC) values for WC and MUAC were not significant (except for children aged 6 years), indicating that both indices performed equally well in predicting obesity</p> <p>Sensitivity was suboptimal through age groups 6–9 years in the boys and sensitivity was suboptimal at 6, 7, 14 and 17 years both in boys and girls</p>

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
122	Candido 2011 ³⁷³	WC, arm circumference, arm fat area, Rohrer Index, conicity Index, WHtR	Children and adolescents	Mixed (stratified)	788	Brazil	Not defined	<p>Comparator: BIA</p> <p>Obesity = excess BF 25% (boys) and 30% (girls)</p> <p>Arm fat area = best for boys</p> <p>Rohrer index = best for girls</p> <p>Based on sensitivity and specificity analysis plus discriminate ability (Youden index)</p> <p>Comparator: Measured BMI (also used 'overfat' from subcutaneous fat)</p> <p>No weight for age cut-off was accurately able to identify overweight with high sensitivity and specificity, or positive predictive value or negative predictive value</p> <p>Comparator: Body density by hydrostatic weighing</p> <p>All measures showed good accuracy (> 88%)</p> <p>The sum of five skinfolds was most sensitive (86.8%) and weight was least sensitive (52%)</p> <p>Weight was most specific (95%) and sum of five skinfolds was least specific (90%)</p> <p>Comparator: Densitometry (%BF)</p> <p>Overall lower specificity and higher sensitivity for all measures [TSF in boys (sensitivity = 24%, specificity = 100%) and BMI (sensitivity = 23%, specificity = 100%) in girls were preferred single anthropometric indicators of obesity]</p>
123	Stettler 2007 ³⁷⁴	Weight for age	Children and adolescents	Mixed (stratified)	12,382	USA	White, African American, Hispanic, other (not defined)	
124	Marshall 1991 ²⁶	Weight, BMI, sum of five skinfolds and triceps skinfold	Children and adolescents	Mixed (stratified)	540	Canada	Race not defined	
125	Himes 1989 ⁷⁵	Weight, BMI, triceps skinfold, subscapular skinfold and %BF estimated from the sum of four skinfolds	Children and adolescents	Mixed (stratified)	316	Canada	Not defined (French)	

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
126	Zheng 2010 ³⁷⁶	Ultrasoundography, BIA, WC and BMI	Children and adolescents	All obese	103	China	Race not defined	<p>Comparator: MRI</p> <p>Correlations with subcutaneous fat MRI are as follows: BMI (0.82), ultrasoundography (0.46), BIA (0.55), WC (0.89)</p> <p>Correlation with VAT MRI are: BMI (0.54), ultrasoundography, (0.35), BIA (0.58), WC (0.61)</p> <p>Conclusion: In both types of fat WC was most associated with MRI</p>
127	Yamborisut 2008 ³⁷⁷	WC	Children and adolescents	Mixed (stratified)	509	Thailand	Race not defined	<p>Comparator: WHZ</p> <p>In ROC analysis, WC risk threshold for predicting the overweight adolescents, using Thai weight-for-height z-score ≥ 1.5 standard deviation as reference, was 73.5 cm for boys (sensitivity 96.8%, specificity 85.7%) and 72.3 cm for girls (sensitivity 96.1%, sensitivity 80.5%)</p> <p>WC threshold was increased to 75.8 cm (sensitivity 96.3%, specificity 86.4%) for boys and 74.6 cm for girls (sensitivity 95.1%, specificity 85.7%) in order to detect the obese children</p> <p>Author's conclusion: WC is a feasible tool</p>
128	Campanozzi 2008 ³⁷⁸	DXA, BIA and SFT	Children and adolescents	All obese	103	France	Race not defined	<p>No gold standard</p> <p>Results from a t-test reveal significant difference between BIA and DXA (-4.37 kg, $p < 0.05$), between DXA and SFT (-1.72 kg, $p < 0.05$) and between BIA and SFT (-2.65 kg, $p < 0.05$)</p> <p>Author's conclusion: In obese children, DXA, BIA and SFT should not be used interchangeably in the assessment of body mass because of an unacceptable lack of agreement between them. The discrepancies between methods increase with the degree of obesity</p>

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
129	Goldfield 2006 ³⁷⁹	BIA	Children	Overweight and obese	17	Canada	Race not defined	<p>Comparator: DXA</p> <p>The correlations for %BF, fat mass and fat-free mass were 0.85, 0.97 and 0.94</p> <p>Bland-Altman tests of agreement showed moderate to large within-subject differences in body composition variables</p> <p>Conclusions: BIA is strongly related to DXA but the two measures may not be used interchangeably. Although BIA may lack the precision to assess small changes in body composition in overweight and obese individuals, it is appropriate for epidemiological use</p> <p>Comparator: BMI</p>
130	Guntsche 2010 ³⁸⁰	WHtR	Children and adolescents	Mixed (stratified)	108	Argentina	Race not defined	<p>WHtR significantly correlated with BMI ($r = 0.95$) and DXA-trunk FMI ($r = 0.93$). The author supports its use in future research</p> <p>Comparators: BMI and WC</p> <p>Neck circumference showed significant positive correlations with BMI (0.78) and WC (0.80)</p> <p>Author's conclusion: NC is not as good as WC in determining overweight and obesity, both providing similar information</p> <p>Comparator: Underwater weighing</p> <p>TSF correctly identified 15 males and four females, and the relative weight identified 16 and 5, respectively</p> <p>Both measures were low in sensitivity (23–50%) but high in specificity (85–100%)</p> <p>Both measures are not advocated</p>
131	Hatipoglu 2010 ³⁸¹	Neck circumference	Children and adolescents	Mixed (stratified)	967	Turkey	Race not defined	
132	Johnston 1985 ³⁸²	TSF and relative weight	Children and adolescents	Mixed (stratified)	235	USA	White, black	

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
133	Kurth 2010 ³⁸³	Self-reported height and weight (BMI)	Children and adolescents	Mixed (stratified)	3436	Germany	Race not defined	<p><i>Comparators:</i> Measured height and weight</p> <p>The bias in the self-reported BMI yielded an underestimation of overweight and obesity prevalence</p> <p>Self-report is not advocated</p> <p><i>Comparator:</i> DXA</p> <p>In healthy children, BIA correlated well with DXA (R = 0.84)</p> <p>In females with PCOS and obesity the correlation was weaker (R = 0.62)</p> <p><i>Author's conclusion:</i> BIA is a useful tool but different prediction equations between black and white children must be determined</p> <p><i>Comparator:</i> BIA</p>
134	Lewy 1999 ³⁸⁴	BIA	Child	Mixed (stratified)	40	USA	African American	<p>Skinfold showed strong correlation with BIA (0.93)</p> <p>The technical error between the two methods was small</p> <p>The ability of the BIA device to categorise into normal and obese categories when compared with the skinfold technique was also impressive (0.95; 95% confidence interval = 0.73 to 0.99)</p> <p>However, the results of the LOA analysis showed that the approximate 95% confidence interval for the differences between methods was wide (-9.1 to 11.4)</p>
135	Moore 1999 ³⁸⁵	SFT	Child	Mixed (stratified)	38	USA	Native Americans, Hispanic, European	<p>The technical error between the two methods was small</p> <p>The ability of the BIA device to categorise into normal and obese categories when compared with the skinfold technique was also impressive (0.95; 95% confidence interval = 0.73 to 0.99)</p> <p>However, the results of the LOA analysis showed that the approximate 95% confidence interval for the differences between methods was wide (-9.1 to 11.4)</p>

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
136	Owens 1999 ³⁸⁶	Weight, BMI, TSF, calf skinfold, sagittal diameter, WC, hip circumference, thigh waist–thigh ratio, sagittal diameter/thigh ratio, and %BF from the sum of calf and triceps skinfolds	Children and adolescents	All obese	76	USA	White, black	<p>Comparator: MRI</p> <p>The highest correlation with VAT from MRI was the sagittal diameter (0.63) and the weakest was calf skinfold (0.41)</p> <p>From this a new prediction equation was created including the anthropometric variables; sagittal diameter and WHR and demographic variable; and ethnicity because of the greatest correlations with MRI</p> <p>The model explained that 63% of the variance in VAT and was associated with a measurement error of 23.9%</p> <p>Although the model seems to lack sufficient explanatory power for routine use in clinical settings with individual patients, it may have some utility in epidemiological studies given its relatively small (<25%) standard error of estimate</p>
137	Tsang 2009 ³⁸⁷	DXA	Children and adolescents	Mixed (stratified)	48	Australia	Race not defined	<p>No comparator. Assessed reliability of several abdominal regions using DXA. All methods had acceptable intra- and inter-rater reliability. Region 1 (android) was most precise in overweight/obese individuals, whereas region 6 (top of iliac crest) was most precise in normal weight individuals</p> <p>In all regions, assessments were less precise in overweight/obese individuals</p>
138	Williams 2007 ³⁸⁸	BIA	Child	Mixed (stratified)	341	Australia	Race not defined	<p>No comparator. Compared different %BF equations derived from BIA with BMI</p> <p>Correlations with BMI are equation 1 (Rush) ($r = 0.43$); equation 2 (Schaefer) ($r = 0.57$); equation 3 (Goran) ($r = 0.33$); and equation 4 (Horlick) ($r = 0.62$). Results support concerns of using BMI and an accurate measure of body fat mass</p>

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
139	Malina 1986 ³⁸⁹	BMI and triceps skinfold	Children and adolescents	Mixed (stratified)	2137	USA	Hispanic	No comparator. Just compared prevalence when using both methods Depending on the method used there was a difference in the prevalence of overweight or obesity Fewer children were classified as overweight or obese when the two criteria were used together than when they were used individually The results suggest that the BMI and the triceps skinfold vary in sensitivity as indicators of overweight and obesity
140	Brambilla 1994 ³⁹⁰	AFA, TFA, WHR	Children and adolescents	Mixed (stratified)	44	Italy	Race not defined	Comparator: MRI AFA was significantly lower, even if significantly correlated with MRI in obese ($r = 0.84$) and normal weight ($r = 0.96$) the agreement between the two methods showed wide LOA TFA was significantly lower, even if significantly correlated with MRI in obese ($r = 0.77$) and normal weight ($r = 0.89$) the agreement between the two methods showed wide LOA Intrabdominal adipose tissue by MRI was not related to WHR in obese ($r = 0.14$) or normal ($r = 0.11$) <i>Author's conclusion:</i> The anthropometric indices do not offer an accurate estimate of adiposity in children

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
141	Pecoraro 2003 ³⁹¹	BMI and TSF, BIA	Child	Mixed (stratified)	228	Italy	Race not defined	No gold standard Comparison between tools. There was no significant difference in prevalence of obesity measured with BMI or tricep skinfold thickness Both measures showed strong correlations with BIA: BMI ($r = 0.92$), tricep skinfold thickness ($r = 0.79$) <i>Author's conclusions:</i> Measurement using tricep skinfold thickness and BIA is similar in different BMI ranges. However, BIA is a useful and alternative method for detecting body composition in children and may be a more precise tool than tricep skinfold thickness for measuring fat mass in epidemiological studies in paediatric populations
142	Mello 2005 ²⁴	ADP, DXA	Adolescents	All obese	88	Brazil	Race not defined	No gold standard Compared two methods No significant correlation between parameters common to both methods [fat-free mass, fat mass (kg) and fat mass (%), $r = 0.88$, $r = 0.92$, $r = 0.75$] was observed <i>Author's conclusions:</i> Our data suggest that for this specific population, plethysmography may be used as an important method of body composition evaluation

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
143	Radley 2003 ²²⁵	ADP	Adolescents	Overweight and obese	69	UK	Race not defined	<p>Comparator = DXA</p> <p>ADP estimates of percentage fat were highly correlated with those of DXA in both male and female subjects ($r = 0.90$ to 0.93)</p> <p>The 95% LOA were relatively similar for all percentage fat estimates, ranging from $\pm 6.73\%$ to $\pm 7.94\%$</p> <p>Also compared with DXA estimates, ADP produced significantly ($p < 0.01$) lower estimates of mean body fat content in boys (-2.85% and -4.64%) and girls (-2.95% and -5.15%)</p> <p><i>Author's conclusion:</i> Siri equation correlated more with DXA than Lohman, but high LOA, using either equation, resulted in percentage fat estimates that were not interchangeable with percentage fat determined by DXA</p>
144	Williams 2006 ¹⁸	DXA	Children and adolescents	Mixed (stratified)	215	UK	Race not defined	<p>Gold standard: 4C model</p> <p>The accuracy of DXA-measured body-composition outcomes differed significantly between groups (obese, normal, cystic fibrosis)</p> <p><i>Author's conclusions:</i> The bias of DXA varies according to the sex, size, fatness and disease state of the subjects, which indicates that DXA is unreliable for patient case-control studies and for longitudinal studies of persons who undergo significant changes in nutritional status between measurements</p>

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
145	Taylor 2008 ³⁹²	WC, WHtR, conicity index	Child	Mixed (stratified)	301	New Zealand	White	<p>Comparator: DXA</p> <p>AUCs indicated that WC correctly discriminated between children with low and high trunk fat mass 87% (for girls) to 90% (for boys) of the time</p> <p>WC performed better than WHtR (AUCs 0.79 in girls and 0.81 in boys) and the conicity index (AUCs: 0.53 in girls and 0.65 in boys)</p> <p>A z-score of 0.55 correctly identified 79% of girls and 81% of boys with high trunk fat mass, and 82% of girls and 84% of boys with low trunk fat mass</p>
146	Freedman 2005 ³⁹³	BMI	Children and adolescents	Mixed (stratified)	1 196	USA	White, black, Hispanic, Asian	<p>Conclusion: WC performs reasonably well as an indicator of high trunk fat mass in preschool-aged children</p> <p>Comparator: DXA</p> <p>Accuracy of BMI as a measure of adiposity varied greatly according to the degree of fatness</p> <p>Among children with a BMI-for-age of > 85th percentile, BMI levels were strongly associated with FMI ($r = 0.85-0.96$ across sex-age categories) but not so for fat-free mass ($r = 0.21-0.70$). In contrast, among children with a BMI-for-age of < 50th percentile, levels of BMI were more strongly associated with fat-free mass ($r = 0.56-0.83$) than with FMI ($r = 0.22-0.65$)</p> <p>Author conclusions: BMI levels among children should be interpreted with caution. Although a high BMI-for-age is a good indicator of excess fat mass, BMI differences among thinner children can be largely due to fat-free mass</p>

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
147	Freedman 2009 ³⁹⁴	BMI	Children and adolescents	Mixed (stratified)	1196	USA	White, black, Hispanic, Asian	<p>Comparator: DXA</p> <p>About 77% of the children who had a BMI for age \geq 95th percentile had an elevated body fatness, but levels of body fatness among children who had a BMI for age between the 85th and 94th percentiles ($n = 200$) were more variable; about one-half of these children had a moderate level of body fatness but 30% had a normal body fatness and 20% had an elevated body fatness</p> <p>The prevalence of normal levels of body fatness among these 200 children was highest among black children (50%) and among those within the 85th–89th percentiles of BMI for age (40%)</p>
148	Kayhan 2009 ³⁹⁵ (Turkish)	BMI, SFT	Adolescents	Mixed (stratified)	713	Turkey	Not defined	<p><i>Author's conclusion:</i> BMI is an appropriate screening test to identify children who should have further evaluation and follow-up but it is not diagnostic of level of adiposity</p> <p>Comparator: BIA (assumed)</p> <p>Correlations between all measurements range between $r = 0.52$ and $r = 0.97$. H correlation found between calf skinfold measurement and %BF using Slaughter's formula ($r = 0.94-0.97$)</p> <p>Note: Information from abstract. The British Library could not obtain a copy</p>
149	Majcher 2008 ³⁹⁶ (Polish)	BMI, WHR, WHR	Children and adolescents	Mixed (stratified)	324	Polish	Not defined	<p>Comparator: BIA</p> <p>%BF by BIA was comparable with results using Slaughter's equation. No correlation observed between %BF and WHR</p> <p>Note: Information from abstract</p>

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
150	Zambon 2003 ³⁹⁷ (Portuguese)	BMI, SFT	Children	Mixed (stratified)	4236	Portugal	Not defined	No gold standard Comparisons between two measures SFT found to be more variable and dependent on weight status Advocates BMI
151	Zaragozano 1998 ³⁹⁸ (Spanish)	Weight: BMI; triceps and submandibular skinfolds; the sum of four skinfolds; body fat and WC; arm circumference	Children and adolescents	Mixed (stratified)	72	Not reported	Not defined	Note: Information from abstract Not clear Highest no. of obese children (12.5%) detected with the submandibular skinfold BMI detected 5.55% Lowest no. of obese children detected with arm circumference (2.77%) Note: Information from abstract
152	Behbahani 2009 ³⁹⁹ (Persian)	BMI	Children	Mixed (stratified)	1800	Iran	Not defined	Comparator: FMI from skinfold (TSF) thickness Determined 'real' obese and 'real' non-obese from FMI BMI identified 43.3% of obese and 0.6% of non-obese children Sensitivity and specificity of the 90th percentile of BMI to identify children as obese were 71.1% and 98%, respectively Conclusion: Efficacy of BMI in determining childhood obesity may be poor and that FMI, in comparison with BMI, is a better indicator of obesity in children Note: Information from abstract

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
153	Chiara 2003 ⁰⁰⁰ (Portuguese)	Weight, stature, BMI and subscapular skinfold	Adolescents	Mixed (stratified)	502	Brazil	Not defined	No gold standard Comparison between tools. Prevalence of risk of obesity = higher with subscapular skinfold measurement ($p < 0.0001$) compared with BMI-based classifications, which showed similar values Specificity was higher than sensitivity in BMI-based classifications BMI able to identify adolescents without obesity but sensitivity was too low for tracking risk of obesity Note: Information from abstract Comparator: %BF from skinfold (TSF) thickness BMI classification showed high sensitivity (83–97%), except for the classification proposed by WHO (65% in males and 48% in females) Specificity was high for all criteria (85–98%) Note: Information from abstract
154	da Silva 2010 ⁰⁰¹ (Portuguese)	BMI	Children	Mixed (stratified)	1570	Brazil	Not defined	Information from abstract
155	Giugliano 2004 ⁰⁰² (Portuguese)	BMI	Children	Mixed (stratified)	528	Brazil	Not defined	Information from abstract Comparators: %Fat from sum of triceps and subscapular, triceps and calf skinfold measurements, and waist and hip circumference %BF, waist and hip circumference were significantly correlated with BMI ($p < 0.01$) Note: Information from abstract

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
156	Jakubowska-Pietkiewicz 2009 ⁰³ (Polish)	BMI, WC, skinfold, BIA	Children and adolescents	Mixed (stratified)	56	Poland	Not defined	<p>Comparator: DXA</p> <p>Correlations with BIA – $r^2 = 0.83$. Correlations with Slaughter's algorithm – $r^2 = 0.83$ ($p < 0.001$)</p> <p>BIA and Slaughter's algorithm were lower than %BF from DXA, which increases with increasing %BF</p> <p>Differences between results obtained by BIA and Slaughter's algorithm in comparison with DXA negatively correlated with BMI-SDS and WC-SDS</p> <p>Note: Information from abstract</p>
157	Perez 2009 ⁰⁴ (Spanish)	BMI, WHtR, conicity index, WC	Children and adolescents	Mixed (stratified)	382	Venezuela	Not defined	<p>Comparator: Unclear, 'the fat area'</p> <p>BMI demonstrated high sensitivity and specificity with ROC AUC at 0.85 ($p < 0.000$)</p> <p>This was not seen in other measures, except for in age 7–9 years with CI [ROC AUC 0.76 ($p < 0.000$)]</p> <p>Note: Information from abstract</p>
158	Ramirez 2010 ¹⁹ (Spanish)	DXA	Children and adolescents	Not reported	32	Mexico	Not defined	<p>Gold standard: 4C model</p> <p>Mean difference between DXA and 4C model was –3.5% body fat ($p = 0.171$)</p> <p>LOA = 5% to –12% body fat</p> <p>Concordance correlation coefficient was $p = 0.85$</p>

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
159	Rodriguez 2008 ⁴⁰⁵ (Spanish)	BMI, WC, BIA, DXA	Children	Not reported	230	Argentina	Not defined	<p>The test of accuracy for coincidence of slope intercepts between DXA and the 4C model showed no coincidence ($p < 0.05$)</p> <p>The precision by R^2 explained 83% of the variance (standard error of the estimate = 4.1%)</p> <p>The individual accuracy assessed by the total error was 5.6%</p> <p>There was an effect of method ($p = 0.043$) in the presence of overweight ($p < 0.001$)</p> <p><i>Author's conclusion:</i> DXA is imprecise compared with the 4C model, but still advocate its use in follow-up comparisons in population analysis</p> <p>Note: Information from abstract</p> <p>No gold standard</p> <p>Comparison between tools</p> <p>BIA measures were lower than DXAs ($p < 0.0001$)</p> <p>Correlations between BIA vs. anthropometric methods and WC vs. DXA were moderate (Pearson's $r = 0.43$ to 0.53), whereas the other correlations were strong ($r = 0.71$ to 0.83)</p> <p>Bland-Altman comparison showed wide LOA between BIA and DXA; BIA significantly underestimated %BF as determined by DXA ($p < 0.0001$)</p> <p>Note: Information from abstract</p>

No.	First author, year (ref. no.)	Name of measure	Age	Weight status	Sample size	Country	Ethnicity	Comments
160	Schonhaut 2004 ⁴⁰⁶ (Spanish)	Height, weight	Children	Mixed (stratified)	416	Chile	Not defined	Inter-rater reliability: Compared measurement between school workers and trained health workers Prevalence of overweight and obesity differed according to whether measured by school worker or health worker ($\kappa = 0.56$)
161	Stein 2006 (German) ⁴⁰⁷	Self-reported height and weight	Children and adolescents	Mixed (stratified)	280	Germany	Not defined	Note: Information from abstract Abstract presents little information, but suggests self-report (by telephone) should not be used in assessment of change in anthropometry
162	Zhang 2004 ⁴⁰⁸ (Chinese)	BMI	Children and adolescents	Mixed (stratified)	1094	China	Not defined	Note: Information from abstract <i>Comparator</i> : DXA Age- and gender-specific correlations range from 0.59 to 0.83 Note: Information from abstract

%BF, per cent body fat; %OB, per cent obese; %OW, per cent overweight; AFA, arm fat area; ANOVA, analysis of variance; BIS, bioelectrical impedance spectroscopy; CDC, Centers for Disease Control and Prevention; CRP, C-reactive protein; ECW, extracellular water; FF, foot to foot; FMI, fat mass index; GRE, gradient recalled echo; HF, hand to foot; HT²Z, height squared/impedance (impedance adjusted for height); IH-MRS, (1H) hydrogen proton magnetic resonance spectroscopy; ILC, intrahepatic lipid content; IMCL, intramyocellular lipid; MRI, magnetic resonance imaging; MRS, magnetic resonance spectroscopy; MUAC, mid-upper arm circumference; PCOS, polycystic ovary syndrome; PWH, per cent height for weight; R_c, regression coefficient; ROC, receiver operating characteristic; SAT, subcutaneous adipose tissue; SDS, standard deviation score; SE, slowness in eating; TAF, total abdominal fat; TEFR, trunk-extremity fat ratio; TR + CA, triceps and calf; TSF, triceps skinfold; TSFT, triceps skinfold thickness; UFA, upper arm fat area; UFE, upper arm fat area estimate; VAT, visceral fat; WC-IC, waist circumference iliac crest; WC-UC, waist circumference umbilicus; WHR/Ht, waist-to-hip ratio/height; WHtR, waist-to-height ratio; WHZ, weight-for-height z-score.

Appendix 7 Dietary assessment evaluation studies: summary table

Dietary assessment methodologies						
Tool information						
No.	Name	First author (type of paper)	Administration	Sample: age; weight status; country (ethnicity)	Evaluation	Comments
FFQs/checklists (16 tools)						
1	Korean Food Frequency Questionnaire (Korean FFQ)	Lee 2007 ⁴⁸ (PDP)	Self-completed	Child; mixed (stratified); Korea (Korean) (n = 153)	TRT (r = 0.37, range = 0.22–0.51)	Developed specifically for obesity-related eating behaviours. Therefore, all items aimed to discriminate
2	Qualitative Dietary Fat Index (QFI)	Yaroch 2000 ⁴¹ (PDP)	Interview administered in person – child Pen and paper	Adolescents (including 11 years); obese and overweight; USA (African American) (TRT n = 22, validity n = 57)	TRT (r = 0.54 full tool) Convergent validity with 24-hour recall (r = 0.23–0.31)	Convergent validity repeated with adjustment for age and BMI with no change (data not shown). Also repeated with five non-fat style items removed, which made relationship with energy significant (r = 0.27). Overall showed significant relationship with total fat, although r-values are low
3	Short-list list Youth Adolescent Questionnaire (Short YAQ)	Rockett 2007 ³⁴ (ModEval)	Self-completed Pen and paper	Children and adolescents; mixed (non-stratified); USA (white) [n = 17,788 (construct validity = 5848 girls)]	Convergent validity with 24-hour recall (r = 0.43, range = 0.05–0.58) and long-version FFQ (r = 0.9)	Items/questionnaire not provided, nor are details of cost or copyright. Web search found no further information for short FFQ. E-mail sent to corresponding author (13/08/12) and received copy of tool – which has 29 food items (not 26)
4	Youth Adolescent Questionnaire (YAQ)	Rockett 1995 ⁴³ (PDP)	Self-completed Pen and paper	Child and adolescents; mixed (non-stratified); USA (white) (n = 179)	Construct validity with screen time (0.55, range = 0.034–0.109) (all significant) TRT (r = 0.41, range: nutrients = 0.26–0.58, foods = 0.39–0.57) Convergent comparisons with other national surveys within 10% (range = 2–25%)	Some information here also taken from an additional paper. ⁴⁰⁹ In these reliability results, absolute comparisons were said to be similar. However, owing to reduction in EI at T2, differences were apparent in results. Linked to further evaluation
5	Youth Adolescent Questionnaire (YAQ)	Rockett 1997 ³⁷ (ModEval)	Self-complete Pen and paper	Child and adolescents; mixed (non-stratified); USA (white) (n = 261)	Convergent validity with 24-hour recall (r = 0.4, range = 0.24–0.75)	Linked to Rockett 1995. ⁴³ Modified to reflect problems with original evaluation (e.g. foods groups as serving units such as burgers; including burger and roll). Number of items reduced

Dietary assessment methodologies					
Tool information					
No.	Name	First author (type of paper)	Type	Administration	Sample: age; weight status; country (ethnicity)
6	Youth Adolescent Questionnaire (YAQ)	Perks 2000 ³⁰ (Eval)	151-item FFQ	Self-complete Pen and paper	Child and adolescents; mixed (stratified); USA (race not defined) (n = 50)
					Criterion validity with DLW EI was similar ($\rho = 0.91$) but with large LOA (-6.30 ml to 6.67 ml) Discrepancy in EI (YAQ-DLW) was related to body fat ($r = 0.25$) and %BF ($r = -0.24$) but not age ($r = 0.07$) or time between measures ($r = 0.00$)
7	Picture sort FFQ	Yaroch 2000 ²⁴² (PDP)	110-item FFQ	Interview in person – child Pen and paper Card sort	Children and adolescents; Obese and overweight; USA (African American) (n = 22)
					TRT ($r = 0.16$, range = 0.02–0.43) Convergent validity with 24-hour recall ($r = 0.66$, range = 0.38–0.84)
8	Children's Eating Habits Questionnaire (CEHQ-FFQ)	Lanfer 2011 ³⁶ (PDP)	43-item FFQ	Parent completed Pen and paper	Child; mixed (non-stratified); seven European countries (race not defined) (n = 258)
					Development paper referenced to previous paper (Suling 2011); some details on development used to complete this extraction. Also, additional paper describes validity (Huybrechts <i>et al.</i>) ³⁴
9	Childs Eating Habits Questionnaire (CEHQ-FFQ)	Huybrechts 2011 ³¹ (PDP)	43-item FFQ	Parent completed Pen and paper	Child; mixed (non-stratified); seven European countries (race not defined) (n = 10,309)
					Criterion validity with urinary calcium (UCa), urinary potassium (UK), creatinine (Cr) $r = \text{Ca/Cr} = 0.01$ – 0.08 $\text{UK/Cr} = 0.09$ – 0.18 ANOVA = UK/Cr = third/highest tertile significantly greater than lowest UCa/Cr = highest tertile significantly greater than lowest
					Primary development is Rockett 1997. ⁴⁰⁹ The author concludes that the YAQ provides accurate estimation of mean EI for a group but not individual. Also boys with greater body fat were more likely to under-report EI than girls with greater body fat Based on Block Health Habits and History Questionnaire (97 items). Without energy adjustment, reliability is considerably higher (ICC range = 0.28–0.42). Validation results are for mean of both administrations
					Results adjusted for age, soft drink consumption and number of meals out of home. Analysis also presented by country

Dietary assessment methodologies						
Tool information						
No.	Name	First author (type of paper)	Type	Administration	Sample: age; weight status; country (ethnicity)	Comments
10	Australian Child and Adolescent Eating Survey (ACAES)	Watson 2009 ⁴⁶ (PDP)	137-item FFQ	Self-completed Pen and paper	Children and adolescents; mixed (non-stratified); Australia (race not defined) ($n = 101$)	Multiple tests performed (including T2 and food records). As suggested by authors, results shown are average of T1 and T2 for validity. Correlations were generally lower for transformed, energy adjusted. Unadjusted reliability = 0.44
11	Australian Child and Adolescent Eating Survey (ACAES)	Burrows 2008 ³² (PDP)	137-item FFQ	Self-completed Pen and paper	Children and adolescents; mixed (non-stratified); Australia (race not defined) ($n = 93$)	Correlation coefficients means include multiple carotene metabolites. Because one (lutein) was less correlated, the overall mean of the tool was lowered. Given high to moderate correlations with other metabolites, the overall scores for robustness were deemed adequate, even although not reflected in mean. Correlations were greatest after adjustment for BMI
12	Brief dietary screener	Nelson 2009 ⁴⁴ (PDP)	21-item food intake checklist	Self-completed Pen and paper	Adolescent; mixed (non-stratified); USA (white) (TRT $n = 33$; convergent validity $n = 59$)	Data only provided for significant results or results in which an adequate number of children reported the event. Thus, means and ranges shown are for available data only and are likely to overestimate actual kappa values (in validity testing). Links to additional validation in Latino children (Davis)
13	Brief dietary screener	Davis 2009 ⁴⁵ (Eval)	21-item food intake checklist	Self-completed Pen and paper	Adolescent; Overweight; USA (Hispanic/Latino) females ($n = 35$)	Additional testing of Nelson tool. ²¹⁰ Results provided are written in a similar to Nelson – with many non-events for fast food restaurant visits. Thus mean and ranges shown for available data

Dietary assessment methodologies

Tool information

First author
(type of
paper)Sample: age;
weight status;
country (ethnicity)

No.	Name	Type	Administration	Evaluation	Comments
14	Intake of fried food away from home (intake of FFA)	1-item food intake checklist	Self-complete Pen and paper (postal)	<p>Convergent validity with fast food checklist ($r = 0.57$, range = 0.56–0.58)</p> <p>Construct validity Generalised estimating equations regression = increased BMI with increase frequency of FFA consumption cross-sectionally (only significant in boys) and longitudinally. LMS showed significant decrease in diet quality based on consumption of 12/13 foods with increased FFA intake</p>	Study-specific tool developed for intervention evaluation
15	Food Intake Questionnaire	66-item FFQ	Self-complete, with parent prompts Pen and paper	<p>Convergent validity with 24-hour recall (agreement = 93%, range = 88–98%; $\kappa = 0.67$, range = 0.64–0.69)</p> <p>IC Time 1 = 0.88; time 2 = 0.87</p> <p>TRT 0.64 (single value only)</p> <p>Convergent validity with weighed food diary: sensitivity (ability to detect high fat) = 81%; specificity (ability to rule out low fat) = 47%; positive predictive value = 79% ($\chi^2 = 4.80$, $df = 1$, $p = 0.028$)</p>	Assessment of daily intake
16	21-item dietary fat screening measure	21-item food intake checklist	Self-competed Pen and paper	<p>Adolescent; mixed (non-stratified); USA [white; African American; Hispanic; Asian] ($n = 239$) (TRT = 231; convergent validity = 59)]</p> <p>Convergent validity with weighed food diary: sensitivity (ability to detect high fat) = 81%; specificity (ability to rule out low fat) = 47%; positive predictive value = 79% ($\chi^2 = 4.80$, $df = 1$, $p = 0.028$)</p>	Results not presented by scale – only overall. No full result presented for validation (r -value) as only data for per cent fat is reported ($r = 0.36$). Paper also presents a four-category screener. However, validity was poor, leading authors to choose to continue testing accuracy on the 21-item tool only

Dietary assessment methodologies							
Tool information							
No.	Name	First author (type of paper)	Type	Administration	Sample: age; weight status; country (ethnicity)	Evaluation	Comments
17	New Zealand Food Frequency Questionnaire (New Zealand FFQ)	Metcalf 2003 ⁴⁷ (PDP)	117-item FFQ	Parent completed Pen and paper	All ages (up to 14 years); mixed (non-stratified); New Zealand (Maori, Pacific Islanders); other (not stated) (n = 130)	IC $\alpha = 0.84$, range = 0.59–0.92 TRT r = 0.72, range = 0.42–0.86, t-test ($p = 0.54$)	TRT analysis also analysed Spearman correlations. Results were similar. Thus, only Pearson correlations are shown here
18	Harvard Service FFQ (HSFFQ)	Blum 1999 ³⁸ (Eval)	84-item FFQ	Parent completed Pen and paper/ electronic entry	Infant and children (< 5 year); mixed (non-stratified); USA (white, Native American) (n = 233)	Convergent validity with 24-hour recall (0.52, range = 0.26–0.63)	Information for extraction on development and description had been supplemented by on-line material [Harvard website and Colditz Women, Infants, and Children (WIC) book] (supplementary material attached to paper). The HSFFQ was developed for (and is implementing in) the WIC programme specifically. Authors advocate its use but there is no information related to its value as an outcome measure
19	5-day food frequency questionnaire (5D FFQ)	Crawford 1994 ³³ (Eval)	42-item FFQ	Interview administered in person – child Pen and paper	Child; all girls; mixed (non-stratified); USA (white; African American) (n = 19)	Criterion validity with direct observation (r = 0.32, range = 0.11–0.50). % absolute error (PAEs = observed foods not reported) median range = 20 (SFAs) 33 (CHOs). 50% of food had quantification errors of > 50%	This paper validates 24-hour recall and food diaries as well (extracted separately). Little information provided on the tool, as the paper focus is on validation of the methods. Overall, FFQ performs least well compared with others

Dietary assessment methodologies

Tool information

No.	Name	First author (type of paper)	Type	Administration	Sample: age; weight status; country (ethnicity)	Evaluation	Comments
20	Dietary Guideline Index for Children and Adolescents (DGI-CA)	Golley 2011 ³⁵ (PDP)	11-Component dietary pattern index	Interview administered in person – parent, child and both; interview administered on phone – parent, child and both	Children and adolescents; mixed (stratified); Australia (race not defined) (n = 3416)	Construct validity with diet quality = regression p-values all significant except PUFA	Results here are for 29 food items. E-mailed author (13 September 2012) for more information. Responded with good groups on 14 September 2012 but no item-level information. Associations with BMI z-score were weak. By DGI-CA score, risk of overweight/obesity was non-significant (Q5 vs. Q1 odds ratio = 0.97, 95% confidence interval 0.76 to 1.24, p = 0.82). Statistics used were appropriate and therefore given 4/4 for robustness, but there is no measure of agreement
21	Familial Influence on Food Intake–Food Frequency Questionnaire (FIFI-FFQ)	Vereecken 2010 ³⁹ (ModEval)	77 item FFQ	Parent completed, pen and paper	Child; mixed (non-stratified); Belgium (race not defined) (n = 216)	Convergent validity with online dietary assessment tool: 'young children nutrition assessment on web' (YCNA-W) (r = 0.47 range 0.22–0.76) (% agreement = 81% range 1–99%)	Adequate correlations but large LOA is a concern. The author concludes that the FFQ was a useful alternative to estimating energy and macronutrient intake at group level, but when used to estimate fibre and calcium intake overestimation and underestimation need to be considered

Bland–Altman identified large LOA

Dietary assessment methodologies					
Tool information					
Sample: age; weight status; country (ethnicity)					
No.	Name	Type	Administration	Evaluation	Comments
Diaries/recalls/observations (three diet histories; 11 diaries; six recalls; one biomarker; one mixed methods; one observation)					
22	Diet history	Diet history	Self-completed; interview administered in person – child	Adolescent: mixed (non-stratified); Scandinavia (race not defined) (n = 35)	Duration of diet history not reported. Two parts: (1) self-complete at school and (2) interviews with nutritionist. Data also analysed for food intake, comparing under- and over-reporters. Significant results = boys adequately reporting have lower intake of energy between meals compared with over-reporters. Girls who under-report consume less energy between meals than accurate reporters. Over-reporting boys consume more soft drinks than adequate reporters
23	2-week Diet History Interview (DHI)	Diet history	Interview administered in person – parent and child	Child; mixed (stratified); Scandinavia (race not defined) (DLW n = 21; SenseWear n = 85)	<p>Criterion validity with DLW [EI vs. TEE $r = 0.59$ ($p < 0.001$)]</p> <p>A 4% difference between EI and TEE ($p > 0.05$)</p> <p>LOA for difference = -5.63–6.45 mJ/day</p> <p>EI/TEE ratio not correlated to body weight/BMI, but BMI was greater in over-reporting boys and greater in under-reporting girls</p> <p>Criterion validity with DLW and SenseWear band = DLW, $r = -0.026$ ($p = 0.0912$); SenseWear $r = 0.08$ ($p = 0.44$)</p> <p>Regression = DLW, $y = -1.33 - 0.0033x$; SenseWear $y = -0.29 - 0.14x$</p> <p>A 14% difference between EI and TEE (DLW), which was not different between those obese and those overweight</p> <p>A 14% difference between EI and TEE (SenseWear) which was greater in those obese (22%) than those</p>

Dietary assessment methodologies

Tool information		Sample: age; weight status; country (ethnicity)	Administration	Evaluation	Comments		
No.	Name					Type	Administration
	First author (type of paper)						
24	3-day weighed food diary	Maffeis 1994 ⁵⁵ (Eval)	Parent completed Pen and paper	Weighed food diary	Child; mixed (stratified); Italy (race not defined) (n = 24)	overweight (11%) Underestimation also negatively associated with BMI (r = -0.38, p < 0.01) Criterion validity with Lusk's formula: PMR x HR for TEE: TEE vs. EI t-test p < 0.001 (obese); p > 0.05 (non-obese) LOA: -5.67-0.01 ml/day (obese); -2.22-2.01 ml/day (non-obese) EI expressed as %TEE negatively associated with body weight (r = -0.80, p < 0.001) and body fat (r = -0.72, p < 0.001)	Same paper also validates a diet history. This diary has poor validity in obese; therefore, robustness point lost for validity results
25	7-day diet history	Maffeis 1994 ⁵⁵ (Eval)	Interview administered in person - parent Pen and paper	Diet history	Child; mixed (stratified); Italy (race not defined) (n = 24)	Criterion validity with Lusk's formula: PMR x HR for TEE TEE vs. EI t-test p < 0.05 (obese); p < 0.05 (non-obese) LOA: -5.44-2.40 ml/day (obese); -1.49-2.91 ml/day (non-obese) EI expressed as %TEE negatively associated with body weight (r = -0.71, p < 0.001) and body fat (r = -0.58, p < 0.01)	Same paper also validates a 3-day diary. Both score 3 out of 4 on robustness of validation as results are less strong in obese

Dietary assessment methodologies

Tool information		Sample: age; weight status; country (ethnicity)	
No.	Name	Type	Administration
26	9-day estimated food diary	Estimated food diary	Self-completed Pen and paper
	First author (type of paper)		
	Singh 2009 ⁵⁷ (Eval)		Adolescent; overweight; USA (race not defined) (n = 34)
			Evaluation
			Criterion validity with DLW = % error = 1065 ± 636 kcal/day
			Relative error = 35% ± 20%
			Dietary fat, BMI and sex explained 86.4% of error variance
			Error positively associated with BMI
			Comments
			Also compared children with low error within ± 500 kcal/day (n = 6) to rest of sample and found these to be more lean
27	3-day estimated dietary intake record	Estimated food diary	Parent completed Pen and paper
	O'Connor 2001 ⁶⁴ (Eval)		Child; mixed (non-stratified); Australia (race not defined) (n = 47)
			Evaluation
			Criterion validity with DLW [r = 0.10 (p = 0.51)]
			Mean percentage of misreporting = 4% ± 23%
			LOA = -3226-3462 kJ
			Significant negative association between misreporting and PA (r = -0.77, p < 0.0001)
			Comments
			Misreporting was overestimation in 55%. One out of three reported within 10%. Not related to sex or body composition. Lost one point in robustness of validity because of poor correlation, although overall misreporting percentage was low compared with other studies
28	2-week weighed food diary	Weighed food diary	Self-completed Pen and paper
	Bandini 1990 ⁵⁸ (Eval)		Adolescent; mixed (stratified); USA (n = 55)
			Evaluation
			Criterion validity with DLW = Correlations compared bias (reported ME/DLW%) and showed negative correlation between weight and reported ME/DLW% (i.e. overweight more likely to under-report (although both under-report)) (t-tests show EI and TEE sign different in obese and non-obese)
			Comments
			Similar reporting with obese and non-obese subjects (both lower than measured), but because of differences in energy expenditure, obese subjects were more likely to be described as under-reporters. Also conducted intra- and inter-variation by day of reporting across 14 days and found similar coefficients between obese and non-obese subjects (0.87 and 0.89, respectively)

Dietary assessment methodologies

Tool information

No.	Name	First author (type of paper)	Type	Administration	Sample: age; weight status; country (ethnicity)	Evaluation	Comments
29	2-week weighed food diary	Bandini 1999 ⁵⁹ (Eval)	Estimated food diary	Self-completed Pen and paper	Children and adolescent; mixed (stratified); USA (race not defined) (n = 43)	Criterion validity with DLW Both groups under-reported EI but obese group under-reported significantly more	Further results showed that high calorie foods were higher in non-obese. The author concludes the 14-day food diary resulted in under-reporting for both obese and non-obese but this was more prominent in obese. The data offers no evidence to support the notion that obese eat more junk food than non-obese
30	3-day tape-recorded estimated food record	Lindquist 2000 ⁶⁰ (PDP)	Estimated food diary	Self-completed Tape recorder	Child; mixed (stratified); USA (white; African American) (n = 30)	Criterion validity with DLW (r = -0.06; regression = r = 0.32 (p > 0.05); t-test p < 0.05 Mean difference = -1.13 mJ/day 61% under-reported; 26% over-reported, with older and fatter children demonstrating more inaccuracy	Also measured energy with 24-hour recall. Good correlation between 24-hour and DLW, but did not analyse correlations between these and tape recorder method. Analysis of misreporting show greater errors in overweight children. Overall poor validity
31	7-day weighed food diary (7-D-WFR)	Bratteby 1998 ⁴¹⁰ (Eval)	Weighed food diary	Self-completed Pen and paper	Adolescent (15 years): mixed (stratified); Sweden (race not defined) (n = 50)	Criterion validity with DLW Only 8/50 reported higher than measured Significant negative correlation between per cent fat mass and EI expressed as %TEE = underestimation with increasing fat mass % [t-test DLW (BMR) vs. EI p > 0.05]	Paper combines this analysis with PAL analysis, but only presents level of activity (therefore not extracted). Results according to body composition are in the discussion and the remaining findings are minimal. The final conclusion by authors is that the 7-D-WDR are underestimated by adolescents, especially those 'toward overweight and increasing body fat'

Dietary assessment methodologies							
Tool information							
No.	Name	First author (type of paper)	Type	Administration	Sample: age; weight status; country (ethnicity)		
					Evaluation		
					Comments		
32	3-day estimated food diary	Crawford 1994 ³³ (Eval)	Estimated food diary	Self-completed Pen and paper	Child; all girls; mixed (non-stratified); USA (white; African American) (n = 25)	<p>Criterion validity with direct observation (lunch only) (r = 0.87, range = 0.78–0.94)</p> <p>Least significant difference range = 1 g (SFA) – 55 kcal (energy)</p> <p>PAE range = 12 (energy) – 22 (cholesterol)</p> <p>36% of correctly reported foods had quantification errors of < 10%</p>	<p>One of the three measures presented by Crawford <i>et al.</i> Overall, although the 3-day diary was worst in terms of feasibility, it was the best in terms of accuracy. As a result, the authors decided to advocate its use in the National Heart, Lung, and Blood Institute Growth and Health Study</p>
33	8-day food record	Champagne 1996 ⁶³ (Eval)	Estimated food diary	Self-completed, interview administered over telephone – parent Pen and paper	Child; mixed (stratified); USA (white; African American) (n = 23)	<p>Criterion validity with DLW; African Americans under-report by 37% and white people under-report by 13%</p> <p>The highest tertile of body fat under-reported EI by 1040 kcal compared with the lowest (420 kcal) and middle (350 kcal)</p>	<p>This study was a pilot before it was undertaken on a larger scale in the 1998 study. The 8-day food record showed to under-report dietary intake. It is clear that African Americans and those with the greatest amount of body fat tend to under-report EI to a greater extent</p>
34	8-day food record	Champagne 1998 ⁶² (Eval)	Estimated food diary	Self-completed (parent assisted) and nutritionist-recorded school lunch Pen and paper	Child; mixed (stratified); USA (white, African American) (n = 118)	<p>Criterion validity with DLW; African Americans under-report by 28%, and whites under-report by 20%</p> <p>With regards to age group, 12-year-olds had the greatest level of under-reporting (33%) and 9-year-olds the least (19%)</p>	<p>The 8-day food record showed to under-report dietary intake, especially among African Americans, girls, those at 12 years of age, and those with central fat</p>

Dietary assessment methodologies							
Tool information							
No.	Name	First author (type of paper)	Type	Administration	Sample: age; weight status; country (ethnicity)		
					Evaluation		
					Comments		
35	Tape-recorded food record	Van Horn 1990 ⁵⁶ (Eval)	Estimated food diary	Self-completed, parent completed Pen and paper	Child; mixed (stratified); USA (white) (n = 32)	Girls under-report more than boys (25% vs. 22%) and, when stratified by weight, the percentage of under-reporting is as follows: central fat (32%), lean (21%), obese (25%) and peripheral fat (17%) Inter rater reliability with parent report of diet (r = 0.84, range 0.68–0.96)	The tape recorded food record produced greater correlations with parent report than the telephone 24-hour diet recall, which was the other method tested in this study
36	24-hour dietary recall (1 day)	Baxter 2006 ⁶⁵ (Eval)	24-hour recall	Interviewed in person – child Pen and paper	Child; mixed (stratified); USA (white; African American) (n = 79)	TRT Inaccuracy, T1 = 7.5 servings/day; T2 = 6.7 servings/day; T3 = 6.2 servings/day Difference between trials = all p-values > 0.05 Other results shown as interaction effect with validity	Paper presents multiple results for omission, intrusion and total inaccuracy by trial by gender and obesity/weight status. Only data of relevance presented – shows accuracy decreased in obese over time and increased over time in normal weight (significant interaction)
						Criterion validity with direct observation = inaccuracy (servings/day) = 6.8 (healthy weight); 8.0 (at risk of obesity); 6.9 (obese) with significance between subject effects (F _{2,72} = 4.5, p = 0.015) Repeated measures (trials) showed significant BMI category by trial interactions (F _{2,72} , p = 0.028)	

Dietary assessment methodologies						
Tool information						
No.	Name	First author (type of paper)	Type	Administration	Sample: age; weight status; country (ethnicity)	Comments
37	24-hour dietary recall (3-day)	Johnson 1996 ⁶⁸ (Eval)	24-hour recall	Interview in person – parent and child Pen and paper	Child; mixed (non-stratified); USA (white) (<i>n</i> = 24)	Disparities between correlations and <i>t</i> -tests mean recall was able to accurately estimate group intakes (non-significant <i>t</i> -test) but not accurate at the individual level (non-significant correlation)
						<p>Criterion validity with DLW (<i>r</i> = 0.25, <i>p</i> = 0.24); <i>t</i>-test (<i>t</i> = 2.07, <i>p</i> = 0.65)</p> <p>LOA = -1102 ± 807 kcal/day</p> <p>Mean difference = -53.8 kcal/day</p> <p>Regression analysis showed no affect of any characteristic (including weight) on under- or over-reporting)</p>
38	24-hour dietary recall (1 day)	Lytle 1998 ⁶⁷ (Eval)	24-hour recall	Interview administered in person – child Pen and paper	Child; mixed (non-stratified); USA (white; African American; Asian) (<i>n</i> = 486)	Added food record prompts (completed day before recall) to determine whether improved accuracy. All analysis repeated with these. Correlations ranged from 0.04 (vitamin C) to 0.69 (vitamin A) but difference between correlations with and without food record prompts were generally non-significant. Authors report that not sufficient to warrant extra resources
						<p>Criterion validity with direct observation (<i>r</i> = 0.5, range = 0.37–0.59)</p> <p>ANOVA: All <i>p</i>-values non-significant except beta carotene (<i>p</i> = 0.008)</p>
39	24-hour recall (1 day)	Crawford 1994 ³³ (Eval)	24-hour recall	Interview administered in person – child Pen and paper	Child; all girls; mixed (non-stratified); USA (white; African American) (<i>n</i> = 30)	This paper validates three methodologies (others = FFQ and 3-day diary). Overall, the 24-hour recall was more accurate than the FFQ but not the 3-day diary
						<p>Criterion validity with direct observation (lunch only) (<i>r</i> = 0.62, range = 0.46–0.79)</p> <p>Least significant difference range = 2 g (SFA) – 120 kcal (energy)</p> <p>PAE range = 19 (energy/protein) – 39 (fat)</p>

Dietary assessment methodologies							
Tool information							
No.	Name	First author (type of paper)	Type	Administration	Sample: age; weight status; country (ethnicity)	Evaluation	Comments
40	Telephone 24-hour diet recall	Van Horn 1990 ⁵⁶ (Eval)	24-hour recall	Interview administered over telephone – parent and child Pen and paper	Child; mixed (stratified); USA (white) (n = 32)	50% of correctly reported foods had quantification errors of < 10% Inter rater reliability with parent report of diet (r = 0.75, range 0.65–0.93)	Further results combined both the telephone diet recall and the tape recorded diet record to compare 10 food groups in parent child. Percentage agreement is in parentheses; beverage (62%), bread (77%), meat/fish (79%), fruit/vegetables (68%), cake (59%), chips (71%), candy (50%), condiment/butter (54%), dairy (59%), mixed dishes (82%). The author concludes that children are able to provide dietary intake data using electronic equipment in a manner that compares favourably with adults
41	Day in the Life Questionnaire (DILQ) (focused on F&V)	Edmunds 2002 ⁶⁶ (PDP)	24-hour recall	Self-complete (in classroom) Pen and paper	Child; mixed (non-stratified); UK (TRI n = 235; inter-rater n = 83; validity n = 255; responsiveness n = 49)	TRT [t-test range $p = 0.188-0.927$ (all non-significant)] Inter rater reliability between coders (κ range = 0.82–0.92)	Results for criterion validity here presented as convergent validity by paper. Responsiveness statistics not clear in paper
						Criterion validity with direct observation = 70% agreement	
						Responsiveness Difference in change all significant (measured fruit only)	

Dietary assessment methodologies						
Tool information						
No.	Name	First author (type of paper)	Type	Administration	Sample: age; weight status; country (ethnicity)	
					Evaluation	
					Comments	
42	Diet Observation at Childcare (DOCC)	Ball 2007 ⁷⁰ (PDP/protocol)	Observation protocol	Researcher conducted/observed Pen and paper	Infant and children; mixed (non-stratified); USA (race not defined) (inter-rater $n = 66$ observations; validity $n = 96$)	Protocol development and reliability paper. Therefore, focus is on training and implementation – not individual level. Direct observation is within child-care centres. Contacted author for more information. Responded with details, including link to another child-care environmental measure (Benjamin <i>et al.</i> ²¹²), which was already picked up by CoOR search
41	Food Behaviour Questionnaire (FBQ)	Vance 2008 ⁷¹ (Eval)	(24-hour recall, FFQ, and nutrition and PA behaviours)	Self-completed Web based	Adolescent; mixed (stratified); Canada (race not defined) ($n = 95$; inter-rater $n = 51$; direct observation $n = 20$; E/BMR $n = 1917$)	<p>Inter-rater reliability between five observers = 100% agreement for 11 items Remaining 10 items ($p > 0.05$) except spaghetti</p> <p>Criterion validity with measured items: $r = 0.96$ (in laboratory testing); $r = 0.88$ (in field); t-test all non-significant except spaghetti)</p> <p>TRT: Presented in abstract (agreement = 77%, range = 62–87%)</p> <p>Inter-rater between self- and dietitian-report ($r = 0.55$–0.70; ICC = 0.51–0.66)</p> <p>Criterion validity with Goldberg cut-off = E/BMR ratio (estimated) and direct observation (direct observation agreement = 87%) E/BMR ratio = 1.4 (50.6) with increased under-reporting in girls</p> <p>The FBQ is a combined tool, including a 24-hour recall, a FFQ and nutrition and PA behaviour questions. Validity shown here compares the 24-hour only. Development information is cited as an abstract. The author has been contacted (13 September 2012) for further information. Note: This tool has been used as the basis to create Web-SPAN (web survey of PA and Nutrition). Further analysis of the 24-hour component is reported in Story <i>et al.</i> (2012) but only gives vague description: 'A subset of students who participated in the current study completed the survey on two days ($n = 379$), and also completed a 3-day food record ($n = 369$). ICC values for the repeat comparisons and between</p>

Dietary assessment methodologies						
Tool information						
No.	Name	Type	Administration	Sample: age; weight status; country (ethnicity)	Evaluation	Comments
42	IGF-1, IGFBP-1, IGFBP-3: biomarkers	Biochemical markers	Self-completed, biochemical	Child; mixed (stratified); Spain (race not defined) ($n = 56$)	<p>EI/BMR ratio decreased with increasing weight status</p> <p>Fisher's post hoc comparisons showed that this was significant in girls ($F = 14.28$, $p < 0.001$) and boys ($F = 33.21$, $p < 0.001$)</p> <p>Note: A ratio of $< 1.74 : 1$ = under-reporting)</p> <p>Construct validity with BMI percentile ($r = 0.38$, range $0.24-0.54$)</p>	<p>the FBQ and the 3-day food record were within ranges reported elsewhere in the adolescent population (ref = PhD dissertation). Furthermore, mean differences of nutrient intakes between the two measurements were small. Managed to locate abstract/poster online. Added information as appropriate. Note: abstract makes same inter-rater comparison ($n = 58$) with slightly different results ($r = 0.57-0.85$; ICC = $0.54-0.84$)</p> <p>Overweight children were found to have higher serum levels of IGF-1 and IGFBP-3, but lower levels of IGFBP-1. The IGF is considered a good biomarker of caloric undernutrition and protein malnutrition. The author advocates the use of biochemical markers of caloric nutritional status in this population</p>

%BF, per cent body fat; ANOVA, analysis of variance; BF, body fat; CHOs, carbohydrates; Eval, evaluated an existing tool without modification; F&V, fruit and vegetables; ICC, intraclass correlation coefficient; LMS, liquid meal supplement; ME, metabolisable energy; ModEval, modified an existing tool and re-evaluated; PAE, percentage absolute error; PAL, physical activity level; PDP, primary development paper; PMR, prone maximum restraint position; PUFA, polyunsaturated fatty acid; SFA, saturated fatty acid.

Appendix 8 Eating behaviour studies: summary table

Eating behaviour questionnaires: tool information

No.	Name	First author (type of paper) (reference)	Administration	Sample: age; weight status; country (ethnicity), (n)	Evaluation	Comments
1	Child Eating Disorder Examination Interview (ChEDE-I), 30 item	Decaluwé 2004 ⁸¹ (Eval)	Interview administered – Child	Children and adolescents; all obese; Belgium (race not defined) (IC/TRT n = 25, inter-rater n = 20, validity n = 138)	IC: $\alpha = 0.65$ (range = 0.53–0.84) TRT: $r = 0.73$ (range = 0.61–0.83) IR: with two interviewers $r = 0.96$ (range = 0.91–0.99)	Concluded that the ChEDE-I interview was necessary to identify eating disorders in obese children, whereas the self-report ChEDE-Q can only be used as a screening measure. Information on tool development from Bryant-Waugh 1996 ⁴¹¹
2	Child Eating Disorder Examination Interview (ChEDE-I), 30 item	Bryant-Waugh 1996 ⁴¹¹ (ModEval)	Interview administered – Child	Children and adolescents; mixed (non-stratified); UK (race not defined)	Convergent validity: with ChEDE-Q (non-interview version) $r = 0.41$ – 0.76 ; agreement = 42–67% Development/face validity (pilot only)	Developed primarily for EDE (eating disorders examination) in adults with few changes, e.g. wording
3	Child Eating Disorder Examination Interview (ChEDE-I)	Goossens 2010 ⁴¹² (Eval)	Self-completed	Children and adolescents; mixed (stratified); Belgium (race not defined) (IC n = 1291, validity n = 235)	IC: $\alpha = 0.84$ (range = 0.77–0.93) Convergent validity: with ChEDE-interview $r = 0.53$ (range = 0.38–0.67)	Included here after being cited as primary development paper in Decaluwe 2004 ⁷² Questionnaire format showed good convergent validity with ChEDE interview and may serve a reliable instrument among overweight youngsters

Eating behaviour questionnaires: tool information					
No.	Name	First author (type of paper) (reference)	Administration	Sample: age; weight status; country (ethnicity), (n)	Comments
4	Child Eating Disorder Examination Interview (ChEDE-I)	Jansen 2007 ²²⁹ (ModEval)	Self-completed	Children and adolescents; mixed (stratified); Europe (race not defined) (IC/validity n = 38)	<p>Tool development is same as Decaluwe (1999)</p> <p>The adjustment of the tool for this study was: modified response options</p> <p>Also inserted definitions of the ambiguous concepts used in ChEDE – (i.e. LOC, binge eating, eating in secret, large amount of food and intense exercising)</p> <p>Authors conclude that adjustment reduced the gap between interview and questionnaire</p>
5	Child Eating Disorder Examination Questionnaire (ChEDE-Q), 30 item	Tanofsky-Kraff 2003 ²³⁰ (Eval)	Self-completed	Child; mixed (stratified); USA (white, African American) (validity n = 87)	<p>IC: $\alpha = 0.65$ (range = 0.53–0.83)</p> <p>Convergent validity: with ChEDE-interview $r = 0.62$ (range = 0.40–0.78); agreement = 82% (73–95%)</p> <p>Convergent validity: with ChEAT and QEWP-A</p> <p>Kendall Tau = 0.31. Sensitivity = 41% specificity = 83% (diagnosis of overeating), sensitivity = 29% specificity = 91% (diagnosis of LOC), sensitivity = 0% specificity = 89% (diagnosis of subjective bulimic episode); sensitivity = 17% specificity = 91% (diagnosis of objective bulimic episode)</p> <p>Type of episodes of eating disorder generated by ChEDE and QEWP were not significantly associated in entire sample or for overweight, except for after excluding 'No episode' ($-0.35, p < 0.01$)</p>

Eating behaviour questionnaires: tool information			
No.	Name	First author (type of paper) (reference)	Administration
6	ChEDE-I, 30 item	Tanofsky-Kraff 2005 ⁴¹³ (Eval)	Self-completed
		Sample: age; weight status; country (ethnicity), (n)	
		Children and adolescents; mixed (stratified); USA [white, African American, other (not defined)] (validity n = 167)	
		Evaluation	Comments
		<p>Convergent validity: with QEW-P $r = 0.38$ (range = 0.16–0.78). QEW-P sensitivity = 30% specificity = 79% (diagnosis of overeating), sensitivity = 50% specificity = 83% (diagnosis of binge eating). Positive predictive value of QEW-P for identification of episodes by ChEDE: detection of overeating 0.29% and detection of binge eating 0.18%</p> <p>IC: $\alpha = 0.54$ (range = 0.24–0.74)</p> <p>FA: 61% total variance; load range = 0.63–0.88</p>	<p>Tool development is same as Bryant-Waugh 1996⁴¹¹</p> <p><i>Conclusion:</i> Generally results of child interview do not accurately correspond with parent report (QEW-P)</p>
7	Infant Feeding Questionnaire (IFQ), 20 item	Baughcum 2001 ⁷⁴ (PDP) (study 2)	Parent completed
		Sample: age; weight status; country (ethnicity), (n)	
		Infant; mixed (stratified); USA [white, African American, Asian, Hispanic, Pacific islander and other (not defined)] (IC/FA n = 453)	
		Evaluation	Comments
		<p>IC: $\alpha = 0.54$ (range = 0.24–0.74)</p> <p>FA: 61% total variance; load range = 0.63–0.88</p>	<p>Citation referenced from Hendy 2009. Overweight children had higher scores on all factors except concern of infant underweight and using food to calm infant. Significant differences were apparent in factor 1 ($p = 0.003$) and factor 4 ($p < 0.001$). Additionally obese mothers scored higher on factors 1, 2, 4 and 5. Significant differences were apparent in factor 1 ($p = 0.0028$) and factor 2 (0.001). Paper has two data extraction forms for two measures [IFQ (study 1) and PFQ (study 2)]</p>

Eating behaviour questionnaires: tool information						
No.	Name	First author (type of paper) (reference)	Administration	Sample: age; weight status; country (ethnicity), (n)	Evaluation	Comments
8	Preschool Feeding Questionnaire (PFQ), 32 item	Baughcum 2001 ⁷⁴ (PDP) (study 2)	Parent completed	Infant and children; mixed (stratified); USA (white, African American, Asian, Hispanic, Pacific islander) (IC/FA n = 633)	IC: $\alpha = 0.6$ (range = 0.18–0.87) FA: 58% total variance; load range = 0.49–0.84	Citation referenced from Hendy 2009 Overweight children had higher scores on factors 2, 4 and 6 Significant differences were apparent in factor 2 ($p < 0.001$) and factor 5 ($p < 0.001$) Additionally, obese mothers scored higher on factors 2, 4, 5 and 8 Significant differences were apparent in factor 2 ($p < 0.001$), factor 7 ($p < 0.001$) and factor 8 ($p = 0.04$) Concluded that this tool is an appropriate measure for screening and prevention of eating disorders. The KEDS is an abbreviated form of the Eating Disorder Symptoms Inventory (ESI) which is for adults
9	Kids Eating Disorder Survey (KEDS), 14 item	Childress 1993 ⁸⁹ (PDP)	Self-completed	Children and adolescents; mixed (non-stratified); USA (race not defined) (IC/FA n = 1883, TRT n = 108)	IC: $\alpha = 0.73$ (range = 0.68–0.77) TRT: $r = 0.8$ (range = 0.68–0.86) FA: 39.8% total variance; load range = 0.17–0.83	

Eating behaviour questionnaires: tool information					
No.	Name	First author (type of paper) (reference)	Administration	Sample: age; weight status; country (ethnicity), (n)	Comments
10	Questionnaire of Eating and Weight Patterns (adolescent reported) (QEWPA), 12 item	Johnson 1999 ⁹⁰ (ModEval)	Self-completed (QEWPA was parent completed)	Children and adolescents; mixed (non-stratified); USA (race not defined) (inter-rater $n = 367$, validity $n = 367$)	Original was in adults but this tool was slightly modified, in particular substituting simpler synonyms from difficult words
				<p>Inter-rater: between parent QEWPA and child QEWPA agreement = 41% (range = 15.5–81.6%), $\kappa = 0.19$</p> <p>Convergent validity: with ChEAT-26 an effect for diagnostic category was found [$F(2,340) = 16.19$, $p < 0.01$]</p> <p>Construct validity: with Child Depression Index (CDI)</p> <p>Symptoms of depression differed over diagnostic categories [$F(2,340) = 18.12$, $p < .001$] (binge eating disorder $R^2 = 18.75$)</p> <p>Non-clinical bingeing $R^2 = 7.92$</p> <p>No diagnosis $R^2 = 5.04$</p>	

Eating behaviour questionnaires: tool information					
No.	Name	First author (type of paper) (reference)	Administration	Sample: age; weight status; country (ethnicity), (n)	Comments
11	Questionnaire of Eating and Weight Patterns (adolescent reported) (QEWP-A), 12 item	Steinberg 2004 ²¹ (Eval)	Self-completed	Child; mixed (stratified); USA [white, African American, other (not defined)] (inter-rater/validity n = 263)	<p>Tool development is same as Johnson (1999).⁹⁰ Child and parent versions are not concordant regarding presence of eating disorders or compensatory behaviours. Frequencies were higher in child reports</p> <p>IR: between parent QEWP-P and child QEWP-A (considered child as criterion): sensitivity = 24%, specificity = 82% for diagnosis of overeating</p> <p>Sensitivity = 20%, specificity = 80% for diagnosis of eating disorders</p> <p>Agreement = $\chi^2 = 4.365$ $p = 0.359$ (obese only)</p> <p>Convergent validity: with ChEAT: ANOVA = all non-significant</p> <p>Construct validity: With Child Depression Index (CDI) ($p =$ non-significant)</p> <p>State-trait anxiety ($p =$ non-significant)</p> <p>Child behaviour checklist (CBCL) ($p =$ non-significant)</p> <p>BMI ($p =$ non-significant) DXA ($p =$ non-significant) Body size dissatisfaction ($p =$ non-significant)</p> <p>(all ANOVA)</p>

Eating behaviour questionnaires: tool information						
No.	Name	First author (type of paper) (reference)	Administration	Sample: age; weight status; country (ethnicity), (n)	Evaluation	Comments
12	Dutch Eating Behaviour Questionnaire (child reported) (DEBQ-C), 20 item	Van Strien 2008 ⁷⁹ (ModEval)	Self-completed	Child: mixed (stratified); the Netherlands (race not defined) (IC/FA study 1 n = 185; study 2 n = 767, validity = 742)	<p>IC: $\alpha = 0.76$ (range = 0.68–0.81)</p> <p>FA: 35.8% total variance; load range = 0.45–0.71</p> <p>Construct validity: with health-related lifestyle measures = restrained eating $r = -0.27$ (snacks) $r = -0.14$ (sports)</p> <p>Emotional eating $r = -0.11$ (sports) $r = -0.17$ (watching TV)</p> <p>External eating $r = -0.1$ (sports) $r = -0.23$ (snacks)</p>	<p>Primary development is in adults; however, this has been modified for use in children</p> <p>Results also showed overweight children had higher scores on restrained eating only ($t = -9.2$ (df = 187.9); $p < 0.01$)</p>
13	Dutch Eating Behaviour Questionnaire (child reported) (DEBQ-C), 20 item	Banos 2011 ⁸³ (Eval)	Self-completed	Children and adolescents; mixed (stratified); Spain (white) (IC n = 392, TRT n = 107, FA/validity n = 292)	<p>IC: $\alpha = 0.72$ (range = 0.69–0.78)</p> <p>TRT: $r = 0.58$ (range = 0.39–0.71)</p> <p>FA: load range = 0.35–0.73</p> <p>Construct validity: with BMI $r = 0.13$ (range = 0.01–0.62)</p>	<p>Tool development is same as Van Strien (2008)⁷⁹</p> <p><i>Conclusions:</i> the DEBQ-C was effective in Spanish children. Although the construct validity was quite poor</p>
14	Dutch Eating Behaviour Questionnaire (child reported) (DEBQ-C), 20 item	Braet 2007 ⁹² (Eval)	Self-completed and parent completed	Children and adolescents; overweight; Belgium (race not defined) (IC/inter-rater n = 498)	<p>IC: $\alpha = 0.84$ (range = 0.81–0.89)</p> <p>IR: between child (DEBQ-C) and parent (DEBQ-P) $r = 0.39$ (range = 0.35–0.45)</p>	<p>Results showed fair correlations with parents, which improved for older children</p>

Eating behaviour questionnaires: tool information						
No.	Name	First author (type of paper) (reference)	Administration	Sample: age; weight status; country (ethnicity), (n)	Evaluation	Comments
15	Dutch Eating Behaviour Questionnaire (parent reported) (DEBQ-P), 33 item	Caccialanza 2004 ²⁸ (Eval)	Parent completed	Children and adolescents; mixed (stratified); Italy (race not defined) (IC/FA n = 312)	<p>IC: $\alpha = 0.87$ (range = 0.81–0.87)</p> <p>FA: three-factor solution accounted for 43.7% of total variance</p> <p>Construct validity: with weight status: obese and OW had higher restrained eating score (1.72 vs. 1.36 $p < 0.001$) and higher emotional eating (1.42 vs. 1.41 $p > 0.05$) but lower external eating (2.77 vs. 2.80 $p > 0.05$) than normal weight</p>	<p>Tool development is same as Braet 1997²⁸</p> <p>Originally in adults but Braet was the first to have modified it in children</p>
16	Dutch Eating Behaviour Questionnaire (parent reported) (DEBQ-P), 33 item	Braet 1997 ²⁸ (ModEval)	Parent completed	Child; mixed (stratified); Belgium (race not defined) (IC/FA validity n = 292)	<p>IC: $\alpha = 0.79$–0.86</p> <p>FA: 42.2% of total variance; load range = 0.32–0.85</p> <p>Construct validity: with diet ($r = 0.04$–0.40) competence ($r = 0.01$–0.31) child behaviour ($r = 0.14$–0.46) and locus ($r = 0.13$)</p>	<p>Only provided a range for IC</p> <p>Obese children had higher scores on the DEBQ-P than normal-weight children</p> <p><i>Conclusion:</i> these findings suggest that DEBQ can be used as an instrument for assessing eating styles of obese children</p>

Eating behaviour questionnaires: tool information					
No.	Name	First author (type of paper) (reference)	Administration	Sample: age; weight status; country (ethnicity), (n)	Comments
17	Children's Eating Attitudes Test (ChEAT), 26 item	Maloney 1988 ⁸⁶ (ModEval)	Self-completed	Child; mixed (non-stratified); USA (white, African American, Hispanic; Oriental) (IC n = 318, TRT n = 68)	<p>Modified from the EAT-26 (development for adults). Primary development paper for EAT-26 FA was conducted and reduced items from 40 to 26; however, because this was conducted in adults, item reduction information has not been excluded here</p> <p>Confirms face validity was completed in discussion but this was vague in main body of text. In addition, 6.8% of children scored with anorectic range of > 20</p>
18	Children's Eating Attitudes Test (ChEAT), 26 item	Smolak 1994 ¹⁰⁰ (Eval)	Self-completed	Children and adolescents; mixed (non-stratified); USA (white) (IC/FA validity n = 306)	<p>Information on tool development from Maloney 1988.²³ IC and construct validity was best with the reduced 23-item questionnaire</p> <p>IC: $\alpha = 26$ item: (0.87), 25 item (0.85) 23 item (0.89) (range = 0.78–0.92)</p> <p>FA: 48% total variance; load range = 0.32–0.83</p> <p>Construct validity: with body dissatisfaction r = 0.4 (range = 0.39–0.42) and weight management behaviour r = 0.38 (range = 0.36–0.38)</p>
19	Children's Eating Attitudes Test (ChEAT), 26 item	Ranzenhofer 2008 ¹⁰¹ (Eval)	Self-completed	Children and adolescents; mixed (stratified); USA [white, African American, Hispanic, other (not defined)] (IC/FA validity n = 265)	<p>Tool development is same as Maloney 1988²³</p> <p>Beta scores were provided only when significant and so the means are biased</p> <p>IC: $\alpha = 0.78$ (range = 0.52–0.78)</p> <p>FA: 0.61 (factor load), total variance 33%; load range = 0.39–0.79</p>

Eating behaviour questionnaires: tool information					
No.	Name	First author (type of paper (reference)	Administration	Sample: age; weight status; country (ethnicity), (n)	Comments
20	Eating Attitudes Test (EAT), 40 item	Wells 1985 ¹¹⁴ (Eval)	Self-completed	Children and adolescents; girls only; mixed (stratified); New Zealand (race not defined) (n = 749)	<p>Authors conclude that subscale generated from school samples are generally supported in overweight child. Body/weight concern and dieting appear to be separate constructs and only total score body/weight concern and diet appear to be associated with body weight and adiposity</p> <p>Convergent validity: with three-factor eating questionnaire $r = 0.25-0.35$</p> <p>Construct validity: with Child behaviour checklist ($\beta = 0.22$), child depression ($\beta = 0.33$), state-trait anxiety ($\beta = 0.37$), BMI z-score ($r = 0.28$) and body fat ($\beta = 0.31$)</p> <p>Internal validity: principal FA with varimax rotation. Four factors emerged with dieting as predominant factor</p> <p>Also compared factors to weight status</p> <p>Factor 1 (diet) is positively related to overweight ($r = 0.29$ for 0-3 scoring and 0.39 for 1-6 scoring)</p> <p>Factor 4 (social pressure to eat) for 0-3 scoring ($r = -0.23$) and factor 3 in 1-6 scoring ($r = -0.34$) were related to underweight</p> <p>Primary Development is in adults (Garner and Garfinkel 1979⁸). Little has been done to make it compatible for children and adolescents. ChEAT was later developed from this and is more specific to children. The author concludes that the FA yielded a major dieting factor. Although this interpretation measures pathology in underweight, its interpretation is ambiguous in normal and overweight girls</p>

Eating behaviour questionnaires: tool information						
No.	Name	First author (type of paper) (reference)	Administration	Sample: age; weight status; country (ethnicity), (n)	Evaluation	Comments
21	Youth Eating Disorder Examination–Questionnaire (YEDE-Q) (#items not stated)	Goldschmidt 2007 ⁹⁹ (ModEval)	Self-completed	Children and adolescents; overweight; USA [white, African American, Hispanic and other (not defined)] (IC/validity n = 35)	<p>IC: $\alpha = 0.75$ (range = 0.63–0.89)</p> <p>Convergent validity: with ChEDE $r = 0.75$ (range = 0.16–0.84)</p> <p>Agreement in identifying bulimic episodes: = 16.91, $p < 0.001$</p> <p>Construct validity: with weight concerns $r = 0.59$ (range = 0.55–0.61)</p> <p>IC: $\alpha = 0.9$ (range = 0.83–0.95)</p> <p>TRT: $r = 0.66$ (range = 0.59–0.74)</p> <p>FA: 67.2% of total variance; load range = 0.50–0.84</p> <p>Convergent validity: with QEWP-A (loss of control)</p>	<p>Primary development in adults: EDE-Q (Goldfein 2005^b). However, this was modified and evaluated in children</p> <p><i>Conclusion:</i> The YEDE-Q seems promising in assessment of eating pathology in overweight adolescents</p>
22	Emotional Eating Scale for Children (EES-C), 26 item	Tanofsky-Kraff 2007 ⁷⁷ (ModEval)	Self-completed	Children and adolescents; mixed (stratified); USA [white, African American, Hispanic, other (not defined)] (IC/FA/construct validity n = 159, TRT = 64; convergent validity n = 155)	<p>IC: $\alpha = 0.9$ (range = 0.83–0.95)</p> <p>TRT: $r = 0.66$ (range = 0.59–0.74)</p> <p>FA: 67.2% of total variance; load range = 0.50–0.84</p> <p>Convergent validity: with QEWP-A (loss of control)</p> <p>Those with LOC from QEWP-A had higher 'eating' in response to anger, anxiety and frustration and higher 'depressive symptoms' compared with people without LOC (analysis = test for difference $p < 0.05$)</p>	<p>Primary development in adults (Arnow 1995^c)</p> <p>Results confirmed inadequate discriminative validity and overweight children were more likely to endorse LOC eating ($p = 0.04$)</p>

Eating behaviour questionnaires: tool information						
No.	Name	First author (type of paper) (reference)	Administration	Sample: age; weight status; country (ethnicity), (n)	Evaluation	Comments
23	Children's Binge Eating Disorder Scale (C-BEDS) (#items not stated)	Shapiro 2007 ²³¹ (PDP)	Interview administered – Child	Children and adolescents; mixed (non-stratified); USA [white, African American, Asian, Hispanic, Native American, other (not defined)] (n = 55)	<p>Construct validity: state-trait anxiety (r = 0.06), Children's Depression Index (r = 0.13), and child behaviour checklist (r = 0.05). BMI z-score: no EES-C subscales were significantly related to or overweight (p > 0.2)</p> <p>Convergent validity: with Structural Clinical Interview for DSM-IV disorders (SCID), Axis-1</p> <p>Fisher's exact test: 40% of those diagnosed with binge eating disorder (per SCID) also diagnosed by C-BEDS; 83% with subsyndromal binge eating disorder (per SCID) diagnosed by C-BEDS (sensitivity = 0.71, specificity = 0.89, κ = 0.61); 89% with no binge eating disorder (per SCID) were no binge eating disorder by C-BEDS (sensitivity = 0.4, specificity = 0.72, Fisher's exact test = 0.62)</p>	<p>Conclusion: There was a significant association between C-BEDS and SCID (item, not scale level)</p>
24	Child Feeding Questionnaire (CFQ), 31 item	Birch 2001 ⁴¹⁵ (PDP)	Parent completed	Child; mixed (non-stratified); USA (white, African American, Hispanic) (IC/FA n = 394)	<p>IC: α = 0.79 (range = 0.70–0.92)</p> <p>FA: load range = 0.37–0.95</p>	<p>Developed based on Constanzo and Woody's model ('1985)^d</p> <p>Conclusions: Confirms that following initial scale development, confirmatory FA revealed that the seven-factor model fitted the data well. In addition the scale showed good IC</p>

Eating behaviour questionnaires: tool information						
No.	Name	First author (type of paper) (reference)	Administration	Sample: age; weight status; country (ethnicity), (n)	Evaluation	Comments
25	Child Feeding Questionnaire (CFQ), 31 item	Haycraft 2008 ⁹³ (Eval)	Parent completed	Infant and children; mixed (stratified); UK (race not defined) (inter-rater/ validity $n = 46$)	<p>IR: between mother and father $r = 0.66$ (range = 0.53 to 0.78)</p> <p>Criterion validity: with direct observation $r = 0.15$ (range = 0.04 to 0.28) (mothers); $r = 0.33$ (range = 0.05 to 0.65) (fathers)</p>	<p>Tool development is same as Birch 2001.⁶² With regards to inter-rater there were no significant differences between mother and father</p> <p>Results confirm that fathers' reporting of child feeding practices appear more valid</p> <p>Further results are strong positive correlations between maternal reports and independent assessment of child height ($r = 0.83$ $p < 0.001$) and weight ($r = 0.94$ $p < 0.001$), and for paternal reports and child height ($r = 0.80$ $p < 0.001$) and weight ($r = 0.86$ $p < 0.001$)</p>
26	Child Feeding Questionnaire (CFQ), 31 item	Anderson 2005 ⁹⁶ (ModEval)	Parent completed	Child; mixed (stratified); USA (African American, Hispanic) (FA/validity $n = 216$)	<p>FA: load range = 0.37 to 0.92</p> <p>Construct validity: with BMI $r = 0.14$ (range = 0.01–0.42)</p>	<p>Tool development is same as Birch (2001)⁶²</p> <p>Problems were identified with the Birch model and so this was adapted to find a better fit for the CFQ (Changed from seven factors to five) and 31 items to 16 even although modified problems remained evident for perceived child weight and restriction</p>
27	Child Feeding Questionnaire (CFQ), 31 item	Corsini 2008 ⁹⁷ (Eval)	Parent completed	Child; mixed (non-stratified); Australia (European heritage) (IC/FA/validity $n = 216$)	<p>IC: $\alpha = 0.82$ (range = 0.69–0.83)</p> <p>FA: eight-factor model accounted for 61.7% of variance, (seven factors had an eigenvalue of > 1); load range = 0.34–0.99</p>	<p>Tool development is same as Birch (2001).⁶² Looked at the seven-factor model used in previous research and compared with new eight-factor model with 'food as reward' as new factor</p> <p>The eight-factor model provided the best fit of data. This highlights the</p>

Eating behaviour questionnaires: tool information					
No.	Name	First author (type of paper) (reference)	Administration	Sample: age; weight status; country (ethnicity), (n)	Comments
					Evaluation
					Construct validity: with BMI $r = 0.23$ (range = 0.01–0.53)
					IC: $\alpha = 0.75$ (range = 0.63–0.76)
					FA: total variance 57.6%; load range = 0.41–0.77
					IC: $\alpha = 0.69$ (range = 0.58–0.81)
					FA (CFA): load range = 0.36–1.39
28	Child Feeding Questionnaire (CFQ), 31 item	Polat 2010 ⁹⁴ (Eval)	Parent completed	Infant and children; mixed (non-stratified); Turkey (race not defined) (IC/FA $n = 158$)	problem in the restriction subscale used by previous research Tool development from Birch (2001), ⁶² <i>Conclusion:</i> Results show good reliability and validity of CFQ in Turkish sample
29	Child Feeding Questionnaire (CFQ), 31 item	Boles 2010 ²³² (Eval)	Parent completed	Infant and children; mixed (non-stratified); USA (African American) (IC/FA $n = 296$)	Tool development from Birch (2001) ⁶² Did not test all scales/domains. <i>Conclusions:</i> The study showed a poor factor structure fit. Also Cronbach's alpha scores were slightly less than optimal
30	McKnight Risk Factor Survey-III (MIRFS-III), 75/79 item	Shisslak 1999 ⁸⁷ (PDP)	Self-completed	Children and adolescents; mixed (non-stratified); USA [white, African American, Hispanic, native American, Asian, other (not defined)] (IC/validity $n = 651$)	TRT, IC and convergent validity suggest this tool is a good measure. However, the tool was large with more than 160 questions so it is likely to be an excessive burden for the children
					TRT: elementary: $r = 0.55$, middle: $r = 0.64$, $r =$ high school: 0.69 (range = 0.01–1.00)
					Convergent validity: with Weight Concerns Scale (WCS) $r = 0.82$, range = 0.74–0.88 Rosenberg self-esteem (RSE) $r = 0.61$, range = 0.46–0.73 Depression scales: (CES-D) $r = 0.70$, range = 0.64–0.76 and (CDI) $r = 0.15$

Eating behaviour questionnaires: tool information				
No.	Name	First author (type of paper) (reference)	Administration	Sample: age; weight status; country (ethnicity), (n)
31	Infant Feeding Style Questionnaire (IFSQ), 83 item	Thompson 2009 ⁷⁶ (PDP)	Parent completed	Infant; mixed (stratified); USA (African American) (IC n = 154, FA n = 149)
				<p>IC: H = 0.84 (range = 0.75–0.94)</p> <p>FA (EFA): load range = 0.22–1.51 (also did confirmatory with good model fit)</p> <p>IC uses a H coefficient</p> <p>Exploratory analysis of difference in infant weight z-score associated with feeding scores documented that WLZ was lower in infants whose mother had higher scores on responsive: satiety (–0.39 $p = 0.03$) and pressuring: cereal (–0.52, $p = 0.03$)</p> <p>Conclusions: the IFSQ is an effective instrument in measuring feeding styles and assessing eating behaviour in infants</p>
32	Child Eating Behaviour Questionnaire (CEBQ), 35 item	Sleddens 2008 ⁷² (Eval)	Parent completed	Child; mixed (stratified); the Netherlands (race not defined) (IC/FA n = 135)
				<p>IC: $\alpha = 0.77$ (range = 0.67–0.91)</p> <p>FA: Seven- factor structure accounted for 62.8% of total variance</p> <p>Interscale correlations = –0.59 (EF vs. SR)–0.61 (SR vs. SE); load range = 0.38–0.88</p> <p>Construct validity: with child BMI z-scores showed a linear increase with food approach subscales (FR, EF, EOE) of CEBQ ($\beta = 0.15$ to 0.22), and a decrease in food avoidant subscales (SR, SE, EUE, food fussiness) ($\beta = -0.09$ to –0.25)</p> <p>Significant relationships were found for FR, EF ($p \leq 0.05$) and SR, SE ($p < 0.01$). Difference</p> <p>Tool development is same as Wardle 2001⁷³</p>

Eating behaviour questionnaires: tool information						
No.	Name	First author (type of paper) (reference)	Administration	Sample: age; weight status; country (ethnicity), (n)	Evaluation	Comments
33	Child Eating Behaviour Questionnaire (CEBQ), 35 item	Wardle 2001 ⁷³ (PDP)	Parent completed	Child; mixed (non-stratified); UK (race not defined) (IC study 1 n = 177, study 2 n = 222, FA n = 208)	<p>between weight categories was found for SR (F = 3.69 p < 0.05) and SE (F = 3.86 p < 0.05)</p> <p>IC: $\alpha = 0.82$ (range = 0.72–0.91)</p> <p>TRT: r = 0.78 (range = 0.52–0.87)</p> <p>FA: all had eigenvalues of > 1 and variance ranging from 50% to 80%</p> <p>Interfactor correlations ranged from –0.70 (SR vs. EF)–0.55 (FR vs. EF)</p>	<p>This study included three samples: a pilot study sample, one for factor structure and IC and one for factor structure, IC and TRT</p>
34	Toddler Snack Food Feeding Questionnaire (TSFFQ), 42 item	Corsini 2010 ⁸² (PDP)	Parent completed	Infant and children; mixed (stratified); Australia (race not defined) (IC/FA/validity study 1 n = 175, study 2 n = 216)	<p>IC: $\alpha = 0.84$ (range = 0.75–0.89)</p> <p>TRT: r = 0.8 (range = 0.67–0.90)</p> <p>FA: the five-factor solution accounted for 46.6% of variance (toddlers) and 40.7% (preschoolers)</p> <p>Convergent validity: with CFQ r = 0.20 (toddlers), r = 0.21 (preschool) (range = 0.02–0.43)</p> <p>Construct validity: with diet r = 0.03–0.52</p>	<p>Included two samples: toddlers and preschool</p> <p>Reliability was good but convergent validity with CFQ was poor</p>

Eating behaviour questionnaires: tool information				
No.	Name	First author (type of paper) (reference)	Administration	Sample: age; weight status; country (ethnicity), (n)
35	Kids' Child Feeding Questionnaire (KCFQ), 16 item	Monnery-Patris 2011 ⁸⁵ (Eval)	Self-completed	Child; mixed (stratified); France (race not defined) (IC/FA validity $n = 240$, TRT $n = 34$)
				<p>IC: $\alpha = 0.69$ (range = 0.64–0.74)</p> <p>TRT: $r = 0.77$ (range = 0.67–0.87)</p> <p>FA: average factor load = 0.56</p> <p>Construct validity: with BMI z-score $r = 0.23$ (range = 0.09–0.36)</p> <p>IC: $\alpha = 0.66$ (range = 0.60–0.71)</p> <p>Convergent validity: with DEBQ: those perceiving parent pressure to eat are more likely to be restrained (OR 3.0, $p < 0.01$) have emotional disinhibition (OR 3.2 $p < 0.01$) and external disinhibition (OR 3.0, $p < 0.01$)</p> <p>CFQ: Daughters are 1.5 times more likely to report parental pressure to eat if parent perception of pressure is high (OR 1.5 $p < 0.05$)</p>
36	Kids' Child Feeding Questionnaire (KCFQ), 28 item	Carper 2000 ²⁵⁰ (PDP)	Self-completed	Child; mixed (non-stratified); USA [white, other (not defined)] (IC/validity $n = 197$)
				<p>Tool development is same as Carper (2000)²⁵¹</p> <p>Conclusions: the scale appears to be a sound tool for highlighting children's perceptions of parental feeding practices and their links to weight status. Children's BMI z-scores were positively related to restriction ($r = 0.36$, $p < 0.001$), but they were not significantly related to pressure-to-eat ($r = 0.09$, $p = 0.24$)</p> <p>Only had two scales/domains. Referenced as primary development from Monnery-Patris 2011⁸²</p> <p>This reports that pressure in child feeding is associated with the emergence of dietary restraint and disinhibition among young girls</p>

Eating behaviour questionnaires: tool information

No.	Name	First author (type of paper) (reference)	Administration	Sample: age; weight status; country (ethnicity), (n)	Evaluation	Comments
37	Un-named, 29 item	Murashima 2011 ⁸⁴ (PDP)	Parent completed	Child; mixed (stratified); USA [white, African American, Asian, Hispanic, mixed, other (not defined)] (IC/validity n = 330, TRT n = 35)	<p>IC: $\alpha = 0.67$ (range = 0.59–0.79)</p> <p>TRT: $r = 0.74$ (range = 0.45–0.85)</p> <p>FA: goodness-of-fit: $\chi^2 = 330$ (df 228), CFI = 0.94, RMSEA = 0.04</p> <p>Interfactor correlations = -0.46 (mealtime vs. high control) – 0.61 (high control vs. high contingency)</p> <p>Construct validity: with child BMI z-score ($r = 0.07$, range = 0.02–0.14) and diet ($r = 0.10$ range = 0.02–0.26)</p>	<p><i>Conclusions:</i> a feeding control instrument with seven factors will allow researchers to quantitatively measure a set of parental control feeding practices</p> <p>Initially three models were constructed, which had poor fit, and thus through restructuring and removal of items came the final model, which is included in results and showed a good fit</p>
38	Eating in the Absence of Hunger–Children (EAH-C), 14 item	Tanofsk–Kraff 2008 ⁸⁰ (PDP)	Self-completed	Children and adolescents; mixed (stratified); USA [white, African American, Hispanic, other (not defined)] (IC/validity n = 226, TRT n = 115)	<p>IC: $\alpha = 0.84$ (range = 0.80–0.88)</p> <p>TRT: $r = 0.68$ (range = 0.65–0.70)</p> <p>FA: three factors accounted for 65.3% of total variance; load range = 0.47 to 0.86</p> <p>Convergent validity: with Emotional Eating Scale (EES-C) $r = 0.45$ (range = 0.27–0.61)</p>	<p>People with LOC had higher negative affect scores ($p < 0.01$), external eating ($p < 0.05$) and fatigue/boredom ($p < 0.01$)</p> <p>In addition obese children had higher negative affect scores $p < 0.05$, and higher fatigue/boredom $p < 0.06$</p> <p>No differences were found for external eating</p> <p><i>Conclusion:</i> the EAH-C subscales showed good TRT, IC and convergent validity but had limited discriminate/construct validity</p>

Eating behaviour questionnaires: tool information				
No.	Name	First author (type of paper) (reference)	Sample: age; weight status; country (ethnicity), (n)	
		Administration	Evaluation	
			Comments	
39	Un-named, 21 item	Kroller 2008 ⁸⁸ (PDP)	Parent completed	<p>Child: mixed (stratified); Germany (race not defined) (IC/TRT n = 163)</p> <p>Construct validity: with children's depression ($r = 0.28$, range = 0.23–0.34) and state-trait anxiety ($r = 0.30$, range = 0.24–0.37)</p> <p>IC: $\alpha = 0.8$ (range = 0.73–0.93)</p> <p>TRT: $r = 0.58$ (range = 0.41–0.78)</p> <p>Results showed that maternal subjective weight category had no significant effect on use of feeding strategies</p> <p>Also children eating more fruit and vegetables had parents who used more child control of feeding and less rewarding with food</p>
40	Child Eating Behaviour Questionnaire (Portuguese)	Viana 2008 ¹⁴ (non-English)	Translation was not possible but this measure has already been included in the review by Sleddens 2008 ⁸⁹ and Wardle 2001 ⁶⁰ (above)	<p>Children eating more snack foods had parents who used more pressure to eat and finally heavier children had parents who used less pressure to eat and allowed less child control of feeding</p>

CFA, confirmatory factor analysis; CFI, comparative fit index; DSM-IV, *Diagnostic and Statistical Manual of Mental Disorders* 4th edition; EF, enjoyment of food; EOE, emotional overeating; EFA, exploratory factor analysis; EUE, emotional undereating; Eval, evaluated an existing tool without modification; FR, food responsiveness; ModEval, modified an existing tool and re-evaluated; OR, odds ratio; PDP, primary development paper; RMSEA, root-mean-square error of approximation; SE, slowness in eating; SR, satiety responsiveness; WLZ, weight-for-length z-score.

a Not linked to bibliography: Gamer DM, Garfinkel PE. The eating attitudes test: an index of the symptoms of anorexia nervosa. *Psychol Med* 1989;**9**:273–9.

b Not linked to bibliography: Goldfein JA, Devlin MJ, Kamenetz C. Eating Disorder Examination-Questionnaire with and without instruction to assess binge eating in patients with binge eating disorder. *Int J Eat Disord* 2005;**37**:107.

c Not linked to bibliography: Arnow B, Kenardy J, Agras WS. The emotional eating scale: the development of a measure to assess coping with negative affect by eating. *Int J Eat Disord* 1995; **18**:79–90.

d Not linked to bibliography: Costanzo PR, Woody EZ. Domain-specific parenting styles and their impact on the child's development of particular deviance: the example of obesity proneness. *J Soc Clin Psychol* 1985;**3**:425.

Appendix 9 Physical activity measurement studies: summary table

No.	Tool information: name	First author and type of paper (reference)	Administration	Sample Age; weight status; country (ethnicity), (n)	Evaluation	Comments
1	Accelerometer	Kelly 2004 ¹⁰⁵ (Eval)	Self-complete; data download	Child; overweight; UK (race not defined) (validity n = 78)	Criterion validity: with direct observation (CPAF) r = 0.72 Convergent validity: with Actiwatch r = 0.36	Also correlated Actiwatch with direct observation (r = 0.16), showing CSA has greater correlation with direct observation Correlations for Actiwatch improved when assessed minute by minute compared with total PA, but accelerometer was better with total PA
2	Accelerometer – Actigraph	Pate 2006 ¹⁰⁷ (Eval)	Self-complete; data download	Child; mixed (non-stratified); USA (white; African American) (validity n = 29)	Criterion validity: with VO ₂ measured by COSMED r = 0.82	Accelerometer counts were highly correlated with VO ₂ in young children
3	Accelerometer – Caltrac monitor	Noland 1990 ¹⁰⁶ (Eval)	Self-complete; data download	Infant and children mixed (stratified); USA (white; African American) (validity n = 48)	Criterion validity: with direct observation r = 0.86 (range = 0.86–0.89)	The Caltrac accelerometer has excellent criterion validity when compared with direct observation
4	Accelerometer – TriTrac Triaxial	Coleman 1997 ¹⁰⁸ (Eval)	Self-complete; data download	Child; all obese; USA (race not defined) (validity n = 35)	Criterion validity: with HR r = 0.71 Convergent validity: with activity diaries r = 0.38	The accelerometer showed good correlation with HR in assessing PA
5	Accelerometer – Actigraph	Guinhouya 2009 ²³⁴ (Eval)	Self-complete; data download	Child; mixed (stratified); France (race not defined) (n = 113)	Construct validity: with BMI r = 0.23 [based on IOTF criteria: cut-off point 3600 pcm had the highest probability of correct decision (0.62), the lowest misclassification errors (0.38), the highest validity coefficient (0.21) and the highest expected maximum utility ^{23]}	<i>Conclusion:</i> when children are classified using BMI based criteria, the threshold at 3600 pcm = appropriate in discriminating normal weight for overweight/obesity

No.	Tool information: name	First author and type of paper (reference)	Administration	Sample Age; weight status; country (ethnicity), (n)	Evaluation	Comments
6	HR monitoring	Maffei 1995 ²³⁷ (Eval)	Self-complete; data download	Child; mixed (stratified); Italy (white) (validity n = 13)	<p>Criterion validity: with DLW = Bland-Altman</p> <p>Level of agreement TEE (HR) vs. TEE (DLW) in obese = 0.04 mJ/day (non-obese -0.59). The <i>t</i>-test shows that the difference between TEE (HR)-TEE (DLW) is 0.48 mJ/day in obese (0.2 mJ/day in non-obese (<i>p</i> = non-significant)</p> <p>Agreement between DLW and HR on individual level ranged = -2.8-9.1% in obese</p>	<p>Results show that the discrepancy between HR and DLW is greater in obese children</p> <p>Note: although added to PA domain, may be considered a measure of fitness</p>
7	Pedometer	Kilanowski 1999 ¹¹⁴ (Eval)	Self-complete; data download	Child; mixed (non-stratified); USA (race not defined) (validity n = 10)	<p>Criterion validity: with accelerometer r = 0.74 and direct observation r = 0.89</p>	<p>Reliability of direct observation was assessed by two observers for 1545 out of 1793 observations</p> <p>%Agreement = 86%</p>
8	Pedometer (SW-200 and NL-2000 models)	Duncan 2007 ²⁴⁸ (Eval)	Self-complete; data download	Child; mixed (stratified); New Zealand [white; Asian, Polynesian; other (not defined)] (validity n = 85)	<p>Criterion validity: with direct observation r = 0.85 SW-2000 (range = 0.77-0.96); r = 0.91 NL-2000 (range = 0.81-0.99)</p>	<p>Results show that the pedometer has good correlation with both direct observation and accelerometer</p> <p>Pedometer slightly under-reports, but precision increases with speed of walking</p> <p><i>Stratification:</i> reduction on mean per cent bias with increasing speeds varies with sex (and age group - NL-2000 only)</p> <p>No significant associations detected between %BIA and BMI, WC or %BF. Pedometer tilt angle was associated with mean per cent bias for both pedometers (SW-200: F = 22.689, <i>p</i> < 0.01; NL-2000: F = 6.310, <i>p</i> = 0.01) regardless of sex, age, speed or body composition</p> <p>SW-200 (< 10°) tilt, per cent bias 5.5% but ≥ 10° = 14% bias NL-2000 (< 10°) tilt, per cent bias 7.1% but ≥ 10° = 10.7% bias</p>

No.	Tool information: name	First author and type of paper (reference)	Administration	Sample Age; weight status; country (ethnicity), (n)	Evaluation	Comments
9	Pedometer	Jago 2006 ¹¹² (Eval)	Self-complete; data download	Children and adolescents; mixed (stratified); USA (Anglo American, African American, Asian, Hispanic) (validity and TRT n = 78)	<p>TRT: r = 0.77 (range 0.51–0.92)</p> <p>Criterion validity: with accelerometer r = 0.60</p>	<p>This study was conducted in boys only. The author concludes that the pedometer provides an accurate assessment of PA and an estimate of 8000 pedometer counts in 60 minutes is equivalent to 60 minutes of MVPA</p> <p>Further results show there was a significant group main effect with number of pedometer steps recorded by varying adiposity status with participant at risk for overweight recording lower counts than normal weight children for the same activity (normal weight had six more counts than overweight in same activity)</p>
10	Pedometer	Mitre 2009 ¹¹⁰ (Eval)	Self-complete; data download	Child; mixed (stratified); USA (white, Asian) (validity and TRT n = 27)	<p>TRT: compared steps counted on both sides of the body at all speeds, mean difference in two measurements was 10% (Omron pedometer) 9% (Yamax pedometer)</p> <p>Criterion validity: with direct observation</p> <p>Assessed per cent error. Error decreased with increasing speed; at 0.5 mph error = ~100% in both pedometers but for 2 mph the error was ~60%. The errors were 92% for under-reporting and close to 8% for over-reporting</p>	<p>Normal weight children showed lower per cent error compared with overweight (Omron $p < 0.0001$) (Yamax $p < 0.0002$)</p> <p>This study also assessed accuracy of the accelerometer per cent error was 24% at 0.5 mph, 5% at 1 mph and 2% at 2 mph. Furthermore when children worked at their own pace average speed was 2.5 mph and this improved per cent error: Omron = 36% and Yamax = 21%</p> <p><i>Author's conclusion:</i> pedometers are inaccurate for children, especially for overweight or obese</p>
11	Pedometer	Treuth 2003 ¹¹³ (Eval)	Self-complete; data download	Child; mixed (non-stratified); USA (African American) (TRT n = 57, validity n = 68)	<p>TRT: r = 0.08</p> <p>Criterion validity: with accelerometer r = 0.47 (range by days = 0.14 in (day 1) –0.64 (day 3))</p>	<p>The pedometer shows extremely poor TRT reliability and adequate criterion validity with CSA accelerometer</p>

No.	Tool information: name	First author and type of paper (reference)	Administration	Sample Age, weight status; country (ethnicity), (n)	Evaluation	Comments
12	SenseWear Pro2 Armband, models 5.1 and 6.1	Backlund 2010 ⁴¹⁶ (Eval)	Self-complete; data download	Child; obese and overweight; Sweden (race not defined) (validity $n = 22$)	Criterion validity: with DLW = t -tests showed model 5.1 SenseWear and DLW were all similar ($p > 0.05$) but model 6.1 were all different ($p < 0.001$) SenseWear underestimated by 1884 kJ/day in girls and 2039 kJ/day in boys Values similar with compliant and non-compliant	Results confirm that the SWA5.1 is an adequate tool, as it does not differ from DLW, whereas the SWA6.1 significantly underestimates when compared with DLW
13	3-day Physical Activity Recall (3DPAR)	Pate 2003 ⁴¹⁷ (Eval)	Self-complete; pen and paper	Adolescent; mixed (non-stratified); USA [white; African American; other (not defined)] (validity $n = 70$)	Convergent validity: with SWA5.1 vs. SWA6.1 found that SWA5.1 estimated higher METs of activity in boys (compared with girls) than the SWA6.1 Statistical differences between genders was greater for SWA5.1 compared with SWA6.1	Results confirm adequate correlations of 3DPAR with CSA accelerometer over 3 days and 7 days
14	Activity Questionnaire for Adults and Adolescents (AQuAA)	Slootmaker 2009 ¹⁷ (Eval)	Self-complete; pen and paper	Adolescent; mixed (stratified); the Netherlands (race not defined) (validity $n = 236$)	Criterion validity: with accelerometer = questionnaire always higher than accelerometer In overweight adolescents, minutes/week of MPA = 480 in AQuAA vs. 162 in accelerometer. VPA is 0 vs. 29 minutes/week, and MVPA is 553 vs. 166 minutes/week	Primary development is based on the SQUASH questionnaire but this is in adults. Results confirm that the questionnaire overestimates PA when compared with accelerometer

No.	Tool information: name	First author and type of paper (reference)	Administration	Sample		Evaluation	Comments
				Age; weight status; country (ethnicity), (n)			
15	Activity rating scale (1 item)	Sallis 1993 ¹²¹ (Eval)	Self-complete; pen and paper	Children and adolescents; mixed (stratified); USA [white; African American; Asian, Hispanic, other (not defined)] (TRT/validity n = 102)	According to AquAA, normal weight = more active in MPA and VPA than overweight, but accelerometer shows normal weight = less active than overweight in MPA (81 vs. 162 minutes, $p = 0.008$) and VPA (12 vs. 29 minutes, $p = 0.05$) TRT: $r = 0.89$ (range = 0.77–0.93) Convergent validity: with Godin–Shephard PA survey and kilocalorie expenditure index $r = 0.32$ (Godin) $r = 0.22$ (kilocalorie expenditure index)	Does not state primary development study. States that it was used in Sallis 1988. Results confirm good reliability but only fair convergent validity	
16	Godin–Shephard Physical Activity Survey (3 item)	Sallis 1993 ¹²¹ (Eval)	Self-complete; pen and paper	Children and adolescents; mixed (stratified); USA [white; African American; Asian, Hispanic, other (not defined)] (TRT/validity n = 102)	TRT: $r = 0.81$ (range = 0.69–0.96) Convergent validity: with Activity Rating Scale ($r = 0.32$) and kilocalorie expenditure index ($r = 0.39$)	Development paper = (Godin 1985 ^a) – but this is adults. Results confirm good reliability, fair convergent validity	
17	7-day recall interview	Sallis 1993 ¹²¹ (Eval)	Interview administered in person – child; pen and paper	Children and adolescents; mixed (stratified); USA [white; African American; Asian, Hispanic, other (not defined)] (TRT/validity n = 102)	TRT: $r = 0.65$ (range = 0.54–0.77) Criterion validity: with HR $r = 0.49$ (range = 0.44–0.53)	Reference = Sallis (1985) ^b for more on development – but regarding adults Results show adequate TRT, and adequate convergent validity when compared with HR	
18	Adolescent Physical Activity Recall Questionnaire (APARQ) (4 item)	Booth 2002 ¹²³ (PDP)	Self-complete; pen and paper	Adolescent; mixed (non-stratified); Australia (race not defined) (TRT n = 226, validity n = 2026)	TRT: $r = 0.69$ (range = 0.52–0.76); ICC = 0.58 (range = 0.52–0.62); agreement = 77% (range = 66–88%) Convergent validity: with multistage fitness test $r = 0.22$ (range = 0.15–0.39)	Gives ethnicity of European, Middle Eastern and Asian Backgrounds for validation study only The APARQ has adequate reliability and poor convergent validity. In addition, the two category measure (active and inactive) was shown to have better reliability than the three-category measure (vigorous, adequate and inactive)	

No.	Tool information: name	First author and type of paper (reference)	Administration	Sample Age; weight status; country (ethnicity), (n)	Evaluation	Comments
19	Children's Leisure Activities Study Survey (CLASS) (30 item)	Telford 2004 ¹¹⁵ (PDP)	Self-complete; parent complete; pen and paper	Child; mixed (non-stratified); Australia (race not defined) (TRT/inter-rater/ validity n = 280)	<p>TRT: child r = 0.24–0.42; parents r = 0.72–0.81</p> <p>IR: r = 0.19; agreement range = 8% (tennis) –85.7% (soccer)</p> <p>Criterion validity: with accelerometer: child r = 0.04 (range = 0.02–0.06); parent r = 0.09 (range = 0.06–0.14)</p>	<p>Neither SR or parent proxy provided an accurate assessment of children PA</p> <p>Results show the CLASS under-reports moderate PA but over-reports vigorous and total PA when compared with accelerometer. Also SR and parent proxy assessment of PA is poorly correlated</p>
20	GEMS Activity Questionnaire (GAQ) (28 item)	Treuth 2003 ¹¹³ (ModEval)	Self-complete; pen and paper	Child; mixed (non-stratified); USA (AA) (TRT/validity n = 67)	<p>TRT: r = 0.59 (range = 0.34–0.82)</p> <p>Criterion validity: with accelerometer r = 0.27 (range = 0.21–0.30)</p>	<p>Modified and evaluated from SAPAC (Sallis 1996).^c The GAQ is an acceptable measure of PA showing good correlations with the CSA accelerometer</p>
21	Activitygram	Treuth 2003 ¹¹³ (ModEval)	Self-complete; web-based tool	Child; mixed (non-stratified); USA (AA) (TRT/validity n = 67)	<p>TRT: r = 0.24</p> <p>Criterion validity: with accelerometer r = 0.37 (range = 0.08–0.43)</p>	<p>Suggests that the Activitygram is based on the PDPAR</p> <p>Results show that the Activitygram has poor TRT reliability and poor to fair correlation with CSA accelerometer</p>
22	Activitygram	Welk 2004 ¹²⁴ (Eval)	Self-complete; data download	Child; mixed (non-stratified); USA (white, AA, Asian, Hispanic, Native American) (criterion n = 28, convergent n = 147)	<p>Criterion validity: with accelerometer r = 0.43 (range = 0.33–0.50)</p> <p>Convergent validity: with PDPAR r = 0.44 (range = 0.35–0.53)</p>	<p>For convergent validity results presented are for both schools combined</p> <p>It is clear that school 1 obtained much higher correlations [mean = 0.72 (range 0.63 to 0.80)] compared with school 2 [mean = 0.30 (range: 0.22 to 0.41)]</p> <p>The author confirms that the large discrepancies between schools could be due to less staff support in school 2</p>

No.	Tool information: name	First author and type of paper (reference)	Administration	Sample		Comments
				Age; weight status; country (ethnicity), (n)	Evaluation	
23	Moderate to vigorous physical activity screening (9 item)	Prochaska 2001 ²³⁵ (study 3) (ModEval)	Self-complete; pen and paper	Adolescent; mixed (non-stratified); USA [white; AA; Asian, Hispanic, Pacific Islander; mixed ethnicity; other (not defined)] (TRT/validity n = 138)	<p>TRT: r = 0.77 (range = 0.53–0.88)</p> <p>Criterion validity: with accelerometer r = 0.40 (range = 0.32–0.42); ICC was 0.77; κ = 61%; correct classification rate = 63%, with 71% sensitivity and 40% false-positive rate</p>	Three studies in one paper (a pilot study was carried out in study 1 n = 6). Study 2 and study 3 evaluated over in two data extraction forms Results show that the moderate to vigorous PA screening measure has adequate correlation with accelerometer
24	Moderate to Vigorous Physical Activity screening (9 item)	Prochaska 2001 ²³⁵ (study 2) (PDP)	Self-complete; pen and paper	Adolescent; mixed (non-stratified); USA [white; AA, Asian Hispanic, Pacific Islander; other (not defined)] (TRT n = 250, validity n = 57)	<p>TRT: r = 0.68 (range = 0.55–0.79); agreement = 52% (range = 47–61%)</p> <p>Criterion validity: With accelerometer r = 0.34 (range = 0.20–0.46); 60 minutes MPA: composite: classification rate = 78%; sensitivity = 80%; false-positive rate = 40% VPA composite: classification rate = 58%; sensitivity = 38%; false-positive rate = 0%</p>	Conducted two studies in one. This is study 1 and is the primary development paper The moderate to vigorous activity screening measure showed adequate inter-rater reliability and on the verge of poor criterion validity when compared with accelerometer
25	National Longitudinal Survey of Children and Youth (4 item)	Sithole 2008 ¹³⁰ (Eval)	Self-complete; parent complete; pen and paper	Child; mixed (stratified); Canada (race not defined) (inter-rater n = 3940)	<p>IR: κ = 0.24 (range = 0.11–0.41)</p> <p>Repeated with obese only and results = organised sports (κ = 0.37), leisure sports (κ = 0.11), television viewing (κ = 0.10), and computer use and video games (κ = 0.25)</p>	Results show that children reporting more PA are more likely to be obese/overweight. The child reporting more organised sport = more chance of being obese (OR = 1.33 p < 0.05) and leisure sports (OR = 1.39 p < 0.05) Parent reporting > 3 hours' day television viewing = more chance of overweight/obese (OR = 0.168 p < 0.05) and computer games OR = 1.23
26	Outdoor Playtime checklist (6 item)	Burdette 2004 ¹²² (study 1) (PDP)	Parent complete; pen and paper	Infant and children; mixed (non-stratified); USA (white; AA) (validity n = 250)	<p>Criterion validity: with accelerometer r = 0.33</p> <p>Convergent validity: with outdoor playtime recall r = 0.57</p>	The parent-reported outdoor playtime checklist showed fair correlation with accelerometer

No.	Tool information: name	First author and type of paper (reference)	Sample Age; weight status; country (ethnicity), (n)	Administration	Evaluation	Comments
27	Outdoor Playtime Recall (2 item)	Burdette 2004 ¹²² (study 2) (PDP)	Infant and children; mixed (non-stratified); USA (white; AA) (criterion validity $n = 214$, convergent validity $n = 250$)	Parent complete; pen and paper	Criterion validity: with accelerometer $r = 0.20$ Convergent validity: with outdoor playtime checklist $r = 0.57$	The parent-reported outdoor playtime recall showed poor correlation with accelerometer
28	Physical Activity Diary	Epstein 1996 ¹¹⁹ (PDP)	Child; all obese; USA (white; AA; Hispanic) (validity $n = 59$)	Self-complete; pen and paper	Criterion validity: with accelerometer $r = 0.46$; self-report energy expenditure = 43% higher than accelerometer Convergent validity: with Child Behaviour Checklist (CBCL) (beta = 0.02), Beck Depression Index (BDI) (beta = 0.02), Bulimia Test (BT) (beta = 0.003), CMI (beta = 0.2), Parent Inventory of Problems (PIP) (beta = -0.2). All p values were NS except for CBCL	Results confirm self-report energy expenditure was 43% higher than accelerometer
29	Physical Activity Questionnaire (PAQ) (8/9 item)	Janz 2008 ¹²⁹ (ModEval)	Children (C)/adolescents (A); mixed (non-stratified); USA (white) (IC/TRT/FA $n = 210$; validity $n = 49$)	Self-complete; pen and paper	IC: PAQ-C $\alpha = 0.74$ (range = 0.72–0.78) (rescaled $\alpha = 0.77$) PAQ-A $\alpha = 0.79$ (range = 0.77–0.88) (rescaled $\alpha = 0.84$) FA: load range PAQ-C = 0.02–0.80, PAQ-A = 0.04–0.74. Only 1 eigenvalue ≥ 1 Criterion validity: with accelerometer Total PA $r = 0.37$ (range = 0.14–0.51); per cent day MVPA $r = 0.42$ (range = 0.18–0.61) (adolescents only)	Conclusion: PAQ-C and PAQ-A show good IC The PAQ-A has acceptable validity. Tool development is same as Crocker 1997 ¹²⁸ as this is primary development paper Questions were slightly modified to adapt to the sample: e.g. snowboarding was included in because it is popular in the sample

No.	Tool information: name	First author and type of paper (reference)	Sample Age; weight status; country (ethnicity), (n)	Administration	Evaluation	Comments
30	Physical Activity Questionnaire for Older Children (9 item)	Kowalski 1997 ¹¹⁸ (Eval)	Children and adolescents; mixed (non-stratified); Canada (race not defined) [criterion validity $n = 97$, validity $n = 89$ (97 in study 2)]	Self-complete; pen and paper	Criterion validity: with accelerometer $r = 0.39$ Convergent validity: with activity rating (study 1 $r = 0.63$, study 2 $r = 0.57$); teachers' rating (study 1 $r = 0.45$); moderate to vigorous PA (study 1 $r = 0.47$); 7 Day Recall Interview (physical activity ratio) (study 2 $r = 0.46$); Leisure Time Exercise Questionnaire (Godin) (study 2 $r = 0.49$)	Kowalski conducted two studies ^{118,125} in the same year. This study is Kowalski 1997. ¹¹⁸ Also this article conducted two studies in one publication with only the activity recall the sole convergent measure to be assessed in both. PAQ-C had moderate correlations with other PA measures Results show that the PAQ-C shows greatest correlation with activity rating
31	Physical Activity Questionnaire for Adolescents (PAQ-A) (8 item)	Kowalski 1997 ¹²⁵ (ModEval)	Adolescent; mixed (non-stratified); Canada (race not defined) (criterion validity $n = 48$, convergent validity $n = 85$)	Self-complete; pen and paper	Construct validity: with Harter's athletic competence ($r = 0.32$) and behavioural conduct ($r = \text{non-significant}$) (supports divergent validity), Canadian Home Fitness Test $r = 0.28$ Criterion validity: with accelerometer $r = 0.33$ Convergent validity: with activity recalls $r = 0.60$ (Godin) $r = 0.59$ (PAR) $r = 0.73$ (Activity rating)	Conclusion: the PAQ-A was moderately correlated to other measures of PA and supports its use in high school students The PAQ-A showed greatest correlation with the activity rating and on the verge of poor correlation with accelerometer
32	Physical Activity Questionnaire for Older Children (PAQ-C) (10 item)	Crocker 1997 ¹²⁸ (study 1) (PDP)	Children and adolescents; mixed (non-stratified); Canada (race not defined) (IC $n = 215$)	Self-complete; pen and paper	IC: $\alpha = 0.83$ (range = 0.80–0.83)	Conducted three studies in one, and so has three data extraction forms. These are the results for study 1. Results show good IC

No.	Tool information: name	First author and type of paper (reference)	Administration	Sample Age; weight status; country (ethnicity), (n)	Evaluation	Comments
33	Physical Activity Questionnaire for Older Children (PAQ-C) (10 item)	Crocker 1997 ¹²⁸ (study 2) (PDP)	Self-complete; pen and paper	Children and adolescents; mixed (non-stratified); Canada (race not defined) (IC/TRT n = 84)	IC: $\alpha = 0.84$ (range = 0.79 to 0.89) TRT: $r = 0.79$ (range = 0.75–0.82)	Study 2: Results show good IC and good TRT
34	Physical Activity Questionnaire for Older Children (9 item)	Crocker 1997 ¹²⁸ (study 3) (PDP)	Self-complete; pen and paper	Children and adolescents; mixed (non-stratified); Canada (race not defined) (IC/TRT n = 200)	IC: α range = 0.81–0.86 TRT: generalisability coefficient for average of three scores (sent out over three seasons) is $G = 0.88$ Generalisation across two scores: sent out over two seasons is $G = 0.83$	Study 3: Results show good IC and good TRT
35	Physical Activity Questionnaire for Older Children (PAQ-C) (8 item)	Moore 2007 ¹²⁶ (study 2) (ModEval)	Self-complete; pen and paper	Child; mixed (non-stratified); USA (AA; European American, Hispanic, Native American, mixed ethnicity) (IC/FA validity n = 414)	IC: $\alpha = 0.66$ (range = 0.56–0.75) FA: Two-factor model goodness of fit: $\chi^2 = 65.71$, RMSEA ≤ 0.05 , NNFI = 0.96, CFI = 0.98 Construct validity: with blood pressure ($r = 0.07$), cardiovascular fitness ($r = 0.16$), BMI ($r = 0.09$), athletic competence ($r = 0.14$), enjoyment of PA ($r = 0.14$), physical appearance (non-significant), global self-worth (non-significant), Task and Ego orientation (non-significant)	Tool development scores same as Crocker 1997. ¹²⁸ Conducted two studies in one and so has two data extraction forms Validity varied by race and so modification may be necessary
36	Physical Activity Questionnaire for Older Children (PAQ-C) (9 item)	Moore 2007 ¹²⁶ (study 1) (Eval)	Self-complete; pen and paper	Children and adolescents; mixed (non-stratified); USA (Hispanic, AA, European American, other (not defined)) (IC/FA n = 991, validity n = 404)	IC: $\alpha = 0.72$ (range = 0.70–0.74) CFA: load range = 0.41–0.74 Factor 3 was a single-item factor (lunch) so analysis was run excluding this CFA = two-factor model = $\chi^2 = 246.11$ RMSEA ≤ 0.01 NNFI = 1.00, CFI = 1.00	Tool development scores same as Crocker 1997. ¹²⁸ Conducted two studies in one and so two data extraction forms are filled out

No.	Tool information: name	First author and type of paper (reference)	Administration	Sample Age; weight status; country (ethnicity), (n)	Evaluation	Comments
37	Physical Activity Questionnaire for Pima Indians (7 item)	Kriska 1990 ²³⁶ (PDP)	Interview administered in person – child; pen and paper	Children and adolescents; mixed (non-stratified); USA (Alaska Native/Native American) (TRT n = 23)	Construct validity: with %BF (BIA) (r = 0.10), cardiovascular fitness (Harvard Step Test and HR) (r = 0.08), BMI (r = NS), glucose (r = NS) TRT: r = 0.36 (range = 0.35–0.37)	Age group was 10–59 years. Validity was tested but not stratified by age so validity was not included Results show that TRT reliability was poor (best in older children and slightly better when recalling the past year)
38	Physical activity Questionnaire for Pima Indians (7 item)	Goran 1997 ¹²⁷ (Eval)	Interview administered in person – child; pen and paper	Adolescent; mixed (non-stratified); Sweden; USA (white; Mohawk) criterion validity n = 166, construct validity n = 83 (study 1) 58 (study 2)	Criterion validity: with DLW r = NS Construct validity: with obesity; fat mass – (BIA) study 1 r = 0.24 (0.32 to 0.33); study 2 r = 0.33 (non-significant to 0.24)	This PA questionnaire is different from the PAQ-C and PAQ-A – it is the same PA questionnaire used and developed by Kriska 1991 in Pima Indians. Results show poor criterion validity with DLW and poor construct validity with BIA
39	Previous Day Physical Activity Recall (PDPAR)	Trost 1999 ⁴¹⁸ (Eval)	Self-complete; pen and paper	Child; mixed (non-stratified); USA [AA; other (not defined)] (validity n = 37)	Criterion validity: with accelerometer r = 0.36 (range = 0.19–0.57)	Full description of PDPAR and scoring protocol is found in Weston (1997). ¹²⁰ The PDPAR shows poor correlation with accelerometer
40	Previous Day Physical Activity Recall (PDPAR)	Weston 1997 ¹²⁰ (Eval)	Self-complete; pen and paper	Children and adolescents; mixed (non-stratified); USA [white; other (not defined)] (TRT = 90; inter-rater n = 112, criterion validity n = 26, convergent validity n = 48)	TRT: r = 0.98 IR: r = 0.99 Criterion validity: with HR r = 0.33 (range = 0.16 to 0.53) Convergent validity: with pedometer (r = 0.88) and CALTRAC Personal Activity Computer (r = 0.77)	Results reveal the PDPAR did not accurately assess PA when compared with HR Greater correlations were evident when compared with pedometer and CALTRAC in convergent validity but this was not reported by scale category

No.	Tool information: name	First author and type of paper (reference)	Administration	Sample Age; weight status; country (ethnicity), (n)	Evaluation	Comments
41	Previous Day Physical Activity Recall (PDPAR)	Welk 2004 ¹²⁴ (Eval)	Self-complete; pen and paper	Child; mixed (non-stratified); USA (white, AA, Asian, Hispanic, Native American) (criterion n = 28, convergent n = 147)	Criterion validity: with accelerometer r = 0.53 (range = 0.22–0.73) Convergent validity: with Activitygram r = 0.44 (range = 0.35–0.53)	For convergent validity, results presented are for both schools combined It is clear that school 1 obtained much higher correlations [mean = 0.72 (range 0.63 to 0.80)] compared with school 2 [mean = 0.30 (range: 0.22 to 0.41)] The author confirms that the large discrepancies between schools could be due to less staff support in school 2
42	Previous Day Physical Activity Recall (PDPAR)	McMurray 2008 ¹¹⁹ (Eval)	Self-complete; pen and paper	Children and adolescents (girls); mixed (stratified); USA (white, AA, other—not stated) (validity n = 691)	Criterion validity: with accelerometer. Compared accelerometer (MVPA minutes/day) vs. PDPAR (MVPA blocks/day) using mixed-model regression analysis For normal weight; 25 minutes/day vs. 1.5 block/day. In at risk 22.5 minutes/day vs. 2 block/day and for overweight 20 minutes/day vs. 1.75 blocks/day With $p < 0.01$ for BMI categories	This study was done in girls only and it was concluded that overweight girls tend to over-report their total PA. Further results from a B ratio analyses showed that those girls at risk obtained 17.7% fewer minutes of MVPA per block and overweight 19.4% fewer when compared with normal weight
43	Youth Risk Behaviour Survey (YRBS)	Troped 2007 ²³⁸ (Eval)	Self-complete; pen and paper	Children and adolescents; mixed (stratified); USA (white, AA, Asian, Hispanic, Native Hawaiian, Alaska Native/Native American) (TRT/validity = 125)	TRT: r = 0.49 (range = 0.46–0.51) Criterion validity: with accelerometer = κ for four measures range = -0.05 to 0.03 Moderate PA: sensitivity range = 0.00–0.23; specificity range = 0.74 to 0.92 Vigorous PA: sensitivity range = 0.75 to 0.92; specificity range = 0.23 to 0.26	The YRBS underestimates the proportion of students attaining recommended levels of moderate PA and overestimates the proportion meeting vigorous recommendations Some information for development was gathered from Kolbe 1993 ^d

No.	Tool information: name	First author and type of paper (reference)	Sample Age; weight status; country (ethnicity), (n)	Administration	Evaluation	Comments
44	System for Observing Children's Activity and Relationships during Play (SOCARP)	Ridgers 2010 ¹⁰² (PDP)	Child; mixed (non-stratified); UK (race not defined) (TRT n = 14, inter-rater n = 2 observers 27 children, validity n = 99)	Researcher conducted/observed; pen and paper	<p>TRT: observer coded 14 children on two occasions within a week</p> <p>Per cent agreement was: activity level (87%), group size (85%), activity type (93%), interactions (87%)</p> <p>IR: agreement = 89% (range 88–90%)</p> <p>Criterion validity: with accelerometer r = 0.67</p>	<p>Was stratified by overweight but not for reliability and validity test</p> <p>There were 42% overweight in entire sample and direct observation was recommended by the collaborators at the CoOR meeting</p> <p>Results also showed that normal-weight children tended to engage in more MVPA and VPA than overweight children, but results were not significant</p> <p>In the whole sample %MVPA was correlated with sport activities (r = 0.28), being in large groups (r = 0.23), frequency of physical conflict (r = 0.27), availability of equipment (r = 0.24), sedentary activities (r = 0.54) and higher temperature (r = 0.21)</p>
45	Observational System for Recording Physical Activity (OSRAC)	Brown 2006 ¹⁰³ (PDP)	Child; mixed (non-stratified); USA (race not defined) (sample size not given)	Researcher conducted/observed; pen and paper	<p>Inter-rater: r = 0.96 (range 0.90–1.0) and κ = 0.87 (range 0.79–0.93)</p>	<p>Preschoolers spent majority of observational intervals as sedentary and MVPA was less frequent (5% or fewer intervals)</p>

CFA, confirmatory factor analysis; CFI, comparative fit index; CMI, Cornell Medical Index; CPAF, Children's Physical Activity Form; Eval, evaluated an existing tool without modification; CSA, CSA/MTI WAM-7164; ICC, intraclass correlation coefficient; mph, miles per hour; ModEval, modified an existing tool and re-evaluated; MPA, moderate physical activity; MVPA, moderate to vigorous physical activity; NNFI, Non-normed Fit Index; PDP, primary development paper; PDPAR, Primary Development Paper; Weston; VPA, vigorous physical activity.

a Not linked to bibliography: Godin G, Shephard RJ. A simple method to assess exercise behavior in the community. *Can J Appl Sport Sci* 1985;**10**:141–6.

b Not linked to bibliography: Sallis JF, Haskell WL, Wood PD, Fortmann SP, Rogers T, Blair SN, et al. Physical activity assessment methodology in the Five City Project. *Am J Epidemiol* 1985;**121**:91–106.

c Not linked to bibliography: Sallis JF, Strikmiller PK, Harsha DW, Feldman HA, Ehlinger S, Stone EJ, et al. Validation of interviewer- and self-administered physical activity checklists for fifth grade students. *Med Sci Sports Exerc* 1996;**28**:840–51.

d Not linked to bibliography: Kolbe LJ, Kann L, Collins JL. Overview of the Youth Risk Behavior Surveillance System. *Public Health Rep* 1993;**108**(Suppl. 1):2–10.

Appendix 10 Sedentary time/behaviour measurement studies: summary table

Sedentary time/behaviour measures						
No.	Tool information: name	First author and type of paper (reference)	Administration	Sample: age; weight status; country (ethnicity), (n)	Evaluation	Comments
1	WAM-7154 accelerometer	Reilly 2003 ¹³¹ (Eval)	Self-complete; data download	Child; mixed (non-stratified); UK (race not defined) (validity $n = 52$)	Criterion validity: with direct observation (CPAF); sensitivity was 83% (438/528 inactive minutes were correctly classified) specificity was 82% (1251/1526 non-inactive minutes correctly classified) was obtained from a cut-off of < 1100 counts/minute	Sedentary behaviour can be quantified objectively in young children using an accelerometer A cut-off point of < 1100 counts/minute established good sensitivity and specificity when compared with direct observation
2	Computer science and Actigraph accelerometer	Puyau 2002 ¹³² (study 1) (Eval)	Self-complete; data download	Children and adolescents; mixed (non-stratified); USA (white; AA; Asian; Hispanic) (validity $n = 26$)	Criterion validity: with energy expenditure from room calorimetry $r = 0.70$ (range = 0.66–0.73) With HR $r = 0.60$ (range = 0.57–0.63) With microwave activity $r = 0.67$ (range = 0.61–0.72)	Can also be extracted for PA domain. Assessed two measures in one study. The CSA showed excellent criterion validity, in particular when compared with room calorimetry
3	Mini-Mitter Actiwatch monitors	Puyau 2002 ¹³² (study 2) (Eval)	Self-complete; data download	Children and adolescents; mixed (non-stratified); USA (white; AA; Asian; Hispanic) (validity $n = 26$)	Convergent validity: with Mini-Mitter Actiwatch monitors $r = 0.86$ (range = 0.82–0.89) Criterion validity: with energy expenditure from room calorimetry $r = 0.79$ (range = 0.78–0.80) With HR $r = 0.67$ (range = 0.66–0.67) With microwave activity $r = 0.80$ (range = 0.76–0.83)	Can also be extracted for PA domain. The Mini Mitter monitor showed excellent criterion validity, particularly when compared with the microwave activity
					Convergent validity: with CSA accelerometer monitors $r = 0.86$ (range = 0.82–0.89)	

Sedentary time/behaviour measures						
No.	Tool information: name	First author and type of paper (reference)	Administration	Sample: age; weight status; country (ethnicity), (n)	Evaluation	Comments
4	Multimedia Activity Recall for Children and Adolescents (MARCA)	Ridley 2006 ¹³³ (PDP)	Self-complete; data download	Children and adolescents; mixed (non-stratified); USA (race not defined) (TRT $n = 32$, validity $n = 66$)	<p>TRT: $r = 0.92$ (range = 0.88 to 0.94)</p> <p>Bland-Altman = PAL, upper LOA = +0.30 and lower LOA was -0.30 with a bias of +0.001. For MVPA, upper LOA = +51.2 and lower LOA = -53.4 with a bias of -1.1. Locomotion minutes had an upper LOA of +79.2 and lower LOA of -65.4 with a bias of +6.9</p> <p>Criterion validity: with accelerometer $r = 0.39$ (range = 0.35-0.45)</p>	<p>Can also be extracted for PA domain. The MARCA had fair correlation with accelerometer. Results indicate females and those > 11 years of age show the greatest correlation with the accelerometer</p>
5	Electronic Momentary Assessment (EMA): self-report survey on mobile phones	Dunton 2011 ¹³⁴ (Eval)	Self-complete; data download	Children and adolescents; mixed (stratified); USA [AA; Asian; Hispanic/Latino; white; mixed race; other (not defined)] (validity $n = 121$)	<p>Criterion validity: with activity (accelerometer): Across both weight status groups, steps were significantly higher for EMA surveys reporting active play, sports or exercise than any other type of activity (adjusted Wald test: $F = 22.16$, $df = 8$, $p < 0.001$). Stratified results were similar. Also children were more likely to engage in at least 5 minutes of MVPA within the 30-minute interval before EMA surveys reporting PA compared with sedentary behaviour as the main activity (adjusted Wald test: $F = 69.18$, $df = 1$, $p < 0.001$)</p>	<p>Can also be extracted for PA domain. Findings support the feasibility, acceptability and construct validity of the EMA</p>

CPAF, Children's Physical Activity Form; Eval, evaluated an existing tool without modification; MVPA, moderate to vigorous physical activity; PAL, physical activity level; PDP, primary development paper.

Appendix 11 Fitness measurement studies: summary table

No.	Tool information: name	First author and type of paper	Sample Age; weight status; country (ethnicity) (n)	Evaluation	Comments
1	6-minute walk test (6MWD) Aerobic capacity	Morinder 2009 ¹³⁸ (Eval)	Children and adolescents; mixed (stratified); Sweden (race not defined) (TRT n = 49; validity n = 250)	TRT: r = 0.84. Bland-Altman: Difference 2.8 m (bias); LOA for bias = -65.3-70.8 m Criterion validity: with cycle ergometry (VO _{2max} : l/minute and ml/kg/minute) r = 0.34	Also did known groups validity (compare obese to non-obese) Found significant difference in distance walked by 6MWD (obese = 57 m; non-obese = 66 m, p < 0.001) Correlated distance by characteristics including BMI (r = -2.27, p < 0.001) and BMI-SDS (r = -0.42, p < 0.001) Responsiveness testing was not discussed in methods or results, but authors report (based on their data) in the discussion, that in order for evaluation in obese children, 6MWD distance would need to change by 68 m to be statistically confident
2	Height-adjustable step test Aerobic capacity	Francis 1991 ¹⁴⁸ (Eval)	Children and adolescents; mixed (non-stratified); USA (race not defined) (n = 93)	Criterion validity: VO _{2max} with open-circuit spirometry with Bruce treadmill test: range r = 0.79-0.81; regression R ² range = 0.61-0.64 ANOVA: No difference between measured VO _{2max} and predicted from step test equation for any of the three frequencies	Overall, authors advocate the 6MWD in this population Paper begins by validating heights of steps based on hip angles (height adjustment avoids early muscle fatigue seen with fixed-height steps) Children then stepped at three difference paces with a metronome set at 120 clicks/minute (30 ascents), 104 clicks/minute (26 ascents) or 88 clicks/minute (22 ascents) As correlations with recovery HR were similar between frequencies, authors advocate lower ascents in younger children (26/22)

No.	Tool information: name	First author and type of paper	Sample Age; weight status; country (ethnicity) (n)	Evaluation	Comments
3	20-m shuttle run Aerobic capacity	Leger 1988 ¹³⁹ (Eval)	Children and adolescents; mixed (non-stratified); Canada (race not defined) (n = 139)	<p>TRT: $r = 0.89$</p> <p>Criterion validity: VO_{2max} with $VO_{2max_Douglas\ bag}$: $r = 0.71$</p> <p>Standard error of 5.9 ml/kg/minute (12.1%) predicted vs. measured</p> <p>Multiple regression showed sex, height and weight were not significant predictors of max speed or efficiency (age was)</p>	<p>The main focus of the paper is the influence of age of speed and efficiency (which are lower in younger children)</p> <p>Authors highlight that a 20-m shuttle run is advantageous over other tests, as it is possible to use the same protocol across age groups (using age-specific equations)</p>
4	International Fitness Scale (IFS), 5 item General fitness	Ortega 2011 ¹³⁶ (Eval)	Adolescent; mixed (stratified); nine European countries (race not defined) (TRT n = 277; convergent validity n = 2405–2727; construct validity n = 855–2728)	<p>TRT: $\kappa = 0.59$ (range = 0.60–0.58)</p> <p>Agreement: perfect agreement (100% same in both) = 65%; acceptable agreement (± 1) = 97%</p> <p>Convergent validity: with 'measured fitness' with 20-m shuttle run (estimated VO_{2max}): those reporting good or very good fitness had better 'measured fitness' than those reporting poor or very poor (ANOVA)</p> <p>Positive linear relationships for all ($p < 0.05$)</p> <p>Construct validity: with obesity and cardiovascular variables</p> <p>Differences found for overall fitness, speed/agility, cardiorespiratory fitness and muscle strength for obesity (negative relationship except muscle strength, which was significantly positive). Waist-to-height ratio and FMI – all significant negative relationship (muscle strength non-significant)</p> <p>Those reporting good overall fitness had healthier levels for most cardiovascular outputs (except muscular strength)</p>	<p>Those reporting very good cardiorespiratory fitness, speed/agility and overall fitness had 80% [OR 0.2 (95% confidence intervals 0.14 to 0.30)], 84% [OR 0.16 (95% confidence intervals 0.11 to 0.24)] and 87% [OR 0.13, (95% confidence intervals 0.08 to 0.19)] lower risk of being overweight/obese than those reporting poor/very poor fitness</p>

No.	Tool information: name	First author and type of paper	Sample Age; weight status; country (ethnicity) (n)	Evaluation	Comments
5	Bioelectrical impedance-derived VO_{2max} aerobic capacity	Roberts 2009 ¹⁴⁷ (Eval)	Adolescent; obese and overweight; USA (AA; white) (n = 134)	<p>Criterion validity: VO_{2max} with cycle ergometry (VO_{2max}: l/minute and ml/kg/minute): VO_{2max} l/minute = 0.48; VO_{2max} ml/kg/minute = 0.03</p> <p>Bland–Altman: 62% had BIA predicted VO_{2max} within 10% of cycle VO_{2max}</p> <p>Significant magnitude bias ($r = 1.0$, $p < 0.002$) but no systematic bias around mean ($r = 0.78$)</p> <p>LOA = -589–574</p> <p>TRT: $r = 0.82$</p> <p>Criterion validity: VO_{2peak} on graded maximal treadmill test $r = 0.57$ (range = 0.55–0.58)</p> <p>The t-test range = $t -0.5$ to -1.5 ($p > 0.05$)</p> <p>Bland–Altman: 20-metre shuttle test values = 0.27 (girls) and 1.07 (boys) lower than measured (i.e. underestimated). But differences are within standard deviation of differences: 85% of measures within 5.9 ml/kg/minute of estimated VO_{2peak}</p>	<p>BIA derived estimates of VO_{2max} differed by sex. Significant (weak) positive relationships between VO_{2max} and resistance, reactance and impedance index in girls ($r = 0.06$–0.15) but not boys</p> <p>Authors conclude that BIA is not a suitable measure of VO_{2max} owing to large variability and magnitude of bias</p>
6	20-m shuttle test Aerobic capacity	Suminski 2004 ¹⁴⁰ (Eval)	Children; mixed (stratified); USA (Hispanic/Latino) (TRT n = 35, validity n = 126)	<p>LOA = -589–574</p> <p>TRT: $r = 0.82$</p> <p>Criterion validity: VO_{2peak} on graded maximal treadmill test $r = 0.57$ (range = 0.55–0.58)</p> <p>The t-test range = $t -0.5$ to -1.5 ($p > 0.05$)</p> <p>Bland–Altman: 20-metre shuttle test values = 0.27 (girls) and 1.07 (boys) lower than measured (i.e. underestimated). But differences are within standard deviation of differences: 85% of measures within 5.9 ml/kg/minute of estimated VO_{2peak}</p>	<p>Validity repeated by weight status</p> <p>Estimated and measured VO_{2peak} did not differ between overweight ($t = -1.20$, $p = 0.51$) and normal weight ($t = -1.42$, $p = 0.17$)</p> <p>Correlations were higher in overweight (0.54) than in normal weight (0.54) with lower standard error of estimate and error</p> <p>Percentage of values within 5.9 m/kg/minute were greater (90.9%) in obese than in normal weight (80%)</p> <p>Note: VO_{2max} and VO_{2peak} are the same measures, but peak is used as cannot assume that this population will reach VO_{2max}</p>

No.	Tool information: name	First author and type of paper	Sample Age; weight status; country (ethnicity) (n)	Evaluation	Comments
7	Fitnessgram Overall fitness	Morrow 2010 ¹³⁷ (Eval)	Children and adolescents: mixed (non-stratified); USA (mixed) (TRT n = 12–467)	<p>TRT: κ: teacher = 0.76 (range = 0.60–0.94); expert = 0.81 (range = 0.61–0.92)</p> <p>Teacher % agreement = 85% (range 74–97%). Expert % agreement = 88% (77–96%)</p> <p>IR: teacher vs. expert = 81% agreement (range 64–96%). Trained teacher vs. expert = 84% agreement (range 64–100%)</p> <p>κ: teacher vs. expert = 0.67 (range = 0.41–0.92); trained teacher vs. expert = 0.73 (range = 0.45–0.90)</p>	<p>Fitnessgram is an educational assessment tool/software. It was not designed (but is used) for intervention assessment</p> <p>Responses on the development section here are based on the manual</p> <p>Confounding variables influence on agreement are generally non-significant</p> <p>Some test reliabilities increased with training (e.g. 20-m pacer, trunk lift and shoulder stretch)</p> <p>Thus, authors advocate training</p>
8	Submaximal Treadmill Test Aerobic capacity	Nemeth 2009 ¹⁴⁹ (Eval)	Adolescents; obese and overweight; USA (race not defined) (n = 27)	<p>Criterion validity: $\dot{V}O_2$ by open circuit spirometry with progressive treadmill r = 0.73</p> <p>Median standard error = 271 ml/minute</p> <p>Mean standard error = 3.36 ml/weight/minute</p> <p>Cross-validity coefficient r = 0.85</p> <p>Predicted deviated < 20% of observed in 96% of tests</p> <p>Median length of 95% confidence interval for predicted was 1073 ml/minute (range 1049 to 1150 ml/minute)</p>	<p>Papers describes two studies. First is a model building study to build the prediction equation (n = 86)</p> <p>Data here are for second validation study</p> <p>Note: Statistics need checking</p>

No.	Tool information: name	First author and type of paper	Sample Age; weight status; country (ethnicity) (n)	Evaluation	Comments
9	BMR with fat-free mass Aerobic capacity	Drinkard 2007 ¹⁴³ (Eval)	Adolescents: mixed (stratified); USA (white; AA) (n = 141)	<p>Criterion validity: measured VO_{2peak} on cycle ergometer $r = 0.48$ (range = 0.35–0.60)</p> <p>Bland–Altman: LOA 478–670 ml/minute (equating up to 30% of the average VO_{2max} in normal weight and 34% in obese)</p> <p>Significant magnitude of bias in obese ($p < 0.0001$) with OUES overestimating VO_{2max}. Similar in normal weight ($p < 0.05$)</p>	<p>Although correlations are high, the LOA were outside acceptable clinical range (defined as > 10%) and there was significant magnitude of bias</p> <p>All results depended on level of intensity and weight status. Thus, authors do not advocate OUES to assess fitness in obese adolescents</p>
10	Estimated maximal oxygen consumption and maximal aerobic power Aerobic capacity Pathway 2	Aucouturier 2009 ¹⁴⁴ (Eval)	Adolescent: all obese; France (race not defined) (n = 20)	<p>Criterion validity: with ml/minute and measured maximum aerobic power (MAPm) (cycle ergometry)</p> <p>Mean difference = %VO_{2max} ACSM vs. %VO_{2max} m = -5.9%; %VO_{2max} W vs. %VO_{2max} m = -13.9%; %MAPth vs. %MAPm = -1.4%</p> <p>Expressed as absolute values, VO_{2max} ACSM overestimated VO_{2max} m (12.1%) and VO_{2max} W overestimated VO_{2max} m (29.3%) both significant</p> <p>Bland–Altman %VO_{2max} ACSM underestimated (5.9%) and %MAPth underestimated (1.4%)</p> <p>Both outside LOA</p>	<p>Data analysis includes only those achieving sufficient respiratory exchange ratio (> 1.02) in measured VO_{2max} test</p> <p>Good correlation but poor agreement with gold standard with submaximal estimation overestimating VO_{2max} (with values underestimated when expressed as %VO_{2max}). Authors suggest estimated values are therefore not valid</p>

No.	Tool information: name	First author and type of paper	Sample Age; weight status; country (ethnicity) (n)	Evaluation	Comments
11	Physical working capacity on cycle ergometer Aerobic capacity	Rowland 1993 ¹⁴⁵ (Eval)	Child; mixed (non-stratified); USA (race not defined) (n = 35)	Criterion validity: with measured VO_{2max} $R = 0.71$ (by body weight: 0.57) range: 0.70–0.71 (by body weight 0.48–0.65)	Mean error from measured VO_{2max} was 3.4 ml/kg/minute for girls and 2.8 ml/kg/minute for boys These findings show that mean predictability of VO_{2max} from physical working capacity is good but the variability is wide with 10–15% error at one standard deviation Author concludes that physical working capacity provides only a crude estimate of VO_{2max} and should not be used to predict individual maximum aerobic power
12	Aerobic cycling power Aerobic capacity	Carrel 2007 ¹⁴⁶ (PDP)	Adolescent (> 11 years); all obese; USA [white (87%); other not defined] (validity n = 35)	Criterion validity: with VO_{2max} progressive treadmill walking $r = 0.39$, $p = 0.03$ Construct validity: with fasting insulin $r = 0.37$, $p < 0.05$	Could also fit within physiological measurement Grouped to fitness domain because of search/review strategy – linked to purpose of measurement in the introduction and title Lost robustness scores for evaluation because of sample size and results Correlations were < 0.4 and no measure of agreement tested Authors still advocate its use, however

No.	Tool information: name	First author and type of paper	Sample Age; weight status; country (ethnicity) (n)	Evaluation	Comments
13	VO _{2peak} Aerobic capacity	Loftin 2004 ¹⁴¹ (Eval)	Children and adolescents; obese and overweight; USA (race not defined) [TRT n = 6 (treadmill); n = 7 (cycle) validity n = 21]	TRT: treadmill r = 0.86 (range 0.76–0.96); cycle r = 0.91 (range = 0.84–0.98) Intraindividual variability: cycle = 5.7% (VO _{2peak}); 6.6% (VO _{2peak} W) Treadmill = 0.5% (VO _{2peak}); 2.5% (VO _{2peak} W) Convergent validity: cycle vs. treadmill VO ₂ : VO _{2peak} r = 0.77; VO _{2peak} W r = 0.72; VCO ₂ r = 0.73; respiratory exchange ratio r = 0.48; HR r = 0.52 The t-test: all indices non-significant except HR	Small sample size for repeatability Results suggest that both cycle and treadmill are similar with regards to evaluation results, but acceptability of cycle in obese sample was greater owing to less perceived exertion
14	Harvard Step Test Aerobic capacity	Meyers 1969 ¹⁴² (Eval)	Adolescents; mixed (non-stratified); USA (race not defined) (n = 119)	TRT: 1-week interval r = 0.65	The sample was boys only and the study was very basic (one page long and one reference)

6MWD, 6-minute walk distance; ACSM, American College of Sports Medicine; ANOVA, analysis of variance; CPAF, Children's Physical Activity Form; Eval, evaluated an existing tool without modification; MAPm, measured maximum aerobic power; MAPth, theoretical maximal aerobic power; ModEval, modified an existing tool and re-evaluated; OR, odds ratio; OUES, oxygen uptake efficiency slope; OR, odds ratio; PAL, physical activity level; PDP, primary development paper; VO_{2max}m, VO_{2max} measured; VO_{2max}W, percentage maximum volume oxygen uptake – estimated from Wasserman equation (Wasserman K, Hansen JE, Sue DY, Whipp B.T. *Principles of Exercise testing and Interpretation*. Philadelphia, PA: Lea and Febiger; 1987. pp. 50–80.) All administered by trained researcher/clinician except Ortega (2011),¹³⁶ which was self-reported on pen and paper.

Appendix 12 Physiology measures studies: summary table

Physiology measures					
No.	Tool information: name	First author and type of paper ^a	Sample: age, weight status, country (ethnicity)	Evaluation	Comments
	Indices of insulin sensitivity	Type			
1	Indices of insulin sensitivity	Yeckel 2004 ¹⁵²	Children and adolescents; all obese; USA (white; AA; Hispanic) (validity $n = 38$)	<p>Criterion validity: with EHC</p> <p>M-value: HOMA-IR vs. M-value $r = -0.57$; WBISI vs. M-value $r = 0.78$; ISI vs. M-value $r = 0.74$</p> <p>Convergent validity: with intramyocellular lipid accumulation: WBISI vs. lipid $r = -0.74$; ISI vs. lipid $r = -0.71$</p>	<p>Authors confident that OGTT can be used as successful markers of insulin</p> <p>Not clear if sample size = 38 or 368 for convergent validity</p>
2	Fasting indices of insulin sensitivity	Conwell 2004 ¹⁵³	Children and adolescents; all obese; Australia (white) ($n = 18$)	<p>Criterion validity: with glucose tolerance test (FSIVGTT): S_i and AIR: study 1 $r = 0.9$; AIR $r = 0.65$ (range: study 1 $r = 0.89-0.91$; AIR $r = 0.60-0.69$)</p>	<p>Test repeated three times, but repeatability not examined</p>
3	Indices of insulin sensitivity	George 2011 ¹⁵⁴	Children and adolescents (≥ 20); overweight and obese; USA (white; AA; mixed) ($n = 188$)	<p>Criterion validity: with EHC test: S_i, $r =$ study 1 = 0.77 (range: study 1 $r = 0.62-0.82$)</p> <p>Range for AUC = 0.89–0.95 (lowest = GluAUC/InsAUC; rest all > 0.94)</p>	<p>Results also stratified by disease state: (1) not glucose intolerant (NGT); (2) glucose intolerant (IGT); (3) type 2 diabetes (T2DM) clinical diabetes, but normal positive antibodies (OB-TDM)</p> <p>Correlations within diseases were all significant except GluAUC/InsAUC vs. study 1</p> <p>Overall 1/F, HOMA-IR and QUICKI were most highly correlated</p>
4	Indices of insulin sensitivity	Gunczler 2006 ¹⁵⁵	Children and adolescents; mixed (stratified); Venezuela (race not defined) ($n = 171$)	<p>Convergent validity: with ISI composition from OGTT $r = 0.60$ (range = 0.45–0.74)</p>	<p>Result available for normal children were higher correlations (mean of four indices = 0.68, range 0.55 to 0.82)</p> <p>Results were also stratified by moderately obese and severely obese</p> <p>Higher correlations were apparent for QUICKI and FGIR in moderately obese participants and for HOMA and FIRI in severely obese participants</p>

Physiology measures				
No.	Tool information: name	First author and type of paper ^a	Type	
		Sample: age; weight status; country (ethnicity)	Evaluation	
			Comments	
5	Indices of insulin sensitivity	Uwaifo 2002 ¹⁵⁶ Child; mixed (stratified); USA (white, black) (n = 31)	Insulin	Author concluded that QUICKI and FGIR had the strongest correlations with ISI composition in normal, moderately obese and severely obese children and adolescents Both fasting insulin and insulinogenic index correlated well with first and second steady phase insulin secretion (r's ranged from 0.79 to 0.86) HOMA-B% was not as highly correlated (0.69 to 0.72) Fasting C-peptide–insulin ratio was not significantly correlated with clamp-derived metabolic clearance rate of insulin ISI-FFA (from Insulin Sensitivity Indices, Free Fatty Acids) was not correlated with degree of free fatty acid suppression obtained from clamps Author's conclusion: QUICKI, fasting insulin and insulinogenic index correlate with corresponding clamp derived indices of insulin sensitivity
6	Insulin sensitivity and pancreatic beta cell function	Gungor 2004 ¹⁵⁸ Children and adolescents; mixed (stratified); USA (white; AA) (n = 156)	Insulin and glucose	Criterion validity: with euglycaemic clamp (IS) hyperglycaemic clamp (beta cell): IS r = 0.83; B cell r = 0.69 (range: S r = 0.82–0.84; beta cell r = 0.61–0.74) Both regressions: slopes sign different to 1 and intercept significantly differ to 0 Multiple regression: IS = BMI contributed significantly and independently to model; beta cell = BMI contributed significantly but not independently

Physiology measures						
No.	Tool information: name	First author and type of paper ^a	Type	Sample: age, weight status, country (ethnicity)	Evaluation	Comments
7	Fasting indices of insulin sensitivity	Atabek 2007 ¹⁵⁹	Insulin	Child; all obese; Turkey (race not defined) (n = 148)	<p>Criterion validity: with OGTT: IR (OGTT) vs. FGIR: r = -0.33; IR (OGTT) vs. HOMA-IR r = 0.34; IR (OGTT) vs. QUICKI r = -0.38; IGT (OGTT) vs. HOMA-IR r = 0.25</p> <p>Sensitivity and specificity of tests to detect whether children were insulin resistant = FGIR sensitivity = 61.8%, specificity = 76.3; HOMA-IR sensitivity = 80%, specificity = 59.1; QUICKI sensitivity = 80%, specificity = 60.2</p>	Includes children with IR and stratifies FGIR, HOMA-IR and QUICKI were all significantly different between groups
8	Homeostasis model assessment of insulin resistance	Keskin 2005 ¹⁶⁰	Insulin	Children and adolescents; all obese; Turkey (race not defined) (n = 57)	<p>Criterion validity: with OGTT: no data for indices apart of means and standard deviation. Only validity shown is with HOMA-IR: significantly lower in children without IR (confirmed by OGTT) compared with those with IR (p < 0.5)</p> <p>Based on cut-off of 3.17 HOMA-IR sensitivity = 76% and specificity = 66%</p>	<p>Specifically discussed utility of measures for use in clinical trials. Also established cut-off points using these data (QUICKI ≤ 0.328; HOMA-IR ≥ 2.7; FGIR ≤ 5.6)</p> <p>Emphasises need for testing in other ethnic groups</p> <p>Advocates indices, especially HOMA-IR and QUICKI</p> <p>Paper presented means values for indices with comparisons between children with and without insulin resistance (defined by OGTT)</p> <p>FGIR did not differ between those with and without IR, and QUICKI was higher in those without IR. Thus, sensitivity and specificity only presented for HOMA</p> <p>Used a data-driven approach to derive the cut-off of 3.16 in adolescents (adults = 2.5)</p> <p>Authors state that HOMA-IR is more reliable than FGIR and QUICKI is based on this</p>

Physiology measures						
No.	Tool information: name	First author and type of paper ^a	Sample: age; weight status; country (ethnicity)	Type	Evaluation	Comments
9	Homeostasis model assessment of insulin resistance	Rossner 2008 ¹⁶¹	Children and adolescents; overweight and obese; Sweden (race not defined) (<i>n</i> = 109)	Insulin	<p>Criterion validity: with FSIVGTT-MMOD: HOMA-IR vs. FSIVGTT $r = -0.53$. Repeated in prepubertal ($r = 0.16$, $p = 0.84$), pubertal ($r = -0.57$, $p < 0.01$) and post pubertal ($r = -0.53$, $p < 0.001$). Further multiple regression found HOMA-IR explained 33.7% of variance in sensitivity index for girls with high insulin sensitivity. But only 3.2% of variance was seen in girls with low insulin sensitivity. No interactions found</p> <p>Convergent validity: with fasting insulin – HOMA-IR vs. FI $r = 0.81$ (girls $r = 0.78$; boys $r = 0.87$)</p>	Sex dependent relationships but overall, poor validity of HOMA-IR. Best validity was with pubertal age (especially boys). Authors discourage use of HOMA-IR, especially in obese children at risk of elevated glucose homeostasis
10	Indices of insulin sensitivity	Schwartz 2008 ¹⁶²	Adolescents; mixed (stratified); USA (white, AA) (<i>n</i> = 323)	Insulin	<p>Criterion validity: with EHC</p> <p>HOMA: 0.42, QUICKI: 0.43, FGIR: 0.33, FI: 0.42, FI + TG: 0.46</p>	<p>Results were stratified by age and correlations were higher in 13-year-olds (mean 0.53 range 0.49 to 0.60) than in 15-year-olds (mean 0.29 range 0.14 to 0.35)</p> <p>Correlations in the > 85th percentile group were higher than those < 85th percentile</p> <p>ROC curves showed only a modest capability to separate true from false-positive values</p> <p>In addition, FI was significantly correlated with HOMA ($r = 0.99$), QUICKI ($r = 0.79$), FGIR ($r = 0.62$) and FI + TG ($r = 0.88$)</p>

Physiology measures					
No.	Tool information: name	First author and type of paper ^a	Type	Sample: age; weight status; country (ethnicity)	Comments
11	Impaired fasting glucose (IFG)	Cambuli 2009 ¹⁶³	Glucose	Children and adolescents; obese and overweight; Italy (race not defined) ($n = 535$)	<p>Paper has multiple objectives, of which one is to assess validity to predict IGT (OGTT) from IFG (fasted)</p> <p>Larger sample for remaining objectives ($n = 535$)</p> <p>Analysis demonstrates poor predictive power of fasting sample to predict 2-hour OGTT</p> <p>Although Si and M measured by both clamps were correlated, absolute values were systematically biased with increased bias in children with increasing insulin sensitivity</p> <p>Euglycaemic clamp is seen as the gold standard, but the hyperglycaemic clamp is easier/preferred</p> <p>Given the bias, authors suggest that hyperglycaemic clamp not be used as a substitute for euglycaemic clamp</p> <p>Reliability repeated stratified by those with IFG and IGT</p> <p>IGF%: positive agreement between 1 and 2 test = 22.2% ($\kappa = 0.17$, $p = 0.17$)</p> <p>IGT%: positive agreement = 27.3% ($\kappa = 0.11$, $p = 0.39$)</p> <p>Noted that although fasting samples are easier, OGTT also enables identification of IGT, which is a risk factor for T2DM and CVD</p>
12	Hyperglycaemic clamp	Uwaifo 2002 ¹⁵⁷	Insulin	Child; mixed (stratified); USA (white; AA) ($n = 31$)	<p>Criterion validity: with euglycaemic clamp: M, Si, GC</p> <p>Range $r = 0.45-0.65$</p> <p>Bland-Altman: Si and M significantly overestimated by hyperglycaemic clamp compared with euglycaemic clamp ($p < 0.001$)</p> <p>Significant increase in difference between measures with increasing Si ($R = -0.91$)</p> <p>Also correlated C-peptide and insulin between measures; range = 0.05-0.53</p> <p>TRT: FG: $r = 0.73$; 2-hour glucose: $r = 0.37$; ICC (FG: $r = 0.72$; 2-hour glucose: $r = 0.34$)</p> <p>Mean difference FG = 0.8 and 2-hour glucose = 0.7</p> <p>Criterion validity: with OGTT: sensitivity and specificity to identify those with glucose tolerance, IGT and IFG</p> <p>First test: 50% with IFG had IGT and 30% with IGT had IFG</p>
13	Oral Glucose Tolerance Test (OGTT)	Libman 2008 ¹⁶⁴	Insulin/glucose	Children and adolescents; overweight and obese; USA (white; AA; Hispanic) (TRT/ validity $n = 60$)	<p>Reliability repeated stratified by those with IFG and IGT</p> <p>IGF%: positive agreement between 1 and 2 test = 22.2% ($\kappa = 0.17$, $p = 0.17$)</p> <p>IGT%: positive agreement = 27.3% ($\kappa = 0.11$, $p = 0.39$)</p> <p>Noted that although fasting samples are easier, OGTT also enables identification of IGT, which is a risk factor for T2DM and CVD</p>

Physiology measures						
No.	Tool information: name	First author and type of paper ^a	Type	Sample: age; weight status; country (ethnicity)	Evaluation	Comments
14	¹³ C-glucose breath test – insulin resistance	Jetha 2009 ¹⁶⁵	Insulin	Child; all obese; Canada (white) ($n = 39$)	<p>Second test: 33% with IFG had IGT and 8% with IGT had IFG</p> <p>Those diagnosed differently at each test (discordant group) were more insulin resistant (HOMA)</p> <p>Criterion validity: with OGTT: $r = 0.44$ (range = 0.22–0.53)</p> <p>LOA range: –3.1 to –3.4 to 3.1–3.5</p> <p>Bland–Altman plots $r = 0.0$ (i.e. C-Glucose breath test of insulin resistance was similar to other indices in lack of bias)</p>	<p>Overall, results show that abnormalities/discordance is higher with an OGTT than a FG, and OGTT had poorer reliability</p> <p>Whole sample were obese but had a good range of BMI – with no differences across the range</p> <p>Correlations with BMI and indices were significant for CG-IR ($r = -0.61$); fasting insulin ($r = 0.44$); 2-hour insulin ($r = 0.42$) HOMA-IR ($r = 0.43$) and sum of insulin (0.44)</p>
15	Ultrasound analysis of liver echogenicity	Soder 2009 ¹⁷⁷	Liver assay	Child; mixed (stratified); Brazil (race not defined) (inter-rater $n = 11$)	<p>IR: three radiologists using three different ultrasound units: $\kappa = \text{all} > 0.8$</p>	<p>This paper has two studies. The first is an evaluation of reliability between administrators and machines and is reported here</p> <p>The second is another sample ($n = 22$) of obese and normal-weight children</p> <p>This is not a validity test, but could be considered discriminant validity</p> <p>In this case, no difference was found for liver parenchyma or kidney cortex echogenicity, but hepatorenal index did differ (greater in obese children)</p> <p>Authors advocate its use to evaluate hepatic steatosis</p>

Physiology measures						
No.	Tool information: name	First author and type of paper ^a	Type	Sample: age, weight status; country (ethnicity)	Evaluation	Comments
16	HbA _{1c}	Nowicka 2011 ¹⁷⁴	Blood cytology	Children and adolescents; all obese; USA (white; AA; Hispanic) (n = 1156)	<p>Convergent validity: with FG: Using ROCAUC, tested ability of HbA_{1c} and fasting glucose to predict IGT and T2DM</p> <p>AIC – IGT AUC = 0.6. Fasting glucose – IGT AUC = 0.7 ($p < 0.05$). HbA_{1c} – T2DM AUC = 0.81. Fasting glucose – T2DM AUC = 0.89 ($p = 0.13$)</p> <p>Construct validity: with diabetic status (pre-diabetes, T2DM, NGT): $\kappa = 0.2$ (95% confidence interval 0.14 to 0.26)</p>	<p>HbA_{1c} differed by BMI (with increasing BMI z-score and BMI seen in increasing HbA_{1c} categories)</p> <p>Analysis would also fit within criterion validity as the construct validity involved ability of both tests to accurately predict diabetes compared with gold standard</p> <p>But, it presented a comparison of results between HbA_{1c} and FG</p> <p>Overall, HbA_{1c} shown to have poor sensitivity</p>
17	Ghrelin	Kelishadi 2008 ¹⁷⁵	Ghrelin	Child; all obese; Canada (race not defined) (validity n = 100; responsiveness n = 100 (baseline) 92 (6 months) 87 (12 months)	<p>Construct validity: with obesity: disease outcome (insulin; blood lipids): BMI $r = -0.2$; other body composition $r = -0.5$; FG $r = -0.2$; total cholesterol $r = -0.3$; insulin $r = -0.5$; HOMA-IR $r = -0.4$; QUICKI $r = 0.3$; BP $r = -0.3$; energy intake $r = 0.1$. OR of predicting metabolic syndrome = 0.79 (95% confidence interval 0.68 to 0.87)</p> <p>Note: correlations significant except for leptin, EI and energy expenditure</p> <p>Responsiveness: change from baseline to 6 month = 417.1 (standard deviation 95.4) $p < 0.05$; change from 6 to 12 months = -278 (89.1), $p < 0.05$. Bivariate regression for change in ghrelin vs. change in body composition, EI, energy expenditure, leptin and insulin = significant correlations for BMI, waist circumference, waist-to-height ratio and total fat mass</p>	<p>Not described or tested as a validation study but shows change after an intervention that was present during the time of negative energy balance, but levelled off during maintenance</p> <p>Thus, if considered as an outcome, would need to be tested immediately following an intervention</p> <p>Others non-significant</p>

Physiology measures						
No.	Tool information: name	First author and type of paper ^a	Type	Sample: age; weight status; country (ethnicity)	Evaluation	Comments
18	Photoplethysmography (PPG)	Russoniello 2010 ⁴²⁰	Pulse rate	Child; all obese; USA (race not defined) ($n = 10$)	Criterion validity: with electrocardiography $r = 0.99$ (range = 0.97–1.0)	Author concludes that the PPG is as effective as ECG in measuring 11 parameters of HR variability
19	Estimated resting metabolic rate	Molnar 1995 ¹⁶⁶	Energy expenditure	Children and adolescents; mixed (stratified); Switzerland (race not defined) ($n = 371$)	Criterion validity: with measured RMR (ventilated hood): all estimated RMR significantly over estimated measured RMR. Range in estimation = underestimate by 16% to overestimate by 35%	Authors created new data driven equations to estimate RMR (stated reason for poor results are than old equations are out of date with today's population) Re-tested with new equation and found no significant differences between estimated and measured for boys, girls and combined (difference = 1%). Thus, final conclusion was that estimated can be a good proxy for measured
20	Predicted REE	Rodriguez 2002 ¹⁶⁷	REE	Children and adolescents; mixed (stratified); Spain (white) ($n = 116$)	Criterion validity: with open-circuit calorimetry-measured REE: range for all equations: $r = 0.73$ – 0.89% Predicted [(predicted REE/measured REE) $\times 100$] FAO = 101.8%; Maffei's = 88.8%; Harris B = 96.7%; Schofield W = 103.2%; Schofield HW = 100.1% LOA: Best = Schofield HW (–293 to 300). Worst = Schofield W (–468 to 391). Schofield HW also best for obese (LOA = –361 to 291)	Data extracted because of relevance to obesity research Equation accuracy differs by characteristics. In obese, in this study, Schofield HW performed best

Physiology measures					
No.	Tool information: name	First author and type of paper ^a	Sample: age, weight status; country (ethnicity)	Evaluation	Comments
	Predicted REE	Type			
21	Predicted REE	Lizzer 2006 ¹⁶⁸	Children and adolescents; all obese; Italy (white) (sample 2 $n = 287$, sample 3 $n = 53$)	<p>Criterion validity: with open-circuit indirect computerised calorimetry with hood: sample 2 $r = 0.8$</p> <p>Bland-Altman: mean difference = 0.14 mJ/day</p> <p>LOA = 2.06–1.77 mJ/day</p> <p>Linear regression: slope significantly different from 1; intercept significantly different from 0</p> <p>Cohort/sample 3 mean difference = 0.08 (equation 1) and 0.11 mJ/day (equation 2)</p>	<p>Three studies presented (1) equation development ($n = 287$ obese); (2) cross-validation of new equation in 50% of sample 1 population; (3) further validation in new sample of 53 obese adolescents</p> <p>Developed 2 new equations (first based on anthropometry easily obtained; second based on fat-free mass (needing BIA/DXA, etc.)</p> <p>Difficult to tease apart results for equations 1 and 2, but discussion reports that they had the same mean difference</p> <p>Authors conclude that these equations are useful for health-care professionals and researchers estimating REE in severely obese subjects</p>
22	Predicted REE	Firouzbakhsh 1993 ¹⁶⁹	Children and adolescents; mixed (stratified); USA (race not defined) ($n = 107, 94$ obese)	<p>Criterion validity: with indirect room calorimetry: ANOVA: no difference between measured and all equations in girls. In boys, measured was significantly higher than estimated by Harris Benedict but non-significant with all other equations</p> <p>Stratified by weight status (defined by > 110% ideal body weight)</p>	<p>Authors used terms BMR BEE and REE interchangeably</p> <p>All results non-sign in obese (showing no difference between measured and predicted)</p> <p>Schofield = closest estimate in obese subjects</p>

Physiology measures			
No.	Tool information: name	First author and type of paper ^a	Sample: age; weight status; country (ethnicity)
	Type		Evaluation
			Comments
23	Predicted REE	Derumeaux-Burel 2004 ¹⁷⁰	Children and adolescents; all obese; France (race not defined) (n = 211)
	REE		<p>Criterion validity: with open-circuit indirect calorimetry: REE (new) vs. measured $r = 0.82$</p> <p>ANOVA: mean measured and estimated were significantly different (not seen with <i>t</i>-test)</p> <p>Regression: Slope = significantly different from 1; significantly different from 0</p> <p>Mean difference: -2.19%</p> <p>Responsiveness: cohort 3 = significant difference between measured and estimated after weight loss</p> <p>Mean difference: 7.45%</p>
24	Predicted REE	Hofsteenge 2010 ¹⁷¹	Adolescent; obese and overweight; the Netherlands (Dutch, non-Dutch) (n = 121)
	REE		<p>Criterion validity: with ventilated hood system—measured REE: range of participants accurately predicted (within 10%) = 12–74%. Most accurate equation = Molnar</p> <p>Bias (% difference between measured and predicted) range = -19.8 to 10.8 (Molnar best)</p>

Physiology measures				
No.	Tool information: name	First author and type of paper ^a	Type	
		Sample: age; weight status; country (ethnicity)	Evaluation	
			Comments	
25	DXA-lean body mass REE	Schmelzle 2004 ¹⁷² Children and adolescents; all obese; Germany (race not defined) (n = 82)	REE	<p>Theory to use LBM (measured by DXA) in prediction equation is based on fact that lean tissue is more metabolically active than whole-body weight</p> <p>Criterion validity: with room calorimetry – measured REE: $r = 0.83$</p> <p>Correlations repeated with specific age and gender equations (range $r = 0.80-0.81$). Compared with other equations without LBM (range $r = 0.76-0.81$)</p> <p>Compared estimated REE using this method to 14 other equations (including six with less precise measure of LBM) and found their method to have the best correlation ($r = 0.83$) (others range = $0.63-0.80$)</p> <p>Bootstrap methods used for extra validation of regression equations</p> <p>Mean per cent deviation for all groups with new LBM equation was 7.7 (between measured and estimated)</p>
26	BMR with fat-free mass	Dietz 1991 ¹⁷³ Adolescent; all obese; USA (race not defined) (study 1 n = 25; study 2 n = 13)	Metabolic rate (BMR)	<p>Two studies: (1) to derive prediction equation (girls only here) and (2) a validation study (girls)</p> <p>BMR and REE reported to be the same thing</p> <p>Criterion validity: with open-circuit calorimetry-measured BMR: ANOVA and GLM</p> <p>Study 1: Harris Benedict and Cunningham significantly different from measured others = non-significant</p> <p>Study 2: remaining equations plus new equation compared with measured. Mayo and FAO1 differed significantly from measured (> 10%). No difference with others</p>

Physiology measures						
No.	Tool information: name	First author and type of paper ^a	Type	Sample: age; weight status; country (ethnicity)	Evaluation	Comments
27	Indices of insulin sensitivity (written in Chinese)	Wang 2005 ¹⁵⁰	Insulin	Children and adolescents; mixed (stratified); China (race not defined) (<i>n</i> = 151)	Convergent validity: comparing tests of HOMA-IR; FBG/FINS; IAI, WBISI; glucose/insulin AUC Results stratified by weight status suggest WBISI is best (most sensitive) in obese children	Note: data not fully extracted via translation
28	Energy expenditure by HR method (EEHF-Flex) (written in German)	Thiel 2007 ¹⁵¹	Energy expenditure	Children; all obese; Germany (race not defined) (<i>n</i> = 12)	Criterion validity: with VO ₂ (treadmill) and by indirect calorimetry (EEIndKal) during field tests doing five different sports Mean differences between EEHF-Flex and Energy Expenditure (indirect) (EEIndKal) for a 6-minute running test, ball games, cycle ergometry (65W) and strength/stability circuit were +3.6 ± 15.4%, +9.4 ± 16.1%, +14.7 ± 20.1% and +28.1 ± 27.8%, respectively. Range <i>r</i> = 0.92 (running, <i>p</i> < 0.001) to <i>r</i> = 0.76 (strength/stability circuit, <i>p</i> = 0.01)	Note: data taken from abstract only Authors conclude accuracy depends on mode of exercise in obese children, with lower accuracy in sports requiring strength

AIr, acute insulin response; ANOVA, analysis of variance; BEE, basal energy expenditure; CVD, cardiovascular disease; ECG, electrocardiogram; FG, fasting glucose; FGIR, fasting glucose-to-insulin ratio; FI, fasting insulin; FINS, fasting plasma insulin; FIRI, fasting insulin resistance index; FSNVGT, frequently sampled intravenous glucose tolerance test; FAO, Food and Agriculture Organization of the United Nations; GC, glucose clearance; GLM, glucose-lowering medication; HOMA, homeostasis model assessment; HOMA-B%, homeostatic model assessment – pancreatic beta cell function; HOMA-IR, homeostatic model assessment – insulin resistance; IFG, impaired fasting glucose; IAI, insulin activity index; IGT, impaired glucose tolerance; IR, insulin resistance; ISI, insulin sensitivity index; LBM, lean body mass; M, M-value; MMOD, minimal model analysis; OGT, oral glucose tolerance; OGTT, oral glucose tolerance test; OR, odds ratio; PCOS, polycystic ovary syndrome; PPG, photoplethysmography; QUICKI, quantitative insulin sensitivity check index; REE, resting energy expenditure; RMR, resting metabolic rate; ROC, receiver operating characteristic; ROCAUC, ROC area under the curve; SI, insulin sensitivity; TG, triglyceride; WBISI, whole-body insulin sensitivity index.

a All evaluated an existing tool without modification.

Appendix 13 Health-related quality-of-life studies: summary table

HRQoL summary table					
Tool information		Sample:	Evaluation	Comments	
No.	Name	First author (type of paper)	Administration ^a	Age; weight status; country; ethnicity; (n)	
1	Child Health Questionnaire (CHQ), 50 item	Waters 2000 ¹⁹² (Eval)	Parent complete	Child and adolescent; mixed (non-stratified); Australia; race not defined; (IC n = 5414)	<p>IC: α range = 0.19–0.87</p> <p>Suggest the primary development is Landgraf (1996),¹⁸⁶ which is a manual and was cited as a validation paper in search 1</p> <p>Also states that construct validity was completed but this is done in another publication (Waters 2000)¹⁹³ – see below)</p> <p>Also did item discriminate validity (%) = classed as: high item-scale correlations (± 2 standard errors) and ranged from 90.09% to 100%</p> <p>In addition, results for per cent total item-scale correlation higher with own scale ranged from 93.9% to 100%</p> <p>Per cent floor effects ranged = 0.0–0.8 and ceiling effects range = 3.7–86.6%</p> <p>Tested in three languages</p> <p>Further analysis looked at per cent scaling success and showed the greatest to be UK (99.4%) and the lowest in Canadian-French (74.2%)</p>
2	Child Health Questionnaire (CHQ), 50 item	Landgraf 1998 ¹⁸⁶ (Eval)	Parent complete	Child and adolescent; mixed (non-stratified); UK, Germany, USA, Canada; white, other (not defined); (IC/convergent validity n = 818)	<p>IC: German $\alpha = 0.75$, UK $\alpha = 0.73$, Canadian English $\alpha = 0.72$, Canadian-French $\alpha = 0.76$, USA $\alpha = 0.79$ (range = 0.43–0.97)</p> <p>Convergent validity: with items and other CHQ scales by country show greatest correlation in Canadian-French (mean full tool correlation: $r = 0.42$ (range 0.09 = 0.83) and lowest in German [mean full tool correlation: $r = 0.26$ (range 0.01 to 0.54)]</p>

HRQoL summary table			
Tool information		Sample:	
No.	Name	First author (type of paper)	Administration ^a
3	Child Health Questionnaire (CHQ), 50 item	Waters 2000 ¹⁹³ (ModEva)	Parent complete
			Age; weight status; country; ethnicity; (n)
			Child and adolescent; mixed (non-stratified); Australia; race not defined; (n = 5,223) American (n = 380)
			IC: α range = 0.60–0.93 (Australian); 0.66–0.94 (American)
			TRT: 2 week $r = 0.54$ – 0.73 (ICC = 0.49–0.78); 6–8 week $r = 0.53$ – 0.78 (ICC = 0.05–0.82)
			Convergent validity: with 'reported health conditions'
			Relationship between mental health scale and 'anxiety problems' $r = -0.35$ and 'depression' $r = -0.31$
			Behaviour scale correlated to 'behavioural problems' $r = -0.50$
			FA: item discriminatory validity: 100% for 8/11 multi-item scales. Varimax rotation analysis also conducted to produce 11 factors
4	DISABKIDS	Ravens-Sieberer 2007 ¹⁹⁴ (study 1) ^b (PDP)	Self and parent complete
			Age; weight status; country; ethnicity; (n)
			Child and adolescent; mixed (non-stratified); Austria, UK, the Netherlands, Sweden, Greece, Germany, France; race not defined; (IC/convergent validity n = 1153)
			IC: $\alpha = 0.8$ (0.74–0.89)
			Convergent validity: with GHP and FS-II-R (all result for FS-II-R in parentheses); $r = 0.33$ (0.30) (range = 0.26–0.42) (0.20–0.35)
			The author does not recommend this tool for population-level analysis
			Also compared results to a predefined US sample: scores on the CHQ were higher in the Australian sample apart from scales: physical functioning and family activities. In addition, discriminant validity was assessed and overall success rates were high with perfect results for 8 out of the 11 multi-item scales
			This paper describes development and testing of two measures
			This tool showed to have relatively poor convergent validity

HRQoL summary table		Sample:		Evaluation	Comments
Tool information		Age, weight status, country, ethnicity; (n)			
No.	Name	First author (type of paper)	Administration ^a		
5	KIDSCREEN, 52 item (long), 27 item (short)	Ravens-Sieberer 2007 ¹⁹⁴ (study 2) ^b (PDP)	Self and parent complete	<p>Child and adolescent: mixed (non-stratified); Austria, UK, Switzerland, the Netherlands, Czech, Sweden, Greece, Poland, Hungary, Germany, France, Spain, Ireland; race not defined; (IC $n = 22,546$, convergent validity $n = 22,830$)</p> <p>IC: $\alpha = 0.84$ (0.77–0.89)</p> <p>Convergent validity: with Child Health and Illness Profile-Adolescent Edition (CHIP-AE), Youth Quality of Life Instrument-Short version (YQOL-S) (all results compared with YQOL shown in parentheses) $r = 0.47$ (0.45) (range = 0.24–0.60 (0.24–0.61))</p>	<p>This measure was shown to be effective for translation in nine different languages with a large sample size</p> <p>Adequate convergent validity was shown with the YQOL and CHIP and excellent IC</p>
6	European Quality of Life-5 Dimensions (youth version) (EQ-5D-Y), 5 item	Burstrom 2011 ²⁴¹ (Eval)	Self-complete	<p>Child; mixed (stratified); Sweden; race not defined ($n = 470$)</p> <p>Construct validity: mean VAS score significantly lower in obese than in non-obese</p> <p>Results for individual scales showed non-significant differences between obese and non-obese except worried/sad</p>	<p>Tool development same as Burststrom 2011²⁴¹ (this is primary development paper)</p> <p>Paper also reports construct validity for other groups (e.g. asthma or rhinitis, severe illness or handicap)</p>
7	European Quality of Life-5 Dimensions (youth version) (EQ-5D-Y), 5 item	Burstrom 2011 ²⁴² (PDP)	Self-complete	<p>Child and adolescent; mixed (non-stratified); Sweden; race not defined</p> <p>Conducted face validity, open response results are: changed adults language from single words to words intelligible and used by children, e.g. depression to sad</p> <p>The second change was related to whole expression using verb form into heading of dimensions</p>	<p>Poor tool development with limited use of psychometric testing</p>

HRQoL summary table				
Tool information				
No.	Name	First author (type of paper)	Administration ^a	Sample: Age; weight status; country; ethnicity; (n)
8	European Quality of Life-5 Dimensions (youth version) (EQ-5D-Y), 5 item	Wille 2010 (PDP) ²⁴³	Self-complete	Child and adolescent; mixed (non-stratified); Sweden, Germany, Italy, Spain, South Africa; race not defined
				<p>Evaluation</p> <p>Convergent validity: with EQ-5D adult version tested in youth</p> <p>Results show that youth tended to report more health problems on EQ-5D-Y the following items: mobility, pain/discomfort, feeling worried, sad or happy</p> <p>EQ-5D-Y was also found to be easier to fill in and yielded fewer missing values</p> <p>Author's conclusion: EQ-5D-Y is a useful tool to measure HRQoL in young people in an age-appropriate manner</p>
9	European Quality of Life-5 Dimensions (youth version) (EQ-5D-Y), 5 item	Ravens-Sieberer 2010 ²⁴⁴	Self-complete	Child and adolescent; mixed (non-stratified); Sweden, Germany, Italy, Spain, South Africa; race not defined
				<p>Evaluation</p> <p>Feasibility was also assessed: complete data for 91–100% of respondents</p> <p>Missing or inappropriate responses ranged from 0 to 2%</p> <p>Known group's validity was assessed and those reporting a medical condition and taking medication reported significantly more problems on EQ-5D-Y for mobility, looking after myself, pain/discomfort and feeling worried, sad or happy when compared with those with no chronic condition and not taking medications</p> <p>Author's conclusion: EQ-5D-Y is a feasible, reliable and valid instrument of HRQoL but needs further testing in population based and clinical studies</p>

HRQoL summary table			
Tool information		Sample:	
No.	Name	First author (type of paper)	Administration ^a
		Age; weight status; country; ethnicity; (n)	
10	Impact of Weight on Quality of Life (IWQoL), 27 item	Kolotkin 2006 ¹⁸¹ (PDP)	Self-complete
		Child and adolescent; mixed (stratified); USA; white, AA, Hispanic, other (not defined); (IC/FA n = 491, convergent validity/construct n = 642; responsiveness = 80)	
		IC: $\alpha = 0.92$ (0.88–0.95) FA: total variance = 71%, interfactor correlations = 0.32–0.65 Convergent validity: with PedsQL $r = 0.75$ (range = 0.70–0.79) Construct: with BMI z-score $r = 0.44$ (range = 0.25–0.51) Responsiveness: SRM = 13.43 ($p < 0.0001$), ES = 0.75	
11	Impact of Weight on Quality of Life (IWQoL), 27 item	Modi 2011 ¹⁸² (Eval)	Self-complete
		Child and adolescent; all obese; USA (white, AA) (IC n = 263, TRT n = 21)	
		IC: full $\alpha = 0.89$ (range 0.87–0.93) TRT: $r = 0.82$ (range = 0.75–0.88)	
12	KINDL-R Questionnaire, 24 item	Erhart 2009 ¹⁸⁷ (Eval)	Self and parent complete
		Child and adolescent; mixed (stratified); Germany; race not defined; (IC/FA/convergent validity n = 7166)	
		IC: self $\alpha = 0.82$ (0.53–0.72), parent $\alpha = 0.86$ (0.62–0.74) FA: load range 0.45–0.78 (self), 0.47–0.87 (parent) Goodness of fit self-report: RMSE = 0.064; CFI = 0.931; AGFI = 0.944 Goodness of fit parent report: RMSE = 0.069; CFI = 0.952; AGFI = 0.965	
		Primary development is Ravens-Sieberer (2003) ⁵ but is in German <i>Conclusion states:</i> the study showed that parent proxy reports and child self-reports on the child's HRQoL differ slightly in perceptions and evaluations Overall, parent reports achieved higher reliability and thus are favoured for small samples In addition, there was a significant difference by weight status for quality of life in both self-report 0.25	
		Results also showed that the IWQoL had greater sensitivity than PedsQL with effect sizes exceeding 1.00 for all scales except family relations, whereas PedsQL effect sizes were 0.47 to 0.95 <i>Conclusion:</i> the IWQoL showed good reliability and validity	
		The study also worked out mean clinically important difference scores for each scale: physical comfort (8.8), body esteem (7.7) social life (8.1), family relations (6.2) and total quality of life (4.8)	
		Interfactor correlations ranged from 0.36 to 0.82 for SR, and 0.36 to 0.78 for parent proxy	

HRQoL summary table			
Tool information		Sample:	
No.	Name	First author (type of paper)	Administration ^a
		Age; weight status; country; ethnicity; (n)	Evaluation
		Comments	
13	Paediatric Cancer Quality of Life Inventory-32, 32 item	Varni 1998 ¹⁸⁸ (Eval)	Self and parent complete
		Child and adolescent; mixed (non-stratified); USA; white, Asian, AA, Hispanic, Native American; (IC n = 281, inter-rater validity n = 271, convergent validity n = 274)	<p>Convergent validity: with strength and difficulties questionnaire (SDQ) r = 0.45 (self), 0.48 (parent); [range = 0.02–0.57 (self), 0.00–0.63 (parent)]</p> <p>IC: self $\alpha = 0.77$ (0.69–0.83), parent $\alpha = 0.79$ (0.64–0.85)</p> <p>Inter-rater: child vs. parent r = 0.45 (0.36–0.59)</p> <p>Convergent validity: with similar scales on CDI, STAI-C, SSSC, SPPC and SPPA, and CBCL range r = 0.03–0.61 (parent with CBCL r = 0.03–0.59)</p>
14	Paediatric Cancer Quality of Life Inventory, 84 item	Varni 1998 ¹⁹⁵ (PDP)	Self (child/ adolescent) and parent complete
		Child and adolescent; mixed (non-stratified); USA; white, Asian, AA, Hispanic, Native American (inter-rater n = 157 ¹¹⁸)	<p>IR: child vs. parent r = 0.30 (range = 0.20–0.33), adolescent vs. parent r = 0.35 (range 0.22–0.44)</p> <p>This tool was used as a basis for construction of the PedsQL</p>
15	Paediatric Quality of Life Inventory V4.0, 23 item	Varni 2001 ¹⁹¹ (ModEval)	Self and parent complete and interview administered over phone to child
		Child and adolescent; mixed (non-stratified); USA; white, AA, Hispanic, Native American, Pacific islander, other (not defined); (IC/inter-rater/FA construct n = 1677)	<p>IC: $\alpha = 0.75$ (0.68–0.83), parent $\alpha = 0.80$ (0.75–0.88)</p> <p>IR: child vs. parent r = 0.41 (0.36–0.50), load range = 0.25–0.84 (child), 0.33–0.90 (parent)</p> <p>FA: load range = 0.25–0.84 (child), 0.33–0.90 (parent)</p> <p>Construct: with illnesses – child r = 0.24 (range = 0.22–0.28), parent r = 0.38 (range = 0.29–0.50)</p> <p>Results show reasonable reliability and validity</p>

HRQoL summary table						
Tool information						
No.	Name	First author (type of paper)	Administration ^a	Sample: Age; weight status; country; ethnicity; (n)	Evaluation	Comments
16	Paediatric Quality of Life Inventory V4.0, 23 item	Varni 2003 ¹⁹⁰ (Eval)	Self and parent complete, and interview administered over phone (parent and child)	Child and adolescent; mixed (non-stratified); USA; white, AA, Hispanic, Asian, Native American, Pacific Islander, other (not defined); (IC/inter-rater/construct n = 5863/6856)	IC: $\alpha = 0.78$ (0.71–0.87), parent $\alpha = 0.82$ (0.74–0.88) IR: child vs. parent $r = 0.61$ (0.44–0.75) Construct: with health: ANOVA: higher quality of life in those reporting 0 days missed from school, days needing care and sick days compared with those reporting > 3 days ($p < 0.001$)	Construct validity shows discriminance between healthy child and chronically ill child Tool development is same as Varni 2001 ¹⁶⁹ The study also assessed feasibility and found missing item responses for self-report was 1.8% and 2.4% in parents
17	Paediatric Quality of Life Inventory V 4.0, 23 item	Hughes 2007 ¹⁹⁶ (Eval)	Self and parent complete	Child; mixed (stratified); UK (race not defined) (n = 126)	Inter-rater: Wilcoxon signed-rank test was done to determine difference in self-report vs. parent report of obese children Results show that self-report score was higher on all scales – mean 71.4 (range 70.2 to 72.6) when compared with parent report: mean 66.3 (range 60.2 to 71.9)	Further tests showing parent proxy and self-report scores in obese clinical group and control group show that in parent proxy all scales were significantly higher in control. However, only physical health was significantly higher in control group when self-reported It is concluded that quality of life scores are different in self-report and parent proxy reports

HRQoL summary table			
Tool information		Sample:	
No.	First author (type of paper)	Administration ^a	Age, weight status, country, ethnicity: (n)
18	Paediatric Quality of Life V1.0, 45 item Varni 1999 ¹⁸⁹ (PDP)	Self and parent complete	Child and adolescent; mixed (non-stratified); USA; white, Asian, AA, Hispanic, Native American; (IC/inter-rater/FA/convergent validity n = 281)
			<p>IC: self $\alpha = 0.75$ (0.67–0.83), parent $\alpha = 0.81$ (0.59–0.89)</p> <p>Inter-rater: child vs. parent $r = 0.41$ (0.13–0.57)</p> <p>FA: total variance = 52% (child) 54% (parent), load range = 0.34–0.84 (child), 0.00–0.88 (parent)</p> <p>Convergent validity: with similar scales on Child Depression Index (CDI), State-trait anxiety (STAIC), Social Support Scale for Adolescents (SSSC), Self Perception Profile for Children (SPPC). Range $r = 0.03$–0.63</p>
19	Sizing Me Up, 22 item Zeller 2009 ¹⁸³ (PDP)	Self-complete and interview administered in person – child	Child and adolescent; all obese; USA; white, AA, mixed ethnicity, other (not defined); (IC/inter-rater/FA/convergent validity/construct n = 141, TRT n = 80)
			<p>IC: $\alpha = 0.76$ (0.68–0.86)</p> <p>TRT: $r = 0.67$ (0.53–0.78)</p> <p>Inter-rater: child vs. parent $r = 0.33$ (0.22–0.44)</p> <p>FA: total variance = 57%, inter-factor correlations range 0.01 (PSA vs. teasing) – 0.79 (emotion vs. total)</p> <p>Convergent validity: with PedsQL $r = 0.45$ (range = 0.35–0.65)</p> <p>Construct: with BMI $r = 0.16$ (range = 0.14–0.20) (only includes significant values)</p>
			<p>Assessed clinical/discriminate validity with on/off treatment: t-test ranged from 0.14 ($p = 0.8$) (communication with nurse) to 5.38 ($p < 0.001$) (nausea) for child and 0.45 ($p = 0.6$) (perceived physical appearance) to 9.30 ($p < 0.001$) (nausea)</p> <p>Also assessed feasibility – missing items was 0.1% for both parent and child</p> <p>Conclusion: parents' proxy report showed better validity and reliability than child</p> <p>Obesity-specific quality-of-life measure</p> <p>Results confirm preliminary evidence of strong reliability and validity properties with the exception of construct, which was fairly poor</p>

HRQoL summary table		Sample:		Evaluation	Comments
Tool information		Age; weight status; country; ethnicity; (n)			
No.	Name	First author (type of paper)	Administration ^a		
20	Sizing Them Up, 22 item	Modi 2008 ¹⁸⁴ (PDP)	Parent complete	<p>IC: $\alpha = 0.74$ (0.59–0.91)</p> <p>TRT: $r = 0.68$ (0.57–0.78)</p> <p>FA: total variance = 66%, interfactor correlation range = 0.08–0.90</p> <p>Convergent validity: with PedsQL ($r = 0.6$, range = 0.31–0.73) and IWQoL ($r = 0.27$, range = 0.24–0.35)</p> <p>Construct: with BMI $r = 0.34$ (no range)</p> <p>Responsiveness: SRM = -5.4 (range = -3.2 to -10.1) (all significant)</p> <p>IC: $\alpha = 0.92$ (0.90–0.95)</p> <p>TRT: $r = 0.74$ (0.71–0.77)</p> <p>FA: total variance = 75%, goodness of fit for final three-factor model = χ^2 9381 (df 231) $p < 0.001$, CFI = 0.90, TLI = 0.89 and RMSEA = 0.10</p> <p>Convergent validity: with YQOL-R $r = 0.54$ (range = 0.48–0.58)</p> <p>Construct: with BMI ($r = 0.39$, range = 0.34–0.43) and depression ($r = 0.53$, range = 0.48–0.59)</p>	Obesity-specific quality-of-life measure finding reliable and valid measurement properties
21	Youth Quality-of-Life Instrument-Weight Module (YQOL-W), 21 item	Morales 2011 ¹⁸⁵ (Eval)	Self-complete	<p>Child and adolescent; mixed (stratified); USA: white, AA, Hispanic; (IC/FA/convergent validity/construct $n = 443$, TRT $n = 30$)</p>	<p>Cites a primary development paper (Skalicky 2010), but this is a conference abstract</p> <p>Conclusion: the YQOL-W shows good reliability and validity for assessing weight-specific quality of life in children and adolescents</p>

HRQoL summary table			
Tool information			
No.	Name	First author (type of paper)	Administration ^a
		Sample:	Age; weight status; country; ethnicity; (n)
		Administration ^a	Evaluation
		Comments	
22	Standardise obesity-related interviews, 29 item (written in German)	Warschburger 2001 ¹⁷⁸ (PDP)	Interview administration
		Children and adolescents; all obese; Germany (race not defined); (n = 15)	Convergent validity: with KINDL-R questionnaire and Aussagen-liste zum Selbstwertgefühl With KINDL-R questionnaire = 0.556 for social questions; r = 0.597 for emotional items With ALS r = 0.48 (social) and r = 0.421 (emotional)
		This was translated from German interview was acceptable by children (more than questionnaires)	
		This study was also the basis for development of the GQ-LQ-KJ (Weight-specific Quality of Life Measure Children and Young) (Warschburger 2005 ¹⁵⁶)	
23	Weight-specific quality-of life measure, children and young (GW-LQ-KJ), 22 item (written in German)	Warschburger 2004 ¹⁷⁹ (ModEval)	Self-complete
		Children and adolescents; mixed (stratified); Germany (race not defined); (n = 936)	Convergent validity: with STA1 (r = -0.51), BIAQ (r = 0.37) and CHQ (r = 0.27-0.56 for multiple scales)
		This was translated from German	
		Discriminate validity also conducted and suggests that GW-LQ-KJ differed by weight status	

HRQoL summary table			
Tool information		Sample:	
No.	Name	First author (type of paper)	Administration ^a
24	Weight-specific quality-of-life measure, children and young (GQ-LQ-KJ), 26 item (written in German)	Watschburger 2005 ¹⁸⁰ (ModEval)	Self-complete
		Children and adolescents; overweight and obese; Germany (race not defined); (n = 448)	
			FA: results not extracted from translated version Convergent validity: with the CHQ (range $r = 0.33-0.62$ for multiple scales); STAI-C ($r = -0.64$); and BIAQ ($r = -0.50$)
			This was translated from German States previous IC of 0.87 (Guttman). Also tested differences in quality of life by weight status as a form of discriminate validity and found increased quality of life in those overweight, decreased quality of life in those obese and further decrease in quality of life in the very obese (although not significant)
25	Impact of Weight on Quality of Life (IWQoL) (written in Dutch)	Wouters 2010 ¹⁵ (ModEval)	It was not possible to translate (and therefore, extract data from) this paper. It has been included here as an additional evaluation of the IWQoL (also evaluated by Kolotkin 2006 ¹⁸¹)

AGFI, adjusted goodness-of-fit index; ANOVA, analysis of variance; BIAQ, Body Image Avoidance Questionnaire; CBCL, Child Behaviour Checklist; CDI, Child Depression Index; CFI, comparative fit index; df, degrees of freedom; Eval, evaluated an existing tool without modification; FS-II-R, Functional Status Questionnaire; GHP, General Health Perceptions; ModEval, modified an existing tool and re-evaluated; PDP, primary development paper; PSA, positive social attributes; RMSE, root-mean-square error; RMSEA, root-mean-square error of approximation; SPPA, Self-Perception Profile for Adolescents; SPPC, Self-Perception Profile for Children; SSSC, Social Support Scale for Adolescents; STAI, State-Trait Anxiety Inventory; STAI-C, State-Trait Anxiety Inventory for Children; TLI, Tucker-Lewis Index; VAS, Visual Analogue Scale.

a All pen and paper unless otherwise stated.

b Note that 'study 1' and 'study 2' are used to indicate manuscripts that report two studies in one paper.

c Not linked to bibliography: Ravens-Sieberer U, Bettge S, Erhart M. Lebensqualität von Kindern und Jugendlichen – Ergebnisse aus der Pilotphase des Kinder- und Jugendgesundheits-surveys. *Bundesgesundheitsbi-Gesundheitsforsch-Gesundheitsschutz* 2003;**46**:340–4.

Appendix 14 Psychological well-being measures: summary table

Psychological well-being measures			
No.	Tool information: name ^a	First author and type of paper	Sample Age; weight status; country (ethnicity), (n)
			Evaluation
			Comments
1	Children's Body Image Scale (CBIS)	Truby 2002 ¹⁹⁸ (PDP)	Child; mixed (stratified); Australia (white, Chinese, Vietnamese); (criterion validity $n = 310$, construct validity $n = 153$)
			Construct validity: with measured BMI $r = 0.43$ (range = 0.08–0.60). ANOVA = significant sex effect, with girls underestimating more than boys
			Convergent validity: with Body Esteem Scale (BES) $r = 0.32$ and DEBQ $r = 0.23$
			Did not receive optimum score for robustness because correlation was poor in boys
2	Body figure perception (pictorial), 5 item	Collins 1991 ²⁰⁵ (PDP)	Child; mixed (stratified); USA (white, AA); (TRT/validity $n = 159$)
			TRT: $r = 0.54$ (range = 0.38–0.71)
			Construct validity: with measured weight ($r = 0.36$) and BMI ($r = 0.37$)
			The body figure perception instrument revealed adequate reliability but showed less than good criterion validity with actual weight and BMI, and thus shows that individual's perceptions of body figure is poor. This value, however, is not necessarily an indication of poor psychometric properties
3	Self-Control Rating Scale (SCRS), 33 item	Kendall 1979 ¹⁹⁷ (PDP)	Children and adolescents; mixed (non-stratified); USA (white, other (not defined)); (IC/FAV validity $n = 110$, TRT $n = 24$)
			IC: $\alpha = 0.98$
			TRT: $r = 0.84$
			FA: load range = 0.03 to 0.91 (only two factors, of which all items loaded on to one)
			Criterion validity: with observation $r = 0.28$
			Convergent validity: with Peabody Picture Vocabulary Test (PPVT) $r = 0.06$, Matching Familiar Figures (MFF) $r = 0.24$, Porteus mazes $r = 0.35$, delay of gratification $r = 0.05$
			Poor criterion and convergent validity. However, the SCRS did show good IC and reliability. Results were for the full scale only and not reported by scale category

Psychological well-being measures			
No.	Tool information: name ^a	First author and type of paper	Sample Age; weight status; country (ethnicity), (n)
			Comments
4	Self-Perception Profile for Children (SPPC), 36 item	Van Dongen-Melman 1993 ²⁰⁹ (ModEval)	Child; mixed (non-stratified); the Netherlands; (race not defined) (IC/FA n = 300) (TRT n = 129)
			<p>IC: $\alpha = 0.76$ (range = 0.65–0.81)</p> <p>TRT: $r = 0.76$ (range = 0.66–0.83)</p> <p>FA: total variance = 50.1%, load range = 0.37–0.88</p> <p>Eigenvalue range = 2.65–4.74. CFA values = similar loadings (0.35 to 0.81)</p> <p>Goodness-of-fit indices = $\chi^2 = 0.959$, adjusted goodness of fit = 0.954, RMR 0.057 (df 395), goodness of fit = 0.96. Interfactor correlations range = 0.29–0.64</p>
5	Perceived competence scale (aka SPPC/Harter), 28 item	Harter 1982 ¹⁹⁹ (PDP)	Child; mixed (non-stratified); USA; (race not defined); (IC n = 2272, TRT n = 208 (3 month)/810 (9 month), FA n = 341, convergent validity n = 2271)
			<p>IC: $\alpha =$ range = 0.73–0.86</p> <p>TRT: $r = 0.79$ (3 month), $r = 0.76$ (9 month) range = 0.70–0.80 (3 month), 0.69–0.80 (9 month)</p> <p>FA: load range = 0.35 to 0.79</p> <p>Convergent validity: with teacher ratings ($r = 0.4$) and sociometric index for social scale ($r = 0.59$)</p> <p>FA: (CFA) goodness-of-fit indices = χ^2 1769.57 (df 451) RMSEA = 0.04, RNI = 0.93 and NNFI = 0.92</p> <p>Interfactor correlations range = 0.19 to 0.45</p>
6	Physical Activity Enjoyment Scale (PACES), 12 item	Motl 2001 ²⁴⁹ (ModEval)	Adolescent; mixed (non-stratified); USA (white, AA, mixed ethnicity), other (not defined); (FA n = 1797)
			<p>Refers to Harter (1982)¹⁷² as primary development (perceived competence scale for children (later changed its name to Self Perception Profile for Children)</p> <p>Results show good internal reliability and validity and good external reliability</p> <p>Problems are identified with the internal validity with only two factors identified and all items have the highest factor loadings in factor 1</p> <p>This tool name was later changed to Self Perception Profile for Children. Results showed that the shorter time period for reliability equates to an improved correlation</p> <p>In addition, this tool showed fair convergent validity</p> <p>Primary development was with university students aged 18–24 years</p> <p>More psychometric testing is required to interpret the appropriateness of this tool</p>

Psychological well-being measures			
No.	Tool information: name ^a	First author and type of paper	Sample Age; weight status; country (ethnicity), (n)
			Evaluation
			Comments
7	Self-Report Depression Symptom Scale (CES-D), 20 item	Radloff 1991 ²⁴⁶ (Eval)	Children and adolescents; mixed (non-stratified); USA; (race not defined); (IC n = 819)
			IC: $\alpha = 0.68$ (range = 0.58–0.85)
			Originally developed in adults This study conducted IC tests across children and adults Only the results for children are included here The tool shows poor IC in children (perhaps because it was developed for adults?) Primary development was by Fox and Corbin (1989) ^b with adults This study presents results after modification for use in children This tool has been rigorously tested and shows good reliability and construct validity
8	Children's Physical Self-Perception Profile (C-PSP), 24 item	Whitehead 1995 ²¹⁰ (study 1) (ModEval)	Child; mixed (non-stratified); USA [white, other (not defined)]; (IC n = 456 + 46, TRT n = 46, FA n = 227, construct validity n = 459)
			IC: $\alpha = 0.89$ (range = 0.79–0.94) TRT: $r = 0.89$ (range = 0.79–0.94) FA: total variance = 60.1% (boys), 64.6% (girls), load range = 0.40 to 0.86 Construct validity: with physical fitness tests (pull-ups, sit-ups, standing long jump, mile run, 50-yard dash and 600-yard run) Internal validity: load range = 0.56–0.82 CFA of six-factor structure showed: $\chi^2 = 1702.35$, $df = 579$, NNFI = 0.90 and CFI = 0.91
9	Children's Physical Self-Perception Profile (C-PSP), 36 item	Eklund 1997 ²⁴⁵ (Eval)	Children and adolescents; mixed (non-stratified); USA; (race not defined); (n = 642)
			Primary development of PSPP was by Fox and Corbin (1989) ^b but this was done in adults and was modified and evaluated for use in children by Whitehead 1995 ¹⁸² The author concludes that the results reported here support the initial evidence of reliability and validity published by Whitehead ²¹⁰ They indicate that the C-PSPP has potential utility for use in appropriate professional and research settings

Psychological well-being measures				
No.	Tool information: name ^a	First author and type of paper	Sample Age; weight status; country (ethnicity), (n)	
			Evaluation	
			Comments	
10	Children's Perceived Importance Profile (C-PIP), 8 item	Whitehead 1995 (study 2) ²¹⁰ (ModEval)	Child; mixed (non-stratified); USA; [white, other (not defined)] (IC/TRT n=46) IC: $\alpha = 0.73$ (range = 0.69–0.75) TRT: $r = 0.82$ (range = 0.75–0.90)	This tool was also modified from Fox and Corbin (1989), ^b which is referenced as the primary development paper The tool shows good IC and good TRT reliability
11	Children's Self Perceptions of Adequacy in and Predislection for Physical Activity (CSAPPA), 20 item	Hay 1992 ²¹¹ (PDP)	Children and adolescents; mixed (non-stratified); Canada; (race not defined); (IC/TRT/validity n = 591, FA n = 543) IC: correlated items with factor subtotals. All items correlated strongly with the appropriate factor. Item partial-total correlations range $r = 0.65$ – 0.85 for appropriate factors/ $r = 0.27$ – 0.59 for inappropriate factors TRT: $r = 0.83$ (range = 0.81–0.85) FA: load range = 0.31 to 0.77	This tool showed good TRT reliability and good construct validity The psychometric results shown to improve with age with best results in children in grade 9 (15 years) compared with grades 4–6 (9–12 years)
12	Body Shape Questionnaire (BSQ), 34 item	Conti 2009 ²⁰⁶ (Eval)	Children and adolescents; mixed (stratified); Brazil; (race not defined); (IC/validity n = 386, TRT n = 366) IC: $\alpha = 0.96$ TRT: $r = 0.91$ Construct validity: with participation questionnaire (PQ) $r = 0.60$, teacher's evaluation (TE) $r = 0.61$ Bruininks–Oseretsky Motor Proficiency test (MPT) $r = 0.76$	Primary development is in adults (Cooper 1987). ^c This questionnaire showed good IC and reliability but the scores were for the overall tool and did not report by scale category
13	Children's Physical Self-Concept Scale (CPSS), 27 item	Stein 1998 ²⁰⁷ (PDP)	Child; mixed (stratified); USA; [white, AA, other (not defined)]; (IC n = 30 + 316, TRT n = 30, validity n = 361 (study 1), 60 (study 2)) IC: $\alpha =$ sample 1 = 0.89 (range = 0.86–0.90), sample 2 = 0.69 (range = 0.60–0.81) TRT: $r = 0.82$ (range = 0.80–0.84) Construct validity: with obesity	Stein conducted two studies in one: a development study (study 1) and evaluation study (study 2) IC was better in sample 1 than sample 2. CPSS distinguished significant differences between overweight and normal-weight children

Psychological well-being measures			
No.	Tool information: name ^a	First author and type of paper	Sample Age; weight status; country (ethnicity), (n)
			Comments
			Evaluation
ANOVA:			
<i>Sample 1</i> : significant differences between normal-weight and overweight children (F = 33.91, $p < 0.001$)			
<i>Sample 2</i> : significant differences between normal weight, overweight and diabetic children (F 8.27, $p < 0.001$)			
14	Pediatric Barriers to a Healthy Diet Scale (PBHDS), 17 item	Janicke 2007 ²⁰⁰ (PDP)	Children and adolescents; obese and overweight; USA; [white, AA, Hispanic, Native American, other (not defined)]; (IC/FA) validity $n = 171$
The PBHDS showed good IC yet had poor convergent and construct validity			
IC : $\alpha = 0.74$ (range = 0.71–0.77)			
FA : total variance = 35.6%, load range = 0.40 to 0.75			
Convergent validity : with multidimensional scale of perceived social support (MSPSS) $r = 0.3$, Child Depression Index (CDI) $r = 0.32$, and Barriers to PA scale (BPA) $r = 0.37$			
Construct validity : with BMI z-score $r = 0.07$			
15	Body Image Avoidance Questionnaire (BIAQ), 13 item	Riva 1998 ²¹ (Eval)	Adolescent; mixed (stratified); Italy; (race not defined) [(IC/FA) $n = 439$ (high school), 142 (obese)]
Primary development of measure is in adults (Rosen 1991 ^d)			
Internal validity tests discarded six items and reduced the questionnaire from 19 items to 13 items			
Scale results are not provided just the total score			
16	Video distortion	Probst 1995 ²⁰⁸ (Eval)	Children and adolescents; mixed (stratified); Belgium; (race not defined) [TRT $n = 41$, validity $n = 83$ (41 obese)]
TRT : $r = 0.52$ (range = 0.80–0.84)			
Construct validity : with measured BMI: full agreement for obese 90.54% (range = 74.09%–98.51%) and for normal weight 90.94% (range = 89.49–93.03%)			
Results indicate adequate reliability and show little difference in obese and normal-weight children in perceived and actual body weight			

Psychological well-being measures			
No.	Tool information: name ^a	First author and type of paper	Sample Age; weight status; country (ethnicity), (n)
			Evaluation
			Comments
17	Social Anxiety Scale for Children—Revised version (SASCR), 26 item	La Greca 1993 ²⁰² (ModEval)	Child; mixed (non-stratified); USA (white, black, Hispanic) (n = 459)
			<p>IC: full $\alpha = 0.78$ (range 0.69–0.86)</p> <p>Internal validity: three factors</p> <p>Load range: 0.45–0.76</p> <p>Total variance: 89.8%</p> <p>CFA of three-factor model: GFI 0.93, RMSEA 0.067</p> <p>The three-factor model produced a significantly better fit than the two-factor model</p> <p>Convergent validity: with self-perception profile for children (SPPC) $r = 0.30$ (range = 0.12–0.47)</p> <p>IC: full $\alpha = 0.73$ (range 0.63–0.83)</p> <p>TRT: $r = 0.55$ (range 0.39–0.70)</p> <p>Internal validity: two factors</p> <p>Load range: 0.34–0.76</p> <p>Total variance: 87.9%</p> <p>Convergent validity: with Children's Manifest Anxiety Scale (CMAS) $r = 0.48$ (range = 0.36–0.57)</p>
18	Social Anxiety Scale for Children (SASC), 10 item	La Greca 1988 (PDP) ²⁰¹	Child; mixed (non-stratified); USA (race not defined) (IC/IV/convergent validity n = 287, TRT = 102)
			<p>Convergent validity was also assessed in the original version and was slightly lower (mean = 0.28 range = 0.09 to 0.41)</p> <p>Further results showed that girls and those in the lower grades reported more social anxiety</p> <p>In addition, author supports the revisions made to questionnaire and further supports the reliability and validity</p> <p>Author concludes that this study provides preliminary support for reliability and validity of SASC</p>

Psychological well-being measures			
No.	Tool information: name ^a	First author and type of paper	Sample Age; weight status; country (ethnicity), (n)
			Evaluation
			Comments
19	Nowicki-Strickland Locus of Control Scale (NS-LOCS), 40 item	Nowicki and Strickland 1973 ²⁰³ (PDP)	Children and adolescents; mixed (non-stratified); USA; (white, black) (IC/TRT n = 1017, convergent validity n = 353 – comparison with IARS and n = 29 – comparison with BCS)
			<p>IC: split R = 0.72 (range: 0.63–0.81)</p> <p>TRT: r = 0.67 (range 0.63–0.71)</p> <p>Convergent validity: with IARS and BCS r = 0.41 (range 0.31–0.51)</p>
20	Body Esteem Scale (BES), 24 item	Mendelson 1982 ²⁰⁴ (PDP)	Child; mixed (stratified); Canada; (Hebrew) (n = 36)
			<p>Convergent validity: with Piers-Harris Child Self-Concept Scale (mean r = 0.66, range = 0.62–0.68)</p> <p>Construct validity: with weight r = 0.55</p>
			<p>Author concludes that the NS-LOCS is an appropriate instrument to measure children's behaviour</p> <p>The correlations with the child concept scale were also compared in obese and normal children: body esteem and self-esteem were related similarly in obese (r = 0.67) and normal (r = 0.73)</p> <p>Within text, author states: 'There was a correlation between odd and even scores on the BES (r = 0.85) which indicates good reliability. I have not put this into the reliability evaluation score because it does not come under a specific type and is not enough to be classed as a reliability test'</p>

AGFI, adjusted goodness-of-fit index; ANOVA, analysis of variance; BCS, Bailler-Cromwell Scale; CFA, confirmatory factor analysis; CFI, comparative fit index; CMI, Cornell Medical Index; df, degrees of freedom; Eval., evaluated an existing tool without modification; IARS, Intellectual Achievement Responsibility Scale; ModEval., modified an existing tool and re-evaluated; NNFI, non-normed fit index; PDP, primary development paper; RMR, resting metabolic rate; RNI, relative non-centrality index; SASC, Social Anxiety Scale for Children; SASC-R, Social Anxiety Scale for Children-Revised version.

a All self-completed (except SCRS, which is teacher completed) and all pen and paper versions (except 'Video distortion', which is data downloaded).

b Not linked to bibliography: Fox KR, Corbin CB. The physical self-perception profile: development and preliminary validation. *J Sport Exerc Psychol* 1989;**11**:408–30.

c Not linked to bibliography: Cooper PJ, Taylor MJ, Cooper M, Fairburn CG. The development and validation of the Body Shape Questionnaire. *Int J Eat Disord* 1987;**6**:485–94.

d Not linked to bibliography: Rosen JC, Srebnik D, Saltzberg E, Wendt S. Development of a body image avoidance questionnaire. *Psychol Assess* 1991;**3**:32–7.

Appendix 15 Environment measures: summary

Environment summary table							
Tool information			Sample				
No.	Name	First author (type of paper)	Type	Administration	Age; weight status; country; ethnicity; (n)	Evaluation	Comments
1	Nutrition and Physical Activity Self-Assessment for Child Care (NAPSACC), 56 item	Benjamin 2007 ²⁴⁷ (PDP)	56 item Child-care environment measure	Child-care centre staff completed	Infant and children (< 5 year); mixed non-stratified; USA (white; AA; Native American) (TRT n = 39 child centres; inter-rater n = 59; validity n = 39 child centres)	TRT: $\kappa = 0.4$ (range = 0.07–1.0), agreement = 60.55% (range = 37.1–1.0%) Inter-rater: $\kappa = 0.57$ (range = 0.2–1.0); agree = 70% (range = 52.6–100.0%) Criterion validity: with researcher observations (Ward 2008 ²¹³) $\kappa = 0.37$ (range = 0.11–0.79)	Item-level results combined in data extraction. Tool developed specifically to evaluate NAPSACC, an intervention for obesity based on child-care centres. Also developed a researcher conducted protocol to use as gold standard (EPAO) (Ward 2008 ²¹³). Questions with poor validity and reliability can be eliminated if needed as no scales were generated. Authors advocate use but are less confident as an outcome measure without sensitivity testing
2	Environment and Policy Assessment and Observation (EPAO)	Ward 2008 ²¹³ (PDP)	192 item Child-care environment measure	Researcher administered	Infant and children (< 5 years); mixed non-stratified; USA (race not defined) (inter-rater n = 17)	Inter-rater: $r = 0.63$ (range = 0.05–1.0)	Direct observation method designed to be an outcome measure and gold standard tool. Items with poorer correlations have now been revised (via observer training manual and item definitions) Note: Although not changeable at the individual level, child-care settings are included in CoOR for potential to use within existing obesity treatment interventions within this setting

Environment summary table					
Tool information			Sample		
No.	Name	First author (type of paper)	Type	Administration	Age; weight status; country; ethnicity; (n)
3	Healthy Home Survey (HHS), 66 item	Bryant 2008 ²¹⁴ (PDP)	66 item Home environment measure	Interview administered – telephone	Child; mixed non-stratified; USA (white; AA) (TRT n = 45; validity n = 82)
					<p>TRT: $r = 0.72$ (range = 0.22–0.93); $\kappa = 0.66$ (range = 0–0.88), agree = 81.5% (range = 42.2–97.8%)</p> <p>Criterion validity: with researcher observations $r = 0.62$ (range = 0.3–0.88); $\kappa = 0.55$ (range = 0–0.96)</p>
4	Environment and Safety Barriers to Youth Physical Activity Questionnaire, 21 item	Durant 2009 ²²⁰ (PDP)	21 item Built environment perception measure	Self and parent completed – in environment	Adolescent; mixed non-stratified; USA (white; other not defined) (IC n = 474; TRT and validity n = 474)
					<p>Prevalence and bias adjusted κ (PABAK) also shown for TRT</p> <p>Although this describes a population-based environment (i.e. built environment), usually targeted by prevention interventions, the tool measures 'perception' to barriers, which could be targeted on at the treatment level. Except for construct validity, all data that have been extracted have been an average across age groups. Results are strong, but authors state caution without criterion validity</p> <p>IC: $\alpha = 0.75$ (range = 0.64–0.87)</p> <p>TRT: $r = 0.6$ (range = 0.48–0.81)</p> <p>FA: per cent variance range = 13.6–46.5, factor loading range 0.45–0.88</p> <p>Construct validity: with PA</p> <p>ANOVA: (1) parental perception of environment and safety barriers in park was not related to activity in park in those aged 5–11 years; (2) parental perception of lower barriers in street PA was related to increased PA in street in those aged 5–11 years; (3) parent perceived environmental barriers were related to activity but safety barriers were not related to PA in those aged 12–18 years; and (4) child perception environmental and safety barriers not related to activity in either area in those aged 12–18 years</p>

Environment summary table					
Tool information			Sample		
No.	Name	First author (type of paper)	Type	Administration	Age; weight status; country; ethnicity; (n)
5	Family Eating and Activity Habits Questionnaire (FEAHQ), 21 item	Golan 1998 ²¹⁵ (PDP)	21 item Home environment measure	Parent completed – outside environment	Child; mixed (stratified); Israel (race not defined) (IC n = 40; TRT n = 40; Responsiveness n = ?)
					<p>IC: $\alpha = 0.83$ (range = 0.78–0.88)</p> <p>TRT: $r = 0.85$ (range = 0.78–0.90)</p> <p>Inter-rater: $r = 0.88$ (range = 0.81–0.84); MANOVA for differences between parents before and after intervention were non-significant</p> <p>Responsiveness: ANOVA (differences in scale scores) were significant for 'exposure' and 'eating style' scales only</p> <p>Fisher's z transformation = similar for both groups (i.e. change related to weight was associated to change in scores in both intervention and control)</p>
					<p>Two studies conducted (1) clarity, TRT and IC (n = 40) and (2) intervention participants = inter-rater reliability and responsiveness</p> <p>Discriminate validity (described in the paper as concurrent) = t-test between obese and normal weight = obese score were significantly higher</p> <p>Sum of parent scores and child scores (family score) was also compared with scores being highest in obese child families [$F(1,37) = 11.5, p < 0.01$]</p> <p>Authors advocate use, but state it should be further evaluated in other samples</p>
6	Parenting Strategies for Eating and Activity Scales (PEAS), 26 item	Larios 2009 ²¹⁶ (PDP)	26 item Home environment measure	Parent completed	Child; mixed non-stratified; USA (Hispanic/Latino) (IC n = 91; convergent validity n = 91, construct validity n = 714)
					<p>IC: $\alpha = 0.82$ (range = 0.81–0.82) (only reported for original version)</p> <p>FA: factor loading range 0.31–0.88 (% variance range = 7.01–24.56); eigenvalue range = 1.4–4.91, goodness of fit (final model = χ^2, $\chi^2 = 1030.81$, $df = 282$, $CFI = 0.89$, incremental fit index = 0.9)</p> <p>Multiple regression: score change explained 27% variance in weight reduction</p>
					<p>Completion in English or Spanish was voluntary for research participants</p> <p>Authors compare PEAS to CFQ and call it construct validity, but it has been extracted under convergent</p>

Environment summary table						
Tool information		Sample		Evaluation		
No.	Name	First author (type of paper)	Type	Administration	Age, weight status; country; ethnicity; (n)	
7	Family Food Behaviour Survey (FFBS), 20 item	McCurdy 2010 ²⁷ (PDP)	20 item Home environment measure	Interview administered in person – parent; Interview administered over the telephone – parent	Child; mixed (stratified); USA (white; AA; Hispanic; other not stated) (n = 38; TRT and validity n = 28)	<p>Construct was, however, also conducted by comparing PEAS to dietary behaviour strategies questionnaire. Results were also correlated with child BMI but findings were poor</p> <p>Paper presents three phases of research with different participants and results which are not clear</p> <p>N extracted here is based on that provided in the methods and may not be the final N</p> <p>Between-scale correlations also measured, finding child choice was negative correlated with maternal control (r = -0.48) and positively correlated with organisation (r = 0.34)</p> <p>Maternal control was positively correlated with maternal presence (r = 0.34)</p> <p>This is a potentially good tool, but requires further evaluation (especially for criterion validity)</p>
	PATHWAY 2					<p>Convergent validity: with CFQ (Birch 2001⁷⁵) (r = 0.22, range = 0.02–0.65)</p> <p>Construct validity: BMI z-score r = 0.03 (range = 0.03–0.21); eating behaviour r = 0.2 (range = 0.06–0.33)</p> <p>IC: $\alpha = 0.78$ (range = 0.73–0.83)</p> <p>TRT: r ≥ 0.7</p> <p>Construct validity: overweight at Time 1 was related to increased maternal control ($p = 0.052$) and normal weight at Time 2 was related to maternal presence ($p = 0.01$)</p>

Environment summary table					
Tool information			Sample		
No.	Name	First author (type of paper)	Type	Administration	Age, weight status; country; ethnicity; (n)
8	Home Environment Survey (HES), 105 item	Gattshall 2008 ²¹⁸ (PDP)	105 item home environment measure	Parent completed – in environment or outside environment	Child; Obese and overweight; USA (white; AA; Hispanic; other not defined) (IC and validity n = 219; TRT n = 156)
					<p>IC: $\alpha = 0.75$ (range = 0.59–0.84)</p> <p>TRT: $r = 0.79$ (range = 0.01–0.99)</p> <p>Inter-rater: $r = 0.47$ (range = 0.02–1.0)</p> <p>Construct validity: with PA range $r = 0.05$–0.36</p>
9	Electronic equipment scale, 21 item	Rosenberg 2010 ²¹⁹ (study 1) (PDP)	21 item home environment measure	Parent completed – in environment and self-reported	Children and adolescent; mixed (stratified); USA (white) (TRT and construct n = 476; inter-rater n = 171)
					<p>TRT: $r = 0.78$ (range = 0.38–0.87) in self-report; $r = 0.75$ (range = 0.26–0.96) in parent proxy</p> <p>Inter-rater: $r = 0.65$ (range = 0.36–0.93)</p> <p>Construct validity: with adolescent proxy report/checklist with adolescent self-report</p> <p>Construct: television viewing time: television in the home positively related to television viewing time in adolescent self-report ($\beta = 0.17$, $p = 0.03$) parent report of adolescents ($\beta = 0.24$, $p = 0.00$) and parent report of children ($\beta = 0.39$, $p = 0.00$) Sedentary composite: high sedentary composition score positively related to adolescent report of electronics in the bedroom ($\beta = 0.22$, $p = 0.005$) and portable electronics ($\beta = 0.16$, $p = 0.05$) BMI z-score: electronics in the bedroom for both adolescent report ($\beta = 0.19$, $p = 0.03$) and parent report of adolescents was positively associated with BMI z-score ($\beta = 0.17$, $p = 0.05$)</p>
					<p>Construct validity is poor, but could be related to low variability in sample (all obese)</p> <p>Authors state that this is a pilot study that 'shows promise'</p>

Environment summary table					
Tool information		Sample			
No.	Name	First author (type of paper)	Type	Administration	Age, weight status, country; ethnicity; (n)
10	Home PA equipment scale, 14 item	Rosenberg 2010 ²¹⁹ (study 2) (ModEval)	14 item home environment measure	Parent completed in environment and self-reported	Children and adolescent; mixed (stratified); USA (white) (TRT and construct n = 476; inter-rater n = 171)
					<p>TRT: $r = 0.59$ (range = 0.48–0.78) in self-report; $r = 0.69$ (range = 0.53–0.85) in parent proxy for child; $r = 0.63$ (range = 0.50–0.76) in parent proxy for adolescent</p> <p>Inter-rater: $r = 0.57$ (range = 0.44–0.70)</p> <p>Construct: television viewing time: activity equipment was negatively associated with television viewing time in adolescent report ($\beta = -0.21$, $p = 0.01$), parent report of adolescents ($\beta = -0.23$, $p = 0.003$) and parent report of child ($\beta = -0.23$, $p = 0.02$)</p> <p>PA: activity equipment was positively related to PA in both adolescent report ($\beta = 0.22$, $p = 0.01$) and parent report of adolescents ($\beta = 0.20$, $p = 0.01$)</p>
					<p>Adapted from the adult version (Sallis 1997⁶). The activity equipment scale was reliable when reported by parents and by adolescents. Home environment attributes were related to multiple obesity-related behaviours and to child weight status, supporting the construct validity of these scales</p>

ANOVA, analysis of variance; CFI, comparative fit index; df, degrees of freedom; MANOVA, multiple analysis of variance; ModEval, modified an existing tool and re-evaluated; PDP, primary development paper.

a Not linked to bibliography: Sallis JF, Johnson MF, Johnson MF, Calfas KJ, Caparosa S, Nichols JF. Assessing perceived physical environmental variables that may influence physical activity. *Res Q Exerc Sport* 1997;**68**:345–5.

Appendix 16 Additional scoping searches for quality-adjusted life-years and clinical cut-offs in physiological measures

Preference-based utility measures (enabling calculation of quality-adjusted life-years)

Run: 9 October 2012.

Ovid MEDLINE(R) <1946 to September Week 4 2012>.

Search strategy:

1. child/ (1,290,311)
2. *obesity/ (76,973)
3. "economic evaluation*" .tw. (5200)
4. program evaluation/ec 323
5. *Cost-Benefit Analysis/ (3917)
6. 3 or 4 or 5 (8933)
7. 1 and 2 and 6 (12)
8. quality-adjusted life years/ (5950)
9. quality adjusted life.tw. (4795)
10. (qaly or qalys or qald or qale or qtime).tw. (3954)
11. 8 or 9 or 10 (8286)
12. obesity/ (112,373)
13. 1 and 11 and 12 (14)
14. 7 or 13 (22)

Clinical meaningfulness of physiological outcome measures in childhood obesity

Run: 30 October 2012.

Ovid MEDLINE(R) <1946 to October Week 3 2012>.

Search strategy:

1. Insulin/ (151,351)
2. Ghrelin/ (4430)
3. Glucose Tolerance Test/ (27,736)
4. Basal Metabolism/ (6360)
5. Blood Pressure/ (228,480)
6. Heart Rate/ (134,007)
7. ((insulin or Ghrelin or HOMA or "Hyperglycemic clamp*" or "Oral Glucose Tolerance Test*" or OGTT or "Haemoglobin A1c" or "Estimated Resting Metabolic Rate*" or "Predicted Resting Energy Expenditure*" or "Basal metabolic rate*" or BMR or "Blood pressure*" or Systolic or Diastolic or "Blood cholesterol*" or "Heart rate*") adj5 "clinical relevanc*").tw. (119)
8. Hemoglobin A, Glycosylated/ (20,012)

9. Cholesterol/bl [Blood] (55,968)
10. 1 or 2 or 3 or 4 or 5 or 6 or 8 or 9 (525,688)
11. (insulin or Ghrelin or HOMA or "Hyperglycemic clamp*" or "Oral Glucose Tolerance Test*" or OGTT or "Haemoglobin A1c" or "Estimated Resting Metabolic Rate*" or "Predicted Resting Energy Expenditure*" or "Basal metabolic rate*" or BMR or "Blood pressure*" or Systolic or Diastolic or "Blood cholesterol*" or "Heart rate*").tw. (568,813)
12. 10 or 11 (803,029)
13. adolescent/ or child/ or child, preschool/ or infant/ (2,444,484)
14. weight gain/ or weight loss/ or overweight/ or exp obesity/ (155,575)
15. 13 and 14 (36,168)
16. (clinical* adj3 (relevan* or meaningful* or useful* or appropriate*)).tw. (105,700)
17. 12 and 16 (7237)
18. 15 and 16 and 17 (49)

Appendix 17 Appraisal decision forms: anthropometry

No.	Type of measure	First author	Author's conclusion: Y (advocate), N (do not advocate), ? (unclear)	Decision of certainty: 1. <i>certain</i> – good evidence, fit for purpose; 2. <i>certain</i> – poor evidence, not fit for purpose; 3. <i>uncertain</i> – requiring further consideration			Expert comments
				CoOR internal decision	3	2	
1	Skinfold measurements	Watts 2006 ³⁵⁹	N				Does not address purpose of review: i.e. change in body composition.
		Rowe 2006 ³⁶⁰	N				Criterion measure in most of these was not actually a criterion (4C or 3C model and TBW)
		Rodriguez 2005 ³⁶¹	N				
		Morrison 2001 ³⁶²	N				
		Ball 2006 ³²⁰	N				
		Marshall 1991 ²⁶	Y				
		Marshall 1990 ²⁷	Y				
		Sardinha 1999 ²⁹	Y				
		Semiz 2007 ²⁷¹	N				
		Elberg 2004 ²²²	N				
		Sampej 2001 ³⁰⁶	?				
		Ayvaz 2011 ²⁸	Y				
		Himes 1989 ³⁷⁵	?				
		Goran 1996 ²⁸³	N				
		Hannon 2006 ²⁸²	?				
		Rolland-Cachera 1997 ²⁷²	?				
		Nuutinen 1991 ³⁰⁹	N				
		^a Campanozzi 2008 ³⁷⁸	N				
		^a Johnston 1985 ³⁹²	N				
		^a Moore 1999 ³⁸⁵	?				
		^a Owens 1999 ¹⁵⁶	N				

No.	Type of measure	First author	Author's conclusion: Y (advocate), N (do not advocate), ? (unclear)	Decision of certainty: 1. <i>certain</i> – good evidence, fit for purpose; 2. <i>certain</i> – poor evidence, not fit for purpose; 3. <i>uncertain</i> – requiring further consideration		Expert consensus decision	Expert comments
				CoOR internal decision	CoOR external decision		
2	BMI	^a Malina 1986 ²⁸⁸	?				
		^a Pecoraro 2003 ³⁹⁰	?				
		Jakubowska-Pietkiewicz 2009 ⁴⁰³ (Polish)	?				
		Chiara 2003 ⁴⁰⁰ (Portuguese)	?				
		Zaragoza 1998 ³⁹⁸ (Spanish)	?				
		Zambon 2003 ³⁹⁵ (Portuguese)	N				
		Kayhan 2009 ³⁹⁵ (Turkish)	Y				
		Himes 1999 ³⁰⁸	N				
		Wickramasinghe 2009 ²³	?		3	1	
		Widhalm 2001 ²⁸⁶	N				
		Gaskin 2003 ²⁸⁷	N				
		Warner 1997 ²⁸⁸	?				
		Ochiai 2010 ²⁹³	Y				
		Potter 2007 ²⁹²	?				
Reilly 2000 ²⁹¹	Y						
Glaner 2005 ²⁹⁰	N						
Pietrobelli 1998 ⁹²	Y						
Himes 1999 ³⁰⁸	?						

For diagnostic purposes BMI is good (e.g. obese) but no one has categorised how much weight needs to be lost in order to change categories. It is particularly useful for large sample sizes and those where large changes are expected (e.g. surgery)

No.	Type of measure	First author	Author's conclusion: Y (advocate), N (do not advocate), ? (unclear)	Decision of certainty: 1. <i>certain</i> – good evidence, fit for purpose; 2. <i>certain</i> – poor evidence, not fit for purpose; 3. <i>uncertain</i> – requiring further consideration		Expert comments
				CoOR internal decision	Expert consensus decision	
		Mast 2002 ²⁹⁷	Y			
		Duncan 2009 ³⁰⁰	Y			
		Ellis 1999 ²⁹⁹	N			
		Malina 1999 ²⁹⁸	?			
		Rush 2003 ²¹	?			
		Bartok 2011 ³⁰¹	Y			
		El Taguri 2009 ³⁰²	Y			
		Yoo 2006 ³⁰³	Y			
		Rolland-Cachera 1982 ³⁰⁵	Y			
		Ayvaz 2011 ²⁸	?			
		Eto 2004 ³⁰⁴	N			
		Reilly 2010 ³⁷¹	Y			
		Wickramasinghe 2005 ²²	?			
		Sampei 2001 ³⁰⁶	?			
		Marshall 1991 ²⁶	Y			
		Goran 1996 ²⁸³	N			
		Adegboye 2010 ³¹³	?			
		Fujita 2011 ³¹⁵	Y			
		Jung 2009 ³¹⁴	Y			
		Glasser 2011 ³¹¹	?			
		Neovius 2005 ³¹²	Y			

No.	Type of measure	First author	Author's conclusion: Y (advocate), N (do not advocate), ? (unclear)	Decision of certainty: 1. <i>certain</i> – good evidence, fit for purpose; 2. <i>certain</i> – poor evidence, not fit for purpose; 3. <i>uncertain</i> – requiring further consideration		Expert consensus decision	Expert comments
				CoOR internal decision	Expert consensus decision		
		Nuutinen 1991 ³⁰⁹	?				
		Mei 2007 ³¹⁰	Y				
		^a Zheng 2010 ³⁷⁶	Y				
		^a Owens 1999 ¹⁵⁶	?				
		^a Malina 1986 ³⁸⁹	?				
		^a Pecoraro 2003 ³⁹¹	?				
		^a Freedman 2005 ³⁹³	N				
		^a Freedman 2009 ³⁹⁴	?				
		Zhang 2004 ⁴⁰⁸ (Chinese)	Y				
		Rodriguez 2008 ⁴⁰⁵ (Spanish)	N				
		Perez 2009 ⁴⁰⁴ (Spanish)	Y				
		Jakubowska-Pietkiewicz 2009 ⁴⁰³ (Polish)	?				
		Giugliano 2004 ⁴⁰² (Portuguese)	Y				
		da Silva 2010 ⁴⁰¹ (Portuguese)	Y				
		Chiara 2003 ⁴⁰⁰ (Portuguese)	?				
		Behbahani 2009 ³⁹⁹ (Persian)	N				
		Zaragoza 1998 ³⁹⁸ (Spanish)	N				

No.	Type of measure	First author	Author's conclusion: Y (advocate), N (do not advocate), ? (unclear)	Decision of certainty: 1. <i>certain</i> – good evidence, fit for purpose; 2. <i>certain</i> – poor evidence, not fit for purpose; 3. <i>uncertain</i> – requiring further consideration		Expert comments
				CoOR internal decision	Expert consensus decision	
		Zambon 2003 ³⁹⁷ (Portuguese)	Y			
		Majcher 2008 ³⁹⁶ (Polish)	?			
		Kayhan 2009 ³⁹⁵ (Turkish)	Y			
		Semiz 2007 ²⁷¹	Y			
3	Weight	Mei 2002 ²⁹⁸	?	3	2	
		Asayama 2002 ⁴²²	?			
		Marshall 1990 ²⁷	Y			
		Ball 2006 ³²⁰	N			
		^a Johnston 1985 ³⁸²	N			
		^a Owens 1999 ¹⁵⁶	?			
		Schonhaut 2004 ⁴⁰⁶ (Spanish – includes height – IR of measures)	N			
		Himes 1989 ³⁷⁵	N			
4	Self-reported height and weight	VanVliet 2009 ³³⁶	N	2	2	
		Goodman 2000 ²²⁶	Y			
		Seghers 2010 ³³⁷	N			
		Jansen 2006 ³³⁸	N			
		Zhou 2010 ³³⁹	N			
		Yan 2009 ³⁴⁰	N			
		Fonseca 2010 ³⁴¹	N			

No.	Type of measure	First author	Author's conclusion: Y (advocate), N (do not advocate), ? (unclear)	Decision of certainty: 1. <i>certain</i> – good evidence, fit for purpose; 2. <i>certain</i> – poor evidence, not fit for purpose; 3. <i>uncertain</i> – requiring further consideration		Expert comments
				CoOR internal decision	Expert consensus decision	
		Enes 2009 ³⁴²	N			
		Crawley 1995 ³⁴³	N			
		Linhart 2010 ³⁴⁴	N			
		Lee 2006 ³⁴⁵	N			
		Wang 2002 ³⁴⁶	N			
		Tsigilis 2006 ³⁴⁷	N			
		Tokmakidis 2007 ³⁴⁸	N			
		Strauss 1999 ²²⁷	Y			
		Rasmussen 2007 ³²²	N			
		Shields 2008 ³⁴⁹	N			
		Abalkhail 2002 ³⁵⁰	N			
		Hauck 1995 ³⁵¹	N			
		Bae 2010 ³⁵²	N			
		De Vriendt 2009 ³⁵³	N			
		Ambrosi-Randic 2007 ³⁵⁴	Y			
		Field 2007 ³⁵⁵	Y			
		Morrissey 2006 ²⁹⁴	N			
		Elgar 2005 ³⁵⁶	N			
		Stein 2006 (German) ⁴⁰⁷	N			
		^a Kurth 2010 ³⁸³	N			
		Brener 2003 ³⁵⁷	Y			

No.	Type of measure	First author	Author's conclusion: Y (advocate), N (do not advocate), ? (unclear)	Decision of certainty: 1. <i>certain</i> – good evidence, fit for purpose; 2. <i>certain</i> – poor evidence, not fit for purpose; 3. <i>uncertain</i> – requiring further consideration		Expert comments
				CoOR internal decision	Expert consensus decision	
5	Self-reported WC	Bekkers 2011 ³⁵⁸ van Vliet 2009 ³³⁶	Y Y	3	2	
6	Parent-reported height and weight (BMI)	Akinbami 2009 ³²⁷ Huybrechts 2006 ³²⁸ Huybrechts 2011 ³¹ Garcia-Marcos 2006 ³³⁰ Dubois 2007 ³²⁴ Jones 2011 ³³¹ O'Connor 2011 ³²¹ Molina 2009 ²⁹⁵ Vuorela 2010 ³³² Tschamler 2010 ³³³ Scholtens 2007 ²²⁸ Wen 2011 ³³⁴ Maynard 2003 ²⁹⁶ Akerman 2007 ³³⁵	N N ? N N N N N N N Y N N N N	2		
7	Neck circumference	Nafiu 2010 ³²⁶ ^a Hatipoglu 2010 ³⁸¹	Y Y	3	2	
8	ADP	Nicholson 2001 ²²¹ Elberg 2004 ²²² Lazzer 2008 ²²³	Y Y Y	1	?2	Not criterion any more, need to re-evaluate based on criterion (4C model and TBW) – Gately 2003 ²⁰ compares with actual criterion

No.	Type of measure	First author	Author's conclusion: Y (advocate), N (do not advocate), ? (unclear)	Decision of certainty: 1. <i>certain</i> – good evidence, fit for purpose; 2. <i>certain</i> – poor evidence, not fit for purpose; 3. <i>uncertain</i> – requiring further consideration		Expert comments
				CoOR internal decision	Expert consensus decision	
		^a Mello 2005 ²²⁴	Y			
		^b Radley 2003 ²²⁵	?			
		Gately 2003 ¹⁷	Y			
9	Arm circumference	Rolland-Cachera 1997 ²⁷²	Y	3	2	
		Sardinha 1999 ²⁹	?			
		Zaragoza 1998 ³⁹⁸ (Spanish)	N			
		Mazicioglu 2010 ³⁷²	Y			
10	TOBEC	Ellis 1996 ²⁸⁴	?	3	2	Not enough evidence and too old
11	WC	Asayama 2000 ²⁷⁹	N	3	2	No better than BMI z-score and observer dependent
		Savagan-Gurol 2010 ²⁷⁰	?			
		Semiz 2007 ²⁷¹	?			
		Himes 1999 ³⁰⁸	?			
		Glasser 2011 ³¹¹	Y			
		Adegboye 2010 ³¹³	Y			
		Neovius 2005 ³¹⁴	Y			
		Mazicioglu 2010 ³⁷²	Y			
		Reilly 2010 ³⁷¹	N			
		Hitze 2008 ³⁷⁰	Y			
		Ayvaz 2011 ²⁸	?			
		Asayama 2002 ⁴²²	?			

No.	Type of measure	First author	Author's conclusion: Y (advocate), N (do not advocate), ? (unclear)	Decision of certainty: 1. <i>certain</i> – good evidence, fit for purpose; 2. <i>certain</i> – poor evidence, not fit for purpose; 3. <i>uncertain</i> – requiring further consideration		Expert comments
				CoOR internal decision	Expert consensus decision	
		Ball 2006 ³²⁰	Y			
		Fujita 2011 ³¹⁵	Y			
		Jung 2009 ³¹⁴	Y			
		^a Zheng 2010 ³⁷⁶	Y			
		^a Yamborisut 2008 ³⁷⁷	?			
		^a Owens 1999 ³⁸⁶	?			
		^a Taylor 2008 ⁴⁰¹	Y			
		Perez 2009 (Spanish) ⁴⁰⁴	N			
		Jakubowska-Pietkiewicz 2009 (Polish) ⁴⁰³	?			
		Garnett 2005 ³⁶⁷	?			
12	BIA	Shaikh 2007 ²⁷³	Y	3	2	Did not address change over time
		Battistini 1992 ³⁶⁵	N			
		Fors 2002 ³¹⁸	?			
		Lu 2003 ³²³	?			
		Lazzer 2008 ²²³	?			
		Rush 2003 ²¹	Y			
		Azcona 2006 ²⁷⁴	?			
		Haroun 2009 ²⁵	N			
		Okasora 1999 ²⁷⁵	Y			
		Loftin 2007 ²⁷⁶	N			

No.	Type of measure	First author	Author's conclusion: Y (advocate), N (do not advocate), ? (unclear)	Decision of certainty: 1. <i>certain</i> – good evidence, fit for purpose; 2. <i>certain</i> – poor evidence, not fit for purpose; 3. <i>uncertain</i> – requiring further consideration		Expert comments
				CoOR internal decision	Expert consensus decision	
		Iwata 1993 ²⁸³	Y			
		Wabitsch 1996 ²⁴	N			
		Guida 2008 ²⁷⁸	?			
		Lazzer 2003 ²²³	N			
		Eisenkolbl 2001 ²⁸¹	?			
		Jakubowska-Pietkiewicz 2009 ⁴⁰³ (Polish)	Y			
		Fernandes 2007 ²⁸⁵	Y			
		Ellis 1996 ²⁸⁴	?			
		^a Goran 1996 ²⁸³	N			
		^a Zheng 2010 ³⁷⁶	?			
		^a Campanozzi 2008 ³⁷⁸	N			
		^a Goldfield 2006 ³⁷⁹	Y			
		^a Lewy 1999 ³⁸⁴	N			
		^a Williams 2007 ³⁸⁸	?			
		^a Pecoraro 2003 ³⁹¹	?			
		Rodriguez 2008 (Spanish) ⁴⁰⁵	N			
		Hannon 2006 ²⁸²	?			
		Asayama 2000 ²⁷⁹	N	3	2	
13	WHR	Savgan-Gurol 2010 ²⁷⁰	Y			
		Ayaz 2011 ²⁸	Y			

No.	Type of measure	First author	Author's conclusion: Y (advocate), N (do not advocate), ? (unclear)	Decision of certainty: 1. <i>certain</i> – good evidence, fit for purpose; 2. <i>certain</i> – poor evidence, not fit for purpose; 3. <i>uncertain</i> – requiring further consideration		Expert comments
				CoOR internal decision	Expert consensus decision	
14	WHtR	Jung 2009 ³¹⁴	N			
		Neovius 2005 ³¹²	N			
		^a Owens 1999 ¹⁵⁶	?			
		^a Brambilla 1994 ³⁹⁰	N			
		Majcher 2008 (Polish) ³⁹⁶	?			
		Serniz 2007 ²⁷¹	N			
		Savgan-Gurol 2010 ²⁷⁰	N	3	2	
		Weili 2007 ³⁶⁹	Y			
		Fujita 2011 ³¹⁵	Y			
		^a Guntsche 2010 ³⁸⁰	Y			
		^a Taylor 2008 ³⁹²	N			
		Perez 2009 (Spanish) ⁴⁰⁴	N			
		Majcher 2008 (Polish) ³⁹⁶	N			
		Adegboye 2010 ³¹³	Y			
15	DXA	Eisenkolbl 2001 ²⁸¹	Y	3	1	Can be precise and recommended for change but is machine dependent
		Wells 2010 ¹⁶	?			
		Gately 2003 ¹⁷	Y			
		^a Campanozzi 2008 ³⁷⁸	?			
		^a Tsang 2009 ³⁸⁷	?			
		^a Mello 2005 ²²⁴	?			
		^a Williams 2006 ¹⁸	N			

No.	Type of measure	First author	Author's conclusion: Y (advocate), N (do not advocate), ? (unclear)	Decision of certainty: 1. <i>certain</i> – good evidence, fit for purpose; 2. <i>certain</i> – poor evidence, not fit for purpose; 3. <i>uncertain</i> – requiring further consideration		Expert comments
				CoOR internal decision	Expert consensus decision	
		Rodriguez 2008 (Spanish) ⁴⁰⁵	?			
		Ramirez 2010 (Spanish) ¹⁹	?			
		Fors 2002 ³¹⁸	?			
16	Weight for age	Stettler 2007 ³⁷⁴	N	3	2	
17	Silhouette rating scales	Killion 2006 ³¹⁷	N	3	2	
		Jorga 2007 ³⁶³	N			
		Hager 2010 ³⁶⁴	Y			
18	WHR/Ht	Asayama 2000 ²⁷⁹	Y	3	2	
19	BIS	Ellis 1996 ²⁹²	?	3	2	Same as BIA
		Fors 2002 ³²⁷	?			
20	Per cent weight for height	Yoo 2006 ³⁰³	Y	3	2	
21	FMI	Eto 2004 ³⁰⁴	?	3	2	Not a tool. It is categorised based on which method to use
22	Rohrer index	Candido 2011 ³⁷³	Y	3	2	Used in very young but not for what we want
		Mei 2002 ³⁰⁷	N			
23	Hip circumference	Ball 2006 ³²⁰	Y	3	2	Not enough evidence
		^a Owens 1990 ¹⁵⁶	?			
24	Predicted thoracic gas volume	Radley 2007 ²⁵¹	?		2	Very rare but may have potential
25	Ultrasound measurement	Pineau 2010 ³⁶⁶	Y	3	2	Not enough evidence
26	NIR	Sampei 2001 ³⁰⁶	?	3	2	Not enough evidence
27	O-Scale	Marshall 1990 ²⁷	Y	3	2	Not enough evidence

No.	Type of measure	First author	Author's conclusion: Y (advocate), N (do not advocate), ? (unclear)	Decision of certainty: 1. <i>certain</i> – good evidence, fit for purpose; 2. <i>certain</i> – poor evidence, not fit for purpose; 3. <i>uncertain</i> – requiring further consideration		Expert consensus decision	Expert comments
				CoOR internal decision	Expert consensus decision		
28	GRE imaging for ILC	Springer 2011 ³¹⁹	Y	3	2		Not enough evidence
29	Mathematical index for assessing changes in body composition	Gillis 2000 ³²⁵	?	3	2		Not enough evidence
30	Conicity index	Taylor 2000 ³⁷⁷	N	3	2		Not enough evidence
		Candido 2011 ³⁷³	?				
		Perez 2009 (Spanish) ⁴⁰⁴	N				
		^a Taylor 2008 ³⁹²	N				
31	Broselow tape measurement	Rosenberg 2011 ³¹⁶	?	3	2		Not enough evidence
32	Ultrasonography	^a Zheng 2010 ³⁷⁶	N	3	2		
33	Waist–thigh ratio	^a Owens 1999 ¹⁵⁶	?	3	2		
34	Sagittal diameter	^a Owens 1999 ¹⁵⁶	Y	3	2		
35	Sagittal diameter–calf ratio	^a Owens 1999 ¹⁵⁶	?	3	2		
36	Thigh circumference	^a Owens 1999 ¹⁵⁶	?	3	2		
37	Arm fat area	^a Brambilla 1994 ³⁹⁰	N	3	2		
38	Thigh fat area	^a Brambilla 1994 ³⁹⁰	N	3	2		

BIS, bioelectrical impedance spectroscopy; FMI, fat mass index; GRE, gradient recalled echo; ILC, intrahepatic lipid content; WHR/Ht, waist-to-hip ratio/height.

^a Manuscripts identified post external meeting (arrival from library loans). Many were linked to measures that had already been discussed by experts. Those that describe measures that had not been discussed already by experts were forwarded to experts, who were asked to make decisions remotely.

Note: Authors often evaluated more than one type of measure in a single manuscript. These author names are repeated in each of the included type of measure.

Appendix 18 Diet methodology studies: development and evaluation scores

No.	Name of tool: diet measures	First author (reference)	Development			Evaluation							
			Clear concept	Underpinned by theory	Description of sample	Sample involved in development	IC rater	TRT/inter-rater	Internal validity	Criterion validity	Convergent validity	Construct validity	Responsiveness
FFQs/checklists													
1	Food Frequency Questionnaire	Lee 2007 ⁴⁸	3	1	4	1		TRT = 3					
2	Qualitative Dietary Fat Index	Yaroch 2000 ⁴²	3	1	4	1		TRT = 2		2			
3	Short-list Youth Adolescent Questionnaire (Short YAQ)	Rockett 2007 ³⁴	4	1	3	1				4		4	
4	Youth Adolescent Questionnaire (YAQ)	Rockett 1995 ⁴³	4	3	3	3		TRT = 3		2			
5	Youth Adolescent Questionnaire (YAQ)	Rockett 1997 ³⁷	4	3	3	3				4			
6	Youth Adolescent Questionnaire (YAQ)	Perks 2000 ³⁰	3	1	2	3				2			
7	Picture sort FFQ	Yaroch 2000 ⁴¹	4	3	4	1		TRT = 2				3	
8	Childs Eating Habits Questionnaire (CEHQ-FFQ)	Lanfer 2011 ³⁶	3	1	2	3		TRT = 4					
9	Childs Eating Habits Questionnaire (CEHQ-FFQ)	Huybrechts 2011 ³¹	3	1	2	3				4			
10	Australian Child and Adolescent Eating Survey (ACAES)	Watson 2009 ⁴⁶	4	2	2	3		TRT = 3				3	

No. diet measures	Name of tool:	First author (reference)	Development			Evaluation						
			Clear concept	Underpinned by theory	Description of sample	Sample involved in development	IC rater	TRT/inter-internal validity	Criterion validity	Convergent validity	Construct validity	Responsiveness
11	Australian Child and Adolescent Eating Survey (ACAES)	Burrows 2008 ³²	See Watson ²²	See Watson ²²	See Watson ²²	See Watson ²²		4				
12	Brief dietary screener	Nelson 2009 ⁴⁴	3	1	3	3	TRT = 3		2			
13	Brief dietary screener	Davis 2009 ⁴⁵	3	1	3	3	TRT = 3		2			
14	Intake of fried food away from home	Taveras 2005 ⁵¹	1	2	2	1			2	2		
15	Food Intake Questionnaire	Epstein 2000 ⁴⁹	1	2	3	1			3			
16	21-item dietary fat screening measure	Prochaska 2001 ⁵⁰	4	1	4	2	3 TRT = 3		3			
17	New Zealand Food Frequency Questionnaire (New Zealand FFQ)	Metcalf 2003 ⁴⁷	4	1	3	3	3 TRT = 4					
18	Harvard Service Food Frequency Questionnaire (HSFFQ)	Blum 1999 ³⁸	3	1	3	3			4			
19	5-day food frequency questionnaire (5D-FFQ)	Crawford 1994 ³³	3	1	3	1		2				
20	Dietary Guideline Index for Children and Adolescents (DGI-CA)	Golley 2011 ³⁵	4	1	3	1					4	

No.	Name of tool: diet measures	First author (reference)	Development			Evaluation							
			Clear concept	Underpinned by theory	Description of sample	Sample involved in development	IC rater	TRT/inter-rater	Internal validity	Criterion validity	Convergent validity	Construct validity	Responsiveness
21	Familial influence on food intake–Food Frequency Questionnaire (FIF-FFQ)	Vereecken 2010 ³⁹	3	1	3	1				4			
Diet/recalls/observations													
22	Diet history	Sjoberg 2003 ⁵³	4	0	2	1				4			
23	2-week Diet History interview (DHI)	Waling 2009 ⁵⁴	2	0	2	1				4			
24	3-day weighed food diary	Maffei 1994 ⁵⁵	3	0	2	1				3			
25	7-day diet history	Maffei 1994 ⁵⁵	3	0	2	1				3			
26	9-day food diary	Singh 2009 ⁵⁷	3	0	2	1				2			
27	3-day dietary intake record	O'Connor 2001 ⁶⁴	4	0	2	1				3			
28	2-week food diary	Bandini 1990 ⁵⁸	3	0	2	1				2			
29	2-week food diary	Bandini 1999 ⁵⁹	3	0	2	1				2			
30	Tape-recorded food record (3 day)	Lindquist 2000 ⁶⁰	4	2	4	1				3			
31	7-day weighed food diary	Bratteby 1998 ⁶¹	3	0	2	1				2			
32	3-day estimated food diary	Crawford 1994 ³³	3	0	3	1				3			

No.	Name of tool: diet measures	First author (reference)	Development				Evaluation						
			Clear concept	Underpinned by theory	Description of sample	Sample involved in development	IC	TRT/inter-rater	Internal validity	Criterion validity	Convergent validity	Construct validity	Responsiveness
33	8-day food record	Champagne 1996 ⁶³	2	0	3	1			2				
34	8-day food record	Champagne 1998 ⁶²	2	0	3	1			2				
35	Tape-recorded food record	Van Horn 1990 ⁵⁶	2	0	4	1	IR = 3						
36	24-hour dietary recall (1 day)	Baxter 2006 ⁶⁵	2	0	4	1	TRT = 3		3				
37	24-hour dietary recall (3 day)	Johnson 1996 ⁶⁸	3	0	1	1			3				
38	24-hour dietary recall (1 day)	Lytle 1998 ⁶⁷	3	1	2	1			4				
39	24-hour recall (1 day)	Crawford 1994 ³³	3	0	3	1			3				
40	Telephone 24-hour diet recall	Van Horn 1990 ⁵⁶	2	0	4	1	IR = 3						
41	Day in the Life Questionnaire (DILQ)	Edmunds 2002 ⁶⁶	4	0	3	3	TRT = 3 IR = 4		3				2
42	Diet Observation at Childcare (DOCC)	Ball 2007 ⁷⁰	4	4	0	1	IR = 4		4				
43	Food Behaviour Questionnaire	Vance 2008 ⁷¹	4	0	2	3	TRT = 4 IR = 4		3				
Biochemical markers													
44	IGF-1, IGFBP-1, IGFBP-3 – biomarkers	Martinez de Icaya 2000 ⁶⁹	4	0	2	0						3	

Note: Scores = 1–4 (0 = N/A) (full details on data extraction form and Excel extraction database).

Appendix 19 Eating behaviour methodology studies: development and evaluation scores

No.	Name of tool: eating behaviour measures	First author	Development				Evaluation					
			Clear concept	Underpinned by theory	Description of sample	Sample involved in development	IC	TRT/inter-rater	Internal validity	Criterion validity	Convergent validity	Construct validity
1	ChEDE-Interview	Decaluwé 2004 ⁸¹	3	1	2	2	3	TRT=3; inter-rater=4		4		
2	ChEDE-Interview	Bryant-Waugh 1996 ⁴¹¹	3	1	2	2						
3	ChEDE-Q	Goossens 2010 ⁴¹²	3	1	2	2	4			4		
4	ChEDE-Q	Jansen 2007 ²²⁹	3	1	2	2	2			3		
5	ChEDE-Q	Tanofsky-Kraff 2003 ²³⁰	3	1	2	2				2		
6	ChEDE-Interview	Tanofsky-Kraff 2005 ⁴¹³	3	1	2	2				3		
7	IFQ	Baughcum 2001 ⁷⁴	3	1	3	2	3		4			
8	PFQ	Baughcum 2001 ⁷⁴	3	1	3	2	3		4			
9	KEDS	Childress 1993 ⁸⁹	3	1	1	2	3	TRT=3	3			
10	QEWPA	Johnson 1999 ⁹⁰	4	1	2	1		Inter-rater=2		2	2	
11	QEWPA	Steinberg 2004 ⁹¹	4	1	2	1		Inter-rater=3		3	3	
12	DEBQ-C	Van Strien 2008 ⁷⁹	4	3	2	2	4		4		3	
13	DEBQ-C	Banos 2011 ⁸³	4	3	2	2	4	TRT=4	3		3	
14	DEBQ-P	Caccialanza 2004 ⁹⁸	3	4	2	2	4		2			
15	DEBQ-P	Braet 1997 ⁷⁸	3	4	2	2	3		4		3	

No.	Name of tool: eating behaviour measures	Development					Evaluation						
		First author	Clear concept	Underpinned by theory	Description of sample	Sample involved in development	IC	TRT/inter-rater	Internal validity	Criterion validity	Convergent validity	Construct validity	Responsiveness
16	DEBQ-C	Braet 2007 ⁹²	4	3	2	2	4	Inter-rater=3					
17	ChEAT	Maloney 1988 ⁸⁶	2	1	4	1	3	TRT=3					
18	ChEAT	Smolak 1994 ¹⁰⁰	2	1	4	1	3		4		3		
19	ChEAT	Ranzenhofer 2008 ¹⁰¹	2	1	4	1	4		4	3			
20	EAT	Wells 1985 ⁴¹⁴	3	1	3	1			4				
21	YEDE-Q	Goldschmidt 2007 ⁹⁹	4	1	2	1	3			3	3		
22	EES-C	Tanofsky-Kraff 2007 ⁷⁷	4	1	4	1	4	TRT=4	4	3	3		
23	C-BEDS	Shapiro 2007 ²³¹	4	1	3	1				2			
24	CFQ	Birch 2001 ⁷⁵	3	4	4	2	4		4				
25	CFQ	Haycraft 2008 ⁹³	3	4	4	2		Inter-rater=3	2				
26	CFQ	Anderson 2005 ⁹⁶	3	4	4	2			4		3		
27	CFQ	Corsini 2008 ⁹⁷	3	4	4	2	4		4		3		
28	CFQ	Polat 2010 ⁹⁴	3	4	4	2	4		4				
29	CFQ	Boles 2010 ²³²	3	4	4	2	3		4				
30	MIRFS-III	Shisslak 1999 ⁸⁷	3	1	4	3	3	TRT=4		4			
31	IFSQ	Thompson 2009 ¹⁶	3	1	4	2	3		3				

No.	Name of tool: eating behaviour measures	First author	Development				Evaluation						
			Clear concept	Underpinned by theory	Description of sample	Sample involved in development	IC	TRT/ inter-rater	Internal validity	Criterion validity	Convergent validity	Construct validity	Responsiveness
32	CEBQ	Sleddens 2008 ⁷²	4	1	2	3		4					
33	CEBQ	Wardle 2001 ⁷³	4	1	2	3	4	4					
34	TSFFQ	Corsini 2010 ⁸²	3	3	2	3	4	4		3		3	
35	KCFQ	Monnery-Patris 2011 ⁸⁵	3	1	4	1	3	4				3	
36	KCFQ	Carper 2000 ²⁵⁰	3	1	4	1				3			
37	Un-named	Murashima 2011 ⁸⁴	4	1	4	2	3	4		3		3	
38	EAH-C	Tanofsky-Kraff 2008 ⁸⁰	4	1	4	1	4	4		4		3	
39	Un-named	Kroller 2008 ⁸⁸	3	1	3	2	4			4			

Note: Scores = 1–4 (0 = N/A).

Appendix 20 Physical activity methodology studies: development and evaluation scores

No.	Name of tool: PA measures	First author	Development				Evaluation						
			Clear concept	Underpinned by theory	Description of sample	Sample involved in development	IC	TRT/inter-rater	Internal validity	Criterion validity	Convergent validity	Construct validity	Responsiveness
1	Accelerometer	Kelly 2004 ¹⁰⁵	3	0	2	1			3		2		
2	Accelerometer – Actigraph	Pate 2006 ¹⁰⁷	4	0	3	1					3		
3	Accelerometer – Caltrac monitor	Noland 1990 ¹⁰⁶	3	0	3	1			3				
4	Accelerometer – TriTrac Triaxial	Coleman 1997 ¹⁰⁸	2	0	2	1			3		2		
5	Accelerometer – Actigraph	Guinhouya 2009 ²³⁴	3	0	2	1						3	
6	HR monitoring	Maffei 1995 ²³⁷	2	0	2	1			2				
7	Pedometer	Kilanowski 1999 ¹¹⁴	2	0	2	1			3				
8	Pedometer	Duncan 2007 ²⁴⁸	3	0	3	1			3				
9	Pedometer	Jago 2006 ¹¹²	3	0	3	1							
10	Pedometer	Mitre 2009 ¹¹⁰	3	0	3	1							
11	SenseWear Pro2 Armband	Backlund 2010 ¹¹¹	4	0	2	1			3		3		
12	3-Day Physical Activity Recall (3DPAR)	Pate 2003 ⁴¹⁷	3	0	3	1			3				
13	Activity Questionnaire for Adolescents and Adults (AQUAA)	Slootmaker 2009 ¹¹⁷	3	1	3	1			3				

No.	Name of tool: PA measures	First author	Development			Evaluation									
			Clear concept	Underpinned by theory	Description of sample	Sample involved in development	IC	TRT/inter-rater	Internal validity	Criterion validity	Convergent validity	Construct validity	Responsiveness		
14	Activity Rating Scale	Sallis 1993 ¹²¹	4	1	3	1			TRT = 4						
15	Godin–Shephard Physical Activity Survey	Sallis 1993 ¹²¹	4	1	3	1			TRT = 4						
16	7-day recall interview	Sallis 1993 ¹²¹	4	0	3	1			TRT = 4	4					
17	Adolescent Physical Activity Recall Questionnaire (APARQ)	Booth 2002 ¹²³	4	3	2	1			TRT = 4				2		
18	Children's Leisure Activities Study Survey (CLASS)	Telford 2004 ¹¹⁵	4	1	3	2			TRT = 3; inter-rater = 2	3					
19	GEMS Activity Questionnaire	Treuth 2003 ¹²³	3	1	4	1			TRT = 4				2		
20	Pedometer	Treuth 2003 ¹²³	3	0	4	1			TRT = 2				3		
21	Activitygram (recall)	Treuth 2003 ¹²³	3	0	4	1			TRT = 3				3		
22	Activitygram	Welk 2004 ¹¹⁶	3	0	4	1							3		
23	Moderate to vigorous physical activity screening (study 3) ²⁵	Prochaska 2001	3	1	2	2			TRT = 4				3		
24	Moderate to vigorous physical activity screening (study 2) ²⁵	Prochaska 2001	3	1	2	2			TRT = 4				3		

No.	Name of tool: PA measures	First author	Development				Evaluation									
			Clear concept	Underpinned by theory	Description of sample	Sample involved in development	IC	TRT/inter-rater	Internal validity	Criterion validity	Convergent validity	Construct validity	Responsiveness			
25	National Longitudinal Survey of Children and Youth	Sithole 2008 ¹³⁰	2	1	1	1					Inter-rater = 4					
26	Outdoor Playtime Checklist	Burdette 2004 (study 1) ¹²²	3	1	3	1						2	3			
27	Outdoor Playtime Recall	Burdette 2004 (study 2) ¹²²	3	0	3	3						2	3			
28	Physical activity diary	Epstein 1996 ¹¹⁹	3	0	4	1						3	1			
29	Physical Activity Questionnaire (PAQ)	Janz 2008 ¹²⁹	3	3	3	2				3						
30	Physical Activity Questionnaire for Older Children (PAQ-C)	Kowalski 1997 ¹¹⁸	3	3	3	2						2	2		2	
31	Physical Activity Questionnaire for Adolescents (PAQ-A)	Kowalski 1997 ¹²⁵	3	3	3	2						2	3			
32	Physical Activity Questionnaire for Older Children (PAC-C)	Crocker 1997 (study 1) ¹²⁸	3	3	3	2			4							
33	Physical Activity Questionnaire for Older Children (PAC-C)	Crocker 1997 (study 2) ¹²⁸	3	3	3	2			3							TRT = 3

No.	Name of tool: PA measures	First author	Development				Evaluation						
			Clear concept	Underpinned by theory	Description of sample	Sample involved in development	IC	TRT/ inter-rater	Internal validity	Criterion validity	Convergent validity	Construct validity	Responsiveness
34	Physical Activity Questionnaire for Older Children (PAC-C)	Crocker 1997 (study 3) ¹²⁸	3	3	3	2	3	TRT = 3					
35	Physical Activity Questionnaire for Older Children (PAC-C)	Moore 2007 (study 2) ¹²⁶	3	3	3	2	2		3			3	
36	Physical Activity Questionnaire for Older Children (PAC-C)	Moore 2007 (study 1) ¹²⁶	3	3	3	2	3		4			3	
37	Physical Activity Questionnaire (PAQ) for Pima Indians	Kriska 1990 ²³⁶	3	1	3	2		TRT = 1					
38	Physical Activity Questionnaire (PAQ) for Pima Indians	Goran 1997 ¹²⁷	3	1	3	2				1		2	
39	Previous Day Physical Activity Recall (PDPAR)	Trost 1999 ⁴¹⁸	3	0	3	1				2			
40	Previous Day Physical Activity Recall (PDPAR)	Weston 1997 ¹²⁰	2	0	2	1				2		2	
41	Previous Day Physical Activity Recall (PDPAR)	Welk 2004 ¹¹⁶	3	0	4	1				3		3	
42	Previous Day Physical Activity Recall (PDPAR)	McMurray 2008 ⁴¹⁹	3	0	3	1				3			

No.	Name of tool: PA measures	First author	Development				Evaluation						
			Clear concept	Underpinned by theory	Description of sample	Sample involved in development	IC	TRT/ inter-rater	Internal validity	Criterion validity	Convergent validity	Construct validity	Responsiveness
43	Youth Risk Behaviour Survey (YRBS)	Troped 2007 ²³⁸	3	1	3	2			3				
44	System for observing children's activity and relationship during play (SOCARP)	Ridgers 2010 ²⁰²	4	0	2	1					2		
45	Observational System for Recording Physical Activity (OSRAC)	Brown 2006 ¹⁰³	3	0	1	2							

Note: Scores = 1-4 (0 = N/A).

Appendix 21 Sedentary time/behaviour methodology studies: development and evaluation scores

No.	Clear concept	Underpinned by theory	Description of sample	Sample involved in development	TRT/IC inter-rater	Internal validity	Criterion validity	Convergent validity	Construct validity	Responsiveness	Construct validity	Responsiveness
1	WAM-7154 Accelerometer	Reilly 2003 ¹³¹	4	0	2 1				4			
2	Computer Science and Actigraph accelerometer	Puyau 2002 ¹³²	3	0	3 1				3	3		
3	Mini-Mitter Actiwatch monitors	Puyau 2002 ¹³²	3	0	3 1				3	3		
4	Multimedia Activity Recall for Children and Adolescents (MARCA)	Ridley 2006 ¹³³	4	1	2 2	3			3			
5	Electronic Momentary Assessment (EMA): self-report survey on mobile phones	Dunton 2011 ¹³⁴	3	1	4 2						3	
6	Habit books with index cards	Epstein 2004 ¹³⁵	3	0	3 1				2			

Note: Scores = 1–4 (0 = N/A).

Appendix 22 Fitness methodology studies: development and evaluation scores

No. fitness measures	Name of tool:	First author	Development			Evaluation						
			Clear concept by theory	Underpinned by theory	Description of sample	Sample involved in development	Internal consistency	TRT/inter-rater	Internal validity	Criterion validity	Convergent validity	Construct validity
1	6-minute walk test (6MWD)	Morinder 2009 ¹³⁸	4	0	2	1		TRT = 3	3			
2	Height-adjustable step test	Francis 1991 ¹⁴⁸	3	0	2	1			4			
3	20-m shuttle run	Leger 1988 ¹³⁹	2	0	1	1		TRT = 3	4			
4	International Fitness Scale (IFS)	Ortega 2011 ¹³⁶	3	1	2	1		TRT = 4	4	4	4	
5	Bioelectrical impedance	Roberts 2009 ¹⁴⁷	3	0	3	1			3			
6	20-minute shuttle test	Suminski 2004 ¹⁴⁰	3	0	3	1		TRT = 3	4			
7	Fitnessgram	Morrow 2010 ¹⁴⁰	4	1	4	2		TRT = 4; inter-rater = 4				
8	Submaximal Treadmill Test	Nemeth 2009 ¹⁴⁹	3	0	2	1			2			
9	BMR with fat-free mass	Drinkard 2007 ¹⁴³	3	0	3	1			3			
10	Estimated maximal oxygen consumption and maximal aerobic power	Aucouturier 2009 ¹⁴⁴	4	0	2	1			3			
11	Physical working capacity in predicting VO _{2max}	Rowland 1993 ¹⁴⁵	3	0	2	1			3			
12	Aerobic cycling power	Carrel 2007 ¹⁴⁶	3	0	3	1			2		2	
13	Measured VO ₂ peak (cycle vs. treadmill)	Loflin 2004 ¹⁴¹	3	0	2	1		TRT = 3		3		
14	Harvard Step Test	Meyers 1969 ¹⁴²	1	0	2	1		TRT = 3				

Note: Scores = 1-4 (0 = N/A).

Appendix 23 Physiology methodology studies: development and evaluation scores

No.	Name of tool: physiological measures	Author	Development			Evaluation						
			Clear concept by theory	Description of sample	Sample involved in development	Internal consistency	TRT/inter-rater	Internal validity	Criterion validity	Convergent validity	Construct validity	Responsiveness
1	Indices of insulin sensitivity	Yeckel 2004 ¹⁵²	3	0	3	1			4	3		
2	Fasting indices of insulin sensitivity	Conwell 2004 ¹⁵³	4	0	3	1			3			
3	Indices of insulin sensitivity	George 2011 ¹⁵⁴	4	0	3	1			4			
4	Indices of insulin sensitivity	Gunczler 2006 ¹⁵⁵	3	0	2	1				4		
5	Indices of insulin sensitivity	Uwaifo 2002 ¹⁵⁶	3	0	3	1			3			
6	Insulin sensitivity and pancreatic beta cell function	Gungor 2004 ¹⁵⁸	3	0	3	1			4			
7	Fasting indices of insulin sensitivity	Atabek 2007 ¹⁵⁹	3	0	2	1			3			
8	Homeostasis model assessment of insulin resistance	Keskin 2005 ¹⁶⁰	4	0	2	1			3			
9	Homeostasis model assessment of insulin resistance	Rosner 2008 ¹⁶¹	2	0	2	1			4	4		
10	Indices of insulin sensitivity	Schwartz 2008 ¹⁶²	3	0	3	1			3			
11	Impaired fasting glucose	Cambuli 2009 ¹⁶³	2	0	2	1			3			
12	Hyperglycaemic clamp	Uwaifo 2002 ¹⁵⁷	3	0	3	1			3			
13	Oral glucose tolerance test (OGTT)	Libman 2008 ¹⁶⁴	4	0	3	1						2
							TRT=3					

No.	Name of tool: physiological measures	Author	Development			Evaluation							
			Clear concept by theory	Underpinned by theory	Description of sample	Sample involved in development	Internal consistency	TRT/inter-rater	Internal validity	Criterion validity	Convergent validity	Construct validity	Responsiveness
14	13C-glucose breath test – insulin resistance	Jetha 2009 ¹⁶⁵	3	0	3	1			4				
15	Ultrasound analysis of liver echogenicity	Soder 2009 ¹⁷⁷	3	0	2	1	Inter-rater = 3						
16	HbA _{1c}	Nowicka 2011 ¹⁷⁴	3	0	3	1				3	2		
17	Ghrelin	Kelishadi 2008 ¹⁷⁵	4	0	2	1					2	4	
18	Photoplethysmography (HR)	Russonello 2010 ¹⁷⁰	4	0	1	1			3				
19	Estimated resting metabolic rate	Molnar 1995 ¹⁶⁶	3	0	2	1			3				
20	Predicted REE	Rodriguez 2002 ¹⁶⁷	3	0	3	1			4				
21	Predicted REE	Lazzer 2006 ¹⁶⁸	4	0	3	1			4				
22	Predicted REE	Firouzbakhsh 1993 ¹⁶⁹	3	0	2	1			4				
23	Predicted REE	Derumeaux-Burel 2004 ¹⁷⁰	3	0	2	1			3			1	
24	Indirect calorimetry for REE	Hofsteenge 2010 ¹⁷¹	3	0	3	1			2				
25	DXA-lean body mass REE	Schmelzle 2004 ¹⁷²	3	0	2	1			4				
26	BMR with fat-free mass	Dietz 1991 ¹⁷³	3	0	2	1			2				

REE, resting energy expenditure.
Note: Scores = 1–4 (0 = N/A).

Appendix 24 Health-related quality-of-life studies: development and evaluation scores

No. measures	Name of tool: health-related quality-of-life measures	First author	Development			Evaluation								
			Clear concept by theory	Underpinned by theory	Description of sample	Sample involved in development	Internal consistency	TRT/inter-rater	Internal validity	Criterion validity	Convergent validity	Construct validity	Responsiveness	
1	Child Health Questionnaire	Waters 2000 ¹⁹²	2	1	1	1	3							
2	Child Health Questionnaire	Landgraf 1998 ¹⁸⁶	3	3	4	2	4				2			
3	Child Health Questionnaire	Waters 2000 ¹⁹³	3	1	2	2	4		TRT = 2	2				
4	DISABKIDS	Ravens-Sieberer 2007 ¹⁹⁴	4	1	4	3	4				3			
5	KIDSCREEN	Ravens-Sieberer 2007 ¹⁹⁴	4	3	4	3	4				4			
6	EQ-5D-Y	Burstrom 2011 ²⁴⁷	3	1	3	3	4					1		
7	EQ-5D-Y	Burstrom 2011 ²⁸	3	1	3	3	4							
8	EQ-5D-Y	Wille 2010 ²⁴³	4	1	3	3	4				2			
9	EQ-5D-Y	Ravens-Sieberer 2010 ²⁴⁴	4	1	3	3	4		3		3			
10	Impact of Weight on Quality of Life	Kolotkin 2006 ¹⁸¹	4	1	3	2	4			4			4	3
11	Impact of Weight on Quality of Life	Modi 2011 ¹⁸²	4	1	3	2	4		TRT = 3					

No.	Name of tool: health-related quality-of-life measures	First author	Development				Evaluation						
			Clear concept	Underpinned by theory	Description of sample	Sample involved in development	Internal consistency	TRT/inter-rater	Internal validity	Criterion validity	Convergent validity	Construct validity	Responsiveness
12	KINDL-R Questionnaire	Erhart 2009 ¹⁸⁹	3	1	3	1	4	3		4			
13	Paediatric Cancer Quality of Life Inventory-32 (short form)	Varni 1998 ¹⁸⁸	4	3	4	2	4	Inter-rater = 4		3			
14	Paediatric Cancer Quality of Life Inventory	Varni 1998 ¹⁹⁵	4	3	4	2		Inter-rater = 3					
15	Paediatric Quality of Life Inventory V4.0	Varni 2001 ¹⁹³	3	3	4	2	4	Inter-rater = 3			3		
16	Paediatric Quality of Life Inventory V4.0	Varni 2003 ¹⁹⁰	3	3	4	2	4	Inter-rater = 4			4		
17	Paediatric Quality of Life Inventory V4.0	Hughes 2007 ¹⁹⁶	3	3	4	2		Inter-rater = 3					
18	Paediatric Quality of Life Inventory V1.0	Varni 1999 ¹⁸³	4	3	4	3	4	Inter-rater = 4		3			
19	Sizing Me Up	Zeller 2009 ¹⁸³	3	1	4	1	4	TRT = 4; inter-rater = 3		4	2		
20	Sizing them up	Modi 2008 ¹⁸⁴	4	1	4	1	4	TRT = 4		4	2	4	
21	Youth Quality of Life Instrument – Weight Module	Morales 2011 ¹⁸⁵	4	4	4	2	4	TRT = 3		4	4		

Note: Scores = 1–4 (0 = N/A).

Appendix 25 Psychological well-being studies: development and evaluation scores

No.	Name of tool: psychological well-being measures	Development			Evaluation								
		First author	Clear concept	Underpinned by theory	Description of sample	Sample involved in development	Internal consistency	TRT/inter-rater	Internal validity	Criterion validity	Convergent validity	Construct validity	Responsiveness
1	Children's Body Image Scale (CBIS)	Truby 2002 ²⁴⁶	3	1	4	2			3			3	
2	Body figure perception (pictorial)	Collins 1991 ²⁰⁵	3	1	3	2	TRT = 4					3	
3	Self-Control rating scale (SCRS)	Kendall 1979 ¹⁹⁷	3	1	4	1	TRT = 2	2	3	3			
4	Self-Perception Profile for Children (SPPC)	Van Dongen-Melman 1993 ²⁰⁹	4	1	2	2	TRT = 4	4					
5	Perceived competence scale (aka SPPC/Harter)	Harter 1982 ¹⁹⁹	3	1	3	2	TRT = 4	4			3		
6	Physical Activity Enjoyment Scale (PACES)	Motl 2001 ²⁴⁹	3	1	3	2					3		
7	Self-Report Depression Symptom Scale (CES-D)	Radloff 1991 ²⁴⁶	3	1	1	1		3			4		
8	Children's Physical Self Perception Profile (C-PSPP)	Whitehead 1995 ²¹²	3	1	3	2	TRT = 3	4				3	
9	Children's Physical Self-Perception Profile (C-PSPP)	Eklund 1997 ²⁴⁸	3	1	2	2							
10	Children's Perceived Importance Profile (C-PIP)	Whitehead 1995 ²¹⁰	3	1	3	1	TRT = 3	4					

No.	Name of tool: psychological well-being measures	First author	Development			Evaluation							
			Clear concept by theory	Underpinned by theory	Description of sample	Sample involved in development	Internal consistency	TRT/ inter-rater	Internal validity	Criterion validity	Convergent validity	Construct validity	Responsiveness
11	Children's Self-Perceptions of Adequacy in and Predislection for Physical Activity (CSAPPA)	Hay 1992 ²¹¹	4	1	2	2	2	4				4	
12	Body Shape Questionnaire (BSQ)	Conti 2009 ²⁰⁶	4	1	1	1	3		TRT = 3			3	
13	Children's Physical Self-Concept Scale (CPSS)	Stein 1998 ²⁰⁷	4	1	4	2	4		TRT = 3			4	
14	Pediatric Barriers to a Healthy Diet Scale (PBHDS)	Janicke 2007 ²⁰⁰	3	1	4	2	4			4		3	3
15	Body Image Avoidance Questionnaire (BIAQ)	Riva 1998 ⁴²¹	2	1	2	1	4			4			
16	Video distortion	Probst 1995 ¹⁵²	3	0	2	1			TRT = 3			4	
17	Social Anxiety Scale for Children—Revised version (SASC-R)	La Greca 1993 ²⁰²	4	4	4	1	4			4		3	
18	Social Anxiety Scale for Children (SASC)	La Greca 1988 ²⁰¹	4	4	2	1	4		TRT = 4			4	
19	Nowicki–Strickland Locus of Control Scale (NS-LOCS)	Nowicki 1973 ²⁰³	4	4	2	2	3		TRT = 3			3	
20	Body Esteem Scale (BES)	Mendelson 1982 ²⁰⁴	3	1	3	1						3	2

Note: Scores = 1–4 (0 = N/A).

Appendix 26 Environment studies: development and evaluation scores

No.	Development				Evaluation								
	Name of tool: environment measures	First author	Development	Clear concept	Underpinned by theory	Sample involved in development	IC	TRT/inter-rater	Internal validity	Criterion validity	Convergent validity	Construct validity	Responsiveness
1	Nutrition and Physical Activity Self-assessment to Child Care (NAPSACC)	Benjamin 2007 ²⁴⁷	4	4	4	2		TRT = 4/ inter-rater = 4		4			
2	Environment and Policy Assessment and Observation (EPAO)	Ward 2008 ²¹³	3	1	1	2		Inter-rater = 4					
3	Healthy Home Survey (HHS)	Bryant 2008 ²¹⁴	4	1	4	2		TRT = 3		4			
4	Environment and Safety barriers to Youth Physical Activity Questionnaire	Durant 2009 ²²⁰	4	4	4	3	4	TRT = 4	4			3	
5	Family Eating and Activity Habits Questionnaire (FEAHQ)	Golan 1998 ²¹⁵	4	2	1	1	3	TRT = 2/ inter-rater = 4					4
6	Parenting Strategies for Eating and Activity Scale (PEAS)	Larios 2009 ²¹⁶	4	1	2	2	3		4		2	2	
7	Family Food Behaviour Survey (FFBS)	McCurdy 2010 ²¹⁹	4	1	3	2	3	TRT = 3		2		2	

		Development					Evaluation						
No. measures	Name of tool: environment	First author	Development	Clear concept	Underpinned by theory	Sample involved in development	IC	TRT/inter-rater	Internal validity	Criterion validity	Convergent validity	Construct validity	Responsiveness
8	Home Environment Survey (HES)	Gattshall 2008 ²¹⁷	4	4	4	1	3	TRT = 4/ inter-rater = 4				3	
9	Electronic equipment scale	Rosenberg 2010 ²¹⁹	4	1	4	2		TRT = 4/ inter-rater = 4				4	
10	Home Physical Activity Equipment scale	Rosenberg 2010 ²¹⁹	4	1	4	2		TRT = 4/ inter-rater = 4				4	

Note: Scores = 1–4 (0 = N/A).

Appendix 27 Non-English manuscripts of search 1 trials (data not extracted)

Childhood obesity treatment trials

1. Alves JG, Galé CR, Souza E, Batty GD. Effect of physical exercise on bodyweight in overweight children: a randomized controlled trial in a Brazilian slum. *Cad Saúde Pública* 2008;**24**:s353–9. URL: www.mrw.interscience.wiley.com/cochrane/clcentral/articles/132/CN-00666132/frame.html
2. Barnow S, Stopsack M, Bernheim D, Schroder C, Fusch C, Lauffer H, *et al.* Results of an outpatient intervention for obese children and adolescents. *Psychother Psychosom Med Psychol* 2007;**57**:353–8.
3. Blaik A, Westphal S, Dierkes J, Aronica S, Luley C. Comparison of two nutritional interventions in obese families. *Ernahrungs-Umschau* 2011;**58**:122–7.
4. Bustos Lozano G, Moreno Martin F, Calderin Marrero MA, Martinez Quesada JJ, Diaz Martinez E, Arana Canedo C. Comparative study of medical advice and cognitive-behavioral group therapy in the treatment of child-adolescent obesity. *An Esp Pediatr* 1997;**47**:135–43.
5. Canlorbe P, Borniche P, Toublanc JE. Controlled trial of an anorectic (An 448) in the treatment of childhood obesity. *Nouv Presse Med* 1976;**5**:1061–2.
6. Dai J, Jiang Z, Zhang B. Exercise and nutrition therapy for simple obesity in children. *Chin J Clin Rehabil* 2006;**10**:20–2.
7. de Mello ED, Luft VC, Meyer F. Individual outpatient care versus group education programs. Which leads to greater change in dietary and physical activity habits for obese children? *J Pediatr (Rio J)* 2004;**80**:468–74.
8. Ebert-Joisten M, Hahnemann B. 'A cheerful magician munch moderately'. The psychomotoric answer to overweight in childhood. *Ernahrungs-Umschau* 2004;**51**:B9–12.
9. Flodmark CE. A family-therapeutic method for the national disease of obesity. Start the treatment already when the children are about 10 years old! *Lakartidningen* 1996;**93**:2347–50.
10. Foger M, Bart G, Rathner G, Jager B, Fischer H, Zollner-Neussl D. Exercise, dietary counselling and psychological support in the treatment of obese children. A controlled study over 6 months. *Monatsschr Kinderh* 1993;**141**:491–7.
11. Golebiowska M, Chlebna-Sokol D, Kobierska I, Konopinska A, Malek M, Mastalska A, *et al.* Clinical evaluation of Teronac (mazindol) in the treatment of obesity in children. Part II. Anorectic properties and side effects. *Przegl Lek* 1981;**38**:355–8.
12. Golebiowska M, Chlebna-Sokol D, Mastalska A, Zwaigzne-Raczynska J. The clinical evaluation of teronac (Mazindol) in the treatment of children with obesity. Part I. Effect of the drug on somatic patterns and exercise capacity. *Przegl Lek* 1981;**38**:311–14.
13. Graf C, Kupfer A, Kurth A, Stutzer H, Koch B, Jaeschke S, *et al.* Effects of an interdisciplinary intervention on the BMI-SDS and the endurance performance capacity of adipose children: the CHILT III project. *Dtsch Z Sportmed* 2005;**56**:353–7.

14. Guzzaloni G, Calo G, Grugni G, Mazzilli G, Tonelli E, Ardizzi A, *et al.* Short term use of dexfenfluramine in a group of obese adolescents. *Clin Dietol* 1993;**20**:363–72.
15. Huang SH, Weng KP, Hsieh KS, Ou SF, Lin CC, Chien KJ, *et al.* Effects of a classroom-based weight-control intervention on cardiovascular disease in elementary-school obese children. *Acta Paediatr Taiwan* 2007;**48**:201–6. URL: www.mrw.interscience.wiley.com/cochrane/clcentral/articles/986/CN-00629986/frame.html
16. Jiang J, Xia X, Hui J, Cheng X. Comprehensive family based behavior modification for obese children. *Chin Ment Health J* 1997;**11**:242–4, 37.
17. Kang S, Kwoun S, Choi Y, Lim Y, Park D. The effects of Monacolin-inoculated rice embryo on the body fat and serum lipid profiles of obese elementary school students. *Korean J Community Nutr* 2005;**10**:565–73.
18. Kim HD, Park JS. The effect of an exercise program on body composition and physical fitness in obese female college students. *Taehan Kanho Hakhoe Chi* 2006;**36**:5–14.
19. Kim H-S. Effects of behavior modification on obesity index, skinfold thickness, body fat, serum lipids, serum leptin in obese elementary school children. *Taehan Kanho Hakhoe Chi* 2003;**33**:405–13.
20. Kwon MS, Hwang KS. Effects of an exercise program on body composition, cardiopulmonary function, and physical fitness for obese children. *Taehan Kanho Hakhoe Chi* 2007;**37**:568–75.
21. Le Q, Wang DX, Xia XH. Clinical observation on effect of heze oral liquid in treating children simple obesity. *Zhongguo Zhong xi yi jie he za zhi Zhongguo Zhongxiyi jiehe zazhi = Chinese J Integr Trad Western Med/Zhongguo Zhong xi yi jie he xue hui, Zhongguo Zhong yi yan jiu yuan zhu ban* 2002;**22**:384–5.
22. Lehrke S, Becker S, Laessle RG. Structured behavioral therapy with obese children: therapeutic effects in nutrition. *Verhaltenstherapie* 2002;**12**:9–16.
23. Lehrke S, Laessle R. Multimodal treatment for obese children: outcome with respect to psychosocial criteria. *Verhaltenstherapie* 2002;**12**:256–66.
24. Leopold K, Wechsler JG. Obesity: gradual-schedule therapy and long-term results. *MMW Fortschr Med* 2001;**143**:I–VIII.
25. Letonturier P. Reducing obesity. *Presse Med* 2006;**35**:77–8.
26. Li L, Wang Z-Y. Clinical therapeutic effects of body acupuncture and ear acupuncture on juvenile simple obesity and effects on metabolism of blood lipids. *Zhongguo zhenjiu* 2006;**26**:173–6.
27. Li WH, Wang JD, Gu LM, Wang YZ. Treatment of simple obesity with electro-acupuncture and auricular acupoint pressing: a report of 177 cases. *Zhong xi yi jie he xue bao* 2004;**2**:449, 58.
28. Liebermeister H, Jahnke K, Voss HJ, Englhardt A, Probst G. Initial and late results of diet therapy in obesity. *Dtsch Med Wochenschr* 1968;**93**:2149–55.
29. Liebermeister H, Probst G, Jahnke K. Experience with the appetite depressant, fenfluramine hydrochloride, in adiposity. *Med Klin* 1969;**64**:1201–7.

30. Lin RD, Lai SP, Cheng PL, Tang FC. Effect of nutrition education intervention on the physical fitness of exercise-induced weight loss children. *Nutr Sci J* 2005;**30**:183–95.
31. Livieri C, Novazi F, Lorini R. The use of highly purified glucomannan-based fibers in childhood obesity. *Pediatr Med Chir* 1992;**14**:195–8.
32. Malecka-Tendera E, Koehler B, Muchacka M, Wazowski R, Trzciakowska A. Efficacy and safety of dexfenfluramine treatment in obese adolescents. *Pediatr Pol* 1996;**71**:431–6.
33. Mulkens S, Fleuren D, Nederkoorn C, Meijers J. RealFit: a multidisciplinary (CBT) group treatment for obese youngsters. *Gedragstherapie* 2007; **40**:27–48.
34. Nagai N, Takekawa A. Assessment of the weight change in the improvement class for obese children. *Japan J Nutr* 1999;**57**:211–20.
35. Rascher W. Hypertension and the metabolic syndrome. Even children and adolescents require treatment. *MMW Fortschr Med* 2003;**145**:43.
36. Sabet-Sarvestani R, Kargar M, Kave MH, Tabatabaee H. The effect of dietary behavior modification on anthropometric indices in obese adolescent female students. *Iran J Pediatr* 2008;**18**(Suppl. 1):71–6.
37. Seo NS, Kim YH, Kang HY. Effects of an obesity control program based on behavior modification and self-efficacy in obese elementary school children. *Taehan Kanho Hakhoe Chi* 2005;**35**:611–20.
38. Sjostrom L, Rissanen A, Andersen T, Boldrin M, Golay A, Koppeschaar H, *et al.* Randomized placebo-controlled trial of orlistat for weight loss and prevention of weight regain in obese patients. *Ter Arkh* 2000;**72**:50–4.
39. Spranger J. Appetite depressants in the management of obesity in children. An expanded double-blind study with chlorphentermin (Avicol). *Munch Med Wochenschr* 1963;**105**:1338–41.
40. Spranger J. Phentermine resinate in obesity. Clinical trial of Mirapront in adipose children. *Munch Med Wochenschr* 1965;**107**:1833–4.
41. Stauber T, Petermann F, Korb U, Bauer A, Hampel P. Cognitive behavioral stress management for training obese children and adolescents. *Monatsschr Kinderheilkd* 2004;**152**:1084–94.
42. Strata A, Cucurachi L, Cucurachi P, Dell'anna A, Zuliani U. Model for clinico-pharmacological experimentation with an appetite depressant. Clinical trial with a delayed-action preparation. *Clin Ter* 1968;**44**:495–516.
43. Tak YR, An JY, Kim YA, Woo HY. The effects of a physical activity-behavior modification combined intervention (PABM-intervention) on metabolic risk factors in overweight and obese elementary school children. *Taehan Kanho Hakhoe Chi* 2007;**37**:902–13.
44. Wang L, Sun MX, Wang MF, Yan Y, Li BW, Zhong WJ, *et al.* Effects of different interventions on body mass index and body fat content in overweight and obese adolescents. *Chin J Clin Nutr* 2011;**19**:16–18.
45. Wu X, Wang J, Dong H. Childhood obesity intervention study in Xuzhou. *Mod Prev Med* 2010;**37**:2225–6.

46. Yang EJ. The effect of dumbbell exercise program training on body composition, blood lipids and cognitive perception in obese high school girls. *Korean Nurse* 1998;**37**:51–67.
47. Yu C, Zhao S, Zhao X. Treatment of simple obesity in children with photo-acupuncture. *Zhongguo Zhong Xi Yi Jie He Za Zhi* 1998;**18**:348–50.
48. Dobe M, Geisler A, Hoffmann D, Kleber M, von Koding P, Lass N, *et al.* The Obeldicks concept. An example for a successful outpatient lifestyle intervention for overweight or obese children and adolescents. *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz* 2011;**54**:628–35.
49. Ferrer Lorente B, Fenollosa Entrena B, Ortega Serrano S, Gonzalez Diaz P, Dalmau Serra J. Multidisciplinary treatment of pediatric obesity. Results in 213 patients. *An Esp Pediatr* 1997;**46**:8–12.
50. He YF, Wang WY, Fu P, Sun Y, Yu SY, Chen R, *et al.* Effects of a comprehensive intervention program on simple obesity of children in kindergarten. *Chin J Pediatr* 2004;**42**:333–6.
51. Korsten-Reck U. Obesity in childhood and adolescence: experiences and results of the intervention programme FITOC (Freiburg Intervention Trial for Obese Children) after 1.5 years. *ZFA* 2006;**82**:111–17.
52. Korsten-Reck U, Bauer S, Keul J. Sports and nutrition: an ambulatory care program for obese children (long-term experiences). *Pediatr Padol* 1993;**28**:145–52.
53. Salas A MI, Gattas Z V, Ceballos S X, Burrows A R. Effects of psychological support as an adjunct to a weight reducing program among obese children. *Rev Med Chil* 2010;**138**:1217–25.

Note: Full eligibility checking of the following citations has not been conducted.

Appendix 28 Childhood obesity Outcomes

Review appraisal decision form: secondary outcomes

			Decision of certainty: 1. <i>certain</i> – good evidence, fit for purpose; 2. <i>certain</i> – poor evidence, not fit for purpose; 3. <i>uncertain</i> – requiring further consideration			
No.	Diet	First author	CoOR internal decision	Expert consensus decision	CoOR internal comments	Expert collaborator comments
1	Korea FFQ	Lee 2007 ⁴⁸	2	2	Very specific to Korean diet and only TRT with poor development	
2	QFQ	Yaroch 2000 ⁴¹	2	2	Poor development with inadequate evaluation robustness scores (TRT and validity = 2) owing to sample size and poor results	
3	Short YAQ	Rockett 2007 ³⁴	1	1	Although development was not strong (although was created from long version – with good development), evaluation is good for this short, much-used tool	
4	YAQ	Rockett 1995 ⁴³	3	1	Well developed, but evaluation not great (validation was comparing with other similar national survey data and TRT had poor results)	This is a long tool, and may not always be feasible in all evaluations
5		Rockett 1997 ³⁷				
6		Perks 2000 ³⁰ (identified from a review post meeting)			(Note: later testing was in slightly different version and evaluation was better)	
7	Picture sort FFQ	Yaroch 2000 ⁴²	3	2	Developed specifically for obese/overweight, but has poor TRT. Validation is strong, but this is a long tool and participants were not involved in development	Might be useful for those with poor English/literacy, learning difficulties or the very young
8	CEHQ-FFQ	Lanfer 2011 ³⁶	3	1	Very strong development with good evaluation for the tests that were conducted (but are limited by criterion of milk consumption only)	
9		Huybrechts 2011 ³¹				
10	ACAES	Watson 2009 ⁴⁶	1	1	Very strong development and good overall validation (analysis needs to be adjusted for BMI for stronger validity)	
11		Burrows 2008 ³²				

Decision of certainty: 1. <i>certain</i> – good evidence, fit for purpose; 2. <i>certain</i> – poor evidence, not fit for purpose; 3. <i>uncertain</i> – requiring further consideration						
No.	Diet	First author	CoOR internal decision	Expert consensus decision	CoOR internal comments	Expert collaborator comments
12	Brief diet screener	Nelson 2009 ⁴⁴	3	2	Very strong development (including participants) but results for robustness were poor based on low sample size and correlations	
13	Brief diet screener	Davis 2009 ⁴⁵				
14	Intake of fried food away from home	Taveras 2005 ⁵¹	2	2	Single item, poor findings with poor development	
15	FIQ	Epstein 2000 ⁴⁹	2	2	Development not great, with convergent validity in only a small sample size	
16	Diet fat screening measure	Prochaska 2001 ⁵⁰	3	1	Used in trial, although developed as a screening tool. Development and results are strong but not stratified by obese (and is not focused on obesity)	Useful tool – but should be used only if the intervention focuses on reduction of dietary fat. Also specifically measured in 14 years only
17	New Zealand FFQ	Metcalfe 2003 ⁴⁷	1	1	Very strong development and reliability testing, but needs further validity testing	
18	HSFFQ	Blum 1999 ³⁸	3	1	Good development, but only tested for convergent validity so far (which was strong)	Note: needs TRT
					Note: at the point of submission of this report, authors contact CoOR to notify that this FFQ has been discontinued due to costs of maintenance	
19	FFQ	Crawford 1994 ³³	2	2	Development not strong/ clear and poor results for the only testing (criterion validity)	No TRT and limited to preschool. More testing required
20	DGI-CA	Golley 2011 ⁴⁵	3	2	Development not strong/ clear but strong results for construct validity	No TRT
21	FIFI-FFQ	Vereecken 2010 ³⁹ (identified after meeting)	2		No external decision, as this arrived (from the library) after involvement from experts. Decision based on those of similar tools	
					Early testing (convergent validity only) of this new tool that has potential in the future	

			Decision of certainty: 1. <i>certain</i> – good evidence, fit for purpose; 2. <i>certain</i> – poor evidence, not fit for purpose; 3. <i>uncertain</i> – requiring further consideration			
No.	Diet	First author	CoOR internal decision	Expert consensus decision	CoOR internal comments	Expert collaborator comments
22	Diet history	Sjoberg 2003 ⁵³	2	2	Decision based on all diet history papers. Although strong correlations in Sjoberg, others (Waling, ⁵⁴ Maffeis ⁵⁵), which were stratified by obese, were not strong and even worse in obese samples	
23		Waling 2009 ⁵⁴				
24		Maffeis 1994 ⁵⁴				
25	3-day food record	Maffeis 1994 ⁵⁵	3	2	All 3-day diaries considered together in decision-making. This has poor validity in obese. Singh ⁵⁷ also shows poor validity and Crawford ³³ has strong – but compares with lunch-time observations only (others = DLW/Lusk's)	
26		O'Connor 2001 ⁶⁴				
27		Crawford 1994 ³³				
28	9-day food diary	Singh 2009 ⁵⁷	2	2	Little development information, with poor validity	Diaries deemed to be explanatory tools, but not valuable as outcome measures
29	2-week food diary	Bandini 1990 ⁵⁸	2	2	Little development information, with poor validity	
30	2-week food diary	Bandini 1999 ⁵⁹ (identified from a review, post meeting)				
31	Tape-recorded food record (3 day)	Lindquist 2000 ⁶⁰	3	2	Although reasonable development and criterion validity robustness, the correlation with DLW was very poor	
32	Tape-recorded food record	Van Horn 1990 ⁵⁶ (same paper as above) (identified after meeting)	3	2		
33	Tape-recorded 240-hour recall	Van Horn 1990 ⁵⁶ (identified after meeting)	3	2		
34	7-day diet record	Bratteby 1998 ⁴¹⁰	2	2	Little development information, with poor validity	

Decision of certainty: 1. <i>certain</i> – good evidence, fit for purpose; 2. <i>certain</i> – poor evidence, not fit for purpose; 3. <i>uncertain</i> – requiring further consideration						
No.	Diet	First author	CoOR internal decision	Expert consensus decision	CoOR internal comments	Expert collaborator comments
35	8-day food record	Champagne 1996 ⁶³ (identified from a review, post meeting)	3	2		
36		Champagne 1998 ⁶² (identified from a review, post meeting)				
37	24-hour	Baxter 2006 ⁶⁵	3	2	Decision for 24-hour recall has been based on all papers, which have varying results Baxter results are strong (compared with observation) but there was a significant effect of obesity on accuracy. Johnson showed poor correlation with DLW. Lytle and Crawford used direct observation and both were well correlated	TRT conducted but showed odd correlations with BMI. Validity studies all have poor findings
38		Johnson 1996 ⁶⁸				
39		Lytle 1998 ⁶⁷				
40		Crawford 1994 ³³				
41	DILQ	Edmunds 2002 ⁶⁶	3	2	Developed for completion in school. Development strong, but statistical tests are not great. Tested responsiveness, but this was not strong	
42	DOCC	Ball 2007 ⁷⁰	3	2	Well developed with strong evaluation, but at child centre level with no description of sample (even though diet is measured on an individual level)	Maybe suitable for prevention/population based research but is high burden (researcher administered)
43	FBQ	Vance 2008 ⁷¹	3	2	Strong development and reliability, but criterion validity results are not clear/strong	
44	Biomarkers	Martinez de Icaya 2000 ⁶⁹ (identified after meeting)	3	2		Added after experts provided feedback. May be appropriate for inclusion but needs to be further considered in future research

Decision of certainty: 1. <i>certain</i> – good evidence, fit for purpose; 2. <i>certain</i> – poor evidence, not fit for purpose; 3. <i>uncertain</i> – requiring further consideration						
No.	Eating behaviours	First author	CoOR internal decision	Expert consensus decision	CoOR internal comments	Expert collaborator comments
1	ChEDE-interview	Decaluwé 2004 ⁸¹	3	2	Evaluation results/robustness = variable	All screening tools for ED (ED diagnosis) and therefore not included on this basis
2		Bryant-Waugh 1996 ⁴¹¹			Development and face validity paper only	
3		Tanofsky-Kraff 2005 ⁴¹³				
4	ChEDE-Q	Goossens 2010 ⁴¹²	3	2		ED diagnosis
5		Jansen 2007 ²²⁹				
6		Tanofsky-Kraff 2003 ²³⁰				
7	IFQ	Baughcum 2001 ⁷⁴	3	1	Moderate development and evaluation	Note: needs TRT
8	PFQ	Baughcum 2001 ⁷⁴	3	1	Evaluation for questionnaire structure only (IC, FA). Stratified by obesity for scores (greater in obese)	Note: needs TRT
9	KEDS	Childress 1993 ⁸⁹	3	2	Moderate development and evaluation	ED diagnosis
10	QEWPA	Johnson 1999 ⁹⁰	2	2	Used by trial in past (cited as Steinberg) but not obesity outcome (ED)	ED diagnosis
11		Steinberg 2004 ⁹¹			As above	
12	DEBQ-C	Van Strien 2008 ⁷⁹	3	1	Reasonably strong tool. No convergent validity or responsiveness	Note: needs TRT
13		Banos 2011 ⁸³				
14		Braet 2007 ⁹²				
15	DEBQ-P	Caccialanza 2004 ⁹⁸	3	1	Good structural validity, little other	
16		Braet 1997 ⁷⁸				
17	ChEAT	Maloney 1988 ⁸⁶	2	2	Variable results and not designed (although has been used in obesity trial): ED	ED diagnosis
18	ChEAT	Smolak 1994 ¹⁰⁰				
19	ChEAT	Ranzenhofer 2008 ¹⁰¹				

Decision of certainty: 1. <i>certain</i> – good evidence, fit for purpose; 2. <i>certain</i> – poor evidence, not fit for purpose; 3. <i>uncertain</i> – requiring further consideration						
No.	Eating behaviours	First author	CoOR internal decision	Expert consensus decision	CoOR internal comments	Expert collaborator comments
20	EAT	Wells 1985 ⁴¹⁴ (identified from a review post meeting)	2	2	Primary Development is in adults (Garner and Garfinkel 1979 ⁹). Little has been done to make it compatible for children and adolescents. ChEAT is later developed from this and is more specific to children	
21	YEDE-Q	Goldschmidt 2007 ⁹⁹	3	2	ED but used in trials. Poor development but strong evaluation	ED diagnosis
22	EES-C	Turnofsky-Kraff 2007 ⁷⁷	1	1	Strong tool, although development did not include participants	
23	C-BEDS	Shapiro 2007 ²³¹	2	2		ED diagnosis
24	CFQ	Birch 2001 ⁷⁵	1	1	Although studies should ensure that it is appropriate for their specific population characteristics, this is a well-used tool with good development and reasonably strong evaluation. Needs responsiveness testing	Haycroft paper needs double checking. Also need to expand search to include other validation papers outside CoOR remit
25		Haycraft 2008 ⁹³				
26		Anderson 2005 ⁹⁶				
27		Corsini 2008 ⁹⁷				
28		Polat 2010 ⁹⁴				
29	Boles 2010 ²³²					
30	MRFS-III	Shisslak 1999 ⁸⁷	2	2	Well developed and robust, but ED – not obesity (even although previously used in a trial)	ED diagnosis

Decision of certainty: 1. <i>certain</i> – good evidence, fit for purpose; 2. <i>certain</i> – poor evidence, not fit for purpose; 3. <i>uncertain</i> – requiring further consideration						
No.	Eating behaviours	First author	CoOR internal decision	Expert consensus decision	CoOR internal comments	Expert collaborator comments
31	IFSQ	Thompson 2009 ⁷⁶	3	1	Well developed but needs more evaluation	Needs TRT
32	CEBQ	Sleddens 2008 ⁷²	1	1	Reasonably well developed, with good robustness scores for evaluation conducted. Would benefit from further criterion/ convergent validity and responsiveness	Also available in other languages [Portuguese version picked up by CoOR search (Viana 2008 ¹⁴)]
33		Wardle 2001 ⁷³				
34	TSFFQ	Corsini 2010 ⁸²	1	1	Well developed, robust tool. Needs responsiveness testing	
35	KCFQ	Monnery-Patris 2011 ⁸⁵	3	1	Development not great, but reasonable evaluation	May be more appropriate in environmental domain
36		Carper 2000 ²⁵⁰				
37	Un-named (control in parental feeding practices)	Murashima 2011 ⁸⁴	3	1	Good evaluation, although construct validity findings were very weak	Need to check relevance to construct
38	EAH-C	Tanofsky-Kraff 2008 ⁸⁰	1	1	Development good, except does not include participants. All evaluation very strong	
39	Un-named (parental feeding strategies)	Kroller 2008 ⁸⁸	2	2	Strong development, but little evaluation and with German population	Poor evaluation

			Decision of certainty: 1. <i>certain</i> – good evidence, fit for purpose; 2. <i>certain</i> – poor evidence, not fit for purpose; 3. <i>uncertain</i> – requiring further consideration			
No.	Physical activity	Author	CoOR internal decision	Expert consensus decision	Comments	Expert comments
40	Accelerometer	Kelly 2004 ¹⁰⁵	1	1	Well-used tool with reasonable validation.	Fit for purpose but often dependent on the model.
41	Accelerometer – Actigraph	Pate 2006 ¹⁰⁷			Would benefit with responsiveness testing	Accelerometers will improve and change with time. The best recommended actigraph instrument is GT31M. For information on the best types of accelerometers please refer to de Vries review paper
42	Accelerometer – Caltrac monitor	Noland 1990 ¹⁰⁶				
43	Accelerometer – TriTrac Triaxial	Coleman 1997 ¹⁰⁸				
44	Accelerometer (Actigraph)	Guinhouya 2009 ²³⁴				
45	HR monitoring	Maffeis 1995 ²³⁷	2	2	(May be more suitable to Fitness domain) Tested against DLW, but found very large variation in agreement in obese (overall poor)	Poor in the individual level and depends on the calibration. More superior when used in combination with accelerometer
46	Pedometer	Kilanowski 1999 ¹¹⁴	3	1	Criterion validity testing reasonably strong, but little else tested	Objective tool so less prone to bias. Again, often depends on type of pedometer
47		Duncan 2007 ²⁴⁸				
48		Jago 2006 ¹¹²				
49		Mitre 2009 ¹¹⁰				
50		Treuth 2003 ¹¹³				
51	SenseWear Pro2 Armband	Backlund 2010 ¹¹¹	3	2	Validity testing strong, but done with small sample. Two models tested, with stronger results for model 5.1	
52	3D-PAR	Pate 2003 ⁴¹⁷	3	2	Criterion validity testing strong, but done with small sample	All self-reports deemed inappropriate
53	AQuAA	Slootmaker 2009 ¹¹⁷	2	2	Criterion validity testing showed questionnaire always overestimated activity in obese	All self-reports deemed inappropriate
54	Activity rating scale	Sallis 1993 ¹²¹	2	2	Poor validation results	All self-reports deemed inappropriate
55	Godin–Shephard Physical Activity Survey	Sallis 1993 ¹²¹	3	2	TRT good, but validity results poor	All self-reports deemed inappropriate

No.	Physical activity	Author	Decision of certainty:		Comments	Expert comments
			CoOR internal decision	Expert consensus decision		
			1. <i>certain</i> – good evidence, fit for purpose; 2. <i>certain</i> – poor evidence, not fit for purpose; 3. <i>uncertain</i> – requiring further consideration			
56	7-day recall interview	Sallis 1993 ¹²¹	1	2	Existing evaluation is strong (better in older children)	All self-reports deemed inappropriate
57	APARQ	Booth 2002 ¹²³	3	2	Good development and strong TRT but poor validation	All self-reports deemed inappropriate
58	CLASS	Telford 2004 ¹¹⁵	3	2	Involved participants in development. Reasonable robustness for evaluation, although criterion validity results were poor. Parent report better than self-report	All self-reports deemed inappropriate
59	GEMS Activity Questionnaire	Treuth 2003 ¹¹³	3	2	Only African American girls. Reasonable development, with good TRT, but poor validation	All self-reports deemed inappropriate
60	Activitygram	Treuth 2003 ¹¹³	2	2	Results of reliability and validity testing were poor (although conducted well)	All self-reports deemed inappropriate
61		Welk 2004 ¹¹⁶ (identified post meeting)				
62	Moderate to vigorous physical activity screening	Prochaska 2001 ²³⁵ (study 3)	1	2	Good development, with involvement of participants and strong criterion evaluation (also did pilot study 1). Needs responsiveness testing	All self-reports deemed inappropriate
63		Prochaska 2001 ²³⁵ (study 2)				
64	National Longitudinal Survey of Children and Youth	Sithole 2008 ¹³⁰	2	2	Items within a National Survey. Only inter-rater reliability testing and poor development	All self-reports deemed inappropriate
65	Outdoor Playtime Checklist – checklist	Burdette 2004 ¹²² (study 1)	2	2	Poor validation results, and convergent validity is (both tools) only marginally better even although were compared with each other	All self-reports deemed inappropriate
66	Outdoor Playtime Checklist – recall	Burdette 2004 ¹²² (study 2)	2	2		
67	Physical Activity Diary	Epstein 1996 ¹¹⁹	2	2	Not great development or evaluation	All self-reports deemed inappropriate

Decision of certainty: 1. <i>certain</i> – good evidence, fit for purpose; 2. <i>certain</i> – poor evidence, not fit for purpose; 3. <i>uncertain</i> – requiring further consideration								
No.	Physical activity	Author	CoOR internal decision	Expert consensus decision	Comments	Expert comments		
68	PAQ	Janz 2008 ¹²⁹	3	2	Strong development, but variable findings in evaluation for all studies (even though a lot of evaluation has been conducted). Criterion validity is poor. [Note: adolescent version is similar in terms of structure (with odd words changed) and finding – which is why they have been grouped together]	All self-reports deemed inappropriate		
69	PAQ-C	Kowalski 1997 ¹¹⁸						
70	PAQ-A	Kowalski 1997 ¹²⁵						
71	PAQ-C	Crocker 1997 ¹²⁸ (study 1)						
72	PAQ-C	Crocker 1997 ¹²⁸ (study 2)						
73	PAQ-C	Crocker 1997 ¹²⁸ (study 3)						
74	PAQ-C	Moore 2007 ¹²⁶ (study 1)						
75	PAQ-C	Moore 2007 ¹²⁶ (study 2)						
76	PAQ for Pima Indians	Kriska 1990 ²⁴¹	1	2			Reasonable development, but very poor evaluation findings	All self-reports deemed inappropriate
77	PAQ for Pima Indians	Goran 1997 ¹²⁷						
78	PDPAR	Trost 1999 ⁴¹⁸	3	2	Development and validity not great, but reliability is good and this is a well-used tool (there are likely to be other papers that have not yet been identified)	All self-reports deemed inappropriate		
79	PDPAR	Weston 1997 ¹²⁰						
80	PDPAR	Welk 2004 ¹¹⁶						
81	PDPAR	McMurray 2008 ⁴¹⁹						
82	YRBS	Troped 2007 ²³⁸	2	2	Items within surveillance tool with reasonable TRT and criterion validity – but only just. Not designed as an outcome measure, even although it was previously used as one	All self-reports deemed inappropriate		

			Decision of certainty: 1. <i>certain</i> – good evidence, fit for purpose; 2. <i>certain</i> – poor evidence, not fit for purpose; 3. <i>uncertain</i> – requiring further consideration			
No.	Physical activity	Author	CoOR internal decision	Expert consensus decision	Comments	Expert comments
83	SOCARP	Ridgers 2010 ¹⁰² [previous Category 4 (not eligible) but recommended by experts]		1		
84	OSRAC	Brown 2006 ¹⁰³ [previous Category 4 (not eligible) but recommended by experts]		1		

ED, eating disorder.
Note that 'study 1' and 'study 2' are used to indicate manuscripts that report two studies in one paper. This is distinct from manuscripts published in the same year by the same lead author, which are distinguished by their individual reference citation numbers.
a Not linked to bibliography: Garner DM, Garfinkel PE. The eating attitudes test: an index of the symptoms of anorexia nervosa. *Psychol Med* 1979;**9**:273–9.

Decision of certainty: 1. <i>certain</i> – good evidence, fit for purpose; 2. <i>certain</i> – poor evidence, not fit for purpose; 3. <i>uncertain</i> – requiring further consideration						
No.	Sedentary behaviour	First author	CoOR internal decision	Expert consensus decision	Comments	Expert comments
1	Accelerometer – WAM-7154	Reilly 2003 ¹³¹	3	1	Only assessed criterion validity, but results were strong	Objective but can often depend on device
2	Accelerometer – Actigraph	Puyau 2002 ¹³²			Strong criterion and convergent validity, but small sample size for both	
3	Mini-Mitter Activwatch monitors	Puyau 2002 ¹³²	3	1	Strong criterion and convergent validity, but small sample size for both	Objective but can often depend on device
4	MARCA	Ridley 2006 ¹³³	3	2	Well developed, using participants	
5	EMA: self-report survey on mobile phones	Dunton 2011 ¹³⁴	3	2	Well developed, using participants	Has potential but needs to be explored further
6	Habit books with index cards	Epstein 2004 ¹³⁵ (identified post meeting)	3	2		

Decision of certainty: 1. <i>certain</i> – good evidence, fit for purpose; 2. <i>certain</i> – poor evidence, not fit for purpose; 3. <i>uncertain</i> – requiring further consideration						
No.	Fitness	Author	CoOR internal decision	Expert consensus decision	Comments	Expert comments
1	6-minute walk test (6MWD)	Morinder 2009 ¹³⁸	3	2	Reasonable evaluation	Body weight dependent
2	Height-adjustable step test	Francis 1991 ¹⁴⁸	3	2	Reasonable evaluation	Body weight dependent
3	20-m shuttle run	Suminski 2004 ¹⁴⁰	3	2	Reasonable evaluation	Body weight dependent
4		Leger 1988 ¹³⁹				Further evaluation required
5	International Fitness Scale (IFS)	Ortega 2011 ¹³⁶	1	2	Although development is not great, evaluation is robust	Self-report should not be used to report CVF. Also not valid for change from baseline to follow-up
6	Bioelectrical impedance	Roberts 2009 ¹⁴⁷	2	2	Large variation in findings (especially by gender) and magnitude of bias	
7	Fitnessgram	Morrow 2010 ¹⁴⁰	2	2	Although developed for obesity research, this is school based (and likely to be for prevention). Good reliability, but no validation conducted	
8	Submaximal Treadmill Test	Nemeth 2009 ¹⁴⁹	3	2	[Stats need checking – not confident that extracted value relates to model building and not validation]	Body weight dependent
9	BMR with fat-free mass	Drinkard 2007 ¹⁴³	2	2	Although significant correlations – limits of agree are outside acceptable range and there was sign magnitude of bias in obese	
10	Estimated maximal oxygen consumption and maximal aerobic power	Aucouturier 2009 ¹⁴⁴	2	2	Although significant correlations = poor agreement and authors suggest the estimated measures are not valid	
11	Physical working capacity on cycle ergometer	Rowland 1993 ¹⁴⁵ (identified post meeting)	3	2		

Decision of certainty: 1. <i>certain</i> – good evidence, fit for purpose; 2. <i>certain</i> – poor evidence, not fit for purpose; 3. <i>uncertain</i> – requiring further consideration						
No.	Fitness	Author	CoOR internal decision	Expert consensus decision	Comments	Expert comments
12	Aerobic cycling power	Carrel 2007 ¹⁴⁶	2	2	Poor results for validation tested in a small sample size	Based on this single study no but for wider evidence it is considered a good tool
13	Measured VO ₂ peak	Loftin 2004 ¹⁴¹	3	1	Good validity for both bike and treadmill, but bike was more acceptable to participants	Measured VO ₂ peak (bike) is often referred to as a criterion measure
14	Harvard Step Test	Meyers 1969 ¹⁴² (identified post meeting)			Body weight dependent	

Decision of certainty: 1. <i>certain</i> – good evidence, fit for purpose; 2. <i>certain</i> – poor evidence, not fit for purpose; 3. <i>uncertain</i> – requiring further consideration							
No.	Physiology	First author	CoOR internal decision	Expert consensus decision	Comments (note: not judged on development, as these measures were not developed specifically for obesity)	Expert comments	
1	Indices of insulin sensitivity	Yeckel 2004 ¹⁵²	1	1	All comparing fasting indices with gold standard (clamp or OGTT). Strong results throughout, indicating the fasting measures are a reasonably good surrogate	Good in epidemiology with large samples as opposed to an individual level	
2	Fasting indices of insulin sensitivity	Conwell 2004 ¹⁵³				A clamp should be used in smaller studies. They are good surrogates for insulin sensitivity but puberty status may affect results	
3	Indices of insulin sensitivity	George 2011 ¹⁵⁴					
4		Gunczler 2006 ¹⁵⁵					
5		Uwaifo 2002 ¹⁵⁶					
6	Insulin sensitivity and pancreatic beta cell function	Gungor 2004 ¹⁵⁸					
7	Fasting indices of insulin sensitivity	Atabek 2007 ¹⁵⁹					
8	Homeostasis model assessment of insulin resistance	Keskin 2005 ¹⁶⁰					
9		Rossner 2008 ¹⁶¹					
10	Indices of insulin sensitivity	Schwartz 2008 ¹⁶²					
11	Impaired fasting glucose	Cambuli 2009 ¹⁶³	3	2	Low sensitivity, but high specificity		
12	Hyperglycaemic clamp	Uwaifo 2002 ¹⁵⁷	3	2	Comparison of two gold standards, basing euglycaemic clamp as the primary. Found hyper to overestimates	Good measure but not appropriate for obese sample	
13	Oral Glucose Tolerance Test (OGTT)	Libman 2008 ¹⁶⁴	3	2	Results for TRT are reasonable, but unclear for validity		
14	¹³ C-glucose breath test – insulin resistance	Jetha 2009 ¹⁶⁵	3	2	Although results are good, they are variable	Diagnostic	

No.	Physiology	First author	Decision of certainty: 1. <i>certain</i> – good evidence, fit for purpose; 2. <i>certain</i> – poor evidence, not fit for purpose; 3. <i>uncertain</i> – requiring further consideration		Comments (note: not judged on development, as these measures were not developed specifically for obesity)	Expert comments
			CoOR internal decision	Expert consensus decision		
15	Ultrasound analysis of liver echogenicity	Soder 2009 ¹⁷⁷	3	2	Good correlation between radiologists using three ultrasound units, but no further testing	
16	HbA _{1c}	Nowicka 2011 ¹⁷⁴	3	2	Overall = poor sensitivity	
17	Ghrelin	Kelishadi 2008 ¹⁷⁵	3	2	Poor construct validity but has tested responsiveness, which was good	
18	PPG	Russoniello 2010 ⁴²⁰ (included post meeting)				Feedback from experts on the provisional CoOR Framework, including consideration of this measure, did not lead to its inclusion
19	Estimated resting metabolic rate	Molnar 1995 ¹⁶⁶	3	2		All compared predicted REE with measured REE. Variable results but all suggest that predictions are adequate. Hofsteenge results are not as good and this is specifically for obese sample
20	Predicted REE	Rodriquez 2002 ¹⁶⁷				
21		Lazzer 2006 ¹⁶⁸				
22		Firouzbakhsh 1993 ¹⁶⁹				
23		Derumeaux-Burel 2004 ¹⁷⁰				
24	Predicted REE	Hofsteenge 2010 ¹⁷¹				
25	DXA-lean body mass for REE	Schmelzle ¹⁷²	1	2		Strong results
26	BMR with fat-free mass	Dietz 1991 ¹⁷³	3	2		Derivation of fat-free mass not clear, therefore comparisons not clear. Results are poor

OGTT, oral glucose tolerance test; REE, resting energy expenditure.

			Decision of certainty: 1. <i>certain</i> – good evidence, fit for purpose; 2. <i>certain</i> – poor evidence, not fit for purpose; 3. <i>uncertain</i> – requiring further consideration			
No.	Health-related quality of life	First author	CoOR internal decision	Expert consensus decision	Comments	Expert comments
1	Child Health Questionnaire (CHQ)	Waters 2000 ¹⁹²	3	1	CoOR appraisal scores for quality are poor. Convergent validity results presented only for significant items	
2		Landgraf 1998 ¹⁸⁶				
3		Waters 2000 ¹⁹³				
4	DISABKIDS	Ravens-Sieberer 2007 ¹⁹⁴ (study 1)	3	1	Strong development and evaluation but only did IC and convergent validity	
5	KIDSCREEN	Ravens-Sieberer 2007 ¹⁹⁴ (study 2)	3	1	Strong development and evaluation but only did IC and convergent validity	
6	EQ-5D-Y	Burstrom 2011 ²⁴¹	2	1	Well-used, historical tool with further testing in sample stratified by obese	
7		Burstrom 2011 ²⁴²			Convergent validity with youth version doing better than original adult version	
8		Wille 2010 ²⁴³				
9		Ravens-Sieberer 2010 ²⁴⁴			Also measure TRT with strong agreement (although kappa less strong)	
10	Impact of Weight on Quality of Life (IWQoL)	Kolotkin 2006 ¹⁸¹	1	1	Strong evaluation – including responsiveness. Also tested in a Dutch study (identified by CoOR), although not able to translate Wouters 2010 ¹⁵	
11		Modi 2011 ¹⁸²				
12	KINDL-R Questionnaire	Erhart 2009 ¹⁸⁷	3	1	Good evaluation of IC, FA and convergent validity, but development not strong	
13	Paediatric Cancer Quality of Life Inventory	Varni 1998 ¹⁸⁸	3	2	This tool was used as a basis for construction of the PedsQL	
14	Paediatric Cancer Quality of Life Inventory (long)	Varni 1998 ¹⁹⁵			Only evaluates inter-rater and not specific to obesity	
15	Paediatric Quality of Life Inventory V4.0	Varni 2001 ¹⁹¹	1	1	Development acceptable and strong evaluation	
16	Paediatric Quality of Life Inventory V4.0	Varni 2003 ¹⁹⁰				

Decision of certainty: 1. <i>certain</i> – good evidence, fit for purpose; 2. <i>certain</i> – poor evidence, not fit for purpose; 3. <i>uncertain</i> – requiring further consideration						
No.	Health-related quality of life	First author	CoOR internal decision	Expert consensus decision	Comments	Expert comments
17		Hughes 2007 ¹⁹⁶				
18	Paediatric Quality of Life V1.0	Varni 1999 ¹⁸⁹	2	2	First version – updated since	
19	Sizing Me Up	Zeller 2009 ¹⁸³	3	1	Overall very good – but no involvement of participants and poor construct validity	
20	Sizing Them Up	Modi 2008 ¹⁸⁴	1	1	Very high evaluation scores but no participant involvement in development	
21	Youth Quality of Life Instrument – Weight Module (YQL-W)	Morales 2011 ¹⁸⁵	1	1	Very good development and strong evaluation specific for obese	

No.	Psychological well-being	First author	Decision of certainty: 1. <i>certain</i> – good evidence, fit for purpose; 2. <i>certain</i> – poor evidence, not fit for purpose; 3. <i>uncertain</i> – requiring further consideration		Decision of certainty: 1. <i>certain</i> – good evidence, fit for purpose; 2. <i>certain</i> – poor evidence, not fit for purpose; 3. <i>uncertain</i> – requiring further consideration	Expert comments
			CoOR internal decision	Expert consensus decision		
1	Children's Body Image Scale (CBIS)	Truby 2002 ¹⁹⁸	3	1	May be more appropriate for ED research, although was stratified by obesity	Developed specifically for children's body image perception for eating disorders. Additional manuscript Truby 2004; <i>Br J Psychol</i> (not identified by CoOR)
2	Body figure perception (pictorial)	Collins 1991 ²⁰⁵	3	1	Good development but evaluation less strong	Needs further evaluation. Is reference population relevant to UK?
3	Self-Control Rating Scale (SCRS)	Kendall 1979 ¹⁹⁷	3	2	Development not strong but has been tested thoroughly. However, robustness score always fails for poor results in validity testing	
4	Self-Perception Profile for Children (SPPC)	Van Dongen-Melman 1993 ²⁰⁹	1	1	Used participants in development and all evaluation tests were strong. Needs responsiveness testing	Experts also noted a version that is used in adolescents (SPPA), which was not identified by the CoOR search
5	Perceived Competence Scale (aka SPPC/Harter)	Harter 1982 ¹⁹⁹			Same tool (name change) as SPPC	The 'Perceived Importance Profile' (PIP) Whitehead 1995, ^{182,210} below) is an add-on to the SPPC and should be used in conjunction to determine the degree to which children feel their perceptions of their selves is important
6	Physical Activity Enjoyment Scale (PACES)	Motl 2001 ²⁵²	3	1	Used participants in development, but only assessed CFA (which was strong)	For use in adolescents only
7	Self-report Depression Symptom Scale (CES-D)	Radloff 1991 ²⁴⁶	2	2	Poor development, and assessed IC only (in which $a < 0.7$). Developed originally for adults	

No.	Psychological well-being	First author	Decision of certainty:		Decision of certainty:	Expert comments
			CoOR internal decision	Expert consensus decision		
			1. <i>certain</i> – good evidence, fit for purpose; 2. <i>certain</i> – poor evidence, not fit for purpose; 3. <i>uncertain</i> – requiring further consideration		1. <i>certain</i> – good evidence, fit for purpose; 2. <i>certain</i> – poor evidence, not fit for purpose; 3. <i>uncertain</i> – requiring further consideration	
8	Children's Physical Self-Perception Profile (C-PSPP)	Whitehead 1995 ²¹⁰	1	1	Good development, including participants. Strong reliability and good structure, but needs further evaluation	
9	Children's Physical Self Perception Profile (C-PSPP)	Eklund 1997 ²⁴⁵ (identified post meeting)				
10	Children's Perceived Importance Profile (C-PIP)	Whitehead 1995 ²¹⁰	3	1	Developed without participants. Tested in small sample, with reasonable results	More testing needed, especially construct validity, but is recommended to use in conjunction with the SPPC
11	Children's Self-perceptions of Adequacy in Predilection for Physical Activity (CSAPPA)	Hay 1992 ²¹¹	1	1	Well developed, with strong results	
12	Body Shape Questionnaire (BSQ)	Conti 2009 ²⁰⁶	3	2	Poor development with moderate results	Developed for adults
13	Children's Physical Self-concept Scale (CPSS)	Stein 1998 ²⁰⁷	1	1	Strong development using participants, with strong evaluation, although needs more testing (showed discriminate validity by obesity)	
14	Pediatric Barriers to a Healthy Diet Scale (PBHDS)	Janicke 2007 ²⁰⁰	3	2	Well developed with participants, with good robustness scores for evaluation. However, all lost scores relate to poor results	Needs much more evaluation; diet focused
15	Body Image Avoidance Questionnaire (BIAQ)	Riva 1998 ⁴²¹	3	2	Development not great, but internal testing on scale is very good. Needs more evaluation	
16	Video distortion	Probst 1995 ²⁰⁸	3	2	Poor development with reasonable evaluation	Technically difficult; developed for disordered eating

No.	Psychological well-being	First author	Decision of certainty:		Decision of certainty:	Expert comments
			CoOR internal decision	Expert consensus decision		
			1. <i>certain</i> – good evidence, fit for purpose; 2. <i>certain</i> – poor evidence, not fit for purpose; 3. <i>uncertain</i> – requiring further consideration		1. <i>certain</i> – good evidence, fit for purpose; 2. <i>certain</i> – poor evidence, not fit for purpose; 3. <i>uncertain</i> – requiring further consideration	
17	Social Anxiety Scale for Children–Revised version (SASC-R)	La Greca 1993 ²⁰² (identified post meeting)	3	1	Both large studies with multiple evaluation (IC, FA, TRT, convergent validity) with fairly strong results, but – not tested for obese. Included as identified in search 1 (already being used)	Basis of development and subsequent use is not child obesity, but social anxiety is an issue in some obese children. Measure is fit for purpose and social anxiety is an issue in some obese children being used
18	Social Anxiety Scale for Children (SASC)	La Greca 1988 ²⁰¹ (identified post meeting)				
19	Nowicki–Strickland Locus of Control Scale (NS-LOCS)	Nowicki 1973 ²⁰³ (identified post meeting)	3	2	Fairly robust testing in large samples. May be dated	Met criterion for eligibility but the basis of development and subsequent use is not child obesity
20	Body Esteem Scale (BES)	Mendelson 1982 ²⁰⁴ (identified post meeting)	3	1	Minimal testing in small sample. Identified through a review and presents results by obesity [construct validity correlation with weight (R = 0.55)]	Long pedigree in child obesity research. It has gone through a few minor modifications and is still the best measure of this construct in the context of child obesity. Fewer people are using a single measure of body esteem, as most measures of dimensional self-esteem and quality of life include some assessment of satisfaction with appearance. However, I would definitely recommend the measure for inclusion and would specify the version in the citation: www.sciencedirect.com/science/article/pii/S0193397396900301

CFA, confirmatory factor analysis.

No.	Environment	Author	Decision of certainty: 1. <i>certain</i> – good evidence, fit for purpose; 2. <i>certain</i> – poor evidence, not fit for purpose; 3. <i>uncertain</i> – requiring further consideration		Internal appraisal comments	Expert appraisal comments
			CoOR internal decision	Expert consensus decision		
1	Nutrition and Physical Activity Self-assessment to Child Care (NAPSACC)	Benjamin 2007 ²⁴⁷	1	1	Very good development and strong criterion validation (but a child-care centre tool)	Has good potential, but may be too intervention specific (may not be generalisable)
2	Environment and Policy Assessment and Observation (EPAO)	Ward 2008 ²¹³	3	2	Development involved users, but has no information on individual level. Needs further assessment (inter-rater very strong)	High degree of burden
3	Healthy Home Survey (HHS)	Bryant 2008 ²¹⁴	2	2	First stage of testing, (second version has been developed – but is in analysis phase)	
4	Environment and Safety Barriers to Youth Physical Activity Questionnaire	Durant 2009 ²²⁰	1	1	Very strong tool but would benefit with criterion validity and responsiveness	
5	Family Eating and Activity Habits Questionnaire (FEAHQ)	Golan 1998 ²¹⁵	3	2	Has potential (and has strong results for responsiveness), but needs further testing (reliability results were poor)	Poor reliability in small sample. Cross-cultural validity not clear
6	Parenting Strategies for Eating and Activity Scale (PEAS)	Larios 2009 ²¹⁶	3	2	Good internal structure but some of the evaluation results are poor	Poor psychometrics
7	Family Food Behaviour Survey (FFBS)	McCurdy 2010 ²¹⁷	3	2	Holds potential, but has poor robustness because of sample size in evaluation	Small sample size
8	Home Environment Survey (HES)	Gattshall 2008 ²¹⁸	1	1	Very well developed with strong evaluation, but is quite long	
9	Electronic equipment scale	Rosenberg 2010 ²¹⁹ (study 1)	1	1	Well developed using participants with strong validation. Needs criterion and responsiveness testing	
10	Home Physical Activity Equipment scale	Rosenberg 2010 ²¹⁹ (study 2)	1	1	Well developed using participants with strong validation. Needs criterion and responsiveness testing	

A decorative graphic consisting of numerous thin, parallel green lines that curve from the left side of the page towards the right, creating a sense of movement and depth.

**EME
HS&DR
HTA
PGfAR
PHR**

Part of the NIHR Journals Library
www.journalslibrary.nihr.ac.uk

This report presents independent research funded by the National Institute for Health Research (NIHR). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health

Published by the NIHR Journals Library